國立臺灣大學生物資源暨農學院農藝學系

碩士論文

Department of Agronomy
College of Bioresources and Agriculture
National Taiwan University

Master Thesis

用於從候選族群中選拔最佳基因型 A-最適與 D-最適訓練集之研究

A-optimal and D-optimal training sets for identifying the best genotypes for a candidate population

宋文修

Wen-Hsiu Sung

指導教授：廖振鐸 博士

Advisor: Chen-Tuo Liao, Ph.D

中華民國 112 年 07 月

July 2023

doi:10.6342/NTU202301253

國立臺灣大學碩士學位論文
口試委員會審定書

MASTER'S THESIS ACCEPTANCE CERTIFICATE
NATIONAL TAIWAN UNIVERSITY

用於從候選族群中選拔最佳基因型 A-最適與 D-最適訓練集之研究

A-optimal and D-optimal training sets for identifying the best genotypes for a candidate population

本論文係 宋文修 (R09621108) 在國立臺灣大學農藝學研究所生物統計組完成之碩士學位論文，於民國 111 年 7 月 21 日承下列考試委員審查通過及口試及格，特此證明。

The undersigned, appointed by the Department of Agronomy_on 21 July 2023 have examined a Master's thesis entitled above presented by Wen-Hsiu Sung (R10621206) candidate and hereby certify that it is worthy of acceptance.

口試委員 Oral examination committee:

中央研究院統計科學研究所
高振宏 研究員 高振宏

國立台灣大學農藝學研究所
蔡欣甫 副教授 蔡欣甫

國立台灣大學農藝學研究所
廖振鐸 教授 (指導教授) 廖振鐸

致謝

首先，我要衷心感謝我的指導教授廖振鐸老師。在我進入研究所之前，他就給了我許多寶貴的人生建議。在整個研究期間，廖老師不辭辛勞地指導我，從我統計學的基礎薄弱到順利完成碩士學業，我都要感謝廖老師一步一步的耐心指導，讓我能順利完成這份畢業論文；我也要感謝農藝系的所有老師，系上的老師真的很盡心盡力，不僅是學業上，在人生規劃方面提供了寶貴的意見和解答。

接著我想特別感謝同實驗室的夥伴們，塗蕙寧、陳思萍跟林寬諺，我向他們尋求許多關於程式和課業的疑問，在彼此的交流中分享了許多寶貴的學習資源。我希望不僅是在現在，未來我們也能繼續互相支持，成為彼此人生道路上的助力。

最後我想感謝朋友跟家人，感謝他們一直以來的支持、鼓勵和理解。感謝大學以及高中時期的朋友們，在假日時一起玩桌遊、爬山、打球，你們的陪伴和理解使我能夠更加努力學習；感謝家人對我的選擇，給予了無限的支持，並且讓我能心無旁騖的專注於學業上。

最後再次向以上所有對我學術和人生道路上給予支持和幫助的人致以最誠摯的謝意。感激之情無法言喻。

doi:10.6342/NTU202301253

中文摘要

隨著分子生物學的進步，基因體選拔 (genomic selection，GS)廣泛用於動物或作物育種計畫中，並成為一項重要的工具。儘管基因型分析 (genotyping)的成本降低，外表型分析 (phenotyping)仍然是要花相對較高的成本以及時間，因此希望透過基因型 (genotype)推測外表型(phenotype)，以此加速育種計畫。基因體選拔透過遍布整個基因體 (genome)的基因標誌 (gene markers)以及已知的連續型性狀外表型，建立統計模型，進而憑藉基因型推測出育種價估計值 (genomic estimated breeding values，GEBVs)，從中選拔出適合的自交系 (inbred lines)或育種計畫中的雜交組合 (hybrids)。

統計模型的建構中，如何只透過基因型資料，選擇適當的個體當作訓練集(training set)進行外表型分析，建構出表現好的預測模型，在基因體選拔是個重要的議題。在本文的研究中，分析兩種方法：A-最適準則 (A-optimality)與 D-最適準則 (D-optimality)兩種判斷方法，原理是試圖挑出最大變異的個體作為適合的訓練集。我們使用四組不同的作物基因資料，分別使用模擬結果與實際資料，並與之前研究的其他方法相比較，兩者相較於隨機訓練集有比較好的表現。

關鍵字：基因體選拔、訓練集選擇、植物育種、基因演算法、混合線性模型

doi:10.6342/NTU202301253

Abstract

Genomic selection (GS) has become a powerful tool in the domains of plant and animal breeding with advanced and cheaper molecular genetic technology. Despite substantial reduction in genotyping costs, phenotyping still remains a time-consuming and expensive process. As a result, phenotype estimation through genotypic information can accelerate the breeding cycle. In GS, markers of the whole genome are used to estimate genomic estimated breeding values (GEBVs) by statistical models, which are built with genotype and phenotype. These GEBVs facilitate the selection of desirable inbred lines or hybrids for further breeding programs.

In the construction of statistical models, selecting appropriate individuals as the training set based on genotype data and building effective prediction models is a crucial topic in genomic selection. In this study, we evaluated two methods: A-optimality and D-optimality, which are criteria aimed at selecting individuals with the highest level of variation. We utilized four different crop genomic datasets and compared the results with previous studies, using both simulated and real data. Both A-optimality and D-optimality demonstrated better performance compared to random training sets.

Keywords: genomic selection; training set selection; plant breeding; genetic algorithm; linear mixed effect model

doi:10.6342/NTU202301253

Contents

doi:10.6342/NTU202301253

doi:10.6342/NTU202301253

List of Figures

doi:10.6342/NTU202301253

doi:10.6342/NTU202301253

List of Tables

doi:10.6342/NTU202301253

Chapter 1 Introduction

Recently, genomic selection (GS) has become a powerful tool in the domains of plant and animal breeding with advanced and cheaper molecular genetic technology. In genomic selection, breeders focus on the quantitative traits, and select the superior breeding lines based on the genomic breeding values (GEBV) instead of traditional phenotypic selection. That is, GS can accelerate the breeding cycle by selecting superior lines before phenotyping. GEBV is estimated by the sum of the effects of dense genetic markers across the whole genome. The dense genetic markers are expected to capture most of the quantitative trait loci (QTL)(Meuwissen et al., 2001). Therefore, relatively small gene effects of QTLs could be included into estimation.

There have been two noteworthy advancements in the field of GS. Firstly, sequencing of the whole genome has led to the identification of DNA markers in the form of single-nucleotide polymorphism (SNP). This has resulted in a substantial reduction in genotyping costs. Secondly, the GS model has been demonstrated to accurately estimate the GEBVs based on the SNP markers (Hayes et al., 2009). There are two common statistical model commonly used in GS, the whole genome regression model and the linear mixed effect model. The former tends to estimate all the marker effects, and then estimate GEBVs. The latter takes the marker effects as random effects, and GEBVs are then estimated by BLUPs (best linear unbiased predictors). The GBLUP model (VanRaden, 2008) has been more commonly used. Besides, some machine-learning and deep-learning algorithms have been utilized in GS. Heslot et al. (2012) conducted a comparison between commonly used methods in the GS and other machine-learning methods including support vector regression, random forests, and neural networks.

doi:10.6342/NTU202301253

González-Camacho et al. (2012) used a neural network method and compared the result with reproducing kernel Hilbert space (RKHS) regression model and a linear regression model.

In spite of lower genotyping costs and progressive estimation methods in GS, phenotyping still remains a time-consuming and expensive process. In a breeding program, there are numerous lines in a germplasm bank or hybrid offspring, and phenotyping every single line is a challenging task. Hence, it becomes crucial to select a good training set based on genotypic information before phenotyping. Some optimality criteria are utilized to select an optimal training set. With the whole genome regression model, Akdemir et al. (2015) used genetic algorithm to minimize the prediction error variance (PEV) for estimating GEBVs based on the ridge regression estimation. Ou and Liao (2019) proposed a criterion which is called r-score. The spirit of r-score method is to find an approximation to the expected value of Pearson's correlation coefficient between GEBVs and phenotypic values. With the GBLUP model, Rincent et al. (2012) conducted a comparison between several optimization criteria. They subsequently used a generalized coefficient of determination (CD)(Laloë, 1993; Laloë et al., 1996) to select an optimal training set.

In our study, we examined two optimality criteria, the A-optimality criterion and the D-optimality criterion, which are based on the variance-covariance matrix of genotypic values. The A-optimality criterion is a relatively intuitive method that does not need an intensive computing algorithm such as genetic algorithm or other exchange algorithms. By contrast, we used genetic algorithm in D-optimality criterion. In summary, our research aimed to offer a comprehensive analysis and comparison of these two different approaches to select an appropriate training set in GS.

doi:10.6342/NTU202301253

Chapter 2 Materials and Methods

2.1 Genome datasets

In this study, four genome datasets were analyzed. The first two datasets were found to be lack of a strong subpopulation structure, while the last two datasets were observed to possess a strong subpopulation structure. A summary of each dataset was presented in table 1.

1. Tropical rice dataset

This dataset, presented in Spindel et al. (2015), consists of 73,147 SNP markers and 363 elite breeding lines belonging to either the indica or indica-admixed group. The dataset includes observations of grain yield (GYD), flowering time (FT), and plant height (PH) measured eight times between 2009 and 2012, with each year having one observation in the dry season and one in the wet season. However, PH data was not available for the wet season of 2009, and 35 phenotypic values were missing out of the 363 individuals. As a result, only the 328 individuals were used in this study.

Because Spindel et al. (2015) suggested that the subset of SNP markers which were efficient enough for GS in this particular collection of rice germplasm, we randomly selected one SNP marker per 0.1-cM interval over each chromosome. The resulting subset included 10,772 out of the 73,147 SNP markers. Each SNP at a given locus was coded as -1, 0, or 1 depending on whether the individual was homozygote of the minor allele, heterozygote, or homozygote of the major allele. When a locus was missing, the imputation method coded it a value of 1 after the SNP coding.

2. Wheat dataset

doi:10.6342/NTU202301253

The dataset, presented in Kristensen et al. (2019), consists of 13,006 SNP markers and 635 F6 winter wheat lines from two breeding cycles. The first cycle included 321 individuals that were harvested in 2014, while the second cycle included 314 individuals that were harvested in 2015. Phenotypic values were recorded on four quality traits: flour yield (FYD), dough tenacity (DT), dough extensibility (DE), and dough strength (DS). For our study, only 313 wheat lines from the second breeding cycle that had complete data on all phenotypic values were utilized.

We filtered out SNPs with a missing rate of less than 0.9 and a minor allele frequency (MAF) of less than 0.05. This resulted in a total of 11,214 SNPs being retained. The SNP coding process was performed using the same approach as described in the tropical rice dataset.

3. Sorghum dataset

This dataset, presented in Fernandes et al. (2018), consists of 56,299 SNP markers and 451 diverse sorghum lines. The dataset also includes best linear unbiased prediction (BLUP) values of plant height (PH), moisture content (MC), and biomass yield (BYD) of each line. BLUP values were estimated to account for variation due to year and spatial effects. Due to the principal component analysis by Fernández-González et al. (2023), the dataset demonstrates a robust subpopulation structure comprising four clusters based on individual classification. This dataset was used to compare methods for training set optimization in genomic selection by Fernández-González et al. (2023).

4. 44k rice dataset

This dataset, presented in Zhao et al. (2011), consists of 44,100 SNP markers and 36 traits of 413 accessions, demonstrating a robust subpopulation structure. We first removed

doi:10.6342/NTU202301253

all SNP markers with a missing rate less than 0.95 and a minor allele frequency (MAF) less than 0.05, resulting in a final set of 34,233 SNP markers. To eliminate redundant markers and calculate genomic relationships between individuals, approximately one-third of these SNP markers were selected (11,043 out of 34,233) evenly distributed over each chromosome. The SNP coding process was performed using the same approach as described in the tropical rice dataset. We further restricted our analysis to a subset of 301 out of 413 accessions that had complete trait data for all three phenotypic values: flowering time in Arkansas (FT-Ark), flowering time in Faridpur (FT-Far), and flowering time in Aberdeen (FT-Abe).

