BMC Genomics

RESEARCH Open Access

Using genotype imputation to integrate Canola populations for genome-wide association and genomic prediction of blackleg resistance

Huanhuan Zhao[1*], Iona M MacLeod[1,2], Gabriel Keeble-Gagnere[1], Denise M Barbulescu[1], Josquin F Tibbits[1], Sukhjiwan Kaur[1] and Matthew Hayden[1,2*]

Abstract

Background Integrating germplasm populations genotyped by different genotyping platforms via genotype imputation is a way to utilize accumulated genetic resources. In this study, we used 278 canola samples genotyped via whole-genome sequencing (WGS) at 10× coverage to evaluate the imputation accuracy of three imputation approaches. The optimal imputation methods were used to impute and integrate two Canola genotype datasets: a diverse canola collection genotyped by genotyping-by-sequencing via transcriptome (GBS-t) and a double haploid (DH) line collection genotyped with low-coverage WGS (skim-WGS). The genomic predictive ability (GP) and detection power of marker–trait association (GWAS) of the combined population for blackleg resistance were evaluated.

Results The empirical imputation accuracy (r^2) measured as the squared correlation between observed and imputed genotypes was moderate for Minimac3 when imputing from the GBS-t density to the WGS. The accuracy dramatically improved from 0.64 to 0.82 by removing SNPs with poor Minimac3-reported Rsq (Rsq < 0.2) quality statistics. The r^2 for GLIMPSE was higher than that for Beagle when imputing from different low-coverage to full-coverage WGS. We imputed and integrated the diverse canola collection and the DH lines, and the combined population showed similar or slightly greater predictive ability (PA) for blackleg resistance traits than did each of the single populations with ~ 921 K SNPs. Higher marker-trait association (MTA) detection powers were indicated with the combined population; however, similar numbers of MTAs were discovered when each single population was combined in a meta-GWAS.

Conclusion It is feasible to impute and integrate germplasms from different sequencing platforms for downstream analyses. However, genetic heterogeneity across populations could add complexity to the analysis. Increasing the sample size by combining datasets showed slightly greater predictive ability and greater detection power in GWASs in the present study.

*Correspondence:
Huanhuan Zhao
huan.zhao@agriculture.vic.gov.au; huan.zhao@adelaide.edu.au
Matthew Hayden
matthew.hayden@agriculture.vic.gov.au

Full list of author information is available at the end of the article

Keywords Imputation, Whole-genome sequencing (WGS), skim-WGS, GBS-t, Canola, GP, GWAS, Blackleg resistance

Introduction

Modern plant breeding programs require accurate and cheap genotyping tools when genotyping a large number of breeding materials. Low-cost genotyping-by-sequencing (GBS) based on reduced representation sequencing and high- or low-density SNP arrays are popular genotyping methods for breeders [1, 2]. With advances in sequencing technology, low-coverage whole-genome sequencing (skim-WGS), which does not require library preparation, has emerged as an alternative approach in genotyping [3]. Thus, it is important to investigate how to integrate the increasing number of genetic and genomic resources obtained from different sequencing platforms and germplasm panels to maximize their utilization in breeding programs.

Genotype imputation is a modern genomic tool developed to infer missing genotypes of individuals. It is widely used to infer sporadic missing genotypes in a dataset (such as in skim sequencing) or to fill in higher-density genotypes for individuals with lower-density genotype data [4]. Schmidt et al.. imputed the missing genotypes from a 9 K SNP chip in barley to assess the genomic prediction of malting quality traits [5]. Shi et al.. conducted genotype imputations in two steps, first from a low SNP density (9 K array) to 90 K and then from 90 K to the exome sequence in wheat [6]. Several genotype imputation methods developed over the past two decades can be broadly clustered into two categories according to whether a reference population is needed. When a published assembly is unavailable and genotype markers are unordered, the map-independent methods of imputation are adopted. This type of imputation deploys linkage disequilibrium (LD) information between SNP markers to predict missing genotypes [7, 8]. He et al. found that map-dependent methods had substantially greater imputation accuracy than map-independent methods using a diverse wheat collection [9]. However, when SNPs are mapped to a high-quality assembly, the accuracy of imputing genotypes without a reference panel for low-coverage skim-WGS reached accuracies comparable to the imputation with a reference panel [10]. Imputation methods with reference populations can be further subdivided into population-based and pedigree-based methods. Pedigree-based imputation requires accurate pedigree information to extract the identity-by-descent (IBD) information for imputation [11], while population-based imputation methods mainly deploy the haplotype information inferred from the reference population [12]. Pedigree-based imputation has been shown to be advantageous for biparental breeding populations, and high imputation accuracy has been achieved for imputing both descendants and parents [13, 14]. Population-based imputing approaches, implemented in popular imputation softwares, such as Minimac, Beagle, and Impute, are generally useful for imputing genotypes in unrelated individuals [15–17]. Torkamaneh & Belzile tested fastPHASE, Beagle, and Impute2 using soybean GBS and SNP array datasets, and reported that an imputation accuracy as high as 90% could be achieved [18].

The use of imputed sequence genotypes is generally reported to improve statistical power in genome-wide association studies (GWAS) [19] and genomic prediction [20]. Sakhale et al.. used imputed genotypes in rice to identify novel candidate genes for adaptation to dry direct seeding in the field [21]. In ryegrass, the imputation of genotypes from low coverage to high coverage resulted in a meaningful increase in genomic prediction accuracy (up to 9%) [22]. Although the use of imputation to integrate datasets from different sequencing platforms for genomic evaluation has been widely studied in human and animal research [23, 24], similar studies in plants are relatively limited [25]. Wang et al.. proposed the first rice imputation pipeline to integrate different rice genetic resources with improved statistical power to identify quantitative trait loci (QTLs) via GWASs [26].

Canola (*Brassica napus* L.) is an oil seed crop cultivated globally, with major growers including the European Union, Canada, China, India and Australia [27]. One of the major diseases affecting this crop is Blackleg, caused by the fungal pathogen *Leptosphaeria maculans*, which can lead yield losses of up to 80% [28]. Developing canola varieties with blackleg resistance is the most efficient and sustainable way to combat this disease [29]. Genomic tools such as GWAS and genomic selection (GS) have been incorporated in plant breeding programs for germplasm enhancement and varieties development [30]. GWAS identifies genomic regions or loci underlying genetic variation in target traits [31], whereas GS uses genome-wide markers to develop genomic prediction models to assist the selection of superior individuals with selected traits that have genotypes but no phenotypes [32]. In canola, a diverse panel of 337 canola/rapeseed accessions was used for GWAS and genomic prediction of sclerotinia stem rot (SSR) resistance, which detected ninety-eight significant SNPs associated with the SSR resistance and achieved medium to high genomic prediction accuracy [33]. GWAS conducted on a set of 213 accessions for canola blackleg resistance revealed eight MTAs distributed among seven chromosomes, and three of these MTAs explained more than 30% of the phenotypic variation [34]. Raman et al.. performed a GWAS on a diverse panel of 179 canola accessions and

discovered a new resistance gene, *Rlm12*, to *L. maculans* located on chromosome A01 and a few additional QTLs [35]. Genomic prediction for canola blackleg resistance achieved low to moderate prediction accuracy within both spring (0.30–0.69) and winter canola (0.19–0.71) populations by using 532 diverse canola accessions [36].

In this study, we collected three canola datasets genotyped on different platforms, the GBS-t dataset, the skim-WGS dataset, and the 10× WGS dataset, and we combined these genomic resources through imputation to explore the benefit of the combined dataset for GS and GWAS in canola blackleg resistance. Therefore, the first objective of this study was to evaluate the imputation performance of three imputation software programs: Minimac, Beagle, and GLIMPSE. The second objective was to impute the datasets with the best-performing imputation approaches and combine the datasets. The third objective was to investigate the performance of the combined dataset for GWAS and GS in canola blackleg resistance.

