frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 21 October 2025
DOI 10.3389/fpls.2025.1675815

OPEN ACCESS

EDITED BY
Mathieu Rouard,
Alliance Bioversity International and CIAT, France

REVIEWED BY
Christopher Cullis,
Case Western Reserve University, United States
Rebecca Grumet,
Michigan State University, United States

*CORRESPONDENCE
Boris Demenou
b.demenou@arvalis.fr

RECEIVED 29 July 2025
ACCEPTED 29 September 2025
PUBLISHED 21 October 2025

CITATION
Gouy M, Bogard M, Mohamadi F and Demenou B (2025) Optimizing core collections for genetic studies: a worldwide flax germplasm case study. *Front. Plant Sci.* 16:1675815. doi: 10.3389/fpls.2025.1675815

Optimizing core collections for genetic studies: a worldwide flax germplasm case study

Matthieu Gouy[1], Matthieu Bogard[1], Faharidine Mohamadi[2] and Boris Demenou[3]*

[1]SAGeP, ARVALIS, Baziège, France, [2]ARVALIS, Station Expérimentale, Boigneville, France, [3]SAGeP, ARVALIS, Ouzouer-le-Marché, France

Core collections provide a strategic approach to reducing population size while retaining genetic diversity and allele frequencies, serving as key resources for genetic research. Although various sampling and selection strategies have been proposed, most of them focused on either diversity or representativeness, rarely both, and none fully integrated these with QTL detection optimization. The first part of our study focuses on a genetic diversity analysis of a flax germplasm (*Linum usitatissimum* L.) maintained by the Arvalis Institute, a prerequisite for the development of core collections. This germplasm is a worldwide flax collection comprising 1,593 accessions originating from 42 countries, encompassing all major flax-growing regions. It includes both spring- and winter-type lines, as well as oilseed and fiber types. The results revealed a pronounced genetic structure within the germplasm with six clusters, strongly influenced by cultivation purposes (fiber *vs.* oilseed flax), growth cycle (winter *vs.* spring), and then geographic origin. Overall genetic diversity was moderate (H_e = 0.22), with oilseed flax clusters displaying greater diversity (H_e from 0.21 to 0.27) than fiber flax (H_e < 0.17). In a second step we evaluated distinct strategies for core-collection development, including approaches -originally developed for core collection construction and others-developed for optimizing genomic-selection calibration panels. We introduced an approach based on QTL detection performance via extensive simulations of QTLs distributed across the genome. We observed a fundamental trade-off between maximizing diversity and ensuring representativeness in core collection design. Diversity-oriented approaches may overemphasize rare or outlier genotypes, compromising representativeness, while representativeness-focused strategies leaded to overlooking rare alleles, thus limiting diversity. In our results we have found that particular combinations of selection criteria achieved a favorable balance between genetic diversity and representativeness, while concurrently maintaining a robust capacity to capture QTL signals across the genome. Finally, the approach using the Shannon index combined with the allelic coverage led to optimal core collection design adapted for GWAS applications in a structured population; and was used to select a core collection of 409 accessions useful for further genetic studies. These results provide knowledge for the development of optimized core collections tailored to GWAS applications.

KEYWORDS

core collection, optimization criteria, quantitative trait loci (QTLs), genetic diversity, flax (*linum usitatissimum* l.), Western European flax breeding

1 Introduction

Plant genetic resources are a crucial source of diversity and are essential for improving crops. Useful plant genetic resources for breeding includes landraces, breeding lines, cultivars and the wild relatives of the target species, offering a broad range of alleles that can be exploited to enhance key agronomic traits. To face climate change, food security challenges, and the need for sustainable agriculture, the management and conservation of genetic resources are fundamental for the development of productive and resilient genetic material (Esquinas-Alcázar, 2005). More than seven million plant accessions are conserved across around 1,750 genebanks, with over 60% belonging to only thirty cultivated species (Hammer et al., 2003; Thormann et al., 2012). The diversity maintained is largely oriented towards human needs. Given the high number of accessions to be conserved, it has become necessary to rationalize the management of these resources, particularly through the selection of core collections.

The core collection concept was formalized in the 1980s to ensure optimal management and use of the genetic resources collected over time. A core collection can be defined as a reduced set of accessions that represents the genetic diversity of a species and its wild relatives with minimal redundancy (Brown and Clegg, 1983; Frankel, 1984). Since its inception, numerous studies have focused on methodologies for creating core collections. Brown et al. (1989) suggested that a core collection should not exceed 10% of the full collection and should never include more than 2,000 entries. In practice, most core collections represent between 5% and 20% of the original germplasm (Van Hintum et al., 2000). The reduced size of a core collection is crucial to ensure its efficient long-term management. The creation of core collections addresses two main objectives: (1) maximizing the genetic diversity, often favored by taxonomists, geneticists, and gene bank curators, and (2) maximizing the representativeness of the germplasm, typically chosen by breeders (Marita et al., 2000). The goal for the former is to maximize diversity criteria and conserve the rarest alleles. The second objective involves faithfully representing the source germplasm by retaining more generalist alleles. Jointly optimizing these two objectives ensures efficient short- and medium-term management of a species' genetic resources, although this remains challenging.

Initially, passport information (i.e., morpho-descriptives data, geographical origins) and other phenotypic traits (e.g. earliness, disease resistance traits) were used to establish core collections. However, it was recognized that environmental factors could influence these variables, leading to inaccurate representations of heritable genetic diversity (Tanksley and McCouch, 1997). Nowadays, the use of molecular markers, such as RAPDs (Marita et al., 2000), SSRs (Soto-Cerda et al., 2013), or SNPs (Bianchi et al., 2020; Fu, 2025) has become standard and essential for studying genetic diversity and developing core collections.

Many approaches to developing core collection (CC) have been described. For these approaches, a comprehensive characterization of the species' genetic diversity and structure is an essential prerequisite, as it is critical to ensure that all genetic clusters are adequately represented within the selected subsets of individuals. This requirement underpins the rationale for employing stratified sampling methods, which offer a more suitable alternative to random sampling by preserving the underlying genetic structure (Charmet and Balfourier, 1995; Gouesnard et al., 2001; Franco et al., 2003). The sampling rate (i.e., allocation) must be defined based on the intended objectives. Several strategies have been suggested: a fixed selection rate, independent from cluster size, a rate proportional to cluster size, a logarithmic-proportional rate (which helps maintaining a manageable collection size), or a rate proportional to intra-cluster genetic distances (or other diversity metrics), also known as the D-method. This latter method has shown significant efficiency compared to alternative approaches (Franco et al., 2006).

The selection of accessions can be based on one or several objective(s) (rarely greater than two) to be optimized, such as genetic distances (Jansen and Van Hintum, 2007), diversity criteria (Franco et al., 2006; Thachuk et al., 2009), or the effective alleles number and their coverage rate (Kim et al., 2023). Some strategies have been developed to simultaneously optimize multiple criteria (Odong et al., 2013; De Beukelaer et al., 2018). These approaches rely on optimization algorithms (e.g., genetic algorithms, simulated annealing) which iteratively optimize an objective function (maximizing or minimizing) by picking a new entry, often randomly, at each iteration.

Similar methodologies have been developed in the genomic selection area. These involve the use of calibration set optimization algorithms, which aim to maximize genomic prediction accuracy based on molecular marker data (Laloë, 1993; Albrecht et al., 2011; Pszczola et al., 2012; Rincent et al., 2012; Akdemir, 2017; Ou and Liao, 2019). While this approach does not directly link genomic selection calibration methods to core collection inception, the optimization techniques used in genomic selection, such as genetic algorithms and diversity-based criteria, could potentially be adapted for core collection creation. The focus on optimizing subsets for prediction accuracy in genomic selection parallels the goal of selecting representative subsets in core collection creation, suggesting a possible methodological crossover. Moreover, core collections are widely used in associations studies for QTL discovery (Nicolas et al., 2016; Berkner et al., 2024). This type of population typically harbors greater genetic diversity than biparental populations and includes a higher number of recombination events. As a result, the resolution of detected QTLs is significantly improved (Breseghello and Sorrells, 2006; Zhao et al., 2007; Huang and Han, 2014; Bandillo et al., 2015).

The quality assessment of a core collection should, whenever possible, be based on data that were not used for its development (Van Hintum et al., 2000). Core collections are often compared to the whole collection (WC) from which they were derived. Various evaluation criteria can be computed to assess the resulting population such as genetic distances, diversity indices (Shannon index, heterozygosity) or even Principal components analysis (Mohammadi and Prasanna, 2003; Reif et al., 2005, and Odong et al., 2013).

The first breeding and improvement flax (*Linum usitatissimum* L.) programs were initiated in the 1920s by Irish and Dutch

researchers (Doré and Varoquaux, 2006). Breeding efforts specifically targeting fiber flax also began during this period, with early hybridizations carried out in 1919 (e.g., with the EGBK or CRGH lines) (Blaringhem, 1926). Genetic improvement priorities in flax vary according to its intended use, fiber or oilseed, and are primarily aimed at addressing current agronomic and climatic constraints. In fiber flax, breeding efforts focus on enhancing resistance to major fungal pathogens, including *Polyspora lini*, *Septoria linicola*, fusarium wilt, flax scorch, and powdery mildew. Improving resistance to lodging is also a key objective, as it contributes to reducing yield losses and facilitating mechanized harvesting. Additionally, the enhancement of fiber quality remains a central goal, along with the development of cultivars with improved tolerance to abiotic stresses such as elevated temperatures, drought, but also cold, particularly for winter-type lines. For oilseed flax, breeding efforts are focused on stabilizing and optimizing yield while accounting for strong genotype-by-environment (G×E) interactions. Disease resistance, particularly against septoria, is another major goal. Lastly, improving oil quality and enhancing cold tolerance for winter-type lines are key breeding targets. The use of extended genetic diversity in breeding programs could help improving flax for resistance/tolerance to biotic and abiotic factors.

The worldwide diversity of cultivated flax and its wild relatives is represented by an estimated 48,000 accessions maintained in 33 genebanks, among which only around 10,000 are considered genetically distinct or truly unique (Diederichsen, 2007). From these resources, many flax core collection have been created (Fu, 2006; Diederichsen et al., 2013; Hoque et al., 2020) in order to investigate for example flowering time (Chandrawati et al., 2017), agronomic, seed and fiber quality, disease resistance traits (You et al., 2017), or even powdery mildew resistance (Speck et al., 2022). The Arvalis Institute, a French institute for applied research in agriculture, maintains a collection of around 1,650 fiber and oilseed flax accessions. This germplasm comprises accessions from countries worldwide where fiber and oilseed flax have been cultivated or are naturally distributed, with a particular focus on recently improved lines from western Europe breeding programs. However, no core collection based on this western European flax genetic resources was available. Then, rare genetic studies in Western Europe have examined a diversity panel including large modern Western European flax. Speck et al. (2022) used a flax panel of 311 lines selected from 38 countries spanning all continents and diverse worldwide climatic regions. However, they did not describe a clear selection methodology to ensure that genetic diversity was adequately represented. This study and others on cultivated flax diversity have revealed a significant genetic structure between fiber and oilseed groups. Further sub-structuring has also been characterized, often related to geographical origins or physiological development (winter vs. spring types) (Hoque et al., 2020; Fu, 2005; Speck et al., 2022). However, the effect of geographic origin is not always significant (Smýkal et al., 2011; Chandrawati et al., 2017; You et al., 2017). This may be attributed to the extensive exchange of genetic material (Soto-Cerda et al., 2013). Developing a core collection of flax germplasm focused on Western European diversity should facilitate genetic studies for flax breeding in Europe, while also allowing comparisons between studies based on this core collection.

