

From core to target collections: genomic prioritization of trait-associated diversity

Alejandro Thérèse Navarro

`alejandro.therese.navarro@rijkszwaa.nl`

Wageningen University & Research

Wenhui Zhou

University of Chinese Academy of Sciences

Giorgio Tumino

Wageningen University & Research

Erik Wijnker

Wageningen University & Research

Tong Wei

BGI Research

Theo van Hintum

Wageningen University & Research

Research Article

Keywords: Genbank, genetic resources, core collections, lettuce, haplotype diversity

Posted Date: October 26th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11175156/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Genebanks preserve extensive genetic diversity, yet their effective utilization is limited by incomplete characterization of accessions. Core collections provide small, manageable subsets intended to represent overall diversity. However, conventional approaches that maximize genome-wide variation may fail to capture diversity at loci underlying traits of interest. In this study, we explore genomic strategies to construct trait-focused *target collections* that prioritize diversity within predefined genomic regions. Using whole-genome sequence data from 1,229 lettuce (*Lactuca spp.*) accessions, we demonstrate that such approaches can effectively prioritize diversity in target regions, albeit at the expense of genome-wide representation. Using a haplotype-based measure of genetic diversity, we reveal substantial heterogeneity in diversity across the genome. Both the number of haplotypes and the size of the target region emerge as key determinants of target collection performance and size requirements. We further show that genetic optimization of target collections preferentially selects accessions carrying rare haplotypes, resulting in an overrepresentation of rare individuals, which may not align with breeding or experimental objectives. Consequently, careful selection of the initial accession set is essential for constructing practically relevant target collections. We propose target collections as a complementary approach to traditional core collections, particularly for trait-focused applications such as allele mining and pre-breeding. By enabling targeted sampling of genetic variation, these methods can improve the efficiency and accessibility of genebank resources and support a broader range of users in exploiting genomic diversity.

Key Message

Target collections complement traditional core collections by trading off genome-wide diversity in favor of increased haplotype completeness in trait-relevant genomic regions, with applications in allele-mining and breeding.

Introduction

Genebank collections are crucial repositories of genetic diversity with a main task of maintaining, characterizing and distributing genetic material to users. Management of such diversity is challenging, partly because germplasm characterization is far more cost-intensive than maintenance, resulting in large collections of poorly characterized material. Furthermore, collection size does not necessarily reflect the amount of genetic diversity conserved, as highly similar and redundant accessions are often included in the collection (Anglin et al., 2025; McCouch et al., 2012). Consequently, genebanks face a substantial challenge in making the genetic and phenotypic diversity within increasingly large collections accessible to users.

One of the earliest responses to this challenge was the development of core collections (CCs), defined as subsets intended to represent the diversity present in the entire collection (Brown, 1989; Frankel, 1984). Because phenotyping large collections is often prohibitively expensive, users can typically

evaluate a limited fraction of the material. A CC provides a pre-defined set of representative material. Since their introduction, substantial effort has been devoted to defining what constitutes a representative subset and how such subsets should be constructed (Bari et al., 2011; Bataillon & David, 1996; Endresen et al., 2012; Jansen & Van Hintum, 2007; Kim et al., 2023; Odong et al., 2012). Depending on their intended use, CCs may be designed to maximize similarity with the original collection, maximize genetic diversity among selected accessions, or maximize allelic richness. While these strategies differ in their interpretation of diversity and representativeness, they share a common objective: providing users with a manageable subset that maximizes the likelihood of identifying useful genetic variation for breeding, research or conservation purposes.

The rapid development of genomic technologies has fundamentally changed the opportunities available for characterization of genebank collections (Anglin et al., 2025; Mascher et al., 2019; McCouch et al., 2012; Odong et al., 2012; Singh et al., 2019; Wambugu et al., 2018). High-density marker sets and whole-genome sequencing enable the identification and classification of material, the detection of duplicate accessions, determination of geographic origin of diversity, highlighting of underrepresented taxa in genebank collections and the identification of loci determining crucial agronomic traits. As costs continue to decrease, genomic characterization data will increasingly become available for genebank accessions.