Materials and methods

Genotype datasets

The three canola genotype datasets used in our study have been described in previous studies: (1) a diverse canola collection of 638 accessions and lines (including spring canola and winter canola types) sourced from the Australian grain GenBank and genotyped by GBS-t [37], (2) a canola double haploid (DH) line sample set of 1500 individuals derived from 84 different 4-way crosses within 97 Australian spring canola varieties and genotyped with skim-WGS [38], and (3) a canola core collection selected from the diverse population and genotyped with WGS at 10× coverage [37]. In addition, we sequenced an additional 89 winter canola, 18 spring canola, and 24 DH lines with 10× coverage WGS using the same protocol described in [37]. Information on these canola individuals is presented in supplementary Table S1.

All canola raw sequences used in the study were realigned to the new 2020 *Darmor-bzh* (*B. napus*) reference assembly [39] using the Burrows–Wheeler Aligner (BWA) [40]. The sequence alignment and variant calling procedures followed the in-house scripts described in Malmberg et al. [37]. We applied a stringent SNP filter for the 10× WGS dataset and removed the sites with a minimum mapping quality score ≤ 30, a read depth ≤ five, SNPs with a missing rate > 30%, a minor allele frequency (MAF) < 0.03, and heterozygosity > 20%. This resulted in 7.9 million SNPs, and sporadic missing genotypes were imputed by Beagle 4.0 with default parameters [16]. The imputed WGS set, which included a total of 278 canola individuals with 7.9 million SNPs, was used as the preliminary imputation reference. We filtered the variants of the GBS-t and skim-WGS datasets with a mapping quality > 20, a mean read depth > 3 and a missing rate of < 80%. The overlapping SNPs (shared SNPs) among GBS-t, skim-WGS, and WGS, including tagging SNPs (LD_r^2 = 0.95) by PLINK 1.9 [41], were generated and used as a confident SNP list for data integration.

Empirical imputation accuracy

There were 160 canola accessions and lines common to both the GBS-t and WGS imputation references. Therefore, we used the 160 canola to evaluate the empirical imputation accuracy of the imputed genotypes from GBS-t to WGS by Minimac3 [17]. We imputed each of the 160 individually by a leave-one-out cross-validation method to maximize the size of the imputation reference. First, we masked the WGS genotypes of the target sample except those that overlapped between the WGS and GBS-t datasets. Then, we used the remaining 277 WGS as the imputation reference to impute the target sample-masked genotypes to WGS by Minimac3 with the default parameters. The GBS-t genotypes were pre-phased by Eagle [42]. The empirical imputation accuracy was measured as the square of the Pearson correlation (r^2) between the imputed and observed genotypes. The concordance rate was calculated as the proportion of the imputed genotypes matching the observed genotypes.

To assess the Minimac3 Rsq quality statistic, we performed an additional validation test by randomly selecting 40 out of the 160 overlapping individuals to form the target set, and the remaining 238 individuals were used as the imputation reference. We repeated the process five times, and each time, we imputed the target set from GBS-t to WGS. The Minimac3 Rsq quality statistic and r^2 for each SNP were averaged across 5 replicates. We grouped the Minimac3 Rsq into different bins and calculated the mean r^2 in each Rsq bin to observe the relationship between those two measurements.

The 24 DH lines that overlapped between WGS and the skim-WGS dataset were used as the target set to evaluate the empirical imputation accuracy from low-coverage skim-WGS to full coverage (10×). The skim coverage level of the DH lines varied; therefore, we tested a range of skim coverage levels ranging from 0.2×, 0.5×, 1×, 2×, to 5× by down sampling the 10× coverage in silico with SAMtools [43]. We randomly selected 16 out of the 24 individuals, reduced their 10× WGS coverage to the defined coverage, and imputed them back to full coverage using the remaining 262 WGS samples as the imputation reference. This process was repeated five times. Two imputation software were compared, Beagle 4.0 [16] and GLIMPSE [44], and the r^2 and concordance rate were averaged across the five cross-validation runs.

We also used the 24 real skim-WGS sequence genotypes as the target set to calculate the r^2 for each SNP.

This was conducted by a 6-fold cross-validation, which divided the 24 individuals randomly into 6 even groups of four and imputed each group at a time using 274 WGS as an imputation reference. We repeated this cross-validation three times and reported the average r^2 across different minor allele frequency (MAF) bins.

Data integration for downstream studies

The diverse canola collection was imputed from GBS-t to WGS using the Eagle + Minimac3 pipeline, while GLIMPSE was used to impute the 1500 DH lines with skim-WGS to full coverage WGS. After imputation, three datasets were combined according to the identified shared SNPs (~ 1 M SNPs). A total of 1867 canola individuals with 921,734 SNP (~ 921 K) genotypes were retained after removing duplicated individuals and SNPs with poor Minimac3 Rsq values (Rsq ≤ 0.4). All samples information was presented in supplementary Table S1. We calculated the genomic relationship matrix (GRM) with the VanRaden method [45]. The first two principal components (PCs) based on GRM were plotted to review the integrated dataset population structure and the relative relatedness of the imputation reference to the target individuals.

Field trials and phenotypes

The diverse canola collection was evaluated for blackleg resistance in 2015, and the sample details and field design were described by Fikere et al. [36]. Briefly, two field sites were established in Wimmera, Victoria, where all canola were planted in canola stubble from the previous year's crop. We sowed 150 seeds per row in a randomized block design with two replications per location. In 2018, 1200 DH lines and 200 diverse spring accessions were tested for blackleg resistance in two locations with the same experimental design and management as in the 2015 experiment. We repeated the blackleg disease trial in 2021 in one location with a total of 600 lines (diverse accessions and DH lines). Three blackleg traits, according to the Spring Blackleg Management Guide, were collected: emergence count (Eme), which is the percentage of the seed resulting in germinated plants 6 weeks after sowing; survival rate (SurvRt), which is the ratio of the surviving plant count at maturity to the emergence count; and average internal infection (AveInf), which is the percentage of the black-rot-affected area of the total stem cross-section. Only SurvRT and AveInf were used in the downstream analysis of this study.

We estimated the best linear unbiased estimates (BLUEs) for SurvRT and AveInf at each field trial by fitting the line as the fixed effect and fitting spatial adjustments in the error using ASReml [46]:

$$\mathbf{y} = \boldsymbol{X}\mathbf{b} + \boldsymbol{Z}_{\boldsymbol{r}}\mathbf{r} + \boldsymbol{Z}_{\boldsymbol{c}}\mathbf{c} + \mathbf{e} \quad (1)$$

where **y** is a vector of phenotypic records for each individual; **b** is a vector of fixed effects, including the mean, individual, and replication; **r** and **c** are vectors of random field design effects for rows and columns; **e** is a vector of random residuals, distributed as N (0, $\boldsymbol{R}\sigma 2$) with $\boldsymbol{R}= \sum_r (\rho_r) \otimes \sum_c (\rho_c)$, which are the row and column two-dimensional covariance; and $\boldsymbol{X}$, $\boldsymbol{Zr}$, and $\boldsymbol{Zc}$ are the corresponding design matrices for b, r, and c, respectively.

High genetic correlations were observed among the trials (supplementary Table S2). Therefore, we further combined the BLUEs by fitting the location and year combination as fixed effects as in Model 1 but without the row and column effects.

Genomic prediction

After matching the genotypes with the phenotypes, 1724 individuals remained, including 675 diverse canola and 1049 DH lines. The BLUEs of 1724 canola individuals are present in supplementary Table S3. We estimated the narrow-sense heritability (h^2) and model predictive ability (PA) for blackleg resistance traits with 921,734 SNPs (~ 921 K) for the DH lines, diverse canola, and combined population. Additionally, we further removed SNPs with heterozygous genotypes in the DH population, and a total of 10,710 SNPs (~ 10 K) remained. We repeated the h^2 and PA estimation with this low-density SNP set.

The h^2 were estimated using BLUEs in the GBLUP model as follows:

$$\mathbf{y} = \boldsymbol{X}\mathbf{b} + \boldsymbol{Z}\mathbf{g} + \mathbf{e} \quad (2)$$

where **y** is the vector of BLUEs, **b** is the vector of fixed effects, **g** is the vector of additive genetic effects with $g \sim N(0, \boldsymbol{G}\sigma_g)$, and **e** is the residual with $e \sim N(0, \boldsymbol{I}\sigma_e)$. Therefore, σ^2_g and σ^2_e are the genetic and residual variances, respectively. $\boldsymbol{I}$ is an incidence matrix, and $\boldsymbol{G}$ is the GRM to account for the population structure. $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the corresponding design matrices for **b** and **g**, respectively.