In this study, we (i) performed genetic diversity analyses of a flax collection, (ii) compared various approaches to identify a core collection for further quantitative genetic studies and (iii) selected a core collection based on the best approach for further genetic studies. We tested and evaluated approaches specifically designed for core collection construction alongside population optimization methods that were originally developed for genomic selection calibration sets. We also proposed a novel criterion to compare approaches to build core collections based on QTL detection performance via extensive simulations of QTLs distributed across the genome. These methods differ in the type of input data used, the nature and number of optimization criteria (diversity indices, representativeness criteria, combination of them), and the algorithms used. The core collection designed will be useful for genome-wide association studies and genomic selection to enhance Western European flax breeding programs.

2 Materials and methods

2.1 Plant materials

The germplasm maintained by Arvalis since 2010 is a collection of 1,650 cultivated flax (fiber, oilseed and dual purposes type) accessions. The initial accessions were collected in 1938 by INRAe from botanical collections and further extended through exchanges with research institutes, international biological resources centers, and breeding companies. The most recent accessions collected are lines originating from breeding programs and obtained in 2021. This diversity panel is predominantly composed of spring-type inbred lines, with 66% belonging to the oilseed group, 22% to the fiber group, and 12% classified as dual-purpose (both fiber and oilseed). Some winter-type lines have been included (oil and fiber) representing a valuable genetic source for low temperature tolerance. This germplasm encompasses the global diversity of cultivated flax, with accessions originating from 42 countries across all continents. It includes 107 common accessions with the PGRC core collection (Canada), the U.S. NPGS core collection, and the composite collection from Guo et al. (2020). The full list of accessions can be found in Supplementary Table S1.

2.2 Phenotypic data

The germplasm has been phenotyped for a set of 22 traits, summarized in Supplementary Table S2. These data are primarily passport data used to describe the accessions, including flower morphology (anther and pollen color, petal shape and color, filament color and winding, style color, ciliation and pigmentation of capsule, corolla size, beak shape), seed morphology (seed color, thousand kernel weight), geographic origin, cultivated group (oilseed *versus* fiber-type), tolerance to low temperatures, lodging tolerance, as well as resistance to

powdery mildew and Fusarium wilt. Prior to analysis, missing values were imputed using the R package missForest v1.5 (Stekhoven and Bühlmann, 2012). No imputation was performed for the country of origin.

2.3 Genotypic data

A seedling was produced for each of the 1650 flax accessions in the growing room at Arvalis Institute site in Boigneville (France) with 20°C/18°C day/night temperature. The Fresh leaves of two-weeks-old seedlings (50–100 mg) were harvested in microtube strips and flash-frozen at -80°C for 24 hours before being freeze-dried for 48 hours and then ground using the MM400 vibro-grinder (Retch). Genomic DNA was extracted from the crushed material using a modified Machery-Nagel NucleoMag Plant kit on the Beckman Coulter Biomek i5 automated workstation. Genomic DNA was then checked for quality on NanoDrop ND8000 (Thermo Fisher Scientific) and quantified on Qubit (Thermo Fisher Scientific) by Picogreen dosage. All accessions were genotyped using the Allegro AT-SNP-30K targeted genotyping tool (Demenou et al., 2025, 2025) at the EPGV platform (INRAe, Evry, France).

The genotyping matrix was generated using the bioinformatics pipeline described in Demenou et al. (2025) and was then filtered. Markers and accessions with more than 50% missing data were discarded. The remaining markers were imputed using Beagle v5.4 (Browning, 2008; Browning and Browning, 2016), applying default parameters. Following this imputation, the genotyping matrix was filtered to remove markers having low minor allele frequency (MAF), retaining only those with MAF > 1% (Supplementary Table S3). This threshold has been chosen to preserve rare alleles that may carry valuable genetic information (Goudet et al., 2018). The distribution of selected imputed markers across the fifteen flax chromosomes was visualized using the R package CMplot v4.5.1 (Yin et al., 2021) to assess the quality and uniformity of the genotyping data.

2.4 Population structure and diversity analysis

Prior to the genetic diversity analysis, the genotyping matrix was intentionally pruned to retain only independent markers, thereby minimizing the confounding effects of collinearity among linked loci (Patterson et al., 2006). Marker pruning was performed using PLINK v1.07 (Purcell et al., 2007) with the following parameters: the 'indep-pairwise' function, a sliding window of 50 SNPs, and a linkage disequilibrium threshold of $R^2 = 0.4$. In other words, pairs of markers within a sliding window of 50 SNPs and an R^2 value greater than 0.4 were pruned, so that only one marker per pair was kept.

We performed a Discriminant Analysis of Principal Components (DAPC) using the R package adegenet v2.1.10 (Jombart et al., 2010). DAPC assigns membership probabilities to predefined genetic clusters, which were inferred via K-means clustering. The number of retained principal components for the DAPC was determined using the Tracy-Widom test (Patterson et al., 2006), which identifies the first axes that significantly explain genetic variation. The optimal number of clusters K was determined by evaluating models with K-values ranging from 1 to 10, using the Bayesian Information Criterion (BIC) to select the best-supported model. Additionally, the most likely K-value was inferred by considering the correspondence between the identified groups and our germplasm knowledge. A Principal Component Analysis (PCA) was conducted on the pruned and standardized matrix using the R package FactoMineR v2.11 (Lê et al., 2008) to visualize the diversity and clustering. Pairwise Fixation indices (Fst) were calculated between genetic clusters using the R package hierfstat v0.5.11 (Goudet, 2005).

To further characterize the diversity hosted by the germplasm, expected mean heterozygosity (Berg and Hamrick, 1997), Shannon's diversity index (Shannon, 1948), average Rogers' genetic distance (Rogers, 1972), and the proportion of rare alleles (considering MAF< 0.10) were computed for each cluster and for all the entire germplasm.

2.5 Establishment of core collections

Two categories of methods have been employed in this study: those specifically dedicated to core collections, and those aimed at building calibration populations, particularly for genomic selection purposes. Core collections were established using the R packages CoreCollection v0.9.5 (Jansen and Van Hintum, 2007; Odong et al., 2013), corehunter III v3.2.3 (Thachuk et al., 2009; De Beukelaer et al., 2018), TrainSel v3.0 (Akdemir et al., 2021), as well as the approach originally proposed by Laloë (1993) and further developed by Rincent et al. (2012) (R code acquired directly from the authors).

The method developed by Jansen and Van Hintum (2007) and later refined by Odong et al. (2013) is based on genetic distances among accessions. Entries are selected using a random descent algorithm, optimizing one of three available criteria: the Average Nearest Entry (A-NE), which minimizes the average distance between each accession and its nearest neighbors, the Nearest Neighboring Entry (E-NE), which maximizes this average distance, and the Entry-Entry (E-EE) criterion, which maximizes the pairwise distance among all accessions in the collection. We optimized the A-NE and E-NE criteria using Rogers' genetic distance (Rogers, 1972). Optimization parameters were kept at their default settings.

The corehunter III R package (Thachuk et al., 2009; De Beukelaer et al., 2018) applies a stochastic local search algorithm based on replica exchange Monte Carlo chains for core collection development. Multiple selection criteria can be combined and weighted. This method can accommodate various input data types, including genetic distance matrices, genotypic and phenotypic datasets. Version III of this package supports the use of the following selection criteria, either individually or jointly: the previously described A-NE, E-NE, and E-EE criteria, expected heterozygosity (He), Shannon diversity index (SH), and allelic

coverage (CV). All available optimization criteria were considered, except for the E-EE criterion. Criteria were applied individually or in pairwise combinations. In the case of bi-objective optimization, equal weights of 0.5 were assigned to each criterion. Additionally, a combination of the following three criteria, A-NE, SH, and CV, was tested, with each criterion assigned an equal weight (~0.33). The execution mode was set to default, and normalization was applied for multi-objective optimizations.

The approach proposed by Laloë (1993) and elaborated by Rincent et al. (2012) aims to select a reference set of individuals for phenotyping that maximizes the reliability of genomic predictions for non-phenotyped individuals based on their genotypes. This method optimizes the generalized coefficient of determination (CDmean), which measures the correlation between predicted and observed values of genetic contrasts. CDmean balances the prediction error variance (PEV) against the genetic variance of the contrasts, accounting for genetic relatedness. The optimization is performed using a hill-climbing algorithm, exchanging one individual at each iteration, with the CDmean recalculated at every step using the individuals' variance-covariance matrix. We use the R-code given by the authors. A total of 3,000 iterations were performed for each of the 10 core collection replicates.

Other available tools for selecting calibration sets include STPGA (Akdemir, 2017), TSDFGS (Ou and Liao, 2019), and more recently, the R package TrainSel v3.0 (Akdemir et al., 2021). TrainSel enables the selection of individuals through mono- or multi-objective optimization, with possible weighting of criteria. It combines a genetic algorithm with simulated annealing. For our study, TrainSel was used with the following objective functions:

- D-optimality criterion (D-opt): aiming to maximize the determinant of the information matrix f(M), corresponding to the principal component transformation of the genotypic matrix, this criterion maximizes the dispersion in the multivariate genetic space
- Avg_GRM_self: aiming to minimize the average relatedness within the calibration population, thus maximizing its genetic variance. The effectiveness of this criterion for calibration population selection has been demonstrated in previous studies (Atanda et al., 2021; Fernández-González et al., 2023)
- The combined optimization of D-opt and Avg_GRM_self.

Optimization algorithm hyperparameters were set as follows: medium population size, low complexity, and unordered sample. The remaining parameters were left at their default settings.

Table 1 summarizes the method × criterion combinations tested. For each combination, ten populations of 350 individuals were generated, with a random selection of the initial set. With such population size, the detection power for QTL studies should be enhanced (Hyne and Kearsey, 1995; Charmet, 2000; Vales et al., 2005).

2.6 Core collection selection and evaluation

2.6.1 Diversity and representativeness criteria

Each combination has been evaluated based on criteria assessing both genetic diversity and representativeness. To quantify the genetic diversity captured by each CC, the following metrics have been computed: the rare alleles ratio (MAF< 10%) (RAR), the mean Rogers' distance (MRD), the expected heterozygosity (He), and the Shannon diversity index (SH). These indices are calculated using the following formulas:

1. Rare alleles ratio (RAR):

$$RAR = \frac{H}{R}$$

where R is the number of rare (i.e. MAF<10%) SNPs identified in the whole collection, and H the number of these rare SNPs founded as heterozygous within the core collection.

2. Mean Rogers' genetic distance (MRD) (Rogers, 1972):

$$MRD = \frac{1}{m}\sum_{i=1}^{m}\sqrt{\frac{1}{2}\sum_{j=1}^{ni}(a_{ij}-b_{ij})^2}$$

where m is the number of loci, n_i is the number of alleles at locus i, a_{ij} and b_{ij}, are the genotype codes for individuals a and b at locus i. This metric can be likened to a Euclidean distance.

3. Expected heterozygosity (He) (Berg and Hamrick, 1997):

$$He = \frac{1}{L}\sum_{l=1}^{L}\left(1-\sum_{i=1}^{nl}p_{li}^2\right)$$

Where L is the number of loci, n_l is the number of alleles at locus l, p_{li} is the relative frequency of the i-th allele at locus l.

4. Shannon diversity index (SH) (Shannon, 1948):

$$SH = -\sum_{i=1}^{n}p_i \cdot log_2(p_i)$$

where n is the number of alleles and p_i is the frequency of the i-th allele.