Despite these opportunities, genomic information does not automatically translate into improved utilization of genebank collections. Generation, storage and analysis of large genomic datasets require specific infrastructure and expertise that are not always available to genebank users. Genebanks are typically well positioned and could perform analyses to integrate the available genomic and phenotypic information for the benefit of the genebank and its users, a transition that would add data stewardship as a new role for genebanks (Mascher et al., 2019). Many breeders, conservationists and applied researchers are primarily interested in efficiently utilizing biological material rather than analysing and processing genomic data. Consequently, approaches that translate genomic information into readily usable germplasm become increasingly important, particularly in the context of ongoing discussions regarding access and benefit sharing of digital sequence information (Brink & van Hintum, 2022).

Genomic information can be readily used to create CCs through the estimation of pair-wise genetic distances. Historically, genetic distances were based on passport or phenotype data, but nowadays high-throughput genotyping is more often used to obtain genetic distances. Regardless of the source of data, different criteria have been proposed to define desirable subsets. They can be divided into 1) maximizing the representation of the original dataset, a strategy known as A-NE (Accession – Nearest Entry) (Odong et al., 2012); 2) maximizing the genetic diversity of selected accessions, the E-NE criterion (Entry – Nearest Entry) (Odong et al., 2012); 3) maximizing the diversity of the original collection found in the subset collection, known as the M strategy (Bataillon & David, 1996). Most current approaches using any of these criteria assume that diversity throughout the genome contributes equally to the diversity of the resulting collection. However, the objectives of many users are not necessarily aligned with maximizing genome-wide diversity.

A target collection is a subset of accessions selected to maximize genetic diversity within predefined genomic regions rather than across the genome as a whole. Such collections can be of considerable value to multiple genebank users. Because many traits are associated with specific genomic loci, researchers and breeders may benefit more from collections enriched for diversity in these regions, than from those optimized for genome-wide diversity. Breeders seeking to broaden variation for a particular trait could access a wider range of relevant alleles while evaluating the same number of individuals. Likewise, allele mining studies would benefit from a minimal set of accessions containing the greatest allelic diversity within regions of interest. Genome-wide association studies may also gain statistical power from increased diversity at loci underlying the investigated trait. Target regions can be defined using prior biological knowledge, including candidate genes, QTL intervals, or coding regions known or suspected to influence the phenotype of interest. This shifts the objective from constructing generally representative collections towards developing diversity sets tailored to specific research or breeding goals. While trait-focused collections have previously been proposed, these approaches primarily relied on phenotypic information for accession selection (Berkner et al., 2024; dos Santos et al., 2023; Pessoa-Filho et al., 2010; Upadhyaya et al., 2011). In contrast, target collections are optimized directly on genomic diversity within biologically relevant regions, aligning collection design with the genetic architecture of the traits being studied.

Genetic diversity can be quantified at different levels, ranging from individual markers to larger combinations of linked variants. In recent years, increasing attention has been given to haplotypes, defined as groups of linked alleles that are inherited together and therefore represent the local genomic context in which genetic variation occurs (Meuwissen et al., 2001; Tewhey et al., 2011). Haplotypes can provide a biologically meaningful representation of diversity beyond individual marker polymorphisms. Haplotype-based approaches have consequently become increasingly common in crop genetics, breeding and genomic prediction studies, particularly for the identification and utilization of favorable genetic variation (Bhat et al., 2021; Jia et al., 2013; Sivabharathi et al., 2024). From the perspective of genebank management, haplotypes also offer an intuitive measure of diversity coverage by enabling the direct estimation of the proportion of genomic variation that is represented in a subset of a collection. Rather than focusing exclusively on genetic distance between accessions, collection subsets can be evaluated according to their ability to capture the haplotypic diversity present within genomic regions of interest. This perspective is particularly relevant for target collections where the objective is not necessarily to maximize genome-wide diversity, but rather to maximize diversity coverage at loci underlying traits of interest.