Table 1. The summary of the datasets

Dataset name	Numbers of SNP markers	Numbers of subpopulation	Sizes of candidate set	Phenotypic data
Tropical Rice	10,772	1	328	GYD: grain yield FT: flowering time PH: plant height
Wheat	11,214	1	314	FYD: flour yield DT: dough tenacity DE: dough extensibility DS: dough strength
Sorghum	56,299	4	451	BYD: biomass yield MC: moisture content PH: plant height
44K Rice	11,047	6	301	FT-Ark: flowering time at Arkansas FT-Far: flowering time at Faridpur FT-Abe: flowering time at Aberdeen

doi:10.6342/NTU202301253

2.2 GBLUP model

We used GBLUP model in our study. The GBLUP model can be described as follows:

$$\boldsymbol{y} = \mu\mathbf{1}_n + \boldsymbol{g} + \boldsymbol{e} \quad (1)$$

where $\boldsymbol{y}$ denotes the vector of phenotypic values, μ the general mean, $\mathbf{1}_n$ the unit vector of length n, $\boldsymbol{g}$ the vector of genotypic values for individuals and $\boldsymbol{e}$ the vector of random errors. $\boldsymbol{g}$ and $\boldsymbol{e}$ are assumed to be mutually independent. $\boldsymbol{g}$ is assumed to follow a multivariate normal distribution, denoted by $\boldsymbol{g} \sim \text{MVN}(\mathbf{0}, \boldsymbol{K}\sigma_g^2)$. $\boldsymbol{e}$ is assumed to follow normal distribution denoted by $\boldsymbol{e} \sim \text{MVN}(\mathbf{0}, \boldsymbol{I_n}\sigma_e^2)$, where $\mathbf{0}$ is a zero vector; σ_g^2 is the genetic variance of additive effects, σ_e^2 is the random error variance, $\boldsymbol{I}_n$ is the identity matrix of order n, and $\boldsymbol{K}$ is a genomic relationship matrix for measuring similarity among individuals. Several forms were employed for $\boldsymbol{K}$ in the context of genomic selection (Forni et al., 2011; Rincent et al., 2012; Tsai et al., 2021; Wu et al., 2023).

$\boldsymbol{X}$ is the original marker scores matrix, which is coded as -1, 0 and 1 for homozygote of the minor allele, heterozygote, and homozygote of the major allele. We let $\boldsymbol{M}$ be normalized in each SNP marker and called it standardized marker score matrix. In other words, $m_{ij} = \frac{x_{ij} - \overline{x}_j}{s_j}$, where m_{ij} and x_{ij} are the (ij)th elements of $\boldsymbol{M}$ and $\boldsymbol{X}$, for $i = 1, 2, \ldots, \text{n}$, $j = 1, 2, \ldots, \text{p}$. We consider the genomic relationship matrix as $\boldsymbol{K} = \boldsymbol{M}\boldsymbol{M}^{\boldsymbol{T}}/\text{p}$.

2.3 A-optimality and D-optimality

In this study, we aim to introduce a novel approach for selecting a training set from a large candidate set with less calculation. The proposed method is based on the concept

doi:10.6342/NTU202301253

of the dispersion or diversity of the training set. Specifically, we seek to identify the training set that exhibits the highest level of variation. To achieve this, we present two different strategies: A-optimal and D-optimal designs.

$$\text{A-opt: } \arg\max\left(Tr(\boldsymbol{K_t})\right)$$

$$\text{D-opt: } \operatorname{argmax}\left(\det(\boldsymbol{K_t})\right)$$

where $\boldsymbol{K}_t$ is the subset of the genomic relationship matrix $\boldsymbol{K}$.

The A-optimality criterion focuses on maximizing the variation of the selected training set by maximizing the trace of the genomic relationship matrix. This approach is based on the intuition that the trace of the genomic relationship matrix provides a measure of the total variance present in the selected training set. The trace of the genomic relationship matrix is calculated as the sum of the diagonal elements of the matrix, which represents the total variance of the sample. To maximize the trace, we can rank the candidate samples according to their variance and select the training set that exhibits the highest level of variation.

As for D-optimality criterion, the determinant is employed to maximize the diversity of the selected training set. Researchers have proposed using the determinant of the genomic relationship matrix as a measure of overall variability. In a recent study by Chung and Liao (2020), whom suggested that the determinant of genomic relationship matrix represents the overall variability of the genotypic values. A higher determinant indicates that the subset spans a larger space in high-dimensional vector space, exhibiting greater genomic diversity. However, maximizing the determinant of genomic relationship matrix is a challenging optimization problem. Exhaustively searching all possible subsets is computationally expensive and infeasible, especially for large datasets. Ou and Liao

doi:10.6342/NTU202301253

(2019) presented a genetic algorithm to identify training set that maximizes the r-score. The genetic algorithm involved generating a population of candidate subsets and evaluating their determinant values using the genomic relationship matrix. The subsets were then evolved through successive generations, with the fittest individuals (i.e., those with the highest determinant values) being selected for the next generation. This process was repeated until a satisfactory solution was obtained. Hence, we utilized genetic algorithm implemented in the R package *Trainsel* (Akdemir et al., 2021) to maximize the determinant of a genomic relationship matrix.

As for the datasets with subpopulation structure, we select training set based on the stratified sampling method, selecting optimal training set by the proportion of each subpopulation. In A-optimality criterion, we selected top variance values in each subpopulation separately. In D-optimality criterion, we presented genetic algorithm and let crossover step in the subpopulation themselves instead of the whole candidate set. The R-code was displayed in Appendix 2.

2.4 Training set evaluation

Training set evaluation is an essential component in selecting the best training set among several possible choices. In genomic selection (GS), researchers have traditionally relied on mean squared error (MSE) and Pearson's correlation as measures. However, recent studies suggested that alternative measures may provide better insights into model performance.

doi:10.6342/NTU202301253

In our study, we used two measures: discounted cumulative gain (DCG) (Jarvelin, 2000) and its normalized version (NDCG), which have been widespread in the Information Retrieval (IR) literature for evaluating the effectiveness of search engines. These measures have also been adopted by Blondel et al.(2015) for model evaluation in GS. Plant breeders usually focus on the top k individuals rather than the entire candidate set in GS, since individuals with low breeding values are ignorable in most cases. Therefore, we use DCG and NDCG measures to evaluate the performance of the training set.

In GBLUP model, the BLUPs of $\boldsymbol{g}$ for all candidates by Henderson's mixed model equations (Dempster et al., 1977) can be estimated as:

$$\hat{\boldsymbol{g}} = \boldsymbol{K}_c(\boldsymbol{K}_t)^{-1}\,\hat{\boldsymbol{g}}_t \tag{2}$$

where $\hat{\boldsymbol{g}}$ devotes the BLUPs of $\boldsymbol{g}$, $\boldsymbol{K}_c$ the genomic relationship matrix between the candidate population and the training set, $\boldsymbol{K}_t$ is the submatrix of genomic relationship matrix $\boldsymbol{K}$ corresponding to the training subset, $\hat{\boldsymbol{g}}_t$ the BLUPs for the genomic values by using training set. Given $g_{(1)} \geq g_{(2)} \geq \cdots \geq g_{(n)}$ represent the actual genotypic values arranged in a decreasing order, in addition, the BLUPs corresponding to the true genotypic values, denoted as $\hat{g}_{(1)}$, $\hat{g}_{(2)}$, … $\hat{g}_{(n)}$, are obtained based on the selected training set. By rearranging these BLUPs, it can be deduced that $\hat{g}_{(\pi_1)} \geq \hat{g}_{(\pi_2)} \ldots \geq \hat{g}_{(\pi_n)}$, where $\pi = (\pi_1, \pi_2, \ldots, \pi_n)$ is a permutation of $\pi_0 = (1, 2, \ldots, n)$. Next, the score of discounted cumulative gain (DCG) at position k in the anticipated ranking obtained using the training set was determined as:

$$DCG@k\big(\boldsymbol{g}, \pi(\hat{\boldsymbol{g}})\big) = \sum_{i=1}^{k} f\big(g_{(\pi_i)}\big)d(i) \tag{3}$$

doi:10.6342/NTU202301253

The score of ideal discounted cumulative gain (IDCG) score at position k in the perfect ranking was determined as:

$$IDCG@k(\boldsymbol{g}, \pi(\hat{\boldsymbol{g}})) = DCG@k(\boldsymbol{g}, \pi_0(\boldsymbol{g})) = \sum_{i=1}^{k} f(g_{(i)})d(i)$$

where $f(g) = g$, the discount function of $d(i) = \frac{1}{log_2(i+1)}$.

Next, NDCG score at position k was determined as:

$$DCG@k(\boldsymbol{g}, \pi(\hat{\boldsymbol{g}})) = \sum_{i=1}^{k} f(g_{(\pi_i)})d(i).$$

NDCG is a measure of model performance that is calculated by dividing the DCG score of the predicted ranking by the IDCG score. Basically, it is the ratio between these two scores. The advantage of using NDCG over DCG is that it is simpler to compare because it always falls between 0 and 1. That is, greater values of NDCG indicate superior model performance. Another score of ranking is the mean of NDCG scores from $k = 1$ to $k = K$:

$$mean\_NDCG@K(\boldsymbol{g}, \hat{\boldsymbol{g}}) = \frac{1}{K}\sum_{k=1}^{K} NDCG@k(\boldsymbol{g}, \hat{\boldsymbol{g}})$$

2.5 Scenarios of this study

In order to assess the effectiveness of A-opt and D-opt, our study comprised the following 3 steps. Firstly, we employed simulated data to evaluate the optimality criteria, thereby establishing a baseline for comparison. Second, we used phenotypic values to validate the optimality criteria, ensuring its applicability to real-world data. Third, we performed a comparative analysis with previously proposed optimality criteria. To

doi:10.6342/NTU202301253

construct GBLUP model and got the BLUPs, we mainly utilized the R package '*sommer*'(Covarrubias-Pazaran, 2016) which is based on restricted maximum likelihood estimation (REML).

2.6 Simulation study for validating A-opt and D-opt methods

Firstly, we set different training set sizes for different datasets, which are shown in Table 2. To determine the training set, we employed three distinct strategies: A-opt, D-opt, and random selection. Second, the simulated data was generated in accordance with GBLUP model, which was given as Eq. (1). The marker score matrix $\boldsymbol{X}$ was considered to be known, and the general mean was set at 100, the genetic variance of additive effects σ_g^2 was set at 25. Furthermore, the heritability h^2 was varied at three levels, specifically low, intermediate and high levels, being set as 0.2, 0.5 and 0.8 respectively. The random error variance σ_e^2 can be calculated by the genetic variance and the heritability, which was given as $\sigma_e^2 = \sigma_g^2(1-h^2)/h^2$. Therefore, the phenotypic values could be generated. Next, the BLUPs was computed by generated phenotypic values and marker score matrix. Subsequently, the NDCG would be calculated. We denoted the generated genotypic values as $\boldsymbol{g}$ and its BLUPs as $\hat{\boldsymbol{g}}$ in Eq. (3) with k=1, 5, 10 and the mean of NDCG with k=10. 3000 times iterations would be employed in the simulation study.

Table 2. The training set size for each dataset.

Dataset name	Training set size					
Tropical Rice	50	75	100	150	200	300
Wheat	50	75	100	150	250	
Sorghum	50	75	100	150	200	350
44K Rice	50	75	100	150	250	

doi:10.6342/NTU202301253

2.7 Analysis of phenotypic values

As for the phenotypic values, we used the same training set in simulation study. Nevertheless, BLUPs here were computed by the observed phenotypic values. In addition, we denoted the normalized phenotypic values as $\boldsymbol{g}$ and BLUPs as $\hat{\boldsymbol{g}}$ in Eq. (3) to calculate the NDCG with k=1, 5, 10 and mean of NDCG with k=10.