To assess the representativeness of each CC relative to the WC, we computed the following metrics: allelic coverage (CV) (Kim et al., 2007), Kullback-Leibler divergence (KL) between allele frequency distributions of CC and WC (Kullback and Leibler, 1951), the average absolute Pearson's correlation of principal component vectors (COR) between CC and WC (Yamamoto et al., 2007), and the Mean Difference ratio (MD) for a set of phenotypic variables (Hu et al., 2000). For MD calculation, independent phenotypic variables were preselected using Cramér's V index (Cramér, 1999) to avoid overrepresentation of specific variable categories.

These representativeness metrics are computed using the following formulas:

TABLE 1 Combinations of twenty methods and selection criteria used in core collection development.

R Package	Algorithm	Criterion	Input data	Objective
CoreCollection [a]	Random descent	A-NE	Genetic distances*	Optimizes representativeness
		E-NE		Maximizes genetic variance
corehunter [b]	Stochastic local search on replica exchange Monte Carlo chains	HE	Genotyping matrix	Maximizes allelic diversity
		CV		Maximizes allelic coverage
		SH		Optimizes genetic diversity while penalizing redundancies
code only [c]	Hill-climbing	CDmean	Variance-covariance matrix	Optimizes representativeness
TrainSel [d]	Genetic algorithm	Avg_GRM_self	Kinship matrix**	Maximizes relatedness among individuals
		D-opt	Principal Components	Optimizes representativeness
corehunter [b]	Stochastic local search on replica exchange Monte Carlo chains	A-NE + E-NE	Genetic distances*	Optimizes representativeness and maximizes genetic variance
		A-NE + HE	Genetic distances* + genotyping matrix	Optimizes representativeness and maximizes allelic diversity
		A-NE + SH		Optimizes representativeness and genetic diversity while penalizing redundancies
		A-NE + CV		Optimizes representativeness and maximizes allelic coverage
		E-NE + HE		Maximizes genetic variance and allelic diversity
		E-NE + SH		Maximizes genetic variance and optimizes genetic diversity while penalizing redundancies
		E-NE + CV		Maximizes genetic variance and allelic coverage
		HE + SH	Genotyping matrix	Maximizes allelic diversity and optimizes genetic diversity while penalizing redundancies
		CV + SH		Maximizes allelic coverage and optimizes genetic diversity while penalizing redundancies
		CV + HE		Maximizes allelic coverage and allelic diversity
		A-NE + SH + CV	Genetic distances* + genotyping matrix	Optimizes representativeness and genetic diversity while maximizing allelic richness
TrainSel [d]	Genetic algorithm	D-opt + Avg_GRM_self	Kinship matrix**	Optimizes representativeness and relatedness among individuals

[a] Jansen and Van Hintum, 2007; Odong et al., 2013; [b] Thachuk et al., 2009, De Beukelaer et al., 2018; [c] Laloë, 1993; Rincent et al., 2012 (code only); [d] Akdemir, 2017; Akdemir et al., 2021; * Rogers's genetic distance; **Computed according Vanraden (2008) formula.
SH, Shannon diversity index; He, Expected heterozygosity; CV, Allelic coverage; CDmean, Mean coefficient of determination; Avg_GRM, Average Genetic Relationship Matrix; A-NE, Average Nearest Entry; D-opt, Optimality of the determinant of the information matrix f(M); E-NE, Nearest Neighboring Entry.

1. Allelic coverage (CV):

$$CV = \left(\frac{1}{L}\sum_{k=1}^{L}\frac{A_{core}}{A_{Wcol}}\right)$$

where L is the number of loci, A_{core} is the number of alleles present in the core collection at locus L and A_{Wcol} is the number of alleles present in the whole collection at the same locus.

2. Kullback-Leibler divergence (KL):

$$D_{KL}(p||q) = \sum_{j=1}^{m} p_j \,.\, \log\left(\frac{p_j}{q_j}\right)$$

where m is the total number of SNPs, p_j is the frequency of the minor allele at SNP j in the core collection and q_j is the corresponding frequency in the whole collection.

3. Mean difference ratio for phenotypic traits (MD):

$$MD = \left(\frac{S_t}{n}\right)$$

where S_t is the number of traits showing a significant difference between the CC and WC and n the total number of traits.

4. Average correlation between Principal Components (COR):

Pearson correlation coefficients ri are computed between principal components of the same rank from the CC and the WC. Due to an asymmetric distribution of correlation coefficients, we apply Fisher's z-transformation (Fisher, 1921) before calculating the mean as follow:

$$z_i = \frac{1}{2}\ln\left(\frac{1+r_i}{1-r_i}\right)$$

The value zi is then transformed to obtain the average correlation:

$$\bar{r} = \frac{e^{2\bar{z}} - 1}{e^{2\bar{z}} + 1}$$

Since not all principal components contribute equally to the genetic variance, each transformed coefficient z_i is weighted by the eigenvalue (inertia) of the corresponding component in the WC. This weighting scheme assigns greater importance to the principal axes. Only the significant axes under the Tracy-Widom test (Patterson et al., 2006) are considered.

2.6.2 QTL simulation and detection

We introduced a novel criterion that aimed to compare the core collections on their ability to detect QTLs. To this end, we simulated two traits for each chromosome using the R package PhenotypeSimulator v0.3.4 (Meyer and Birney, 2018), using the genotypic matrix of the whole collection. The obtained simulated QTLs thus leverage the existing linkage disequilibrium. QTLs were simulated separately on each chromosome. To obtain QTLs evenly distributed along the genome, QTLs were thus simulated separately on each chromosome. In total, 940 QTLs distributed across the 15 flax chromosomes were obtained for the whole collection.

QTL detection was carried out for each of the 200 core collections generated, using a mixed linear model (MLM) accounting for both population structure and relatedness (Yu et al., 2006). The model used was the following:

$Y = \mu 1 + Q_c + b_x + g + \varepsilon$where Y is the vector of phenotypic simulated values, μ the overall mean, Q the matrix of covariates derived from the DAPC to capture population structure, c the vector of fixed effects associated with these covariates, b the additive fixed effect of the SNP, x the vector of SNP genotypes coded as 0, 1, or 2, g the vector of polygenic random effects, and ε the vector of residuals. Residuals were assumed to follow a normal distribution ~ N (0, I σ_e^2), and polygenic effects were assumed to follow a normal distribution ~ N (0, K σ_g^2), with K being the kinship matrix computed using the Vanraden (2008) method, as implemented in the R package AGHmatrix v2.1.4 (Amadeu et al., 2023). Mixed linear models were fitted using the R package GMMAT v1.4.2 (Chen et al., 2020). To assess the effectiveness of population structure correction, the genomic control inflation factor λ (Devlin and Roeder, 1999) was calculated for each trait. Values of λ below 1.05 were considered indicative of appropriate control for population structure effects (Price et al., 2010). SNP–trait associations with p-values below the significance threshold determined using Gao's method (2008) were considered statistically significant and interpreted as putative QTLs. The proportion of QTLs identified within each core collection that were previously detected in the whole collection (considered as common QTLs) was calculated.

2.6.3 Synthetic index for an appropriate comparison

To facilitate comparison among core collection construction methods, we computed an index from the standardized values of our evaluation criteria. Criteria were pre-selected to balance representativeness and diversity. A preliminary analysis of the correlations between the indices was conducted to avoid including those that were highly redundant. The index I was defined accordingly as:

$$I = \frac{1}{7}(SH^* + He^* + RAR^* + CV^* + COR^* - KL^* + QTL^*)$$

where SH is the Shannon index, He is the expected heterozygosity, RAR is the rare allele rate, CV is the allelic coverage, COR is the mean correlation coefficient, -KL is the negative Kullback-Leibler divergence and QTL is the proportion of shared QTLs between core collection and whole collection. Each criterion has been standardized via normalization (i.e. centered and scaled).

2.6.4 Designing a core collection for flax germplasm

The objective was to select a core collection of c.a. 400 flax accessions, a desirable size allowing both diversity conservation and QTL discovery (Hyne and Kearsey, 1995; Charmet, 2000; Vales et al., 2005).To achieve this, we used a mixed approach: i) preselecting a part of the core collection based on the breeder's expertise and ii) used the best-identified core collection method to select the remaining accessions. For the first accessions selection step, a list of important accessions according to the breeder's expertise of two flax breeding companies in France (Linéa and Terre de Lin) was retained. Finally, we used the best core collection method identified in this study to select the remaining accessions. This core collection will be maintained by Arvalis institute and used for further genetic studies. It will be considered as a flax reference collection.

3 Results

3.1 Genetic diversity of the flax germplasm

A total of 30,893 Single Nucleotide Polymorphism (SNP) markers were obtained after genotyping the whole collection. Following filtering for missing data, the dataset comprised 29,007 SNP markers for 1,593 accessions with a minor allele frequency (MAF) greater than 1%. These markers are evenly distributed along the chromosomes, providing a significant genome-wide coverage (Figure 1). The imputed genotyping matrix was then pruned to retain only a set of independent markers more adapted for structure analyses. The resulting matrix contained 17,368 markers. The distribution of Tracy-Widom test statistics (Supplementary Figure S1) indicated that the first 203 principal components significantly contributed to the genetic variance. These components were retained for the DAPC and subsequent analyses.

Based on the Bayesian Information Criterion (BIC), the most likely number of genetic clusters was determined to be 6 (Figure 2A). At K = 2, the structuring of the germplasm followed the cultivation type, distinguishing oilseed from fiber flax

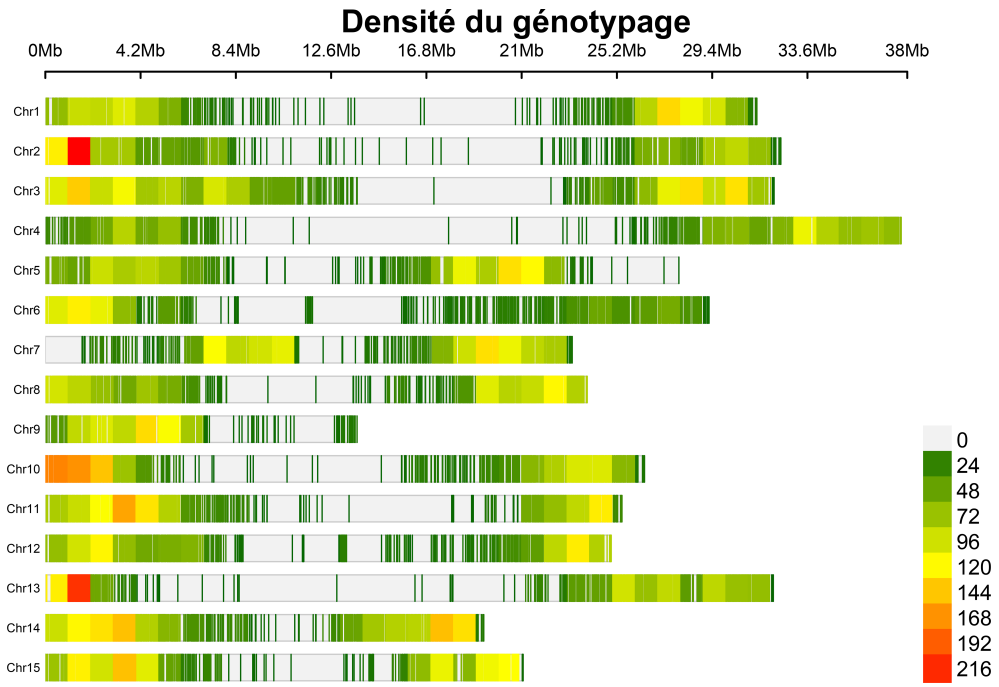

FIGURE 1
Distribution of SNP density along the 15 chromosomes of the flax genome, based on 28,475 mapped SNPs. Scale bar on the right provides numerical values for color density from zero (white) to two hundred sixteen (red).