In this study, we have explored several methodologies that utilize genomic information and prior knowledge on regions of interest to create target collections from a large, genomically characterized lettuce collection from the Center for Genomic Resources of the Netherlands (CGN). We compared collections selected using genetic distances based on the entire genome or trait-linked loci (i.e. target collections), as well as haplotype-maximization strategies designed to maximize haplotype coverage within the whole genome or at target loci. All collections were evaluated based on their haplotype coverage at the whole genome and target loci. Through these analyses, we aim to develop approaches

that make use of genomic data to produce usable collections that incorporate insights from genomic analyses, thereby supporting genebanks in their role as providers of diversity for breeding and research.

Materials and Methods

Plant Material and Genotyping

A total of 1229 lettuce accessions from the genebank collection of the Center for Genetic Resources (CGN) were used for this study. These accessions represent a wide range of genetic diversity within cultivated lettuce. This collection comprised 1014 *Lactuca sativa* accessions, 196 *Lactuca serriola* accessions, as well as 15 hybrid accessions labelled as “sativa x serriola”. Note that in several of our analyses only the sequence information of cultivated accessions was used (*L. sativa*) as this is a much more homogeneous group. Cultivated lettuce (*L. sativa*) includes several lettuce types: 284 butterhead, 201 romaine or cos lettuce, 250 Batavian or crisp, 183 cutting, 53 latin, 15 oilseed and 29 stalk. A large portion of cultivated accessions represent cultivars (70%) while the rest are landraces, research lines and other types of material. Most material originated from Europe (69%), complemented with accessions sampled from Asia (18%), the Americas (9%), Africa (4%) and Oceania (0.3%). A complete overview of used accessions can be found in Supplementary Data 2.

Plant material was collected in 2019 and DNA extraction and sequencing was performed at the Beijing Genomics Institute (BGI). The genotyping procedure was based on variant detection against the reference genome *Lactuca sativa* Salinas v8 (Reyes-Chin-Wo et al., 2017) which produced ~ 18M polymorphic positions. A hard filter was applied to retain variants only if their mapping quality was above 50 (MQ > 50) and quality-by-depth was above 30 (QD > 30). The obtained SNPs were further down-sampled by selecting homogeneously distributed markers across the genome, obtaining a set of 2M markers that was used for subsequent analysis (Supplementary Data 1).

Phenotypic data

The accessions used in this study have been phenotyped throughout several years and experimental settings, information that has been collected and unified at the CGN. Each phenotype in this study represents a curated score between 1 and 5 of one or more measurements on one or more trials. For instance, disease resistance assays to different *Bremia lactucae* races were combined by scoring each tested accession as 1 to 5 depending on whether they were the most resistant or most susceptible within their trial. After this, scores across trials were unified based on the interpretation of all the resistance scores: if an accession is resistant against all races it is ranked 5, if it is never resistant 1 and if it is resistant only to some races 2–4 depending on the number of resistant assays. This approach to unify experimental data enabled the generation of a large comprehensive phenotype dataset with 30 different traits (Supplementary Data 2). Of those, only 16 were phenotyped for > 40% of the collection.

Haplotype inference

Haplotypes were defined through a two-step approach using first Beagle 5.4 (version 06Aug24) (Browning et al., 2021) followed by a custom haplotype-grouping approach.