2.8 Comparison with other optimality criteria

We incorporated three other optimality criteria in our study: r-score (Ou & Liao, 2019), the prediction error variance (PEV) (Akdemir et al., 2015), and the generalized coefficient of determination (CD) (Laloë, 1993). Similarly, the simulation study as mentioned above was employed. We used R package '*TSDFGS*' (Ou & Liao, 2019) to select optimal training sets.

doi:10.6342/NTU202301253

Chapter 3 Results

3.1 The variance pattern of a candidate set in A-optimality

The variances of individuals in each dataset are displayed in Figure 1 in a bar chart form. As for datasets without a strong subpopulation structure, tropical rice and wheat datasets, a clear decrease in variance is observed in both cases. As for datasets with a strong subpopulation structure, sorghum and 44K rice dataset, there is an ambiguous demarcation in the top part of the sorghum dataset, while there is an evident and clear classification based on the subpopulation structure throughout the entire 44K rice dataset. The variances within each subpopulation are found to be similar.

doi:10.6342/NTU202301253

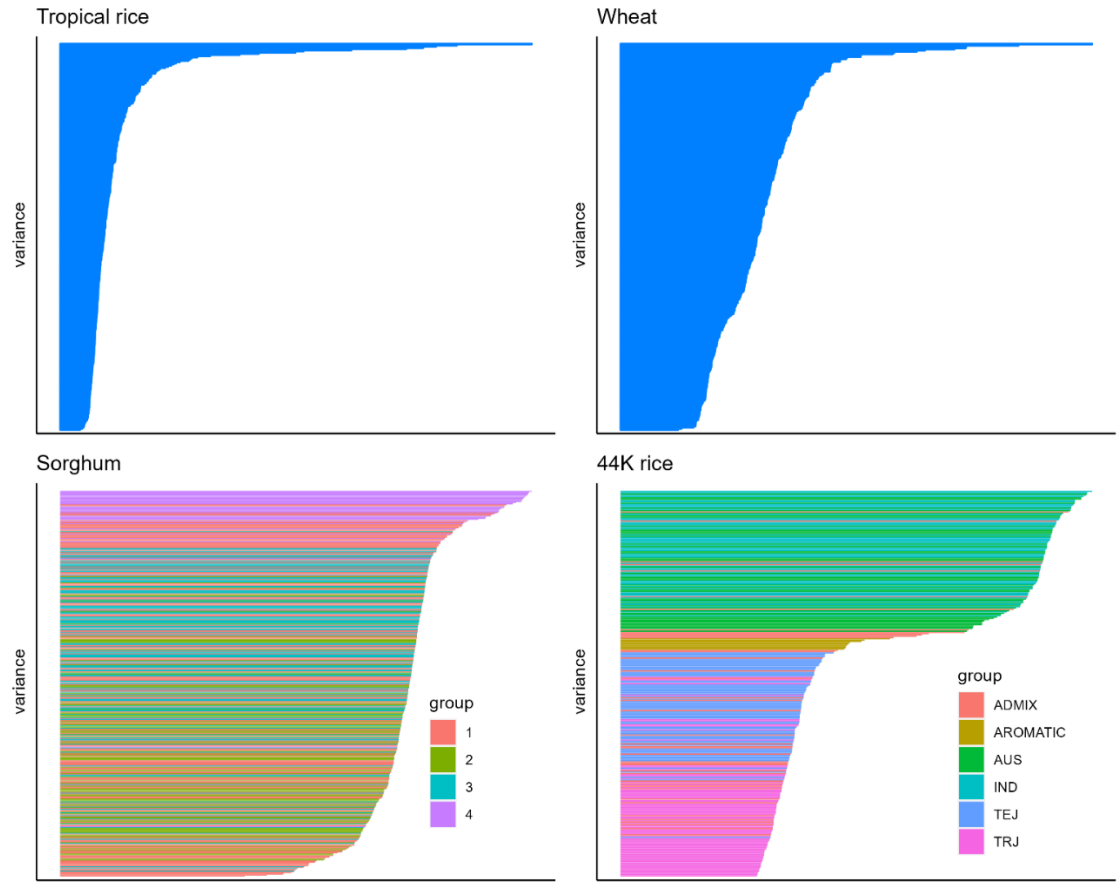

Figure 1 Bar chart representing the variance in A-opt.

The length of each bar presents the variance of each candidate set individual, which is the trace of the normalized genomic relationship matrix. The individuals are ordered in descending order. In the sorghum and 44K rice dataset, the subpopulations are distinguished in different colors.

doi:10.6342/NTU202301253

3.2 Simulation results

The simulation results were displayed in Figure 2-5, some common results could be observed in each simulation test across all four datasets: (1) The trend of NDCG values remains similar across different values of k (k = 1, 5, 10) as well as the mean NDCG values, but the NDCG values and mean NDCG values with k = 10 are seem to be more stable. (2) The NDCG values are significantly higher in the simulations with the high level of heritability compared with those with the low level of heritability. (3) The NDCG values are increasing but the rate of increase slow down as the training set size increases. Probably the estimation methods reach the limitation.

With regards to the optimal training set, the A-opt and D-opt led to higher NDCG values significantly compared to the random training sets across most of scenarios, especially for the dataset without a strong subpopulation. For the tropical rice dataset in Figure 2, A-opt and D-opt perform well in a similar way and reach the performance limitation with lower sample size (100 out of 328) compared to the other training sets. For the wheat dataset in Figure 3, the trends perform similarly, A-opt performed better than D-opt. For the sorghum dataset in Figure 4, A-opt consistently performed better compared to the random training set over all scenarios, however D-opt did not perform in the same way. D-opt only outperformed in the high level of heritability, but did not perform better compared to the random training set in the low and medium levels of heritability. For the 44K rice dataset in Figure 5, both A-opt and D-opt outperformed in the medium and high level of heritability, however D-opt did not perform well in the low level of heritability.

doi:10.6342/NTU202301253

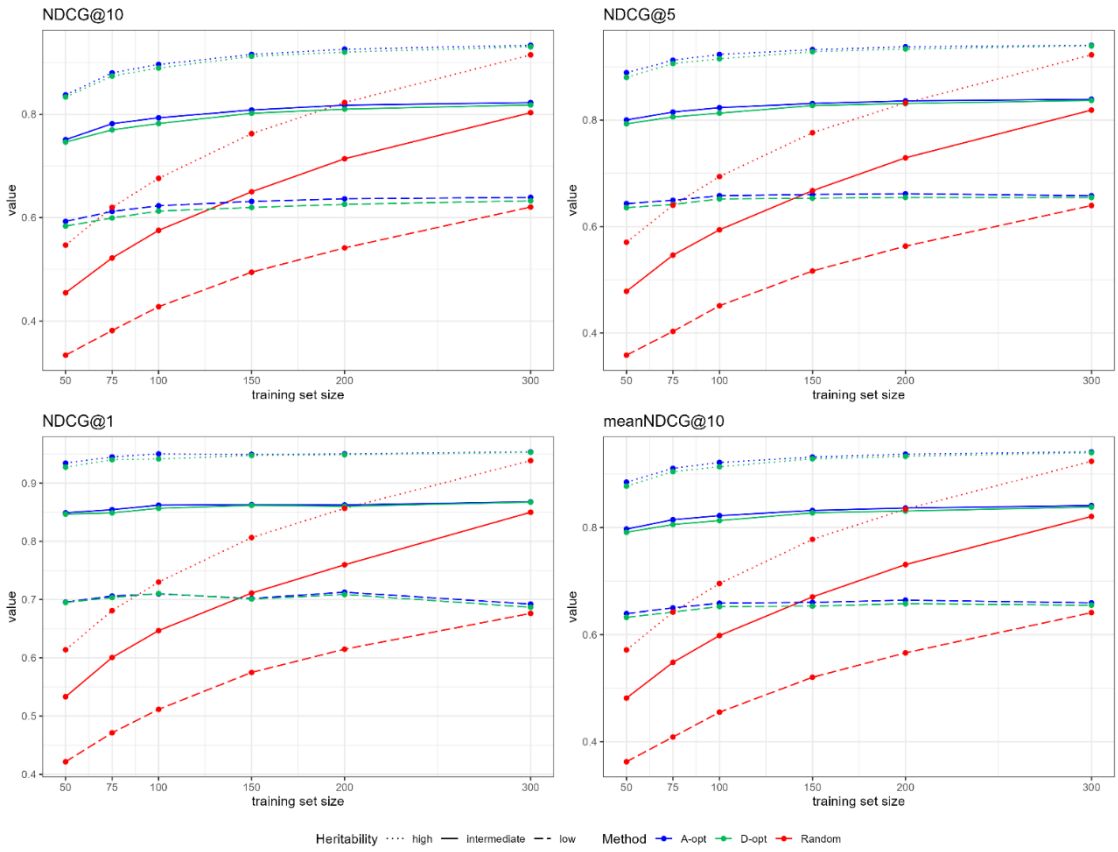

Figure 2. The average NDCG values for the tropical rice dataset across three heritability levels and various values of k.

Horizontal axis represents different training set size. Vertical axis represents the average values of NDCG with various values of k. Types of line represent heritability levels and Colors of line represent A-opt, D-opt and simple random method for selecting training set.

doi:10.6342/NTU202301253

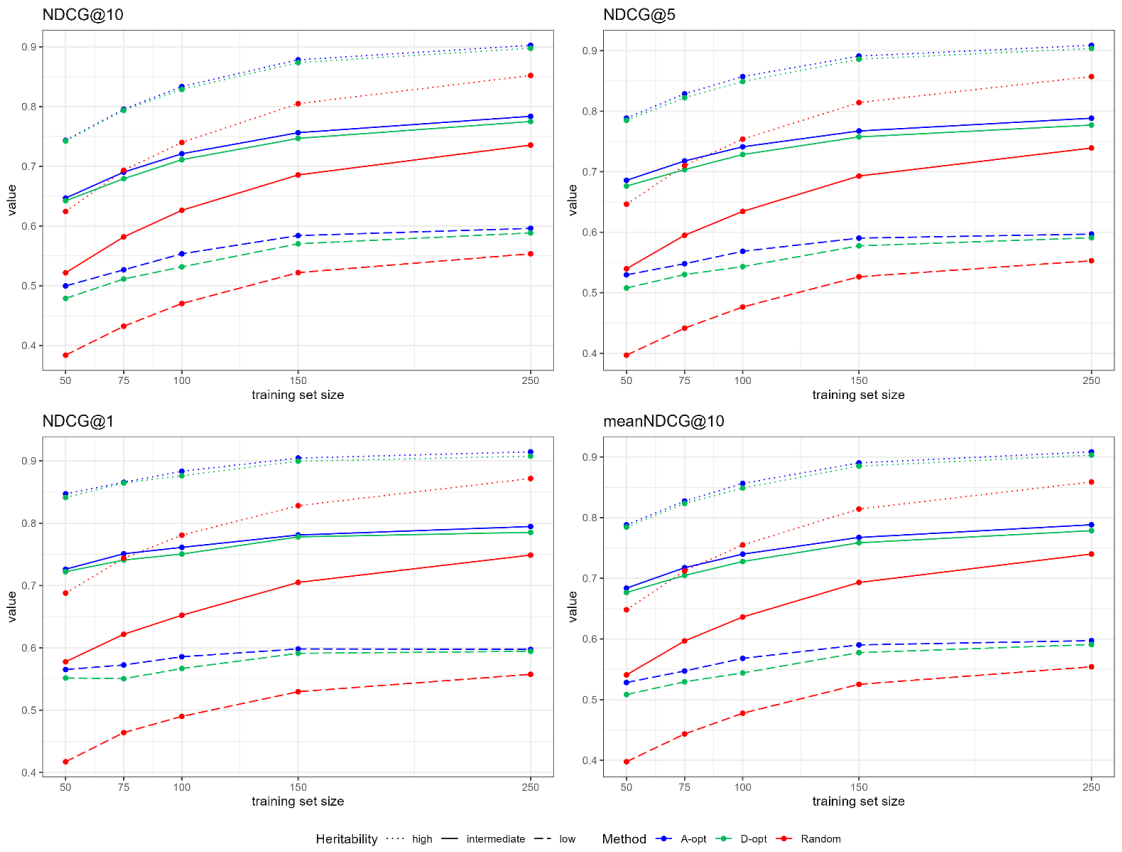

Figure 3. The average NDCG values for the wheat dataset across three heritability levels and various values of k.