accessions. At K = 3, a winter flax cluster emerged, characterized by enhanced tolerance to low temperatures (data not shown). From K = 4 onwards, the structuring primarily reflected the geographical origin of accessions. For instance, at K = 4, a new cluster was identified within the oilseed group, separating South American flax accessions from the rest. At K = 5, another cluster was found within the fiber group, separating Western European fiber flax from Eastern European fiber flax. At K = 6, the oilseed group further subdivided into three sub-clusters: South American, North American, and Eastern European oilseed flax. Beyond K = 6,

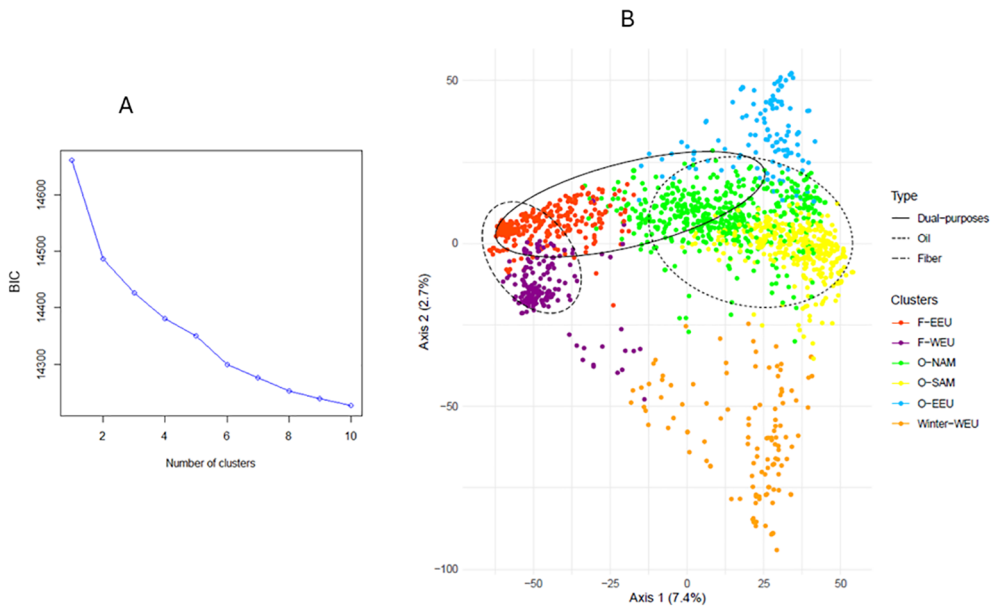

FIGURE 2
A two-part figure includes: **(A)** Bayesian Information Criterion (BIC) values as a function of the number of clusters defined in the flax germplasm; **(B)** Scatter plot of the first two principal components of the PCA of the Arvalis flax germplasm (1,593 accessions), with color-coded clusters identified by DAPC.

10.3389/fpls.2025.1675815

TABLE 2 Genetic diversity and differentiation indices of the six identified flax clusters.

Cluster	N	Ratio (%)	Diversity indices				Fixation index				
			Shannon index (SH)	Expected heterozygosity (He)	Mean rogers distances (MRD)	Rare alleles ratio (RAR)	O-SAM	O-NAM	Winter WEU	O-EEU	F-EEU
O-SAM	357	23%	0.64	0.26	0.40	0.95	–	–	–	–	–
O-NAM	498	31%	0.68	0.27	0.43	0.99	0.05	–	–	–	–
Winter-WEU	117	7%	0.55	0.21	0.34	0.91	0.16	0.14	–	–	–
O-EEU	121	8%	0.61	0.24	0.39	0.85	0.14	0.08	0.22	–	–
F-EEU	321	20%	0.45	0.17	0.28	0.80	0.24	0.14	0.28	0.24	–
F-WEU	179	11%	0.43	0.15	0.27	0.73	0.25	0.17	0.27	0.26	0.13
Whole Collection	1593	100%	0.67	0.22	0.42	1.00	0.19				

O-SAM, South American oilseed flax; O-NAM, Northern American oilseed flax; Winter-WEU, Winter Western European flax; O-EEU, Eastern European oilseed flax; F-EEU, Eastern European fiber flax and F-WEU, Western European fiber flax; N, Cluster size.

further differentiation occurred within the oilseed group, notably separating Eastern European lines from those of South American origin. Genetic diversity for K = 6 is illustrated via a principal component analysis (Figure 2B).

Table 2 summarizes the main features of the six previously identified clusters. Oilseed lines are overrepresented relative to fiber lines (69% vs 31%). Cluster O-NAM (Northern American oilseed flax) contained the largest number of accessions (31%), mostly composed of North American spring-type oilseed flax. Cluster Winter-WEU (Winter Western European flax) contained the fewest accessions (7%), primarily consisting of winter flax originated from Western Europe.

The pairwise fixation index Fst values revealed relatively high and significant genetic differentiation between clusters, especially between oilseed and fiber clusters (Fst up to 0.28, Table 2). The highest pairwise Fst value (Fst = 0.28) was found between clusters Winter-WEU and F-EEU (Eastern European fiber flax). The lowest Fst value (Fst = 0.05) was observed between oilseed clusters O-SAM (South American oilseed flax) and O-NAM (Northern American oilseed flax), both comprising American oilseed lines. The Fst value between all clusters was significantly larger than zero (Fst =0.19).

Genetic diversity indices were generally moderate, with the oilseed clusters being the most diverse. (Table 2). The clusters exhibiting the greatest genetic diversity were the American oilseed lines (clusters O-SAM and O-NAM), with Shannon diversity indices ranging from 0.64 to 0.68 and expected heterozygosity ranging from 0.21 to 0.27. These clusters also harbored the highest rate of rare alleles. The cluster O-EEU (Eastern European oilseed flax) and Winter-WEU showed a genetic diversity slightly lower than O-SAM and O-NAM (He of 0.24 for O-EEU and 0.21 for Winter-WEU). Conversely to oilseed clusters, the fiber clusters F-EEU and F-WEU (Western European fiber flax) exhibited the lowest level of diversity (He<0.17 and SH<0.45).

3.2 Core collection methods comparison

We evaluated 20 *methods x selection* criteria combinations for core collections establishment. In total, 200 core collections, each consisting of 350 accessions, were generated and assessed. For each core collection, representativeness and diversity indices were computed. Additionally, the proportion of shared QTLs with the whole collection was measured. Table 3 summarizes the mean values (calculated from ten replicates) and standard deviation of the evaluation criteria for the twenty tested combinations (non-normalized data).

The evaluation criteria showed different levels of variability between the methods. The Rare Allele Ratio (RAR) exhibited the lowest variability across methods. Regardless of the method used, nearly all rare alleles were consistently captured. Similarly, the allelic coverage rate (CV) displayed limited variation across methods (ranging from 0.965 to 0.994), suggesting that these two metrics were not strongly discriminative. In contrast, diversity-related indices such as Shannon's index (SH), expected heterozygosity (He), and Mean Rogers 'Distances (MRD)

TABLE 3 Evaluation criteria and standard errors computed for all core collection construction methods.

Methods	Diversity								Representativeness								QTL	
	SH	sd (SH)	He	sd (He)	MRD	sd (MRD)	RAR	sd (RAR)	CV	sd (CV)	KL	sd (KL)	COR	sd (COR)	MD	sd (MD)	QTL	sd (QTL)
CDmean	0.68	1.E-03	0.21	6.E-04	0.43	9.E-04	1.000	0.E+00	0.981	2.E-03	3.E-04	3.E-05	0.65	2.E-02	0.11	0.06	27%	0.04
CV	0.67	3.E-03	0.21	3.E-03	0.42	2.E-03	1.000	0.E+00	0.992	1.E-03	4.E-04	4.E-05	0.59	2.E-02	0.12	0.06	27%	0.02
Avg_GRM	0.67	3.E-03	0.21	2.E-03	0.43	2.E-03	1.000	2.E-04	0.991	1.E-03	4.E-03	1.E-02	0.58	4.E-02	0.17	0.05	27%	0.03
ANE_CV	0.67	2.E-03	0.21	3.E-03	0.42	2.E-03	1.000	8.E-05	0.992	1.E-03	4.E-04	5.E-05	0.57	4.E-02	0.14	0.07	28%	0.03
ANE	0.68	3.E-03	0.21	2.E-03	0.43	2.E-03	1.000	5.E-05	0.991	1.E-03	6.E-04	2.E-04	0.56	3.E-02	0.13	0.05	26%	0.04
D-opt+Avg_GRM	0.68	2.E-03	0.21	1.E-03	0.43	2.E-03	1.000	4.E-04	0.985	2.E-03	2.E-02	4.E-02	0.51	3.E-02	0.16	0.06	26%	0.02
D-opt	0.69	1.E-03	0.21	1.E-03	0.44	8.E-04	1.000	1.E-04	0.965	2.E-03	1.E-03	1.E-04	0.50	1.E-02	0.22	0.08	28%	0.02
ENE	0.69	1.E-03	0.22	4.E-03	0.44	1.E-03	1.000	2.E-04	0.994	1.E-03	2.E-03	3.E-04	0.44	4.E-02	0.34	0.12	24%	0.02
ANE_ENE	0.69	4.E-04	0.21	7.E-04	0.44	3.E-04	0.999	2.E-04	0.974	1.E-03	3.E-03	1.E-04	0.41	7.E-03	0.49	0.08	18%	0.01
SH_CV	0.71	1.E-04	0.22	9.E-04	0.46	1.E-04	1.000	0.E+00	0.991	5.E-04	6.E-03	6.E-05	0.41	9.E-03	0.49	0.04	27%	0.02
ANE_SH	0.71	1.E-04	0.22	1.E-03	0.45	1.E-04	0.999	4.E-04	0.991	4.E-04	2.E-02	4.E-02	0.48	6.E-03	0.43	0.04	24%	0.04
ANE_SH_CV	0.71	4.E-04	0.22	9.E-04	0.45	3.E-04	1.000	0.E+00	0.991	7.E-04	4.E-03	1.E-04	0.48	2.E-03	0.39	0.04	20%	0.01
ANE_HE	0.71	1.E-04	0.22	1.E-03	0.45	8.E-05	0.998	3.E-04	0.991	5.E-04	1.E-01	1.E-02	0.46	3.E-03	0.46	0.03	20%	0.03
SH	0.71	2.E-05	0.22	7.E-04	0.46	1.E-04	0.997	4.E-04	0.990	5.E-04	1.E-01	3.E-03	0.41	6.E-03	0.51	0.06	27%	0.02
HE_SH	0.71	5.E-05	0.22	9.E-04	0.46	5.E-05	0.996	5.E-04	0.990	2.E-04	1.E-01	5.E-03	0.39	1.E-02	0.51	0.04	27%	0.01
HE_CV	0.71	7.E-05	0.22	2.E-03	0.46	1.E-04	1.000	0.E+00	0.991	8.E-04	7.E-03	8.E-05	0.38	8.E-03	0.54	0.05	26%	0.01
ENE_SH	0.71	1.E-04	0.22	3.E-04	0.45	1.E-04	0.998	2.E-04	0.982	1.E-03	1.E-01	2.E-04	0.37	9.E-03	0.53	0.03	21%	0.02
HE	0.71	7.E-05	0.22	1.E-03	0.46	1.E-04	0.995	7.E-04	0.990	5.E-04	1.E-01	4.E-03	0.35	3.E-02	0.51	0.05	25%	0.02
ENE_HE	0.71	1.E-04	0.23	1.E-03	0.45	1.E-04	0.995	8.E-04	0.982	9.E-04	1.E-01	8.E-04	0.18	2.E-02	0.57	0.05	19%	0.02
ENE_CV	0.69	3.E-04	0.22	2.E-03	0.44	2.E-04	1.000	0.E+00	0.975	1.E-03	5.E-03	2.E-04	0.32	9.E-03	0.49	0.07	21%	0.02

SH, Shannon diversity index; He, Expected heterozygosity; MRD, Mean Rogers's distances; RAR, Rare allele ratio (considering MAF<10%); CV, Allelic coverage; KL, Kullback-Leibler divergence of allelic frequencies; COR, Average correlation between Principal components; MD, Mean phenotypic differences; QTL, Ratio of common QTLs simulated.