The 5-fold cross-validation method was used to evaluate the PA for blackleg resistance with model 2, and the PA, as a measure of model prediction accuracy was defined as 'Pearson's correlation coefficient between genomic estimated breeding values (GEBVs) and BLUEs for the validation set for each blackleg trait. Briefly, the DH lines were randomly divided into five equal subsets, and each subset was, in turn, chosen as the validation set and subsequently predicted by the remaining individuals in either the DH population or the combined population. We applied the same prediction processes to the diverse population, the prediction was repeated five times, and the mean PA and standard deviation (SD) were averaged across all 25 validation sets for each trait.

Genome-wide association study (GWAS)

We initially investigated the LD patterns of 675 diverse canola, 1049 DH lines, and the combined 1724 individuals with 921,734 SNPs (~921 K). The LD between all pairs of SNP markers within 1 Mb of physical distance was calculated using PLINK 1.9 [41], and LD decay was plotted in R [47].

We performed GWAS with three populations (a diverse population with 921,734 SNPs, a DH population with 10,710 SNPs, and a combined population with 10,710 SNPs) for the SurvRT and AveInf traits. The mixed linear model (MLM) in the GCTA [48] can be described as follows:

$$\mathbf{y} = \mathbf{a} + \mathbf{bx} + \mathbf{g} + \mathbf{e} \quad (3)$$

where **y** is the combined BLUE for each canola line, **a** is the mean, **b** is the allele substitution effect of the test SNP, **x** is the SNP genotype coded as 0, 1, or 2, and **g** and **e** are the same as those for the GBLUP model (model 2). In the combined population GWAS, we fitted the population as a fixed effect in the model to account for population differences. In addition, we also conducted a meta-GWAS by using diverse and DH single-population GWASs with ~10 K SNPs to compare the detection power between the combined population GWAS and the meta-GWAS. The meta-GWAS was implemented with the weighted z score method in METAL [49], where the weighted z score is proportional to the square root of the sample size for each of the populations. Manhattan plots were generated to visualize all the GWAS results in R [50]. Quantile–quantile (QQ) plots and the false discovery rate (FDR) [51] were used to determine the SNP significance thresholds. A single QTL region for each trait was defined as the position of the most significant SNP and the 200 kb flanking region on either side.

Results

Empirical imputation accuracy

The empirical accuracy of imputation was assessed either as the squared Pearson's correlation (r2) or the concordance between imputed and observed genotypes. The r^2 values of the WGS genotypes imputed from GBS-t density by Minimac 3 are shown in Fig. 1a. The average r^2 was 0.64, with a broad range from 0.44 to 0.83, indicating that the r^2 varied among imputed individuals. The average concordance was 0.89, with a narrow range between the minimum and maximum concordance (0.83 to 0.92).

The r^2 for each SNP in different Minimac Rsq bins showed a positive trend, although the two values were not directly equivalent (Fig. 1b). The Minimac Rsq values of 0.2, 0.4, and 0.8 corresponded to r^2 values of 0.58, 0.78, and 0.92, respectively. After removing the SNPs with Rsq < 0.2, the r^2 improved from 0.64 to 0.82, and when we set the Rsq acceptable threshold to 0.4, the r^2 increased to 0.87 (supplementary Table S4). This indicated that filtering SNPs with the Minimac Rsq could improve r^2 by discarding poorly imputed data.

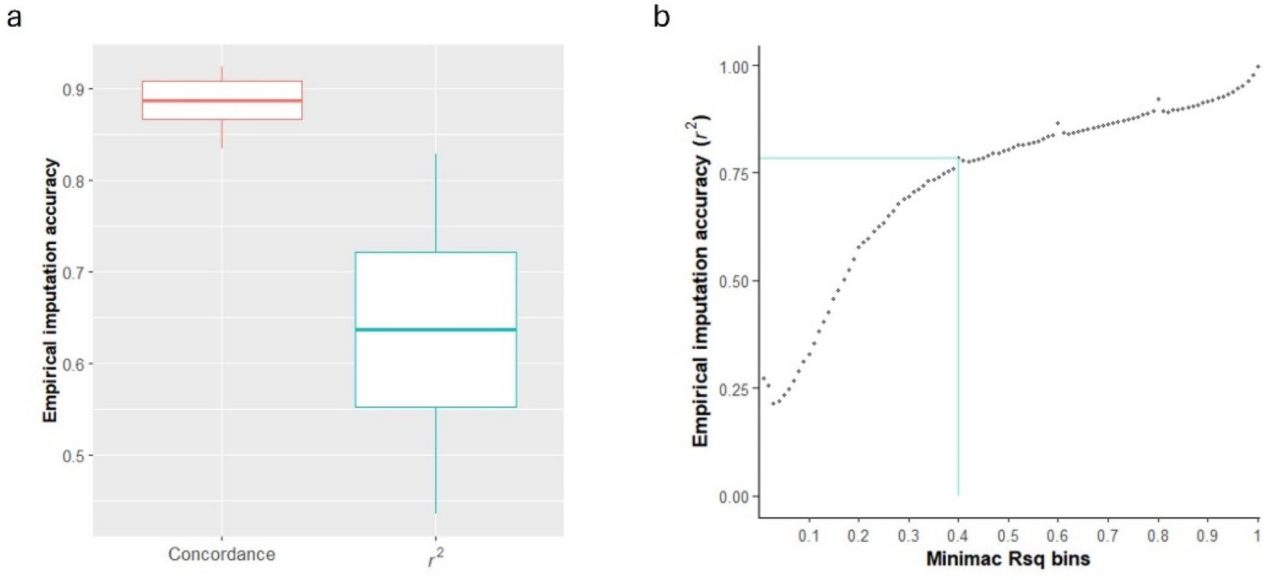

Fig. 1 Boxplot of the empirical imputation accuracy (represented by Concordance and r^2) (**a**) and the mean r^2 values across different Minimac Rsq quality statistic bins (**b**). Both panels are used to visualize the performance of imputing from GBS-t density to WGS using Minimac3. We further compared the r^2 in Beagle and CLIMPSE by imputing from different skim coverages to 10× WGS (Fig. 2a). GLIMPSE outperformed Beagle in all low-coverage datasets, and the greatest differences were at 0.5× and 1.0×, where GLIMPSE (0.88 and 0.9) achieved 16% greater accuracy than Beagle (0.74 and 0.66). Additionally, GLIMPSE also had a slightly greater r^2 for all MAF bins than did Beagle (Fig. 2b). The r^2 was lower for SNPs with low MAFs (MAF < = 0.06) for both methods; however, the concordance did not change dramatically across different MAF bins in either Beagle or GLIMPSE

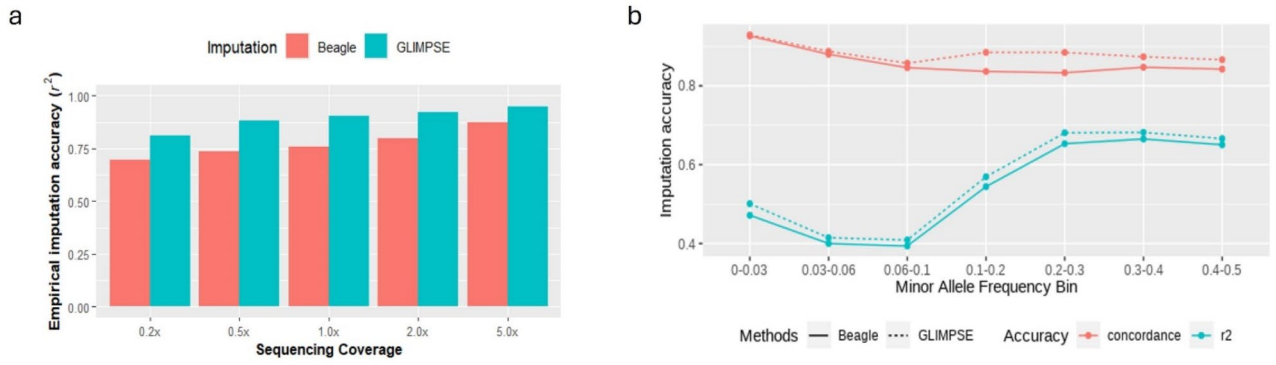

Fig. 2 The average empirical imputation accuracy (r^2) for imputing from different sequencing coverage to 10× WGS using Beagle and GLIMPSE (**a**) and the average r^2 at different MAF bins with the full SNP dataset using Beagle and GLIMPSE (**b**)