Horizontal axis represents different training set size. Vertical axis represents the average values of NDCG with various values of k. Types of line represent heritability levels and Colors of line represent A-opt, D-opt and simple random method for selecting training set.

doi:10.6342/NTU202301253

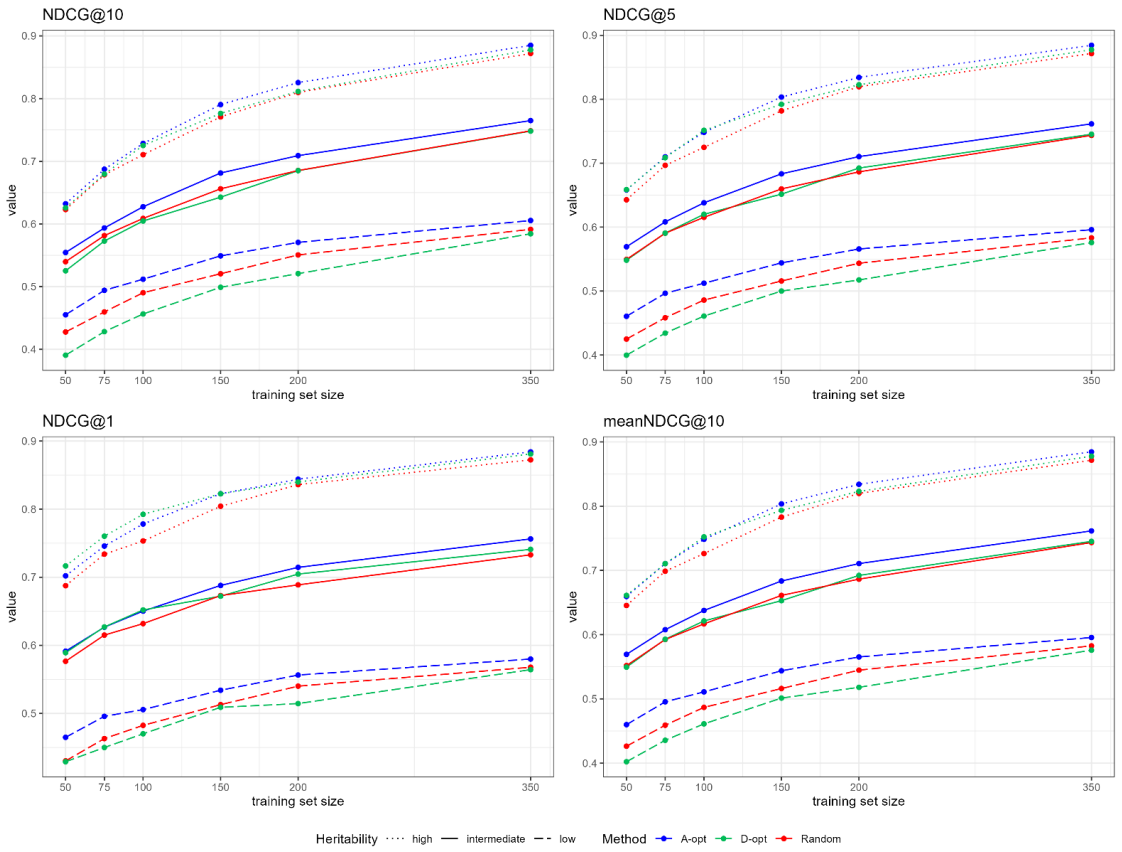

Figure 4. The average NDCG values for the sorghum dataset across three heritability levels and various values of k.

Horizontal axis represents different training set size. Vertical axis represents the average values of NDCG with various values of k. Types of line represent heritability levels and Colors of line represent A-opt, D-opt and stratified random method for selecting training set.

doi:10.6342/NTU202301253

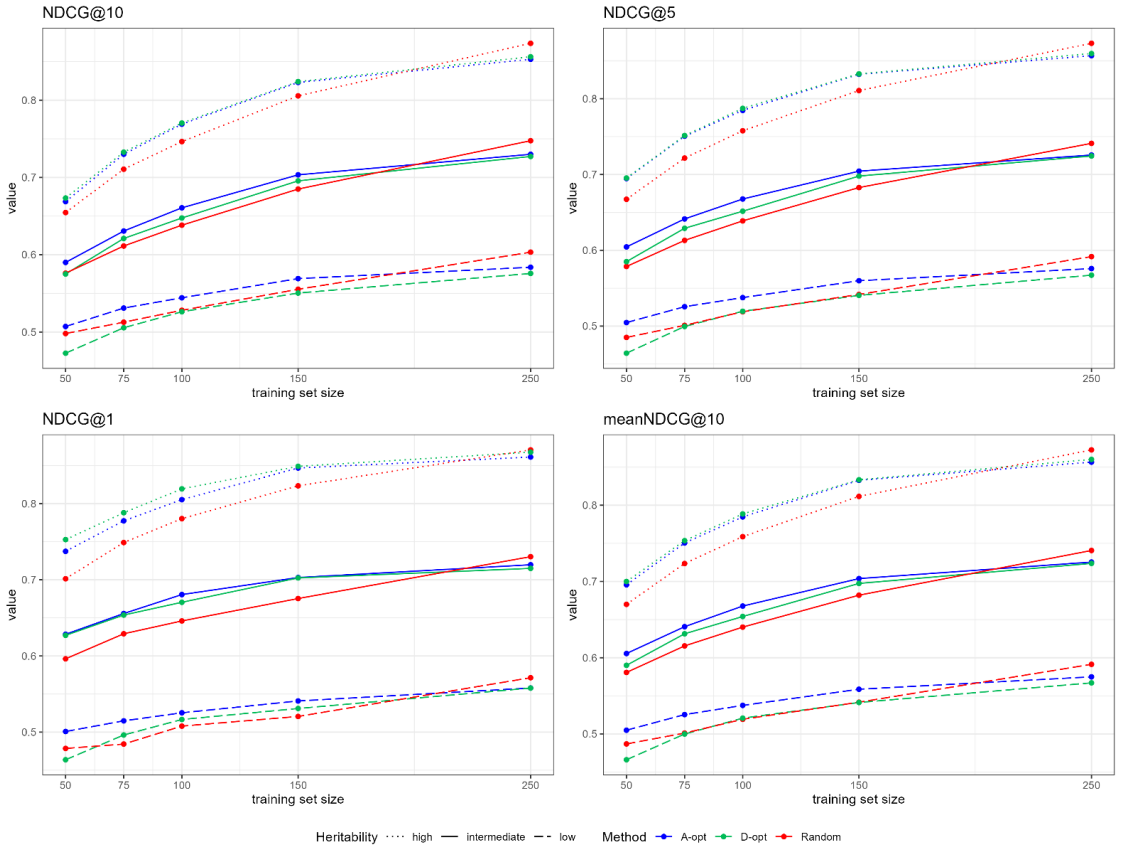

Figure 5. The average NDCG values for the 44K rice dataset across three heritability levels and various values of k.

Horizontal axis represents different training set size. Vertical axis represents the average values of NDCG with various values of k. Types of line represent heritability levels and Colors of line represent A-opt, D-opt and stratified random method for selecting training set.

doi:10.6342/NTU202301253

3.3 Results of phenotypic value analysis

The results of phenotypic value analysis are displayed in Figures 6-9, and only mean_NDCG values are used, since the NDCG values trend is similar across values of k. According to Figures 6-9, the performance is similar to the simulation study as mentioned above. The NDCG values is growing with the increase of sample size, and A-opt and D-opt outperform random training set in most scenarios.

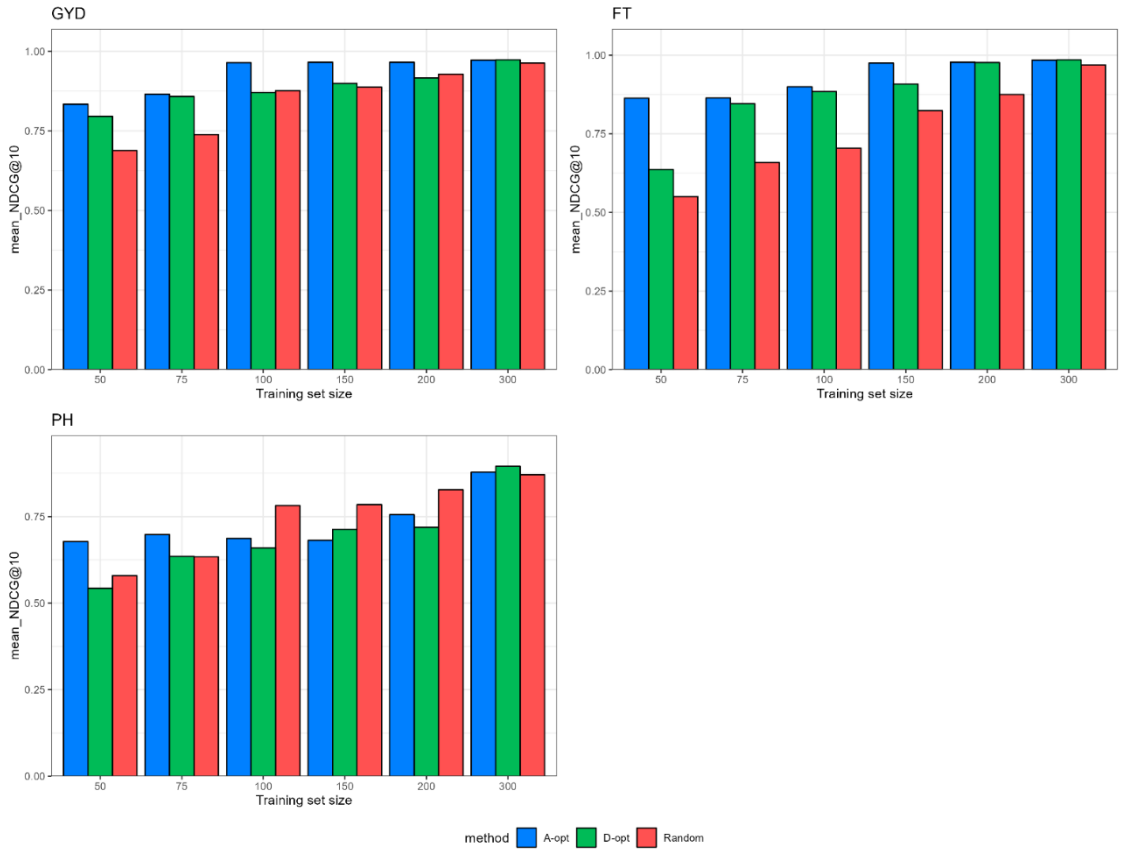

Figure 6. The average mean of NDCGk@10 for the phenotypic data in the tropical rice dataset.

Colors of line represent A-opt, D-opt and simple random method for selecting training set.

doi:10.6342/NTU202301253

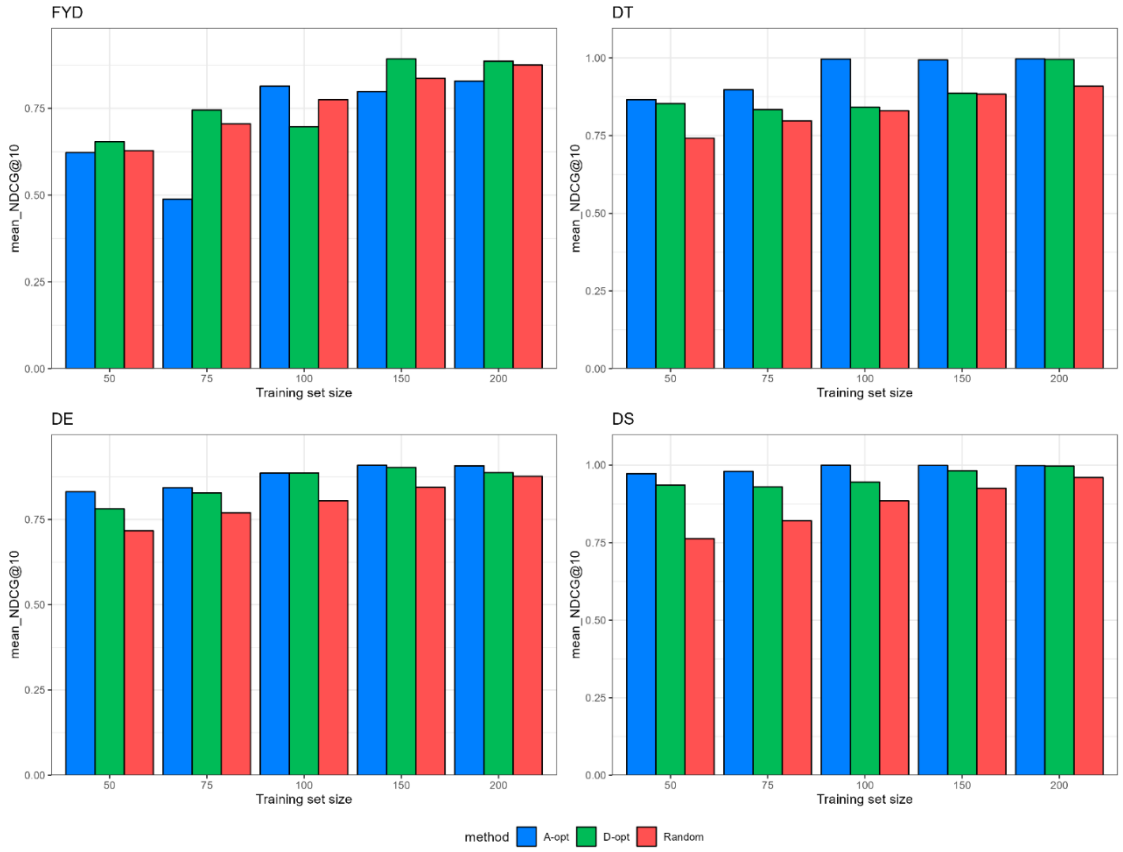

Figure 7. The average mean of NDCGk@10 for the phenotypic data in the wheat dataset.