CDmean, Mean coefficient of determination; Avg_GRM, Average Genetic Relationship Matrix; ANE, Average Nearest Entry; D-opt, Optimality of the determinant of the information matrix f(M); ENE, Nearest Neighboring Entry.

10.3389/fpls.2025.1675815

demonstrated more substantial variation and followed similar trends across methods (Table 3). As expected, methods that maximized these indices tended to yield higher overall genetic diversity. Representativeness indices such as Kullback-Leibler divergence (KL), correlation coefficient (COR), and Mean Differences (MD) revealed significant differences between methods. The CDmean method consistently achieved the lowest KL divergence, the highest COR, and the lowest MD values. Other methods showing high representativeness included the CV, Avg_GRM_self, and ANE-based approaches. In general, representativeness indices exhibited greater variability compared to diversity indices. The highest MD values (greatest difference between core and whole collection reflecting low phenotypic representativeness) were observed for methods such as ENE_HE, He_CV, and ENE_SH.

The novel criterion based on the proportion of shared QTLs between the core and whole collections, was evaluated. All QTL detection models successfully controlled the inflation of test statistics (see Supplementary Figures S2A, B). QTL detection on core collections generated, varied between methods, ranging from 18% to 28% of the 940 QTL simulated in the whole collection (Supplementary Figure S3). Methods that detected the highest average number of simulated QTLs included D-opt, ANE_CV, and SH_CV. CDmean also showed a high median catching rate, ranking second only to D-opt. The ANE_CV, ANE_He and ANE_SH methods were more subject to sampling variations, exhibiting greater variability in QTL detection rate. Generally, methods ensuring high representativeness tended to catch more simulated QTLs. The method with the lowest QTL rate is ANE_ENE, and methods that prioritized the ENE index tended to bring fewer QTLs overall. A clear trade-off was observed between maximizing diversity and maximizing representativeness. Methods that were most effective at enhancing diversity generally performed less in terms of representativeness, and vice versa. Nonetheless, certain methods, such as ANE_SH, ANE_SH_CV, and SH_CV, provided a balanced compromise between the two objectives.

3.3 Composite score index for core collection evaluation

For a simplified cross-method comparison, a composite index integrating some evaluation criteria was calculated. Prior to index construction, inter-criteria correlations were assessed to avoid the inclusion of highly collinear metrics. Concurrently, a balance between representativeness and diversity was sought. The MD index was intentionally excluded from this composite index because it relies primarily on passport data that cannot accurately reflect the full extent of phenotypic diversity. Correlation analysis (Supplementary Figure S4) revealed that diversity criteria (SH, MRD, and He) were mutually and significantly correlated; moreover, MRD and SH exhibited redundancy, warranting the inclusion of only one of these metrics in the composite index. Although KL and COR were correlated, they conveyed distinct information and were therefore both retained. The distributions of KL, RAR, and CV were found to be highly skewed (Supplementary Figure S4). Notably, COR exhibited the strongest correlation with the QTL criterion (R = 0.45). All the seven selected criteria were normalized before integration into the composite score. The resulting composite index values are summarized in the boxplot Figure 3.

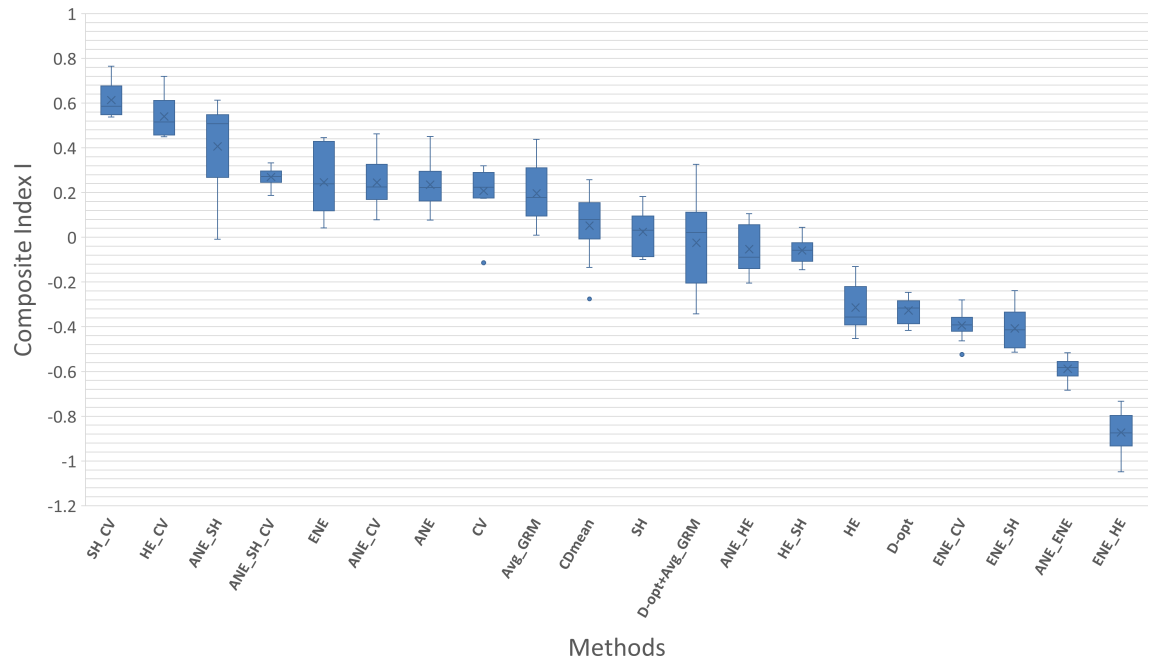

FIGURE 3
Boxplot showing the Composite Index I, calculated across ten repetitions of each tested method. The x-axis lists the various methods, and the y-axis represents the index values. Each method's box plot illustrates the median (the cross X), the interquartile range, and potential outliers.

The method yielding the highest index value, and thus the best overall compromise was the one that simultaneously optimized both the SH and CV criteria (SH_CV). In general, methods that integrated multiple criteria tended to produce higher index scores. Conversely, the method with the lowest index was the one that combined ENE and HE optimization (ENE_HE). More broadly, methods that optimized the ENE criterion, either alone or in combination with other criteria, consistently resulted in lower index values.

3.4 Core collection based on breeder's expertise, Shannon index and allelic coverage

A preliminary list of 98 essential accessions was preselected; including 30 rare winter flax accessions and 68 accessions (33 oilseed and 35 fiber flax) recommended by the two French flax breeders, Linéa and Terre de Lin. The breeder's list included breeding lines (modern lines) and selection material. Most of these were used in Western European breeding programs. Then, to build bridges between our core collection and existing collections around the world, and encourage international collaboration; we retained all the 107 accessions that are common to our collection with PGRC core collection (Canada), U.S.NPGS (USA) core collection, as well as the composite collection from Guo et al. (2020).

A complementary list of 204 new accessions was selected using SH_CV method by fixing the 205 preselected accessions. This core collection of 409 accessions included 300 oilseed flax, 69 fiber flax and 40 mixed types (Table 4; Figure 4).

We then compared genetic diversity and representativity parameters between the whole germplasm and the core collection (Table 4; Figure 4). The core collection captured almost all the genetic diversity of the whole collection (Table 4) and was fully representative of the whole collection (Figure 4). For example, we obtained He values of 0.216 for this core collection compared to 0.22 for the whole collection (Table 4).

TABLE 4 Comparison of the core collection and the whole collection.

Panel	Size	Oilseed	Fiber	Dual purposes	Shannon index	Expected Heterozygosity (He)	Mean Rogers' distance (MRD)
Germplasm	1593	1053 (66%)	354 (22%)	186 (12%)	0.672	0.220	0.424
Core collection	409	300 (73%)	69 (17%)	40 (10%)	0.692	0.216	0.440

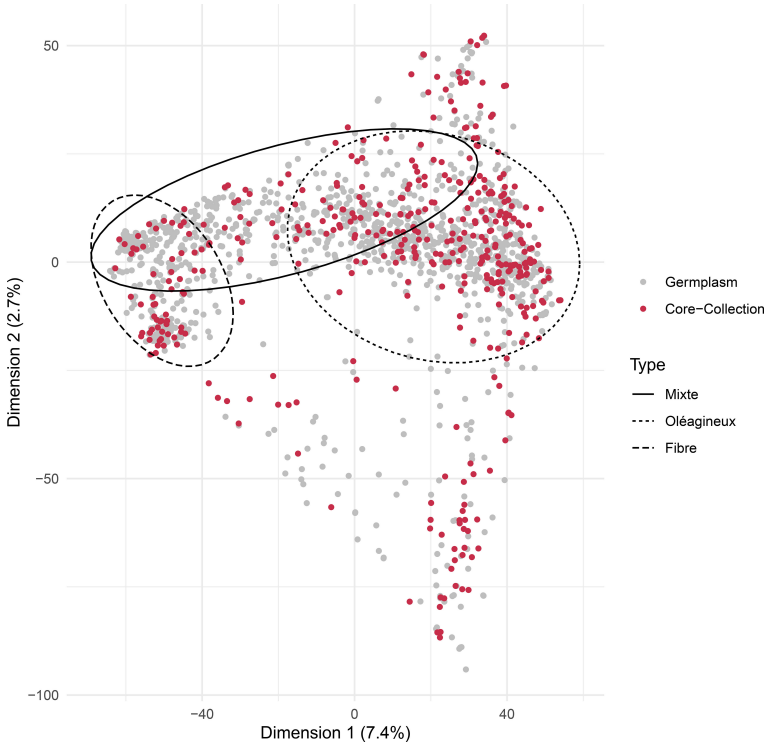

FIGURE 4
Scatter plot of the first two principal components of the PCA of the Arvalis flax germplasm (1,593 accessions) highlighting the 409 selected accessions that represent the core collection.