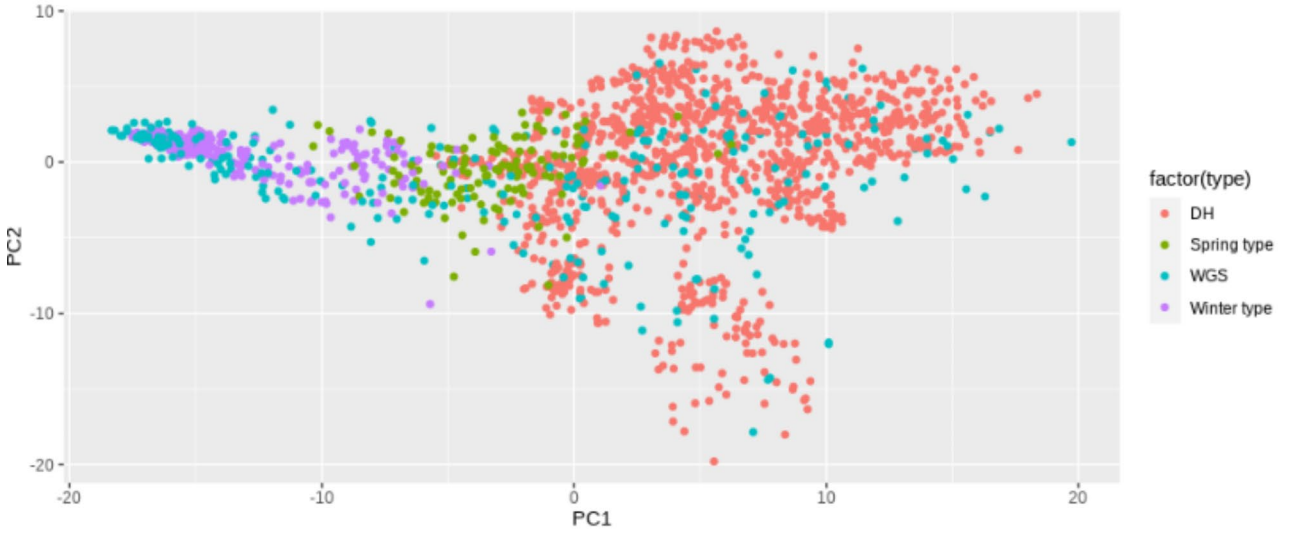

Fig. 3 Dot plots of the first two principal components (PC1, PC2) of the genomic relationship matrix (GRM) with all canola lines included in the study, with color-coded groups

Population structure of the integrated dataset

We integrated three genotype datasets after imputation with 921,734 shared SNPs (~921 K), which were more evenly distributed in the A genome than in the C genome (supplementary Fig S1). The population structure of 1867 individuals was assessed based on PC1 and PC2 of the GRM (Fig. 3). The diverse canola collection was further grouped into winter canola and spring canola according to seasonality, and PC1 clearly differentiated the two subgroups (spring and winter type) and the DH lines. The DH lines overlapped with the spring canola but also showed more diversity according to PC2 than did the spring and winter lines. The WGS imputation reference (the blue dot) distributed across the diverse canola and DH lines indicated that it fully covered the population diversity.

Genomic prediction

After matching canola with genotypes and phenotypes, 1724 individuals remained. The summary statistics showed that the DH lines had a slightly greater mean SurvRT and lower mean AveInf than did the Diverse population (Table 1). The estimated narrow-sense heritability (h^2) by the GBLUP model (model 2) was moderate, ranging from 0.3 to 0.53 for both blackleg traits. In the diverse population, the estimated h^2 using ~921 K SNPs were 0.5 and 0.53 for AveInf and SurvRT, respectively and they were 0.39 and 0.3 in the DH population. The estimated h^2 were lower in the combined population than in the diverse population. In addition to using the ~921 K SNPs for genomic prediction, we also used the ~10 K SNPs, which resulted from removing all the SNPs with heterozygous genotypes in DH lines. The estimated h^2 in the combined population for both traits was greater for ~921 K SNPs than for 10 K SNPs, which was the same trend as that in the diverse population; however, in the

Table 1 Summary statistics, heritability (h^2), and predictive ability (PA) for blackleg traits of average internal infection (AveInf) and survival rate (SurvRt)

Trait	Population	Sample size	summary statistic				No. of SNPs	h^2	SE	Predictive Ability (SD)	
			Min	Max	Mean	SD				DH_PA	Diverse PA
AveInf	Diverse	675	-3.82	88.61	58.02	16.91	921,734	0.5	0.06	-	0.71 (0.0)
	DH	1049	1.623	87.3	52.12	17.28	921,734	0.39	0.05	0.68 (0.0)	-
	combined	1724	-3.82	88.61	54.52	17.37	921,734	0.45	0.04	0.68 (0.0)	0.72 (0.0)
	Diverse						10,710	0.43	0.06	-	0.63 (0.01)
	DH						10,710	0.38	0.05	0.5 (0.01)	-
	combined						10,710	0.33	0.04	0.44 (0.0)	0.2 (0.02)
SurvRT	Diverse	675	12.69	118.7	55.28	19.95	921,734	0.53	0.06	-	0.69 (0.01)
	DH	1049	14.47	114.7	65.13	21.18	921,734	0.3	0.04	0.66 (0.0)	-
	combined	1724	12.69	118.7	61.45	21.29	921,734	0.4	0.04	0.66 (0.0)	0.71 (0.0)
	Diverse						10,710	0.43	0.06	-	0.60 (0.01)
	DH						10,710	0.42	0.05	0.52 (0.01)	-
	combined						10,710	0.35	0.04	0.45 (0.0)	0.3 (0.04)

DH population, the estimated h^2 using 10 K SNP density was approximately 0.38 and 0.42 for AveInf and SurvRT, respectively.

The PA did not follow the trends of estimated h^2. Moderate to high PAs were observed for both blackleg traits, ranging from 0.66 to 0.72 in all populations with ~921 K SNPs. The combined population showed a slightly greater PA when predicting diverse individuals than did the diverse population, while no improvement was observed when predicting DH lines. When using the ~10 K SNP dataset, the PAs to DH individuals decreased to 0.5 and 0.52 for the AveInf and SurvRT traits, respectively, in the DH population but were still greater than the PAs in the combined population (0.44 and 0.45, respectively). The PAs to diverse individuals were around 0.6 for both traits within the diverse population; however, they decreased dramatically to 0.2 (AveInf) and 0.3 (SurvRT) when predicted by the combined population (Table 1).

Genome-wide association study

The LD decay pattern differed among the diverse population, the DH population, and the combined population with ~921 K SNPs (supplementary Fig S2). LD decayed at the slowest rate in the DH lines, indicating that strong LD and long haplotypes existed, and it decayed moderately in the combined population between the LD decay rates in the DH lines and the diverse populations.

The Manhattan plots of GWAS with different populations and methods for both blackleg traits are shown in Fig. 4. According to the QQ plot (supplementary s3, s4) and the varied SNP numbers used in the GWAS, we applied different *P* value thresholds to identify the significant SNPs in each population (Fig. 4; Table 2). We found that most of the significant SNPs were common between SurvRT and AveInf, which was due to the strong phenotypic correlation between the two blackleg traits (-0.93). The SNPs in red are the significant SNPs from all four GWAS. We observed that most of the significant SNPs identified in the diverse population were absent in the DH, combined, and meta-GWASs after SNP number decreased from 921 K to ~10 K. Several SNPs that were not significant in diverse populations showed significance in the DH, combined, or meta-GWAS. Although the number of significant SNPs did not increase dramatically in the combined population, the -log10 (*P* value) of some SNPs increased from 8 to 14 compared to those in the DH population.