Beagle was used to phase the curated SNP dataset containing ~2M markers in 1229 plant accessions (cultivated and wild material) (Supplementary Data 3). Effective population size was set to 100 ($n_e = 100$), error rate was set to 0.05 ($err = 0.05$), the em parameter was set to false ($em = false$). These parameters were chosen to steer the hidden Markov Model used by Beagle to consider relatively low number of haplotypes (thanks to a low effective population size) with a high number of small variations in haplotypes (thanks to a high error rate). This aligns with our expectation of a relatively small genetic diversity and differentiation, particularly in the *L. sativa* portion of the collection. Nevertheless, haplotype phasing performed using the estimated effective population size and error rate estimated by beagle did not change substantially the results. Additionally, we set the window parameter to 10 ($window = 10$) for computational efficiency and because our interest lies in the local haplotype accuracy rather than entire chromosome phasing.

After obtaining phased markers from Beagle, we divided markers in haploblocks of 50Kb containing a median of 201.8 markers. Within each haploblock, each unique combination of marker genotypes was considered a unique *variant*. Each haploblock contained a median of 186 variants, with a broad variation (from 5 to 1821). Nevertheless, a large majority of variants were unique, only found once, while a few variants were found in most of the material. Moreover, these unique variants were often highly similar to frequent variants, indicating that unique variants were the result of point mutation events or even genotyping errors. For this reason, we used a variant grouping approach based on the expected number of mismatches given a genotyping error rate e and the number of markers in a haploblock l . Given this, the number of expected mismatches can be calculated as follows:

$$p = 2e(1 - e)$$
$$m = p \cdot l + 1.96\sqrt{l \cdot p \cdot (1 - p)}$$

Where p is the probability of a single-base mismatch between two sequences and m is the upper boundary for expected number of mismatches given a 95% confidence interval. Assuming an error rate of 1% and for an average number of markers per haploblock of 166 this corresponds to 6.8 mismatches. After calculating all mismatches between all observed variants, the threshold m was chosen to define groups of linked variants, a haplogroup. For simplicity, in this study we refer to haplogroups as haplotypes, considering all observed variants within a haplogroup as being represented by the most common variant. The resulting haplogroups identified can be found in Supplementary Data 4).

The implementation of this approach in R can be found in the github repository <https://github.com/alethere/hapgroup>

Detection of trait-specific markers

To identify markers and genomic regions related to specific traits we used an approach based on genome-wide association study (GWAS).

GWAS-based selection of genomic regions was performed using the R package *statgenGWAS* (van Rossum & Kruijer, 2026). This package implements a mixed-model framework, with population structure accounted for using the leave-one-chromosome-out approach. For all available phenotypes a GWAS was conducted using the curated SNP set and significance was established at 3.1×10^{-8} using the Bonferroni adjustment for multiple tests. As not all individuals were phenotyped for all traits, the exact number of markers used per trait differed somewhat, based on which markers segregated within the phenotyped subset. Quantile-quantile plots were also used to assess whether p-values were not inflated. To define genomic regions associated with traits, a linkage disequilibrium (LD) decay study was performed which showed in all chromosomes that 90% of LD decay was achieved at $\sim 1.3\text{Mb}$ in all chromosomes (see Sup. Figure 1). This relatively large LD decay range is surprising when compared to other crops, but has been previously reported in cultivated lettuce (Zhang et al., 2017).

Based on this, whenever two significant markers were at less than 1.3Mb they were considered as part of the same QTL region, then defining all QTL regions for each trait. The large LD decay range meant that in many cases associated regions may have merged multiple associated signals. When a single significant marker was not surrounded by other markers, this was discarded. All markers within a QTL region (significant and non-significant), as well as markers within a 1.3Mb range of the left and right edges of a QTL region were selected as phenotype-related markers and later used to define GWAS-based, trait-specific core collections. For those traits where no significant markers were detected, no set of markers could be defined.

All regions used in this study can be found in Supplementary Data 5.

Collection definition methods

Two main approaches to build collections were used in this study: one based on genomic distances and the other based on maximization of haplotype sampling. For comparison, a randomly chosen collection has been included in the analyses. In all cases, collections were defined only within the *L. sativa* subset of the collection, since including the wild material heavily skewed collection methods to prioritize wild material selection.