4 Discussion

4.1 Population structure and genetic diversity

The genetic diversity study of the worldwide flax collection revealed a pronounced genetic structure within this germplasm. The primary axis of differentiation was between oilseed flax and fiber flax types as expected, a pattern consistent with previous findings reported in the literature (Fu, 2005; Hoque et al., 2020; Speck et al., 2022). At K = 3, the emergence of a new cluster composed predominantly of winter-type flax lines highlights the differentiation between spring and winter types. This type of structuring has also been observed in previous studies (Uysal et al., 2010; Fu, 2012; Hoque et al., 2020). This cluster predominantly comprised lines tolerant to low temperatures and thus represents a valuable reservoir of alleles for the improvement of other clusters, particularly given its composition of both fiber and oilseed genotypes. For K = 4, the population structure became more refined, with subdivisions reflecting the geographical origin of the accessions. A new cluster composed primarily by south American oilseed lines appeared. In contrast, the fiber flax group, only began to subdivide from K = 5 onwards, distinguishing between Western and Eastern European origins. At K = 6 and for higher K values, further subdivision occurred within the oilseed clusters, notably distinguishing Eastern European flax from Western European and South American oilseed flax. In the K-means clustering procedure, although the Bayesian Information Criterion exhibited a peak at K = 2, it ultimately designated six clusters as the optimal configuration. The investigation of cluster structure could be supplemented by the Bayesian framework of Pritchard et al. (2000) using the STRUCTURE software, thereby reinforcing our results.

Genetic diversity analyses for the whole collection revealed moderate diversity (H e= 0.22), a value similar to that reported by Hoque et al. (2020) who used 6200 SNP markers to analyze the diversity of 350 genotypes. This level of diversity is expected, given that flax possesses an autogamous reproductive system (Hoque et al., 2020). The expected heterozygosity value of the clusters revealed that oilseed flax clusters harbored greater diversity than fiber flax clusters (Table 2). These findings are consistent with those reported in the literature. Hoque et al., 2020 reported seven genetic clusters with only one cluster for fiber type flax. This difference is probably due to the history of domestication and selective breeding focused on specific traits in each type. Oil flax is considered the ancestral form from which fiber flax was derived. During domestication, fiber flax underwent strong selection for traits like stem length and fiber quality, which reduced its genetic diversity compared to oil flax (Xie et al., 2018). Selective breeding for specific fiber traits in fiber flax led to narrowing the gene pool. Genomic studies confirmed that many genes associated with fiber traits in fiber flax showed strong selection signals, further supporting the idea of a genetic bottleneck (Povkhova et al., 2021). Furthermore, fiber flax is cultivated in a more restricted geographic area compared to oilseed flax, thus leading to high selective pressure to adapt varieties to the specific agro-climatic conditions of this growing area. In contrast, oil flax retained more of the original genetic variation because it was selected for a broader range of traits, including oil content and seed characteristics (Jiang et al., 2021). This finding further underscores that oilseed flax lines harbor a more substantial diversity reservoir, a factor that likely accounts for their predominance over fiber flax lines in conservation collections.

4.2 Core collection assessment

In this study, we evaluated a comprehensive suite of core collection development approaches (twenty method x selection-criterion combinations) resulting in 200 core collections of 350 accessions each. Dedicated core collection methods were compared both among themselves and against calibration-population optimization approaches (as for genomic selection calibration methods). Our aim was to assess outcomes from approaches that differ in their input data (e.g., diversity indices, genetic-distance matrices, or kinship matrices) and in their optimization criteria. The underlying algorithms also varied between methods.

First, we observed that the evaluation criteria did not exhibit the same level of variability. Both the CV and RAR criteria showed low variability across methods. All approaches managed to capture most alleles, including the rarest. This limited variability may be attributed to the core collections size. Indeed, with 350 individuals selected, it is more likely to encompass the full allelic diversity of the initial collection of 1,593 individuals. It would be relevant to compare the tested methods using smaller core collections, ranging from 50 to 100 individuals for example. With such reduced sample sizes, differences between methods might become more pronounced. The choice of using 350 individuals was based on studies assessing the statistical power for QTL detection (Hyne and Kearsey, 1995; Charmet, 2000; Vales et al., 2005) but also for practical reasons with the perspective of testing this CC in field experiments.

Secondly, we observed that diversity-related criteria exhibited lower variability across methods compared to representativeness-related criteria. With core collections composed of 350 individuals, genetic diversity is rapidly captured. This sample size corresponds to approximately 22% of the total population, which exceeds the proportion of 10% generally recommended in the literature (Van Hintum et al., 2000). Nevertheless, methods that jointly optimized He with CV, or SH with CV, significantly increased genetic diversity within the core collections compared to other methods.

Among the methods that best preserved representativeness, CV, ANE, ANE_CV, Avg_GRM_self, and CDmean produced core collections that closely mirrored the initial collection. These methods effectively maintained the overall genetic structure. As expected, the methods originally developed for optimizing calibration populations (CDmean and Avg_GRM_self) yielded core collections that were highly representative of the whole collection. Notably, high levels of representativeness can be achieved through different strategies, by maximizing allelic coverage, minimizing the average distance between an accession

and its nearest neighbors (ANE approach), or by optimizing pairwise relatedness between individuals.

Regardless of the approach, a clear trade-off emerged between optimizing representativeness and maximizing the intrinsic diversity of the core collection. Optimizing for diversity can lead to over-representation of rare alleles, which may not reflect the typical characteristics of the full germplasm, thus reducing representativeness (Marita et al., 2000; Franco-Duran et al., 2019). Conversely, optimizing for representativeness may result in a CC that misses rare alleles or unique genotypes, thus underrepresenting the full spectrum of the diversity (De Beukelaer et al., 2018). Some advanced algorithms attempt to balance both objectives, but improving one often comes at the expense of the other, requiring a compromise based on the intended use of the core collection.

In our study, we also used a novel approach to assess the capacity of core collections to capture QTLs. Core collections are particularly valuable for QTL discovery because they harbor extensive genetic diversity and thus represent a rich source of QTLs of interest (Soto-Cerda et al., 2014; McLeod et al., 2023). It is therefore important to determine which optimization criteria yield a core collection that maximizes QTL detection power. We detected on average 225 QTLs out of the 940 simulated on the whole collection, corresponding to approximately 24% overlap in detected QTLs. This reduction in detection, observed regardless of the method employed, can be attributed to the smaller size of the core collections, which diminishes power to detect QTLs with smaller effect sizes. Indeed, numerous studies have demonstrated that QTL detection power is strongly influenced by the population size: larger populations consistently achieve higher detection power revealing more QTLs, especially those with minor effects, whereas small populations often fail to detect these minor QTLs (Vales et al., 2005; Wang et al., 2012; Wang and Xu, 2019; Nwogwugwu et al., 2022). Among the methods tested, the highest number of QTLs detected within a core collection was obtained using the ANE_CV method, where one core collection allowed detecting a total of 299 QTLs. However, this method exhibited high variability in QTL detection rates. Such variability arised because each random seed initiates the selection with a different individual, leading to a distinct ensemble of cluster centers. The ANE method's combination of cluster center representativeness (ensuring thorough coverage of each region in genotype space) and randomized starting points (inducing different traversals of that space) naturally produces subpopulations with unique allele frequency landscapes and linkage patterns. Since QTL detection critically depends on these landscapes and patterns, ANE yields high variability in the sets of QTLs discovered across different core collections. The D optimality criterion method (Dopt) create CC that capture on average the highest number of QTLs, with relatively low variability across replicates giving therefore a more stable QTL detection. The Dopt method selects a subset of individuals that optimally represents the genetic diversity and structure of the whole collection. By maximizing the determinant of the feature matrix (typically the first q principal components of the marker matrix), Dopt ensures that the selected subset spans the broadest possible range of genetic variation. This is crucial for QTL detection because a training set encompassing the full spectrum of genetic diversity increases the likelihood that QTL alleles segregate within the subset, thereby enhancing detection power. Moreover, maximizing the determinant reduces multicollinearity in the design matrix, yielding more precise estimates of marker effects. In general, methods that maximized representativeness, such as Dopt, ANE_CV, and Avg_GRM_self, captured the largest number of simulated QTLs. Any subpopulation that preserves the same underlying genetic structure of the whole collection necessarily has a higher probability of including those QTLs.

Evaluating methods based on multiple indices can make selecting the best approach challenging. To facilitate the identification of an optimal trade-off between diversity, representativeness, and QTL detection efficiency, we computed a composite index. This index integrated the criteria in a comprehensive and balanced manner. Significant differences in index values were observed across methods. In our study, the combined optimization of the Shannon index and the allelic coverage (with equal weighting) yielded the highest average composite index. The Shannon index favors both high allelic richness and evenness in allele frequencies, thereby promoting allelic balance. Balanced allele frequencies increase the likelihood that causal variants (QTLs) segregate at detectable frequencies. The allelic coverage criterion tends to generate populations that are structurally similar to the whole collection, thus enhancing representativeness. These two criteria appear to be complementary. We also observed that methods combining the ENE criterion with other criteria tended to yield lower average composite index values. The ENE criterion maximizes the genetic distances between the selected entry and its neighboring accessions. This tends to select individuals located at the extremes of the diversity cloud. Combinations including ENE did not produce good compromises in our evaluations.

Evaluating the quality of a core collection should, whenever possible, be based on data that were not used in its construction (Van Hintum et al., 2000). In the present study, we employed the full set of SNP markers both to assemble the core collections and to assess their performance. A more impartial evaluation could be achieved by partitioning the marker dataset: one subset of independent, evenly spaced SNPs, would be used to define the core collections, and the remaining marker set would serve exclusively for their validation. This two-step approach would reduce circularity and provide a more rigorous assessment of core collection construction methods.

4.3 A core collection representative of the whole collection for future genetic studies

We selected a core collection of 409 accessions from the 1,593 accessions of the whole collection using a mixed approach based on breeder's expertise and optimization of Shannon diversity index and allelic coverage. Its size would allow it to be used in field experiments, making it suitable for achieving high statistical power in QTL detection studies (Hyne and Kearsey, 1995; Charmet, 2000; Vales et al., 2005). This core collection will allow us to build bridges between our core collection and existing collections; as it includes 107 accessions that are

common with PGRC core collection (Canada), U.S.NPGS (USA) core collection and composite Collection from Guo et al. (2020). Apart from this list of common accessions identified by accession name, we were unable to compare our core collection with others, as genotypic data for the latter was unavailable. We would indeed like to identify whether there are any genetically similar individuals among the 302 other accessions. The availability of this core collection is an important step in the development of new projects aimed at improving marker-assisted selection breeding of new lines in the context of climate change.

5 Conclusion

The diversity analysis of the Arvalis flax germplasm revealed a moderate genetic diversity and a clear genetic split between oilseed and fiber types, with additional clusters reflecting seasonal and geographical variation. When reducing the germplasm to 350 accessions across twenty sampling strategies, most methods captured nearly all alleles but differed substantially in representativeness and QTL detection power. While ANE_CV detected the most QTLs, it showed high variability, and D-optimality offered a more stable and significant recovery. By integrating diversity, representativeness, and QTL-detection into a composite index, the Shannon-index plus allelic coverage (SH + CV) combination emerged as the superior compromise for our case study, maximally balancing genetic richness, representativeness, and trait-discovery potential for GWAS applications. A mixed approach, which included fixing a list of accessions recommended by breeders and selecting with the SH+CV method, allowed us to select a list of 409 accessions that are representative of the whole collection.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

Author contributions

MG: Data curation, Formal Analysis, Methodology, Validation, Writing – original draft. MB: Data curation, Validation, Writing – review & editing, Methodology. FM: Data curation, Validation, Writing – original draft, Methodology. BD: Conceptualization, Data curation, Funding acquisition, Methodology, Supervision, Validation, Writing – original draft.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work has been granted by Plant2Pro® Carnot Institute in the frame of its 2023 call for projects.Plant2Pro® is supported by ANR (agreement #23-CARN-0024-01). It is also part of the "GenoFLAX" project, which is mainly funded by CIPALIN (France) and the "Filière Lin fibre" led by the Arvalis Institute. Open Access funding was provided thanks to ARVALIS Institute.