A large number of significant SNPs were detected in the diverse population with ~921 K SNPs, and 35 putative QTL regions were identified for AveInf, while 29 QTL regions were identified for SurvRT (Table 2). These putative QTLs were distributed on nearly all 19 chromosomes, except chromosomes A10, C02, C04, and C08 (Fig. 4). The DH population with ~10 K SNPs had fewer significant SNPs for both traits, and they were distributed across 8 chromosomes: A03, A05, A06, C01, C03, C04, C07, and C09 (Fig. 4). There were 17–25 significant SNPs for both blackleg traits in the combined population GWAS and meta-GWAS, most of which overlapped with significant SNPs in the DH population (supplementary Table S5). No significant overlapping SNPs were detected between the diverse and DH populations for either trait; however, a single putative QTL for AveInf was identified on chromosome C04, which was shared among the diverse, DH, and combined populations. The putative QTL regions related to both traits are listed in supplementary Table S5.

Discussion

A larger sample size will benefit genomic prediction and GWAS. In our study, we showed that combining canola individuals genotyped by different sequencing platforms is feasible via genotype imputation. However,

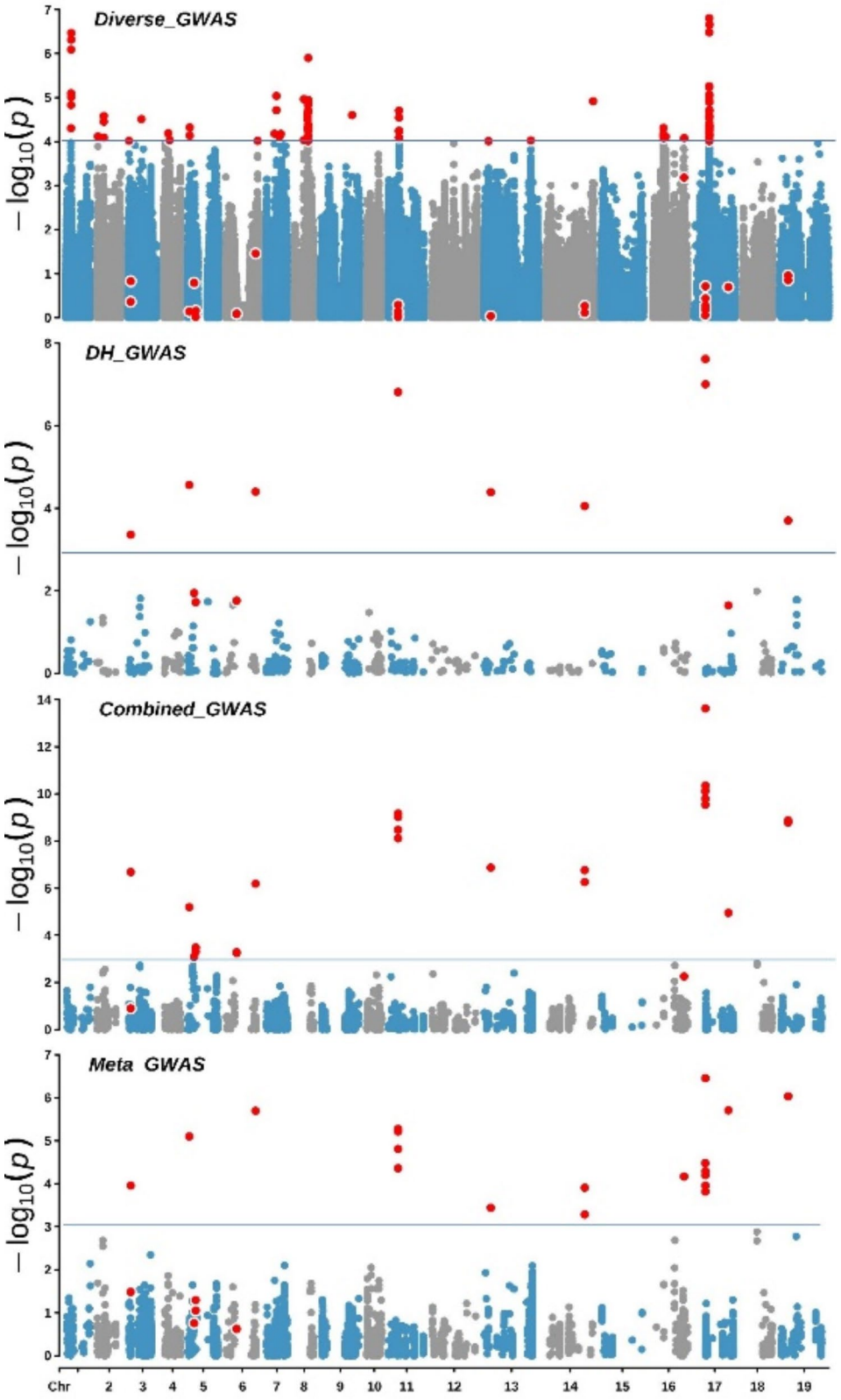

Fig. 4 Manhattan plots of the diverse, DH, and combined population and meta-GWAS for survival rate (SurvRt, left) and average internal infection (AveInf, right). The SNPs shown in red are all significant SNPs from all four GWAS

Table 2 Number of significant SNPs and putative quantitative trait loci (QTLs) detected by GWAS in different populations for two blackleg resistance traits, survival rate (SurvRt) and average internal infection (AveInf)

Trait	Population	Sample	SNP	P_value	No. of Sig_SNP	No. of Putative_QTL
SurvRT	Diverse	675	921,734	$P<10^{-4}$	217	29
	DH	1049	10,710	$P<10^{-3}$	20	8
	combined	1724	10,710	$P<10^{-3}$	25	11
	meta_GWAS		10,710	$P<10^{-3}$	21	10
AveInf	Diverse	675	921,734	$P<10^{-4}$	188	35
	DH	1049	10,710	$P<10^{-3}$	18	8
	combined	1724	10,710	$P<10^{-3}$	22	11
	meta_GWAS		10,710	$P<10^{-3}$	17	9

the combined population with an increased sample size showed limited benefit for GS and GWAS.

Genotype imputation

The r^2 was moderate to high for imputing to full-coverage WGS using Minimac and GLIMPSE, suggesting that imputing from low-density / coverage genotypes to high density / coverage WGS is feasible in canola. The diverse genetic background and relatively lower LD between SNPs could partly explain the lower accuracy observed for Minimac than reported for human or animal studies [52, 53]. A similar imputation accuracy of 0.71 for imputing from the 90 K SNP chip to the exome sequence was reported with different imputation methods in a diverse wheat collection [54]. When the Minimac Rsq was 0.4, the corresponding empirical r^2 was around 0.78 ($r=0.883$), which was very close to the value (0.865) reported for sheep [55]. We found a very strong relationship between the empirical r^2 and the Minimac Rsq. Hence, the Minimac Rsq can be used to set a threshold to filter SNPs that are likely poorly imputed, which is useful when it is not feasible to test the empirical imputation accuracy.

Skim-WGS genotyping is becoming more cost-competitive and is bias-free compared with SNP chips or complexity-reduced GBS. GLIMPSE is designed specifically for the imputation of skim-WGS datasets, and it integrates the haplotype information of both reference and target sets by computing a matrix of genotype likelihoods and updating these likelihoods by iteratively running genotype imputation and haplotype phasing with a Gibbs sampling procedure [44]. Our results confirmed the advantage of GLIMPSE over Beagle for skim-WGS imputation, where the Beagle imputation of missing genotypes was largely based on phased haplotype information in the reference set. Similar methods have been developed to increase the imputation accuracy of multiparent advanced generation intercrossing (MAGIC) populations or biparental crossing populations [56, 57]. In addition, several new imputation methods, such as practical haplotype graph pangenome databases, have been used to impute low-coverage WGS datasets [58]. The whole-genome sequence is the gold standard for genomic study because all the variants underlying the target traits are expected to be present in the genotypes. However, GLIMPSE and Beagle both had lower imputation accuracy at minor allele frequencies < 0.06 for the skim-WGS dataset. Improving the imputation accuracy for SNPs with very low MAFs by developing novel imputation methods requires further investigation.