Colors of line represent A-opt, D-opt and simple random method for selecting training set.

doi:10.6342/NTU202301253

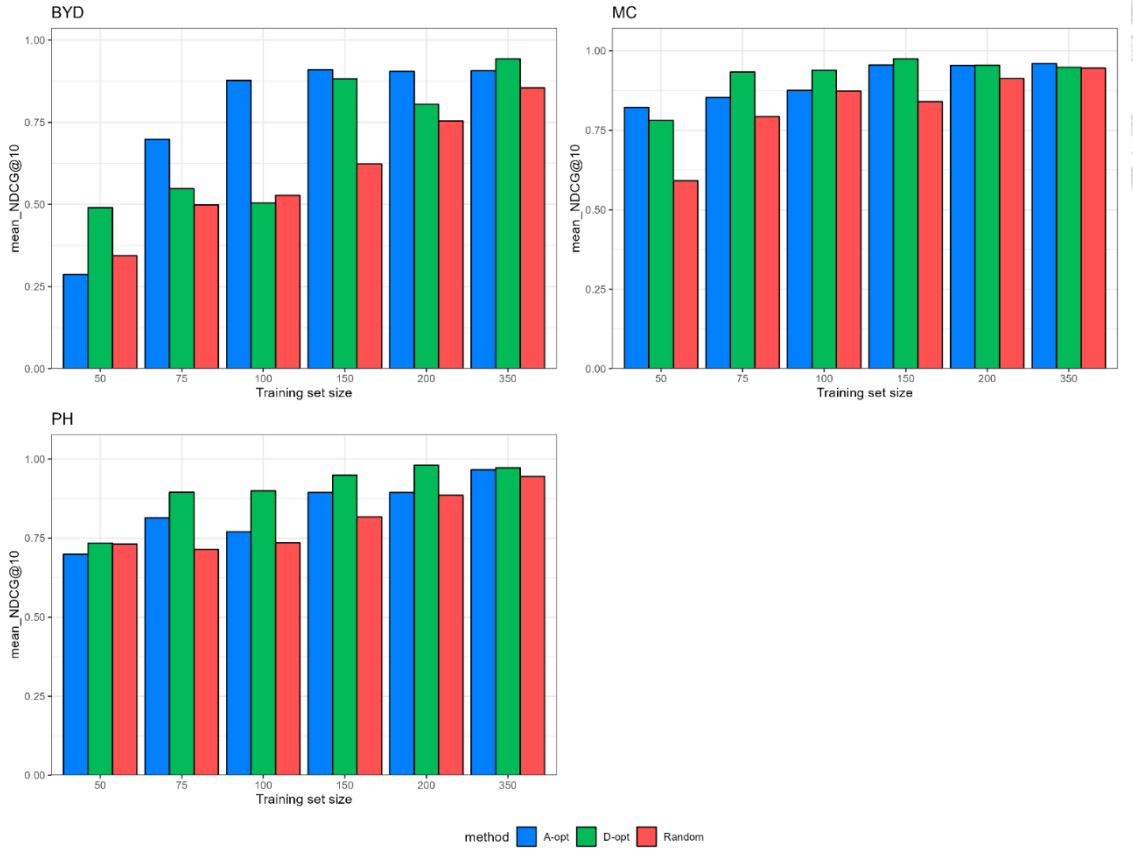

Figure 8. The average mean of NDCGk@10 for the phenotypic data in the sorghum dataset.

Colors of line represent A-opt, D-opt and stratified random method for selecting training set.

doi:10.6342/NTU202301253

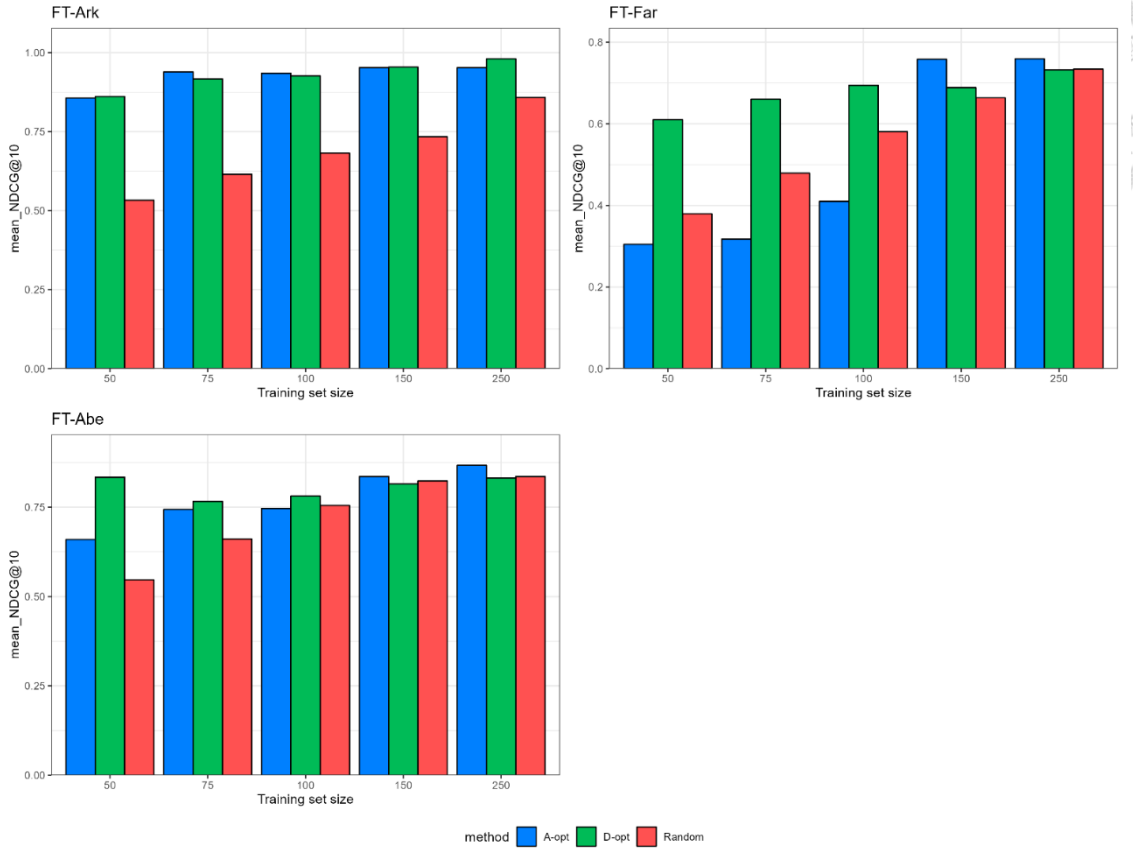

Figure 9. The average mean of NDCGk@10 for the phenotypic data in the 44K rice dataset.

Colors of line represent A-opt, D-opt and stratified random method for selecting training set.

doi:10.6342/NTU202301253

3.4 Comparison results of different training set optimality criteria

The comparison results among different optimality criteria are displayed in Table 3. It can be observed that all optimality criteria outperform the random training set, particularly in the datasets without a strong subpopulation structure such as the tropical rice and wheat datasets. Additionally, A-opt outperforms the other optimality criteria in most of the scenarios.

doi:10.6342/NTU202301253

Table 3. The comparison between optimality criteria with different training set size across each dataset.

The values of this table presented the averaged mean of NDCGk@10 with the simulation study conducted 3000 times and $h^2 = 0.5$. The colored values highlight the best performance for each training set size and dataset combination.

Dataset	Criterion	Training set size					
		50	75	100	150	200	300
Tropical rice	A-opt	0.8005	0.8126	0.8227	0.8325	0.8337	0.8434
	D-opt	0.7810	0.8110	0.8155	0.8281	0.8348	0.8416
	R-score	0.7765	0.7952	0.8049	0.8170	0.8238	0.8426
	PEV	0.7882	0.8087	0.8205	0.8327	0.8329	0.8433
	CD	0.7800	0.7931	0.8057	0.8223	0.8235	0.8426
	Random	0.4873	0.5502	0.5982	0.6751	0.7340	0.8239
		50	75	100	150	250	
Wheat	A-opt	0.6851	0.7198	0.7392	0.7688	0.7882	
	D-opt	0.6766	0.7047	0.7277	0.7586	0.7785	
	R-score	0.6262	0.6760	0.7004	0.7477	0.7713	
	PEV	0.6641	0.6961	0.7141	0.7556	0.7786	
	CD	0.6296	0.6683	0.6996	0.7440	0.7725	
	Random	0.5462	0.5971	0.6401	0.6984	0.7390	
		50	75	100	150	200	350
Sorghum	A-opt	0.5759	0.6134	0.6434	0.6831	0.7130	0.7561
	D-opt	0.5494	0.5928	0.6214	0.6529	0.6922	0.7452
	R-score	0.5877	0.6134	0.6373	0.6754	0.6992	0.7437
	PEV	0.5671	0.5941	0.6264	0.6656	0.7045	0.7507
	CD	0.5867	0.6149	0.6360	0.6734	0.6995	0.7446
	Random	0.5533	0.5903	0.6197	0.6597	0.6874	0.7461
		50	75	100	150	250	
44K rice	A-opt	0.6041	0.6432	0.6666	0.7015	0.7236	
	D-opt	0.5900	0.6314	0.6540	0.6974	0.7237	
	R-score	0.6020	0.6401	0.6616	0.7008	0.7229	
	PEV	0.5896	0.6162	0.6390	0.6839	0.7155	
	CD	0.6066	0.6428	0.6605	0.6997	0.7230	
	Random	0.5847	0.6116	0.6407	0.6853	0.7167	

doi:10.6342/NTU202301253

Chapter 4 Discussion

4.1 Coding of marker score matrix

Consider an equally-spaced setting for the marker scores:

$X_A = a$ for minor homozygote AA;

$X_H = \frac{a+b}{2}$ for heterozygote AB;

$X_B = b$ for major homozygote BB.

The above setting can be easily transformed to be -1, 0, 1 system as follows.

$(X_A - c) \times d = -1$;

$(X_H - c) \times d = 0$;

$(X_B - c) \times d = 1$

where

$$c = \frac{a+b}{2},\ d = \frac{2}{b-a}$$

Therefore, it may be sufficient to use the coding system of -1, 0 and 1.

4.2 Normalization of the marker score matrix

Normalization of the marker score matrix is a crucial step in GS process. Without normalization of each SNP marker, the information from each SNP marker may appear to be equal, disregarding the variation in their contributions to the phenotype. With the proper normalization, the individual markers' contributions can be appropriately weighted.

doi:10.6342/NTU202301253

According to Appendix 1, the normalized marker score matrix can be described as follows. Let *P* denote the frequency of the locus of one single SNP marker, then the normalized marker scores can be obtained as

$$M_A = \frac{-1-\overline{x}}{s} = \frac{-P_H - 2P_B}{s};$$

$$M_H = \frac{0-\overline{x}}{s} = \frac{-P_A + P_B}{s};$$

$$M_B = \frac{1-\overline{x}}{s} = \frac{2P_A + P_H}{s}$$

where

$s = \sqrt{(1-P_H)P_H + 4P_A P_B}$.

$P_A = \frac{n_A}{\mathrm{n}}$ devotes the frequency of homozygote AA in one SNP marker.

$P_B = \frac{n_B}{\mathrm{n}}$ devotes the frequency of homozygote BB in one SNP marker.

$P_H = \frac{n_H}{\mathrm{n}}$ devotes the frequency of heterozygote AB in one SNP marker.

If $n_H = 0$, which is highly homogenous genome, then

$$M_A = \frac{-1-\overline{x}}{s} = \frac{-2P_B}{\sqrt{4P_A P_B}} = \frac{-\sqrt{P_B}}{\sqrt{P_A}}$$

$$M_B = \frac{1-\overline{x}}{s} = \frac{2P_A}{\sqrt{4P_A P_B}} = \frac{\sqrt{P_A}}{\sqrt{P_B}}$$

Consequently, the standardized marker score matrix is based on the frequency of each SNP marker.