Acknowledgments

The authors wish to thank, Isabelle CHAILLET (Arvalis Institute) and Christophe PINEAU (Linéa) for their help conceiving the CoreFLAX projet (Genetic diversity analysis of cultivated flax genetic resources and definition of flax core collection). They would also like to thank Adriane ROLLAND and Caroline LAFFRAY from the GenoPAV laboratory (Arvalis Institute) for producing the seedlings and extracting the DNA for the entire collection. The authors would also like to thank Patricia FAIVRE-RAMPANT, Damien HINSINGER and the entire team at the INRAE Unit 'Étude du Polymorphisme des Génomes Végétaux' (EPGV) for their help with the genotyping of the entire collection. Finally, they would like to thank Guillaume BAUCHET (Terre de Lin), Christophe PINEAU (Linéa), and all their teams for helping to select a list of recommended accessions, and for their constructive contributions to discussions about the CoreFLAX project.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2025.1675815/full#supplementary-material

References

Akdemir, D. (2017). Selection of training populations (and other subset selection problems) with an accelerated genetic algorithm (STPGA: an R-package for selection of training populations with a genetic algorithm). *arXiv preprint*. arXiv:1702.08088. doi: 10.48550/arXiv.1702.08088

Akdemir, D., Rio, S., and Isidro y Sánchez, J. (2021). TrainSel: an R package for selection of training populations. *Front. Genet.* 12. doi: 10.3389/fgene.2021.655287

Albrecht, T., Wimmer, V., Auinger, H. J., Erbe, M., Knaak, C., Ouzunova, M., et al. (2011). Genome-based prediction of testcross values in maize. *Theor. Appl. Genet.* 123, 339–350. doi: 10.1007/s00122-011-1587-7

Amadeu, R., Garcia, A. A. F., Munoz, P. R., and Ferrão, L. F. V. (2023). AGHmatrix: genetic relationship matrices in R. *Bioinformatics* 39, 7. doi: 10.1093/bioinformatics/btad445

Atanda, S. A., Olsen, M., Burgueño, J., Crossa, J., Dzidzienyo, D., Beyeneet, Y., et al. (2021). Maximizing efficiency of genomic selection in CIMMYT's tropical maize breeding program. *Theor. Appl. Genet.* 134, 279–294. doi: 10.1007/s00122-020-03696-9

Bandillo, N., Jarquin, D., Song, Q., Nelson, R., Cregan, P., Specht, J., et al. (2015). A population structure and genome-wide association analysis on the USDA soybean germplasm collection. *Plant Genome*. 8 (3), eplantgenome2015.04.0024. doi: 10.3835/plantgenome2015.04.0024

Berg, E. E., and Hamrick, J. L. (1997). Quantification of genetic diversity at allozyme loci. *Can. J. For. Res.* 27, 415–424. doi: 10.1139/x96-195

Berkner, M. O., Jiang, Y., Reif, J. C., and Schulthess, A. W. (2024). Trait-customized sampling of core collections from a winter wheat genebank collection supports association studies. *Front. Plant Sci.* 15. doi: 10.3389/fpls.2024.1451749

Bianchi, D., Brancadoro, L., and De Lorenzis, G. (2020). Genetic diversity and population structure in a Vitis spp. Core collection investigated by SNP markers. *Diversity* 12, 103. doi: 10.3390/d12030103

Blaringhem, L. (1926). Études sur la Sélection du Lin. *Rev. botanique appliquée d'agriculture coloniale*, 193–204. doi: 10.3406/jatba.1926.4400

Breseghello, F., and Sorrells, M. E. (2006). Association mapping of kernel size and milling quality in wheat (Triticum aestivum L.) cultivars. *Genetics* 172, 1165–1177. doi: 10.1534/genetics.105.044586

Brown, A. H., and Clegg, M. T. (1983). Isozyme assessment of plant genetic resources. *Isozymes* 11, 285–295.

Brown, A. H. D., Frankel, O. H., and Marshall, D. R. (1989). "The case for core collections," in *The use of plant genetic resources*. Ed. J. T. Williams (Cambridge University Press, Cambridge), 136–156. 19901610664, English, Book chapterUK, 9780521368865.

Browning, B. L., and Browning, S. R. (2016). Genotype imputation with millions of reference samples. *Am. J. Hum. Genet.* 98, 116–126. doi: 10.1016/j.ajhg.2015.11.020

Browning, S. R. (2008). Missing data imputation and haplotype phase inference for genome-wide association studies. *Hum. Genet.* 124, 439–450. doi: 10.1007/s00439-008-0568-7

Chandrawati,, Singh, N., Kumar, R., Kumar, S., Singh, P. K., Yadav, V. K, et al. (2017). Genetic diversity, population structure and association analysis in linseed (*Linum usitatissimum* L.). *Physiol. Mol. Biol. Plants* 23, 207–219. doi: 10.1007/s12298-016-0408-5

Charmet, G. (2000). Power and accuracy of QTL detection: simulation studies of one-QTL models. *Agronomie* 20, 309–323. doi: 10.1051/agro:2000129

Charmet, G., and Balfourier, F. (1995). The use of geostatistics for sampling a core collection of perennial ryegrass populations. *Genet. Resour Crop Evol.* 42, 303–309. doi: 10.1007/BF02432134

Chen, H., Matthew, P., and Duy, T. (2020). Gmmat: Generalized linear mixed modelassociation tests version 1.3. 2 (Houston, TX). Available online at: https://github.com/cran/GMMAT.

Cramér, H. (1999). Mathematical Methods of Statistics (PMS-9). Princeton University Press. Available online at: http://www.jstor.org/stable/j.ctt1bpm9r4.

De Beukelaer, H., Davenport, G. F., and Fack, V. (2018). Core Hunter 3: flexible core subset selection. *BMC Bioinf.* 19, 203. doi: 10.1186/s12859-018-2209-z

Demenou, B. B., Ndar, A., Pineau, C. P., Hinsinger, D. D., Marande, W., Hourcade, D., et al. (2025b). Chromosome-scale assembly of European flax (*Linum usitatissimum* L.) genotypes and pangenomic analysis provide genomic tools to improve breeding (Research Square). Preprint (Version 1).doi: 10.21203/rs.3.rs-6065803/v1

Demenou, B. B., Pineau, C. P., Le Clainche, I., Berard, B., Faivre-Rampant, P., and Hinsinger, D. D. (2025). High-throughput SNP discovery, development and validation of a 30 K target SNP genotyping tool for cultivated flax (Linum usitatissimum) breeding and germplasm characterization. *bioRxiv*. doi: 10.1101/2025.09.24.678307

Devlin, B., and Roeder, K. (1999). Genomic control for association studies. *Biometrics*. 55 (4), 997–1004. doi: 10.1111/j.0006-341X.1999.00997.x

Diederichsen, A. (2007). *Ex situ* collections of cultivated flax (*Linum usitatissimum* L.) and other species of the genus *Linum* L. *Genet. Resour Crop Evol.* 54, 661–678. doi: 10.1007/s10722-006-9119-z

Diederichsen, A., Kusters, P. M., Kessler, D., Bainas, Z., and Gugel, R. K. (2013). Assembling a core collection from the flax world collection maintained by Plant Gene Resources of Canada. *Genet. Resour Crop Evol.* 60, 1479–1485. doi: 10.1007/s10722-012-9936-1

Doré, C., and Varoquaux, F. (2006). Histoire et amélioration de cinquante plantes cultivées. Institut national de la recherche agronomique, Collection Savoir Faire. Edition Quae, 812 p.

Esquinas-Alcázar, J. (2005). Protecting crop genetic diversity for food security: political, ethical and technical challenges. *Nat. Rev. Genet.* 6, 946–953. doi: 10.1038/nrg1729

Fernández-González, J., Akdemir, D., and Isidro y Sánchez, J. (2023). A comparison of methods for training population optimization in genomic selection. *Theor. Appl. Genet.* 136, 3. doi: 10.1007/s00122-023-04265-6

Fisher, R. A.. (1921). On the "Probable Error" of a Coefficient of Correlation Deduced from a Small Sample. *Metron*. 1, 3–32.

Franco, J., Crossa, J., Taba, S., and Shands, H. (2003). A multivariate method for classifying cultivars and studying group× environment× trait interaction. *Crop Sci.* 43, 1249–1258. doi: 10.2135/cropsci2003.1249

Franco, J., Crossa, J., Warburton, M. L., and Taba, S. (2006). Sampling strategies for conserving maize diversity when forming core subsets using genetic markers. *Crop Sci.* 46, 854–864. doi: 10.2135/cropsci2005.07-0201

Franco-Duran, J., Crossa, J., Chen, J., and Hearne, S. J. (2019). The impact of sample selection strategies on genetic diversity and representativeness in germplasm bank collections. *BMC Plant Biol.* 19, 520. doi: 10.1186/s12870-019-2142-y

Frankel, O. H. (1984) Genetic Perspectives of Germplasm Conservation. In: Arber, W., Illmensee, K., Peacock, W.J. and Starlinger, P., Eds., Genetic Manipulation: Impact on Man and Society. (Cambridge, England: Cambridge University Press), 161–170.

Fu, Y.-B. (2005). Geographic patterns of RAPD variation in cultivated flax. *Crop Sci.* 45, 1084–1091. doi: 10.2135/cropsci2004.0345

Fu, Y.-B. (2006). Redundancy and distinctness in flax germplasm as revealed by RAPD dissimilarity. *Plant Genet. Resour.* 4, 117–124. doi: 10.1079/PGR2005106

Fu, Y.-B. (2012). Population-based resequencing revealed an ancestral winter group of cultivated flax: implication for flax domestication processes. *Ecol. Evol.* 2, 622–635. doi: 10.1002/ece3.101

Fu, Y.-B. (2025). Flax domestication processes as inferred from genome-wide SNP data. *Sci. Rep.* 15, 8731. doi: 10.1038/s41598-025-89498-9

Gao, X., Starmer, J., and Martin, E. R. (2008). A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genet. Epidemiol.* 32, 361–369. doi: 10.1002/gepi.20310

Goudet, J. (2005). Hierfstat, a package for R to compute and test hierarchical F-statistics. *Mol. Ecol. Notes* 5, 184–186. doi: 10.1111/j.1471-8286.2004.00828.x

Goudet, J., Kay, T., and Weir, B. S. (2018). How to estimate kinship. *Mol. Ecol.* 27, 4121–4135. doi: 10.1111/mec.14833

Gouesnard, B., Batallion, T. M., Decoux, G., Rozale, C., Schoen, D. J., and David, J. L. (2001). MSTRAT: an algorithm for building germ plasm core collections by maximizing allelic or phenotypic richness. *J. Heredity* 92, 93–94. doi: 10.1093/jhered/92.1.93

Guo, D., Jiang, H., Yan, W., Yang, L., Ye, J., Wang, Y., et al. (2020). Resequencing 200 flax cultivated accessions identifies candidate genes related to seed size and weight and reveals signatures of artificial selection. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.01682

Hammer, K., Arrowsmith, N., and Gladis, T. (2003). Agrobiodiversity with emphasis on plant genetic resources. *Naturwissenschaften* 90, 241–250. doi: 10.1007/s00114-003-0433-4

Hoque, A., Fiedler, J. D., and Rahman, M. (2020). Genetic diversity analysis of a flax (*Linum usitatissimum* L.) global collection. *BMC Genomics* 21, 557. doi: 10.1186/s12864-020-06922-2