To evaluate random target windows, we generated collections including 50 individuals. For the trait-specific collections, we used instead 100 individuals. Less individuals were used for the target windows to reduce the computational burden. All collections of 100 individuals can be found in Supplementary Data 6.

Genomic distance collections

The first approach, which we refer to as the **general distance**, took the entirety of the 2M markers to calculate pairwise distance matrices using Euclidean distance. This can be considered as a core

collection (CC) of the entire genotyped set of accessions.

In the second approach, which we refer to as **target distance**, all markers from trait-specific regions as identified during our GWAS study were used to calculate Euclidean distance. Consequently, it was only possible to obtain such distances for traits that showed a positive GWAS signal (9 traits). Collections defined using this distance method are target collections.

Lastly, we generated an additional set of collections based on **random windows**, which computed genetic distance using all markers within a randomly-chosen genomic window to compute an Euclidean distance. This is equivalent to the GWAS distance, except for the fact that windows are randomly chosen. This is therefore also a target collection.

All distance matrices calculated by any method were afterwards used to compute CCs using the R package Core Hunter 3 (De Beukelaer et al., 2018). In all cases, the average entry-to-nearest-entry algorithm was used with precomputed distances. The collections are named after the distance matrix used to generate them, i.e. general CC, GWAS target collection, and random target collection.

Haplotype maximisation collection

We also used an algorithm to obtain collections by maximizing the number of population haplotypes represented in the collection. Our implementation is based on a maximum set coverage greedy algorithm, using a measure of haplotype contribution we name rarity (r).

Rarity is computed per haplotype first, and then for each individual, by adding the haplotype rarity scores for all its haplotypes. The measure of rarity is also relative to an existing collection, where $r = 1$ if the haplotype is *not* found in the collection, and $r = 0$ if the haplotype is already present in the collection. As such, an individual's rarity will be higher if it contains many haplotypes that are not yet present in the collection.

An algorithm to implement this approach can be described as follows:

1. Set all haplotype rarity scores across all haploblocks as 1.
2. Compute individual rarity scores for all individuals. Initialize the collection by selecting the individual with the highest rarity score (or a random individual if all have equal rarity).
3. Update haplotype rarity scores based on selected individuals in the collection.
4. Re-compute individual rarity scores.
5. Add individual with maximum rarity score to the collection.
6. Repeat steps 3 to 5 n times to achieve a CC of size n .

This algorithm ensures that in every iteration the individual with the most unsampled haplotypes is added to the collection, thus finding a greedy solution to a maximum set coverage problem. It is important to realize that this greedy algorithm is computationally very intensive as it has complexity

$O(nHI)$ where n is the CC size, H is the total number of haplotypes across haploblocks and I is the number of individuals. Such was the case when estimating our optimal collections.

We computed a collection using this algorithm which we refer to as **genome optimal** collection, meaning it is optimal in its haplotype completeness across the entire genome. This collection represents the theoretical maximum of haplotypes that can be achieved in a collection. Similarly to the classical distance-based methods, this approach can also be adapted to perform target collections. Instead of defining rarity scores based on all haplotypes across all haploblocks in the genome, a subset of haploblocks covering a region of interest can be used. Thus we have obtained a haplotype-maximized target collection, which we term the **optimal** collections.

An implementation of this algorithm in R can be found in Supplementary Data 7.

Collection evaluation

All collections generated (Table 1) were evaluated on haplotype completeness. Per haploblock, the total number of haplotypes in the *L. sativa* subset was obtained, excluding unique haplotypes only observed once. The completeness for one haploblock is then the percentage of those population haplotypes found within a collection. The average completeness of a collection was the average of all haploblock completeness measures.

Table 1
Summary of collections made within this study. The following methods were used to produce collections of 50 (in the random-window study) or 100 individuals (in the trait-based analysis).