4.3 The influence of subpopulation

Figures 4 and 5 showed those results with considering the possible influence of subpopulation structure. However, in the beginning of the breeding program, considering

doi:10.6342/NTU202301253

the subpopulation into GS process still remains a crucial problem. Therefore, we here conducted 3000 times simulations to evaluate the performance regarding whether the subpopulation structure is considered or not.

As for A-opt in Figure 10, considering subpopulation performs significantly better in both the sorghum and 44K rice datasets with a small training set size. The reason can be observed in Figure 1, which shows the clear demarcation in each dataset with subpopulation structure. As a result, when selecting training set using the A-opt criterion without considering subpopulation structure, it tends to select the individuals from only one or two subpopulations. However, with bigger training set size, subpopulation structure seems to be a limit of criteria. Methods without considering population structure performs better for the 44K rice dataset, for training set size above 150.

In contrast, in Figure 10 for the 44K rice dataset, D-opt methods without considering subpopulation structure outperform those with subpopulation structure. This is probably because that we restricted the crossover of individuals in the genetic algorithm, the global optimal training set couldn’t be obtained beyond this restriction.

doi:10.6342/NTU202301253

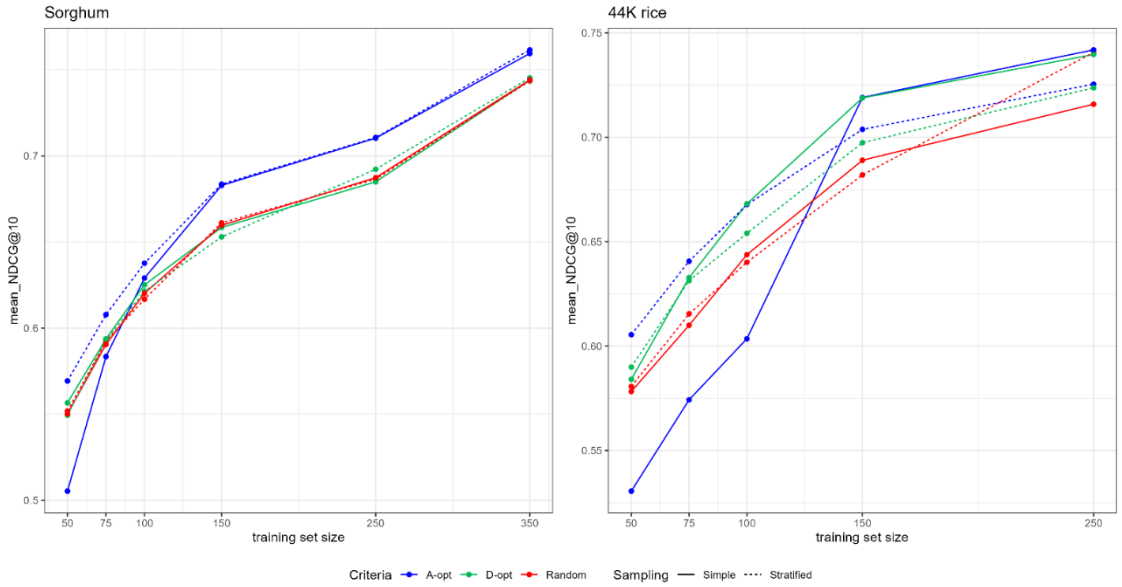

Figure 10. The average mean of NDCGk@10 for the dataset with subpopulation structure, sorghum dataset and 44K rice dataset.

Be simulated 3000 times of generating with h = 0.5. Horizontal axis represents different training set size. Vertical axis represents the average values of mean of NDCGk@10. Types of line represent whether the subpopulations were considered and colors of line represent A-opt, D-opt and random method for selecting training set.

doi:10.6342/NTU202301253

4.4 The influence of heritability in phenotypic analysis

Table 4 represents the heritability of each phenotypic data. For those phenotypic data with lower level of heritability, like FYD in wheat dataset, BYD in sorghum dataset, FT-Far in 44K dataset. In Figures 7,8 and 9, the results of these traits show that the NDCG values doesn't always increase with an increase of the training set size. Besides, the optimality criteria do not outperform the random training set in certain scenarios for a lower heritability trait. That is, in phenotypic analysis, there are various uncontrolled factors, making it impossible to reflect a consistent result with the simulated study. This is especially evident in scenarios in low level of heritability.

Table 4. The heritability of each phenotypic data

Dataset name	Phenotypic data	Heritability
Tropical Rice	GYD: grain yield	0.7586
	FT: flowering time	0.8416
	PH: plant height	0.7518
Wheat	FYD: flour yield	0.6140
	DT: dough tenacity	0.8197
	DE: dough extensibility	0.5982
	DS: dough strength	0.9123
Sorghum	BYD: biomass yield	0.4123
	MC: moisture content	0.6974
	PH: plant height	0.8014
44K Rice	FT-Ark: flowering time at Arkansas	0.7637
	FT-Far: flowering time at Faridpur	0.4096
	FT-Abe: flowering time at Aberdeen	0.5907

doi:10.6342/NTU202301253

4.5 Robustness in different estimation methods

We want to assess the robustness of the optimality criteria by examining their performance using other estimation methods. For this purpose, we conducted a comparative analysis using three more other methods: RKHS regression, random forest, and Ordinal Mcrank, which have been shown to have satisfactory performance in analyzing specific datasets (Blondel et al., 2015). For model construction, we utilized phenotypic values as inputs. and calculated mean of NDCGk@10 to evaluate the performance and compare different models. For RKHS regression, the R package "rrBLUP" (Endelman, 2011) was used. For random forest, the Python "scikit-learn" package (Fabian, 2011) was used. For ordinal Mcrank, the Python source code was used at https://github.com/mblondel/ivalice.

Although A-opt and D-opt is based on the GBLUP model, we can observe roughly that in Figures 11-14, the same optimality criterion exhibits a similar trend with the GBLUP model for the other three methods in most scenarios. In addition, the performance of different estimation methods is dependent on the specific case, and it is hard to indicate clearly that which method is superior in our study. Nevertheless, the statistic model: GBLUP model and RKHS regression, and the machine-learning based method: Random forests and Ordinal Mcrank, it is observed that their values are close and exhibit a similar trend in most of the situations.

doi:10.6342/NTU202301253

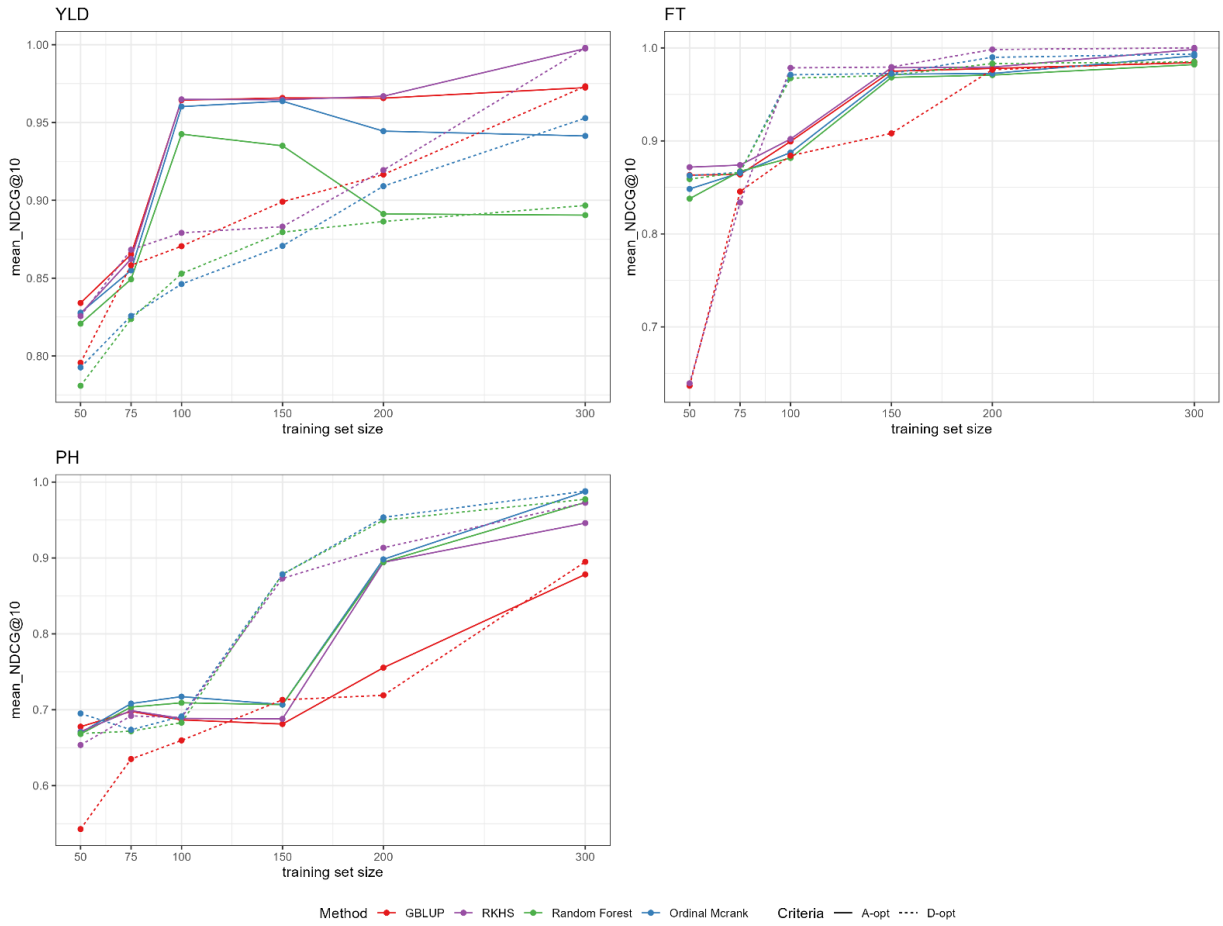

Figure 11. The comparison between different estimation methods for the phenotypic data of the tropical rice dataset.

Vertical axis represents the average mean of NDCGk@10. Types of line represent A-opt and D-opt optimality criteria for selecting training set and colors of line represent various estimation methods.

doi:10.6342/NTU202301253

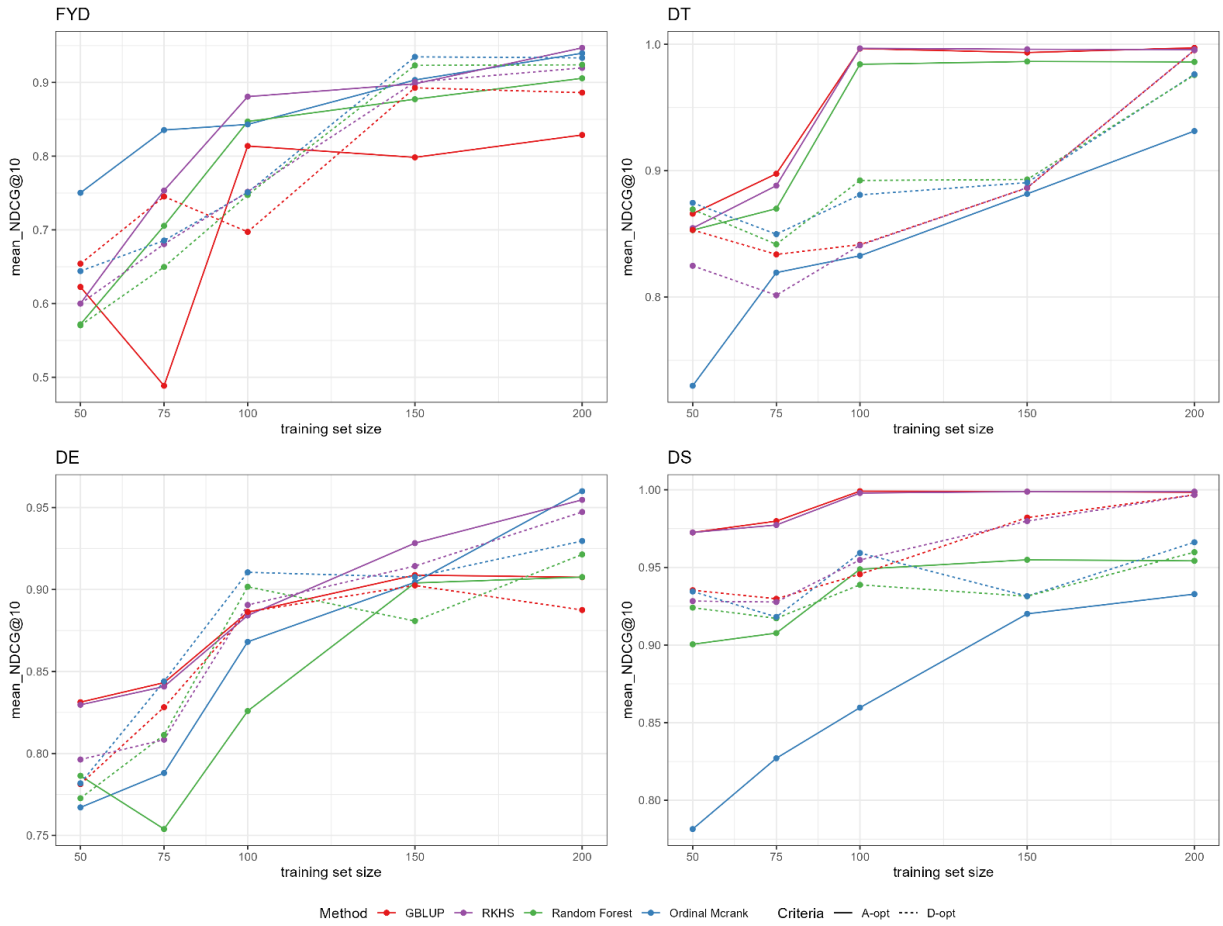

Figure 12. The comparison between different estimation methods for the phenotypic data of the wheat dataset.