Hu, J., Zhu, J., and Xu, H. (2000). Methods of constructing core collections by stepwise clustering with three sampling strategies based on the genotypic values of crops. *Theor. Appl. Genet.* 101, 264–268. doi: 10.1007/s001220051478

Huang, X., and Han, B. (2014). Natural variations and genome-wide association studies in crop plants. *Annu. Rev. Plant Biol.* 65, 531–551. doi: 10.1146/annurev-arplant-050213-035715

Hyne, V., and Kearsey, M. J. (1995). QTL analysis: further uses of 'marker regression'. *Theoret. Appl. Genet.* 91, 471–476. doi: 10.1007/BF00222975

Jansen, J., and Van Hintum, T. (2007). Genetic distance sampling: a novel sampling method for obtaining core collections using genetic distances with an application to cultivated lettuce. *Theor. Appl. Genet.* 114, 421–428. doi: 10.1007/s00122-006-0433-9

Jiang, H., Guo, D., Ye, J., Gao, Y., Liu, H., Wang, Y., et al. (2021). Genome-wide analysis of genomic imprinting in the endosperm and allelic variation in flax. *Plant J.* 107, 1697–1710. doi: 10.1111/tpj.15411

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

Kim, K. W., Chung, H. K., Cho, G. T., Ma, K. H., Chandrabalan, D., Gwag, J. G., et al. (2007). PowerCore: a program applying the advancedMstrategy with a heuristic search for establishing core sets. *Bioinformatics*. 23 (16), 2155–2162. doi: 10.1093/bioinformatics/btm313

Kim, S., Kim, D. S., Moyle, H., and Heo, S. (2023). ShinyCore: An R/Shiny program for establishing core collection based on single nucleotide polymorphism data. *Plant Methods* 19, 106. doi: 10.1186/s13007-023-01084-0

Kullback, S., and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statistics Ann. Math. Statist* 22, 79–86. doi: 10.1214/aoms/1177729694

Laloë, D. (1993). Precision and information in linear models of genetic evaluation. *Genet. Sel Evol.* 25, 557. doi: 10.1186/1297-9686-25-6-557

Lê, S., Josse, J., and Husson, F. (2008). FactoMineR: an R package for multivariate analysis. *J. Stat. Software* 25, 1–18. doi: 10.18637/jss.v025.i01

Marita, J. M., Rodriguez, J. M., and Nienhuis, J. (2000). Development of an algorithm identifying maximally diverse core collections. *Genet. Resour. Crop Evol.* 47, 515–526. doi: 10.1023/A:1008784610962

McLeod, L., Barchi, L., Tumino, G., Tripodi, P., Salinier, J., Gros, C., et al. (2023). Multi-environment association study highlights candidate genes for robust agronomic quantitative trait loci in a novel worldwide Capsicum core collection. *Plant J.* 116, 1508–1528. doi: 10.1111/tpj.16425

Meyer, H. V., and Birney, E. (2018). PhenotypeSimulator: A comprehensive framework for simulating multi-trait, multi-locus genotype to phenotype relationships. *Bioinformatics* 34, 2951–2956. doi: 10.1093/bioinformatics/bty197

Mohammadi, S. A., and Prasanna, B. M. (2003). Analysis of genetic diversity in crop plants—salient statistical tools and considerations. *Crop Sci.* 43, 1235–1248. doi: 10.2135/cropsci2003.1235

Nicolas, S. D., Péros, J. P., Lacombe, T., Launay, A., Le Paslier, M. C., Bérard, A., et al. (2016). Genetic diversity, linkage disequilibrium and power of a large grapevine (*Vitis vinifera* L) diversity panel newly designed for association studies. *BMC Plant Biol.* 16, 74. doi: 10.1186/s12870-016-0754-z

Nwogwugwu, C. P., Kim, Y., Cho, S., Roh, H. J., Cha, J., Lee, S. H., et al. (2022). Optimal population size to detect quantitative trait loci in Korean native chicken: a simulation study. *Anim. Biosci.* 35, 511–516. doi: 10.5713/ab.21.0195

Odong, T. L., Jansen, J., van Eeuwijk, F. A., and van Hintum, T. J. L. (2013). Quality of core collections for effective utilization of genetic resources review, discussion and interpretation. *Theor. Appl. Genet.* 126, 289–305. doi: 10.1007/s00122-012-1971-y

Ou, J. H., and Liao, C. T. (2019). Training set determination for genomic selection. *Theor. Appl. Genet.* 132, 2781–2792. doi: 10.1007/s00122-019-03387-0

Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PloS Genet.* 2, e190. doi: 10.1371/journal.pgen.0020190

Povkhova, L. V., Melnikova, N. V., Rozhmina, T. A., Novakovskiy, R. O., Pushkova, E. N., Dvorianinova, E. M., et al. (2021). Genes associated with the flax plant type (Oil or fiber) identified based on genome and transcriptome sequencing data. *Plants* 10, 2616. doi: 10.3390/plants10122616

Price, A., Zaitlen, N., Reich, D., and Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.* 11, 459–463. doi: 10.1038/nrg2813

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959. doi: 10.1093/genetics/155.2.945

Pszczola, M., Strabel, T., Mulder, H. A., and Calus, M. P. L. (2012). Reliability of direct genomic values for animals with different relationships within and to the reference population. *J. Dairy Sci.* 95, 389–400. doi: 10.3168/jds.2011-4338

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Reif, J. C., Melchinger, A. E., and Frisch, M. (2005). Genetical and mathematical properties of similarity and dissimilarity coefficients applied in plant breeding and seed bank management. *Crop Sci.* 45, 1–7. doi: 10.2135/cropsci2005.0001

Rincent, R., Laloë, D., Nicolas, S., Altmann, T., Brunel, D., Revilla, P., et al. (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, 715–728. doi: 10.1534/genetics.112.141473

Rogers, J. S. (1972). "Measures of genetic similarity and genetic distance," in *Studies in Genetics VII* (University of Texas Publication, Austin), 145–153. 7213.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Tech. J.* 27, 379–342. doi: 10.1002/j.1538-7305.1948.tb01338.x

Soto-Cerda, B. J., Diederichsen, A., Ragupathy, R., and Cloutier, S. (2013). Genetic characterization of a core collection of flax (Linum usitatissimum L.) suitable for association mapping studies and evidence of divergent selection between fiber and linseed types. *BMC Plant Biol.* 13, 78. doi: 10.1186/1471-2229-13-78

Soto-Cerda, B. J., Duguid, S., Booker, H., Rowland, G., Diederichsen, A., and Cloutier, S.. (2014). Association mapping of seed quality traits using the Canadian flax (*Linum usitatissimum* L.) core collection. *Theor. Appl. Genet.* 127, 881–896. doi: 10.1007/s00122-014-2264-4

Speck, A., Trouvé, J. P., Enjalbert, J., Geffroy, V., Joets, J., and Moreau, L. (2022). Genetic architecture of powdery mildew resistance revealed by a genome-wide association study of a worldwide collection of flax (*Linum usitatissimum L.*). *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.871633

Stekhoven, D. J., and Bühlmann, P. (2012). MissForest, non-parametric missing value imputation for mixed-type data, (2012). *Bioinformatics* 28, 112–118. doi: 10.1093/bioinformatics/btr597

Smýkal, P., Bačová-Kerteszová, N., Kalendar, R., Corander, J., Schulman, A. H., Pavelek, M., et al. (2011). Genetic diversity of cultivated flax (Linum usitatissimum L.) germplasm assessed by retrotransposon-based markers. *Theor Appl Genet*, 122, 1385–1397. doi: 10.1007/s00122-011-1539-2

Tanksley, S. D., and McCouch, S. R. (1997). Seed banks and molecular maps: unlocking genetic potential from the wild. *Science* 277, 1063–1066. doi: 10.1126/science.277.5329.1063

Thachuk, C., Crossa, J., Franco, J., Dreisigacker, S., Warburton, M., and Davenport, G. F.. (2009). Core Hunter: an algorithm for sampling genetic resources based on multiple genetic measures. *BMC Bioinf.* 10, 243. doi: 10.1186/1471-2105-10-243

Thormann, I., Gaisberger, H., Mattei, F., Snook, L., and Arnaud, E. (2012). Digitization and online availability of original collecting mission data to improve data quality and enhance the conservation and use of plant genetic resources. *Genet. Resour Crop Evol.* 59, 635–644. doi: 10.1007/s10722-012-9804-z

Uysal, H., Fu, Y. B., Kurt, O., Peterson, G. W., Diederichsen, A., and Kusters, P. (2010). Genetic diversity of cultivated flax (*Linum usitatissimum* L.) and its wild progenitor pale flax (*Linum bienne* Mill.) as revealed by ISSR markers. *Genet. Resour Crop Evol.* 57, 1109–1119. doi: 10.1007/s10722-010-9551-y

Vales, M. I., Schön, C. C., Capettini, F., Chen, X. M., Corey, A. E., and Mather, D. E.. (2005). Effect of population size on the estimation of QTL: a test using resistance to barley stripe rust. *Theor. Appl. Genet.* 111, 1260–1270. doi: 10.1007/s00122-005-0043-y

van Hintum, T. J. L., Brown, A. H. D., Spillane, C., and Hodgkin, T. (2000). Core collections of plant genetic resources. In IPGRI Technical Bulletin No.3. (Rome, Italy: International Plant Genetic Resources Institute). 48.

Vanraden, P. M. (2008). Efficient methods to compute genomic predictions. *J. dairy Sci.* 91, 4414–4423. doi: 10.3168/jds.2007-0980

Wang, H., Smith, K. P., Combs, E., Blake, T., Horsley, R. D., and Muehlbauer, G. J.. (2012). Effect of population size and unbalanced data sets on QTL detection using genome-wide association mapping in barley breeding germplasm. *Theor. Appl. Genet.* 124, 111–124. doi: 10.1007/s00122-011-1691-8

Wang, M., and Xu, S. (2019). Statistical power in genome-wide association studies and quantitative trait locus mapping. *Heredity* 123, 287–306. doi: 10.1038/s41437-019-0205-3

Xie, D., Dai, Z., Yang, Z., Tang, Q., Sun, J., Yang, X., et al. (2018). Genomic variations and association study of agronomic traits in flax. *BMC Genomics* 19, 512. doi: 10.1186/s12864-018-4899-z

Yamamoto, M., Sugiyama, T., Murakami, H., and Sakaori, F. (2007). Correlation analysis of principal components from two populations. *Comput. Stat Data Anal.* 51, 4707–4716. doi: 10.1016/j.csda.2006.08.034

Yin, L., Zhang, H., Tang, Z., Xu, J., Yin, D., Zhang, Z., et al. (2021). rMVP: A memory-efficient, visualization-enhanced, and parallel-accelerated tool for genomewide association study. *Genomics Proteomics Bioinf.* 19, 619–628. doi: 10.1016/j.gpb.2020.10.007

You, F. M., Jia, G., Xiao, J., Duguid, S. D., Rashid, K. Y., and Booker HM and Cloutier, S. (2017). Genetic variability of 27 traits in a core collection of flax (*Linum usitatissimum* L.). *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.01636

Yu, J., Pressoir, G., Briggs, W., Bi, I. V., Yamasaki, M., Doebley, J. F., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* 38, 203–208. doi: 10.1038/ng1702

Zhao, K., Aranzana, M. J., Kim, S., Lister, C., Shindo, C., Tang, C., et al. (2007). An arabidopsis example of association mapping in structured samples. *PloS Genet.* 3, e4. doi: 10.1371/journal.pgen.0030004