Collection name	Sampling method	Is a target collection?
Random	Random selection of individuals	No
General collection	Distance-based with CoreHunter using whole genome information	No
Target collection	Distance-based with CoreHunter using only markers within region of interest	Yes
Genome Optimal	Haplotype maximization algorithm for the entire genome	No
Optimal	Haplotype maximization algorithm using only haploblocks within region of interest	Yes

Results

Lactuca haplotype diversity

Genetic diversity within the *Lactuca* collection was largely stratified between *L. sativa* and the wild species (Fig. 1). Such can be seen by the high amount of explained variance (29,46%) by the first

principal component in the PCA in Fig. 1 left. This indicates that *L. serriola* and *L. sativa* populations are clearly genetically differentiated and can be clearly distinguished based on genomic data. It is also clear that *L. serriola* harbors a much larger diversity than cultivated lettuce as illustrated by their differentiation based on PC2. For this reason, we chose to exclude *L. serriola* from the target collection generation, as otherwise *L. serriola* accessions had a large impact on the selection procedure.

Haplotypes were estimated in consecutive windows of 50Kb resulting in 43,366 haploblocks across the genome with an average of 201 markers per haploblock. Several haplotypes were detected per haploblock, but their frequencies varied greatly. In the entire collection (including *L. sativa* and wild species) up to 334 haplotypes were detected per haploblock, with an average of 25.2. Of those, on average 17.2 were confident haplotypes (observed more than once). Analysing haplotype diversity within the *L. sativa* accessions (i.e. excluding wild relatives) reveals a greatly reduced haplotype diversity, with up to 288 haplotypes detected per haploblock, an average of 9.9 haplotypes, of which 5.3 were confident (present in 2 or more alleles). This clearly confirms the expectation that wild species harbour much larger genetic diversity than cultivated *L. sativa*. It is relevant to highlight that of the total detected haplotypes, on average only 2.8 haplotypes per haploblock had a haplotype frequency > 5% in the total collection (2 in the *L. sativa* set). This highlights that typically a few haplotypes are shared across most of the germplasm, while most haplotypes are rare or very rare. Nevertheless, rare haplotypes (especially those observed only once) are less confidently phased, thus it cannot be excluded that they are the result of phasing or genotyping artifacts.

Haplotype diversity was not homogeneously distributed across the genome (Fig. 2), with some regions containing much higher numbers of haplotypes, resulting in a broad spread of haplotype counts across haploblocks. Especially on chromosome 1 and 2 large areas of increased haplotype diversity are present. Smaller areas can also be found on chromosomes 3 and 4, as well as small sections on chromosomes 6 and 8. The regions on chromosomes 1, 2 and 3 coincide with major resistance clusters (MRC) that have been previously identified (Christopoulou, McHale, et al., 2015; Christopoulou, Wo, et al., 2015).

Target collections increase haplotype completeness on target regions

To obtain target collections (TC) that are enriched for genetic diversity at specific regions, we propose computing distance matrices using only markers within those regions. To evaluate this proposed approach, we selected random regions of varying sizes (20Kb, 2Mb, 10Mb) which were then used to calculate 100 TC per window size. This simulates the procedure of selecting genomic regions of interest throughout the genome. Each TC was evaluated by the average haplotype completeness within the target region as well as the haplotype completeness across the entire genome. Additionally, a randomly selected collection, and a general CC using markers from the entire genome were evaluated for haplotype completeness within the selected windows (Fig. 3).

Our analysis shows that TCs have increased average haplotype completeness in the target regions when compared with CCs using the general distance matrix. We compared the within-window completeness of

a random collection, the general CC, and target collections (Fig. 3). In all cases, the target collections showed increased haplotype completeness within windows, with improved completeness with the smallest window sizes. Median completeness for 0.2Mb windows was 0.69, 0.87 and 1 for random, general and target collections respectively. A similar trend was seen with larger window sizes, although the completeness achieved by the target collections was reduced with medians 0.95 and 0.93 for 2Mb and 10Mb respectively, which was nevertheless higher than the random or general collections. On the other hand, TCs produce lower completeness at the whole-genome level. A general CC achieves a completeness of 0.85, and a random CC achieves a completeness of 0.76. TCs reach 0.75, 0.77 and 0.79 for 0.2Mb, 2Mb and 10Mb respectively.