Vertical axis represents the average mean of NDCGk@10. Types of line represent A-opt and D-opt optimality criteria for selecting training set and colors of line represent various estimation methods.

doi:10.6342/NTU202301253

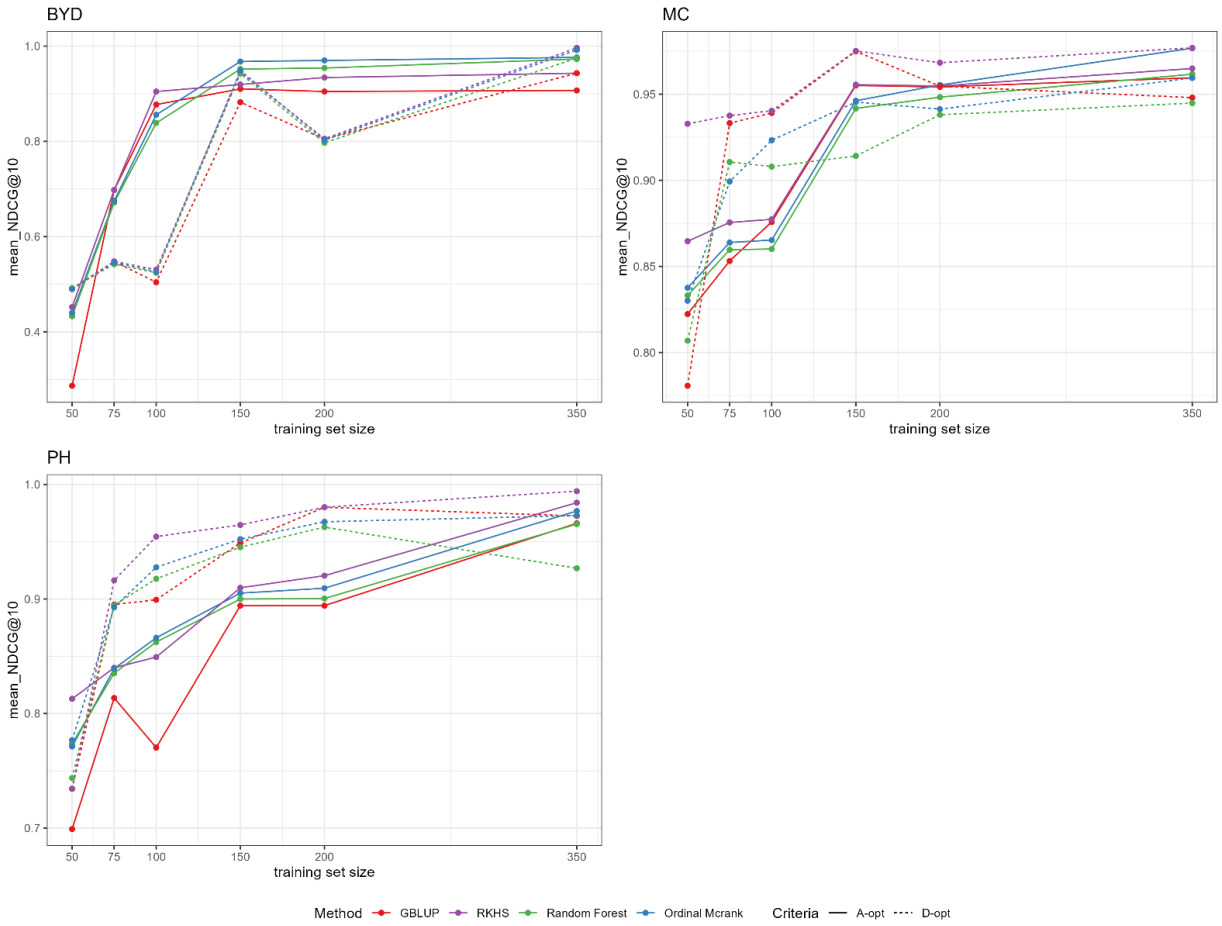

Figure 13. The comparison between different estimation methods for the phenotypic data of the sorghum dataset.

Vertical axis represents the average mean of NDCGk@10. Types of line represent A-opt and D-opt optimality criteria for selecting training set and colors of line represent various estimation methods.

doi:10.6342/NTU202301253

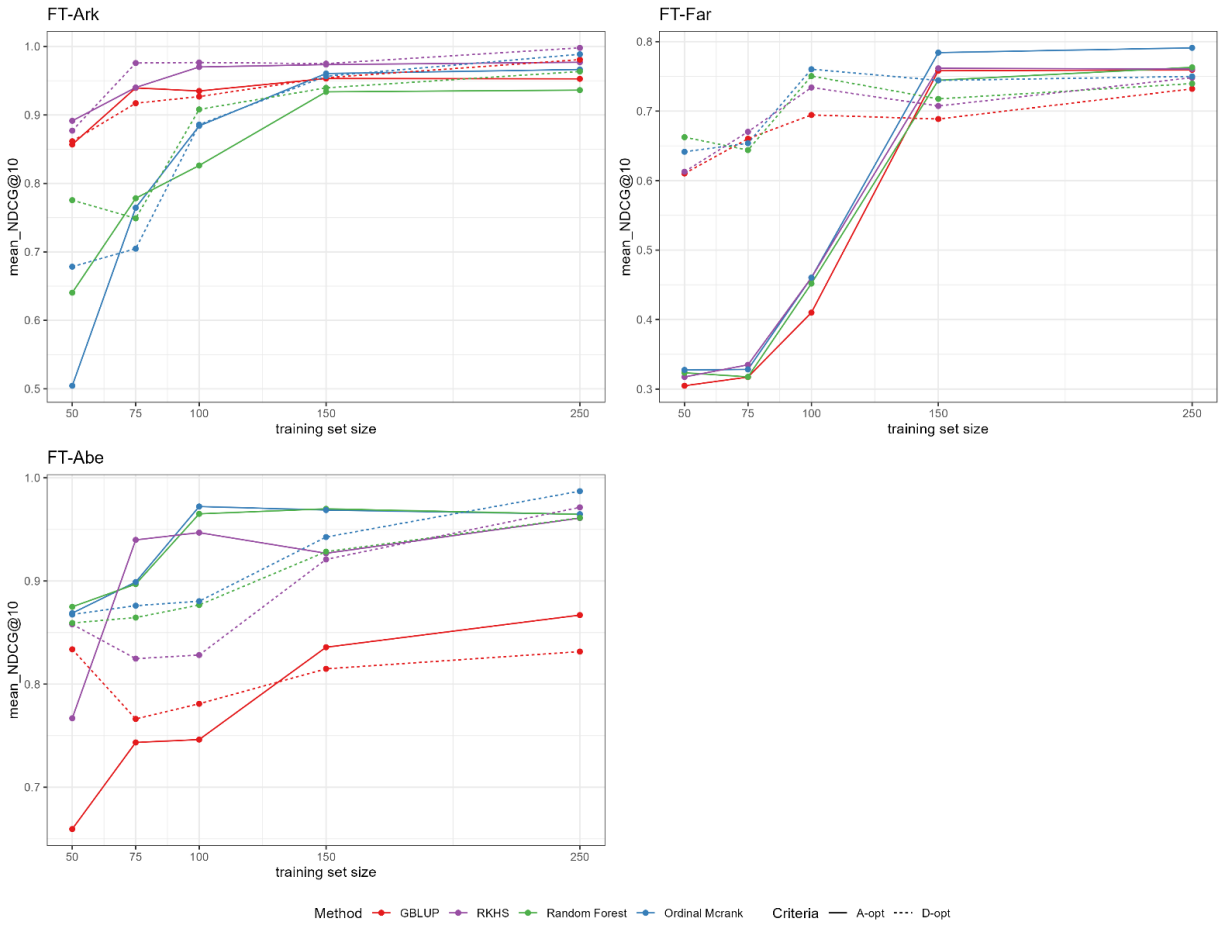

Figure 14. The comparison between different estimation methods for the phenotypic data of the 44K rice dataset.

Vertical axis represents the average mean of NDCGk@10. Types of line represent A-opt and D-opt optimality criteria for selecting training set and colors of line represent various estimation methods.

doi:10.6342/NTU202301253

Chapter 5 Conclusion

In our study, we aimed to compare the performance of two optimality criteria, A-optimality and D-optimality, with random training sets in GS. Both A-optimality and D-optimality demonstrated better performance compared to random training sets in most cases.

Initially, we hypothesized that D-optimality, which considers covariances between individuals, was supposed to outperform A-optimality. However, interestingly, A-optimality demonstrated superior performance in a greater number of situations. We presumed that the utilization of the genetic algorithm in D-optimality may have led to the identification of only local optima rather than global optima.

Overall, our study contributes to the understanding of the performance of A-optimality and D-optimality, providing breeders with a smart approach to selecting training sets in breeding programs.

doi:10.6342/NTU202301253

Appendix 1

Normalization for SNP data

For a particular SNP, there are $n_A, n_H \; and \; n_B$ individuals, with AA, AB and BB respectively. The standardized marker scores are defined as:

$$M_A = \frac{-1-\overline{x}}{s};$$

$$M_H = \frac{0-\overline{x}}{s};$$

$$M_B = \frac{1-\overline{x}}{s}$$

where

$$\overline{x} = \frac{n_A \times (-1) + n_H \times 0 + n_B \times 1}{n} = -P_A + P_B;$$

$P_A = \frac{n_A}{\text{n}}$ devotes the frequency of homozygote AA in one SNP marker.

$P_B = \frac{n_B}{\text{n}}$ devotes the frequency of homozygote BB in one SNP marker.