Genomic prediction

The successful implementation of GS in a breeding program depends on the availability of a large reference population. The diverse gene-bank collections and the advanced breeding lines formed the basis of the reference populations; however, the reference size of such populations is limited [59]. In our study, the combined population included a large number of individuals (1724 canola individuals), which formed the GS training population. However, PA was similar or slightly greater in the combined population than in each of the populations with ~ 921 K SNPs. In a multibreed reference in dairy cattle, a minimal advantage for genomic evaluations of multibreed references over single-breed was indicated [60]. A study of a purebred population combined with crossbred animals also showed that a purebred population had greater prediction accuracy than a combined population [61].

Several factors could impact the multi-population genomic predictive ability in our study. First, the LD between markers and QTLs associated with blackleg resistance may vary across subpopulations, or QTLs could segregate only in a specific subpopulation. We found a large number of putative QTL regions in the diverse population, which was in line with the findings of the previous study [62]. The overlap between the DH lines and the diverse spring canola in the PCA plot indicated that the DH lines were related to the diverse populations. Therefore, the genomic prediction of diverse individuals using the combined population could benefit from the extra genomic information added by the DH population [63]. In a Japanese pear tree study, a high PA was reported when the parent population was combined

with the full-sib breeding population [64]. When we reduced the SNP density from ~921 K to ~10 K, most of the significant signals in the Manhattan plot of the diverse population disappeared, which was an indication of low LD between SNPs and the underlying causal variants. The common significant SNPs observed between the combined population and DH population suggested that the QTLs in the combined population were mainly associated with the DH population. Therefore, the PA in the combined population of diverse individuals decreased significantly. The DH lines in our study showed a very strong LD with ~921 K SNPs. Although the combined population showed faster LD decay than did the DH lines, the predicted effects in the combined population could be dominated by the LD of the DH lines. Consequently, we did not observe an increased PA for DH individuals in the combined population. However, when the SNP density decreased to ~10 K by removing all the heterozygous SNPs in the DH lines, the added heterogeneous diverse population could violate the LD in the combined population, and we detected decreased PA for DH individuals in the combined population.

Second, SNP density needs to be sufficient for multi-population genomic prediction. Studies in cattle suggested that a large number of SNPs are required in multi-bred populations to ensure that the causal mutations are in high linkage disequilibrium with at least one SNP [65]. The SNP density should be sufficient to account for the relationships between all breeds [66]. In our study, we used the shared SNP list instead of the whole genome sequence SNP for data integration. The estimated h^2 and PA for the diverse population with the shared SNP list were consistent with a previous study indicating that the ~921 K SNP set was able to capture the genetic variation within the diverse population [36]. The estimated h^2 and PA of the diverse population decreased when the ~10 K SNP was used, indicating that the SNP density was not sufficient to predict the performance of the diverse lines, although it captured a reasonable amount of genetic variation within the DH population. The combined population showed greater predictive ability with ~921 K SNPs than with ~10 K SNPs for blackleg traits, indicating that ~921 K SNPs were sufficient for genomic prediction in the combined population.

Third, model selection has an impact on the predictive ability for multiple populations. In a previous study on canola blackleg resistance, the GBLUP model performed slightly better than a Bayesian model (BayesR) within populations [36]. However, the simple assumption that all molecular markers shared the same effects in the GBLUP model may not be viable for the combined population. Studies in cattle have shown that the GBLUP model has limited or no benefit when applied to a multibreed population [67, 68]. In our study, the haplotypes observed in the diverse population could be different from those in the DH population through recombination, and the rare variant in the diverse population could have a greater frequency in the DH lines due to crossings and selection. Hence, the genetic heterogeneity of the two subpopulations increased the genomic complexity of the combined population. Lehermeier et al.. tested the heterogeneity of marker effects across naturally diverse populations and plant breeding populations and agreed that models considering population structure and admixture have greater predictive power [69, 70]. However, in our study, without sharing QTLs between the subpopulations, PA improvement by model choice was limited.

Combined population GWAS

We performed GWAS in the combined population with 10 K SNPs to identify the genomic regions associated with canola blackleg resistance. The increase in the significance level of SNPs (i.e., lower *P* values) in the combined population suggested that a larger population size improved the ability to detect some of the marker–trait associations. This is consistent with previous studies, e.g., in dairy cattle, multibreed GWAS could improve the precision of mapping causative variants underlying milk production, although a significant proportion of QTLs are segregated within rather than across breeds [71]. The meta-GWASs shared the most putative QTL regions with the combined population in our study, indicating that meta-GWASs could be an alternative way to combine marker–trait associations of populations, especially where the raw genotypes and phenotypes cannot easily be shared.

The canola diverse population has been previously used for the GWAS of blackleg resistance, and a total of 79 genomic regions were reported to confer potential blackleg resistance [62]. Although the putative QTL regions observed in our study showed different physical positions in the diverse population, it could be due to the alignment of 2020 *Darmor-bzh* (*B. napus*) [39]. The putative QTL regions discovered in DH lines in our study could be particularly valuable, as DH lines containing these QTLs can be directly utilized in breeding programs. For example, the first known *R* gene for blackleg resistance on the C subgenome of *B napus*, *Rlm13*, located within the homoeologous A03/C03 region, was discovered from a mapping population derived from the cross CB-Telfer/ATR-Cobbler [72]. Those two varieties were also the parental lines used to produce the DH lines in this study. We detected two putative QTLs on chromosomes A03 and C03, which differed from the QTLs regions detected by the diverse population. Further annotation will be required to determine if these represent novel QTLs for blackleg resistance. The other six putative QTL regions located on chromosomes A05, A06, C01, C04, C07, and

C09 detected in the DH line population didn't harbor any known R genes [73], however, GWAS with different diverse population have revealed significantly associated QTLs on those chromosomes [74, 75].

Conclusions

Our results confirmed that the Minimac3 quality statistic (Rsq) was useful for filtering out poorly imputed genotypes to improve imputation accuracy. GLIMPSE was the preferred imputation approach compared to Beagle for imputing skim-WGS. We imputed and integrated two genotype datasets sequenced by different platforms. The combined population showed similar or slightly greater predictive ability and greater detection power for the marker–trait associations.

Abbreviations

WGS	Whole-genome sequencing
GBS-t	Genotyping by sequencing via transcriptome
GBS	Genotyping-by-sequencing
DH	Double haploid
Skim-WGS	Low-coverage whole-genome sequencing
MTA	Marker trait association
SNP	Single-nucleotide polymorphism
GP	Genomic prediction
QTLs	Quantitative trait loci
IBD	Identity-by-descent
GS	Genomic selection
GWAS	Genome-wide association study
LD	Linkage disequilibrium
r^2	Imputation accuracy - squared Pearson correlation
Rsq	Minimac3 quality statistic
MAF	Minor allele frequency
GRM	Genomic relationship matrix
PC	Principal component
SurvRt	Survival rate
AveInf	Average internal infection
BLUE	The best linear unbiased estimate
h^2	Narrow-sense heritability
PA	Genomic predictive ability
GEBV	Estimated genomic breeding value
MLM	Mixed linear model
Q–Q plots	Quantile–quantile

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-025-11250-4.

Supplementary Material 1

Supplementary Material 2

Acknowledgements

This article has benefited from the kind support and thorough review of Dr. Majid Khansefid. Many thanks go to the canola project AVR staff of both laboratory and field operations for their hard work.

Author contributions

HZ, SK, and MH conceived and designed the experiment; HZ performed the statistical analysis and wrote the draft; IM assisted with GWAS results interpretation and draft writing; GK and JT assisted with genotype imputation; DB assisted with phenotyping; and SK and MH secured funds and supported the study. All authors revised the manuscript.

Funding

This study was funded by Agriculture Victoria Research, Victoria State Government, Australia.