$P_H = \frac{n_H}{\text{n}}$devotes the frequency of heterozygote AB in one SNP marker.

$$s^2 = \frac{n_A \times (-1-\overline{x})^2 + n_H \times (0-\overline{x})^2 + n_B \times (1-\overline{x})^2}{n}$$

$$= \frac{1}{n}\left(n_A + n_B - n\overline{x}^2\right)$$

$$= \frac{1}{n}(n_A + n_B - n(-P_A + P_B)^2)$$

$$= P_A + P_B - (-P_A + P_B)^2$$

$$= P_A + P_B - P_A^2 + 2P_AP_B - P_B^2$$

$$= (P_A + P_B) - (P_A^2 + P_B^2) + 2P_AP_B$$

$$= (1 - P_H) - (P_A + P_B)^2 + 4P_AP_B$$

$$= (1 - P_H) - (1 - P_H)^2 + 4P_AP_B$$

doi:10.6342/NTU202301253

$$= (1 - P_H)P_H + 4P_A P_B.$$

Thus, we have that

$$M_A = \frac{-1 - \overline{x}}{s} = \frac{-P_H - 2P_B}{s};$$

$$M_H = \frac{0 - \overline{x}}{s} = \frac{-P_A + P_B}{s};$$

$$M_B = \frac{1 - \overline{x}}{s} = \frac{2P_A + P_H}{s}$$

where

$$s = \sqrt{(1 - P_H)P_H + 4P_A P_B}.$$

If $n_H = 0$, which is highly homogenous genome, then

$$M_A = \frac{-1 - \overline{x}}{s} = \frac{-2P_B}{\sqrt{4P_A P_B}} = \frac{-\sqrt{P_B}}{\sqrt{P_A}};$$

$$M_B = \frac{1-\overline{x}}{s} = \frac{2P_A}{\sqrt{4P_A P_B}} = \frac{\sqrt{P_A}}{\sqrt{P_B}}.$$

doi:10.6342/NTU202301253

Appendix 2 Source code in R

```
### source code for master thesis
### Wen Hsiu
### 2022.7.3

###package
library('dplyr')
library('devtools')
install_github("TheRocinante-lab/TrainSel")
library('devtools')
library("TrainSel")
###############

###function###
###DCG values###
get_dcg <- function(true,pred,k){
df = data.frame(y_true=true,y_pred=pred)
df = df[order(df[,2],decreasing = T),]
dcg= 0
for (i in 1:k){
a=df[i,1]/log2(i+1)
dcg = dcg + a
}
return(dcg)
}

###NDCG value###
get_ndcg <- function(y_true,y_pred,k){
dcg = get_dcg(y_true,y_pred,k)
idcg = get_dcg(y_true,y_true,k)
ndcg = dcg/idcg
return(ndcg)
}


###mean NDCG value###
```


doi:10.6342/NTU202301253

```
get_ndcg_mean=function(y_true,y_pred,k){
nmean=c()
for(i in 1:k){
nmean[i]=get_ndcg(y_true,y_pred,i)
}
return(mean(nmean))
}

###stratified numbers of subpopulation###
##cluster=subpopulation data
##p=proportion
##N=training set size
sub_number=function(cluster,N){
p=table(clusters)/length(clusters)
max=table(cluster)
sub=round(N*p)
stop1=0
while(stop1==0){
for (i in 1:length(p)){
if (sub[i]>max[i]){sub[i]=max[i]}
}
stop=0
while(stop==0){
if (sum(sub)>N){
a=sample(length(p),1)
sub[a]=sub[a]-1
}else if(sum(sub)<N){
a=sample(length(p),1)
sub[a]=sub[a]+1
}else{stop=1}
}
a=c()
for (i in 1:length(p)){
a[i]=sub[i]>max[i]
}
if (sum(a)==0 && sum(sub)==N){stop1=1}
}
```


doi:10.6342/NTU202301253

```
return(sub)
}


###A-opt###
##kin=normalized kinship matrix
##N=training set size
##cluster=sub population cluster
get_a_opt=function(kin,N,cluster=0){
##without subpopulation
if (sum(cluster)==0){
trace=diag(kin)
o=order(trace,decreasing = T)
kin1=trace[o]
return(names(kin1)[1:N])
}

##with subpopulation
else{
trace=diag(kin)
o=order(trace,decreasing = T)
n=sub_number(cluster,N)
trace_o=trace[o]
a=c()
for (i in 1:length(unique(cluster))){
trace_sub=trace[cluster==unique(cluster)[i]]
trace_sub_order=trace_sub[order(trace_sub,decreasing = T)]
      a=c(a,trace_sub_order[1:n[i]])
      }
      return(names(a))
      }
      }

      ##GA function
      #cross over
      ####2 chromosome a1,a2,chromosome length=n
      crossover = function(a1,a2,n){
```


doi:10.6342/NTU202301253

```
    x = sample(1:(length(a1)-1),1)
    cross = c(a1[1:x],a2[(x+1):n])
    cross = sort(cross)
    while (length(unique(cross))<n){
    cross=unique(cross)
    cross=sample(setdiff(union(a1,a2),cross),1) %>% c(cross,.) %>% sort
    }
    return(cross)
    }

    ##mutation
    ##1 chromosome a1
    ##all candidate a2
    ##mutation rate=p
    mutation = function(a1,a2,p){
    n=length(a1) #向量長度
    m=sample(c(1,0),n,replace = T,prob=c(1-p,p)) #mutated loci
    m.loc = which(m==0)
    m.number = length(m.loc) ##loci number
    if (m.number!=0){
    m.pool=setdiff(a2,a1[-m.loc]) ##delete those be chose
    a1[m.loc]=sample(m.pool,m.number) ##mutate
    }
    return(sort(a1))
    }

    ###D-opt###
    ##kin=normalized kinship matrix
    ##N=training set size
    ##cluster=sub population cluster
    get_d_opt=function(kin,N,cluster=0){
    cluster=clusters
    n=sub_number(cluster,N)
    ##without subpopulation
    if (sum(cluster)==0){
    dataDopt = list(d.matrix=kin)
    DOPT = function(soln,Data){
```


doi:10.6342/NTU202301253

```
147.      Fmat=Data[["d.matrix"]]
148.      return(det(Fmat[soln,soln]))
149.      }
150.
151.      ##GA parameter
152.      TSC = TrainSelControl()
153.      TSC$niterations=1000
154.      TSC$npop=nrow(kin)
155.      TSC$nelite=20
156.
157.      TSOUT=TrainSel(Data = dataDopt,
158.      Candidates = list(1:nrow(kin)),
159.      setsizes = c(N),
160.      settypes = "UOS",
161.      Stat=DOPT,control = TSC)
162.      d_opt=rownames(kin)[TSOUT[["BestSol_int"]]]
163.      return(d_opt)
164.
165.      #with subpopulation
166.      }else{
167.
168.      sublist=list()
169.      for (i in 1:length(unique(cluster))){
170.      sublist[[i]]=names(clusters[clusters==unique(cluster)[i]])
171.      }
172.      #create one chromosome
173.      get_1chro=function(n){
174.      chro=c()
175.      for (i in 1:length(max)){
176.      a=sort(sample(sublist[[i]],n[i],replace=F))
177.      chro=append(chro,a)
178.      }
179.      return(chro)
180.      }
181.      ##create 20 chromosome
182.      x = replicate(20,get_1chro(n)) %>% data.frame()
183.      ###存 result
```


doi:10.6342/NTU202301253

```
      lit=100000
      result=c()

      litnow=1; stop=0
      while (stop==0){
      cat(litnow,"50")
      deter=apply(x,2,function(x)
  kin[rownames(kin)%in%x,rownames(kin)%in%x] %>% det)

      top = which.max(deter)
      p =(max(deter)-deter[-top])/sum(max(deter)-deter[-top]) ##eliminate
  probability
      del=setdiff(1:20,top) %>% sample(.,8,prob=p) ##8 eliminate
      sel=c(1:20)[-del]
      ##create 7 cross over chro
      cross7=data.frame(matrix(rep(NA,50*7),nrow=50,ncol=7))

      #crossover seperately by subpopulation
      for (k in 1:7){
      ch_part_cross=c()
      for (i in 1:length(unique(cluster))){
      sub=cluster %>%
      .[.== unique(cluster)[i]] %>%
      names()
      cho=sample(sel,2)
      a=x[,cho[1]] %>% intersect(.,sub)
      b=x[,cho[2]] %>% intersect(.,sub)
      ch_part= data.frame(a,b)
      a=crossover(ch_part[,1],ch_part[,2],
      nrow(ch_part))
      ch_part_cross=append(ch_part_cross,a)
      }
      cross7[,k]=ch_part_cross
      }

      off=data.frame(x[,sel],cross7)

```


doi:10.6342/NTU202301253

```
219.      ##mutate seperately by subpopulation
220.
221.      mut=data.frame(matrix(rep(NA,N*19),nrow=N,ncol=19))
222.
223.      for (i in 1:length(unique(cluster))){
224.      sub=cluster %>%
225.      .[.== unique(cluster)[i]] %>%
226.      names()
227.      before=off[off[,1]%in%sub,]
228.      after=before
229.
230.      for (k in 1:19){
231.      after[,k]=mutation(before[,k],sublist[[i]],0.05)
232.      }
233.      mut[off[,1]%in%sub,]=after
234.      }
235.
236.      ##add the best
237.      new.x = data.frame(mut,top=x[,top])
238.      x = new.x
239.      result[litnow]=max(deter)
240.      if ((litnow-which.max(result))>=20000){stop=1}
241.      litnow=litnow+1
242.      }
243.      }
244.      return(x[,20])
245.      }
```


doi:10.6342/NTU202301253

Bibliography

Akdemir, D., Rio, S., & Isidro y Sánchez, J. (2021). Trainsel: an r package for selection of training populations. *Frontiers in genetics*, *12*, 655287.

Akdemir, D., Sanchez, J. I., & Jannink, J.-L. (2015). Optimization of genomic selection training populations with a genetic algorithm. *Genetics Selection Evolution*, *47*, 1-10.

Blondel, M., Onogi, A., Iwata, H., & Ueda, N. (2015). A ranking approach to genomic selection. *Plos one*, *10*(6), e0128570.

Chung, P.-Y., & Liao, C.-T. (2020). Identification of superior parental lines for biparental crossing via genomic prediction. *Plos one*, *15*(12), e0243159.

Covarrubias-Pazaran, G. (2016). Genome-assisted prediction of quantitative traits using the R package sommer. *Plos one*, *11*(6), e0156744.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society: series B (methodological)*, *39*(1), 1-22.

Fernandes, S. B., Dias, K. O., Ferreira, D. F., & Brown, P. J. (2018). Efficiency of multi-trait, indirect, and trait-assisted genomic selection for improvement of biomass sorghum. *Theoretical and Applied Genetics*, *131*, 747-755.

Fernández-González, J., Akdemir, D., & Isidro y Sánchez, J. (2023). A comparison of methods for training population optimization in genomic selection. *Theoretical and Applied Genetics*, *136*(3), 30.

Forni, S., Aguilar, I., & Misztal, I. (2011). Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. *Genetics Selection Evolution*, *43*, 1-7.

doi:10.6342/NTU202301253

González-Camacho, J., de Los Campos, G., Pérez, P., Gianola, D., Cairns, J., Mahuku, G., Babu, R., & Crossa, J. (2012). Genome-enabled prediction of genetic values using radial basis function neural networks. *Theoretical and Applied Genetics*, *125*, 759-771.

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., & Goddard, M. E. (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of dairy science*, *92*(2), 433-443.

Heslot, N., Yang, H. P., Sorrells, M. E., & Jannink, J. L. (2012). Genomic selection in plant breeding: a comparison of models. *Crop science*, *52*(1), 146-160.

Jarvelin, K. (2000). IR evaluation methods for retrieving highly relevant documents. Proc. International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), July 2000,

Kristensen, P. S., Jensen, J., Andersen, J. R., Guzmán, C., Orabi, J., & Jahoor, A. (2019). Genomic prediction and genome-wide association studies of flour yield and alveograph quality traits using advanced winter wheat breeding material. *Genes*, *10*(9), 669.

Laloë, D. (1993). Precision and information in linear models of genetic evaluation. *Genetics Selection Evolution*, *25*(6), 557-576.

Laloë, D., Phocas, F., & Menissier, F. (1996). Considerations on measures of precision and connectedness in mixed linear models of genetic evaluation. *Genetics Selection Evolution*, *28*(4), 359-378.

Meuwissen, T. H., Hayes, B. J., & Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*(4), 1819-1829.

Ou, J.-H., & Liao, C.-T. (2019). Training set determination for genomic selection. *Theoretical and Applied Genetics*, *132*, 2781-2792.

doi:10.6342/NTU202301253

Rincent, R., Laloë, D., Nicolas, S., Altmann, T., Brunel, D., Revilla, P., Rodriguez, V. M., Moreno-Gonzalez, J., Melchinger, A., & Bauer, E. (2012). Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (Zea mays L.). *Genetics*, *192*(2), 715-728.

Spindel, J., Begum, H., Akdemir, D., Virk, P., Collard, B., Redona, E., Atlin, G., Jannink, J.-L., & McCouch, S. R. (2015). Genomic selection and association mapping in rice (Oryza sativa): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. *PLoS genetics*, *11*(2), e1004982.

Tsai, S.-F., Shen, C.-C., & Liao, C.-T. (2021). Bayesian optimization approaches for identifying the best genotype from a candidate population. *Journal of Agricultural, Biological and Environmental Statistics*, *26*, 519-537.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of dairy science*, *91*(11), 4414-4423.

Wu, P.-Y., Ou, J.-H., & Liao, C.-T. (2023). Sample size determination for training set optimization in genomic prediction. *Theoretical and Applied Genetics*, *136*(3), 57.

Zhao, K., Tung, C.-W., Eizenga, G. C., Wright, M. H., Ali, M. L., Price, A. H., Norton, G. J., Islam, M. R., Reynolds, A., & Mezey, J. (2011). Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa. *Nature communications*, *2*(1), 467.

doi:10.6342/NTU202301253