Data availability

The datasets analyzed during the current study are available in the figshare repository, https://doi.org/10.6084/m9.figshare.25661574.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Agriculture Victoria, AgriBio, Centre for AgriBioscience, Bundoora, VIC 3083, Australia
[2]School of Applied Systems Biology, La Trobe University, Bundoora, VIC 3083, Australia

Received: 22 April 2024 / Accepted: 16 January 2025
Published online: 04 March 2025

References

1. Kim C, Guo H, Kong W, Chandnani R, Shuang L-S, Paterson AH. Application of genotyping by sequencing technology to a variety of crop breeding programs. Plant Sci. 2016;242:14–22.
2. Rasheed A, Hao Y, Xia X, Khan A, Xu Y, Varshney RK, He Z. Crop breeding chips and genotyping platforms: Progress, challenges, and perspectives. Mol Plant. 2017;10(8):1047–64.
3. Kumar P, Choudhary M, Jat BS, Kumar B, Singh V, Kumar V, Singla D, Rakshit S. Skim sequencing: an advanced NGS technology for crop improvement. J Genet 2021, 100.
4. Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. Annu Rev Genomics Hum Genet. 2009;10:387–406.
5. Schmidt M, Kollers S, Maasberg-Prelle A, Großer J, Schinkel B, Tomerius A, Graner A, Korzun V. Prediction of malting quality traits in barley based on genome-wide marker data to assess the potential of genomic selection. Theor Appl Genet. 2016;129(2):203–13.
6. Shi F, Tibbits J, Pasam RK, Kay P, Wong D, Petkowski J, Forrest KL, Hayes BJ, Akhunova A, Davies J, et al. Exome sequence genotype imputation in globally diverse hexaploid wheat accessions. Theor Appl Genet. 2017;130(7):1393–404.
7. Money D, Gardner K, Migicovsky Z, Schwaninger H, Zhong G-Y, Myles S. LinkImpute: fast and accurate genotype imputation for Nonmodel organisms. G3 Genes|Genomes|Genetics. 2015;5(11):2383–90.
8. Zhao H, Li Y, Petkowski J, Kant S, Hayden MJ, Daetwyler HD. Genomic prediction and genomic heritability of grain yield and its related traits in a safflower genebank collection. Plant Genome. 2021;14(1):e20064.
9. He S, Zhao Y, Mette MF, Bothe R, Ebmeyer E, Sharbel TF, Reif JC, Jiang Y. Prospects and limits of marker imputation in quantitative genetic studies in European elite wheat (Triticum aestivum L). BMC Genomics. 2015;16(1):168.
10. Davies RW, Flint J, Myers S, Mott R. Rapid genotype imputation from sequence without reference panels. Nat Genet. 2016;48(8):965–9.
11. Sargolzaei M, Chesnais JP, Schenkel FS. A new approach for efficient genotype imputation using information from relatives. BMC Genomics. 2014;15(1):478.
12. Brandariz SP, González Reymúndez A, Lado B, Malosetti M, Garcia AAF, Quincke M, von Zitzewitz J, Castro M, Matus I, del Pozo A, et al. Ascertainment bias from imputation methods evaluation in wheat. BMC Genomics. 2016;17(1):773.
13. Liu C-T, Deng X, Fisher V, Heard-Costa N, Xu H, Zhou Y, Vasan RS, Cupples LA. Revisit Population-based and family-based Genotype Imputation. Sci Rep. 2019;9(1):1800.

14. Gonen S, Wimmer V, Gaynor RC, Byrne E, Gorjanc G, Hickey JM. Phasing and imputation of single nucleotide polymorphism data of missing parents of biparental plant populations. Crop Sci. 2021;61(4):2243–53.
15. De Marino A, Mahmoud AA, Bose M, Bircan KO, Terpolovsky A, Bamunusinghe V, Bohn S, Khan U, Novković B, Yazdi PG. A comparative analysis of current phasing and imputation software. PLoS ONE. 2022;17(10):e0260177.
16. Browning BL, Browning SR. Genotype imputation with millions of reference samples. Am J Hum Genet. 2016;98(1):116–26.
17. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, Vrieze SI, Chew EY, Levy S, McGue M, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016;48(10):1284–7.
18. Torkamaneh D, Belzile F. Scanning and Filling: ultra-dense SNP genotyping combining Genotyping-By-Sequencing, SNP array and whole-genome Resequencing Data. PLoS ONE. 2015;10(7):e0131533.
19. Treccani M, Locatelli E, Patuzzo C, Malerba G. A broad overview of genotype imputation: standard guidelines, approaches, and future investigations in genomic association studies. BIOCELL. 2023;47(6):1225–41.
20. Calus MP, Bouwman AC, Hickey JM, Veerkamp RF, Mulder HA. Evaluation of measures of correctness of genotype imputation in the context of genomic prediction: a review of livestock applications. Animal. 2014;8(11):1743–53.
21. Sakhale SA, Yadav S, Clark LV, Lipka AE, Kumar A, Sacks EJ. Genome-wide association analysis for emergence of deeply sown rice (Oryza sativa) reveals novel aus-specific phytohormone candidate genes for adaptation to dry-direct seeding in the field. Front Plant Sci 2023, 14.
22. Cericola F, Lenk I, Fè D, Byrne S, Jensen CS, Pedersen MG, Asp T, Jensen J, Janss L. Optimized use of low-depth genotyping-by-sequencing for genomic prediction among Multi-parental Family pools and single plants in Perennial Ryegrass (Lolium perenne L). Front Plant Sci 2018, 9.
23. Mathur R, Fang F, Gaddis N, Hancock DB, Cho MH, Hokanson JE, Bierut LJ, Lutz SM, Young K, Smith AV, et al. GAWMerge expands GWAS sample size and diversity by combining array-based genotyping and whole-genome sequencing. Commun Biol. 2022;5(1):806.
24. Lund MS, Su G, Janss L, Guldbrandtsen B, Brøndum RF. Genomic evaluation of cattle in a multi-breed context. Livest Sci. 2014;166:101–10.
25. Torkamaneh D, Boyle B, Belzile F. Efficient genome-wide genotyping strategies and data integration in crop plants. Theor Appl Genet. 2018;131(3):499–511.
26. Wang DR, Agosto-Pérez FJ, Chebotarov D, Shi Y, Marchini J, Fitzgerald M, McNally KL, Alexandrov N, McCouch SR. An imputation platform to enhance integration of rice genetic resources. Nat Commun. 2018;9(1):3519.
27. USDA. USDA_FAS. In.; 2023.
28. West JS, Kharbanda PD, Barbetti MJ, Fitt BDL. Epidemiology and management of Leptosphaeria maculans (phoma stem canker) on oilseed rape in Australia, Canada and Europe. Plant Pathol. 2001;50(1):10–27.
29. Salisbury P, Ballinger D, Wratten N, Plummer K, Howlett B. Blackleg disease on oilseed < I > Brassica in Australia: a review. Aust J Exp Agric. 1995;35(5):665–72.
30. Varshney RK, Bohra A, Yu J, Graner A, Zhang Q, Sorrells ME. Designing Future crops: Genomics-assisted breeding comes of age. Trends Plant Sci. 2021;26(6):631–49.
31. Uffelmann E, Huang QQ, Munung NS, de Vries J, Okada Y, Martin AR, Martin HC, Lappalainen T, Posthuma D. Genome-wide association studies. Nat Reviews Methods Primers. 2021;1(1):59.
32. Gizachew Haile G. Genomic Selection: A Faster Strategy for Plant Breeding. In: *Case Studies of Breeding Strategies in Major Plant Species.* Edited by Haiping W. Rijeka: IntechOpen; 2022: Ch. 2.
33. Roy J, del Río Mendoza LE, Bandillo N, McClean PE, Rahman M. Genetic mapping and genomic prediction of sclerotinia stem rot resistance to rapeseed/canola (Brassica napus L.) at seedling stage. Theor Appl Genet. 2022;135(6):2167–84.
34. Mansouripour S, Oladzad A, Shahoveisi F, Rahman MM, del Río Mendoza LE, Mamidi S, Moghaddam SM. Identification of genomic regions associated with resistance to blackleg (Leptosphaeria Maculans) in canola using genome wide association study. Eur J Plant Pathol. 2021;161(3):693–707.
35. Raman H, Raman R, Coombes N, Song J, Diffey S, Kilian A, Lindbeck K, Barbulescu DM, Batley J, Edwards D et al. Genome-Wide Association Study Identifies New Loci for Resistance to Leptosphaeria maculans in Canola. Front Plant Sci 2016, 7.
36. Fikere M, Barbulescu DM, Malmberg MM, Shi F, Koh JCO, Slater AT, MacLeod IM, Bowman PJ, Salisbury PA, Spangenberg GC, et al. Genomic prediction using prior quantitative trait loci information reveals a large Reservoir of Underutilised Blackleg Resistance in Diverse Canola (Brassica napus L.) lines. Plant Genome. 2018;11(2):170100.
37. Malmberg MM, Shi F, Spangenberg GC, Daetwyler HD, Cogan NOI. Diversity and Genome Analysis of Australian and global oilseed Brassica napus L. Germplasm using transcriptomics and whole genome re-sequencing. Front Plant Sci 2018, 9.
38. Malmberg MM, Barbulescu DM, Drayton MC, Shinozuka M, Thakur P, Ogaji YO, Spangenberg GC, Daetwyler HD, Cogan NOI. Evaluation and recommendations for routine genotyping using Skim Whole Genome re-sequencing in Canola. Front Plant Sci 2018, 9.
39. Rousseau-Gueutin M, Belser C, Da Silva C, Richard G, Istace B, Cruaud C, Falentin C, Boideau F, Boutte J, Delourme R et al. Long-read assembly of the Brassica napus reference genome Darmor-Bzh. Gigascience 2020, 9(12).
40. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv: Genomics* 2013.
41. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75.
42. Loh P-R, Danecek P, Palamara PF, Fuchsberger C, Reshef A, K Finucane Y, Schoenherr H, Forer S, McCarthy L, Abecasis S. Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet. 2016;48(11):1443–8.
43. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup GPDP. The sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
44. Rubinacci S, Ribeiro DM, Hofmeister RJ, Delaneau O. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. Nat Genet. 2021;53(1):120–6.
45. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91(11):4414–23.
46. Gilmour AR, Gogel BJ, Cullis BR, Welham SJ, Thompson R. ASReml User Guide Release 4.1 Functional Specification. In. VSN International Ltd, Hemel Hempstead, HP1 1ES, UK; 2015.
47. Zhang C, Dong SS, Xu JY, He WM, Yang TL. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. Bioinformatics. 2019;35(10):1786–8.
48. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88(1):76–82.
49. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinf (Oxford England). 2010;26(17):2190–1.
50. Turner SD. Qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. J Open Source Softw 2018, 3(25).
51. Benjamini Y, Hochberg Y. Controlling the false Discovery rate: a practical and powerful Approach to multiple testing. J Roy Stat Soc: Ser B (Methodol). 1995;57(1):289–300.
52. Fernandes Júnior GA, Carvalheiro R, de Oliveira HN, Sargolzaei M, Costilla R, Ventura RV, Fonseca LFS, Neves HHR, Hayes BJ, de Albuquerque LG. Imputation accuracy to whole-genome sequence in Nellore cattle. Genet Selection Evol. 2021;53(1):27.
53. Shi S, Yuan N, Yang M, Du Z, Wang J, Sheng X, Wu J, Xiao J. Comprehensive Assessment of Genotype Imputation Performance. Hum Hered. 2018;83(3):107–16.
54. Hickey JM, Crossa J, Babu R, de los Campos G. Factors affecting the Accuracy of Genotype Imputation in populations from several maize breeding programs. Crop Sci. 2012;52(2):654–63.
55. Bolormaa S, Chamberlain AJ, Khansefid M, Stothard P, Swan AA, Mason B, Prowse-Wilkins CP, Duijvesteijn N, Moghaddar N, van der Werf JH, et al. Accuracy of imputation to whole-genome sequence in sheep. Genet Selection Evol. 2019;51(1):1.
56. Pook T, Mayer M, Geibel J, Weigend S, Cavero D, Schoen CC, Simianer H. Improving Imputation Quality in BEAGLE for Crop and Livestock Data. *G3 Genes|Genomes|Genetics* 2020, 10(1):177–188.
57. Huang BE, Raghavan C, Mauleon R, Broman KW, Leung H. Efficient imputation of missing markers in Low-Coverage genotyping-by-sequencing data from Multiparental crosses. Genetics. 2014;197(1):401–4.
58. Bradbury PJ, Casstevens T, Jensen SE, Johnson LC, Miller ZR, Monier B, Romay MC, Song B, Buckler ES. The practical haplotype graph, a platform for storing and using pangenomes for imputation. Bioinformatics. 2022;38(15):3698–702.
59. Tibbs Cortes L, Zhang Z, Yu J. Status and prospects of genome-wide association studies in plants. Plant Genome. 2021;14(1):e20077.
60. Pryce JE, Gredler B, Bolormaa S, Bowman PJ, Egger-Danner C, Fuerst C, Emmerling R, Sölkner J, Goddard ME, Hayes BJ. Short communication:

genomic selection using a multi-breed, across-country reference population. J Dairy Sci. 2011;94(5):2625–30.
61. Karaman E, Su G, Croue I, Lund MS. Genomic prediction using a reference population of multiple pure breeds and admixed individuals. Genet Sel Evol. 2021;53(1):46.
62. Fikere M, Barbulescu DM, Malmberg MM, Spangenberg GC, Cogan NOI, Daetwyler HD. Meta-analysis of GWAS in Canola Blackleg (Leptosphaeria Maculans) disease traits demonstrates increased power from imputed whole-genome sequence. Sci Rep. 2020;10(1):14300.
63. Edwards SM, Buntjer JB, Jackson R, Bentley AR, Lage J, Byrne E, Burt C, Jack P, Berry S, Flatman E, et al. The effects of training population design on genomic prediction accuracy in wheat. Theor Appl Genet. 2019;132(7):1943–52.
64. Minamikawa MF, Takada N, Terakami S, Saito T, Onogi A, Kajiya-Kanegae H, Hayashi T, Yamamoto T, Iwata H. Genome-wide association study and genomic prediction using parental and breeding populations of Japanese pear (Pyrus pyrifolia Nakai). Sci Rep. 2018;8(1):11994.
65. Hayes BJ, Corbet NJ, Allen JM, Laing AR, Fordyce G, Lyons R, McGowan MR, Burns BM. Towards multi-breed genomic evaluations for female fertility of tropical beef cattle1. J Anim Sci. 2018;97(1):55–62.
66. Steyn Y, Lourenco DAL, Misztal I. Genomic predictions in purebreds with a multibreed genomic relationship matrix1. J Anim Sci. 2019;97(11):4418–27.
67. van den Berg I, MacLeod IM, Reich CM, Breen EJ, Pryce JE. Optimizing genomic prediction for Australian red dairy cattle. J Dairy Sci. 2020;103(7):6276–98.
68. Khansefid M, Goddard ME, Haile-Mariam M, Konstantinov KV, Schrooten C, de Jong G, Jewell EG, O'Connor E, Pryce JE, Daetwyler HD et al. Improving genomic prediction of crossbred and purebred dairy cattle. Front Genet 2020, 11.
69. Lehermeier C, Schön C-C, de los Campos G. Assessment of genetic heterogeneity in structured plant populations using Multivariate whole-genome regression models. Genetics. 2015;201(1):323–37.
70. Misztal I, Steyn Y, Lourenc DAL. Genomic evaluation with multibreed and crossbred data. JDS Commun. 2022;3:156–9.
71. Raven L-A, Cocks BG, Hayes BJ. Multibreed genome wide association can improve precision of mapping causative variants underlying milk production in dairy cattle. BMC Genomics vol. 2014;15:62.
72. Raman H, Raman R, Qiu Y, Zhang Y, Batley J, Liu S. The Rlm13 gene, a New Player of Brassica napus–Leptosphaeria maculans Interaction maps on chromosome C03 in Canola. Front Plant Sci 2021, 12.
73. Vasquez-Teuber P, Rouxel T, Mason AS, Soyer JL. Breeding and management of major resistance genes to stem canker/blackleg in Brassica crops. Theor Appl Genet. 2024;137(8):192.
74. Kumar V, Paillard S, Fopa-Fomeju B, Falentin C, Deniot G, Baron C, Vallée P, Manzanares-Dauleux MJ, Delourme R. Multi-year linkage and association mapping confirm the high number of genomic regions involved in oilseed rape quantitative resistance to blackleg. Theor Appl Genet. 2018;131(8):1627–43.
75. Raman H, Raman R, Diffey S, Qiu Y, McVittie B, Barbulescu DM, Salisbury PA, Marcroft S, Delourme R. Stable Quantitative Resistance Loci to Blackleg Disease in Canola (Brassica napus L.) over continents. Front Plant Sci 2018, 9.

Publisher's note