When considering very small target windows, we see a broad variability in haplotype completeness using any method. This variability stems from the fact that target window approaches haploblock size, meaning that completeness is mostly defined by 2 to 3 haploblocks. Similarly, these windows offer the least amount of genetic information about the entire genome, leading to whole-genome haplotype completeness in the same range as a random collection.

Regions selected for trait-specific core collections

We performed a genome-wide association study (GWAS) on 30 traits using the entire *Lactuca* collection. The study revealed significant associations for 14 of 30 traits. Of those 14, 3 were discarded since only one marker was significant and 2 were discarded (*Bremia lactucae* 2–9 and leaf shape wild material) due to QQ-plots indicating inflated p-values. As a result, we could define trait-associated regions using GWAS for 9 traits (Table 2). For each trait there were between 2.2K and 51.6K markers (average 18.7K) belonging to the trait-associated regions.

For each trait, we have evaluated the haplotype completeness within the trait-associated regions of five collections: random, general, target, genome optimal and optimal (Table 1). Of those, only target and optimal are focused on sampling within-window diversity. In all cases, the target method produced equal or better collections than the general method (Fig. 4), confirming the usefulness of region-targeting to achieve higher haplotype completeness. The random collection always showed the lowest haplotype completeness, with median completeness ranging between 0.25 and 0.5. The general collection was somewhat better (0.28 to 0.6), followed by genome optimal (0.32 to 1). The targeted collections using the distance method (target) and the haplotype optimization method (optimal), were better in most cases with median completeness between 0.5-1 and 0.38-1 respectively. In three cases (anthocyanin presence, leaf color and plant diameter) the target collection also outperformed the haplotype-optimal collection. Although unexpected, this is not entirely surprising given that the greedy haplotype optimization algorithm implemented can fail to reach the best possible collection depending on which individual is chosen as the first individual.

Table 2

Summary of trait-associated markers. For each trait (phenotype) the table shows the number of phenotyped individuals (n), the number of regions identified as significantly associated in the GWAS analysis, the number of significant markers (significant) and the number of markers within the significant regions of the GWAS (markers).

Phenotype	n	regions	significant	markers
anthocyanin presence	1046	4	867	9846
<i>Bremia lactucae</i> 20	476	9	455	42982
leaf anthocyanin content	955	22	1165	51582
leaf anthocyanin pattern	168	3	19	5731
leaf blistering	847	1	4	2246
leaf color	842	6	498	11688
leaf division	979	10	379	22054
plant diameter	846	2	101	4055
seed color	938	5	2137	18199

Bremia-specific collection

As an example, we can show in more detail how the different collections capture (haplotype/genetic) diversity for a region that harbors loci conferring resistance to *Bremia*. In this case, two main QTL regions were identified by GWAS and used to select a set of trait-specific markers which were then used to obtain a target collection. We show the completeness along the genome of the three non-target collections (random, general, genomeOptimal) and of the two targeted collections (target, optimal). In all cases, the main driver of haplotype completeness is the underlying number of haplotypes in a given genomic region (Fig. 5). For example, all collections have very high completeness at the beginning of chromosome 3, due to the low general diversity in that region. Similarly, the highly diverse region in the center of chromosome 4 leads to reduced completeness in all collections in the same region.

Within the windows of interests (chr2 and chr4) we can see that the target collection only improves completeness in chr2. The same also happens with the optimal collection when compared with the genome optimal collection. This is due to the size of the two target regions. Since the chr2 region is much larger, and both target and optimal methods count each marker or haploblock with equal weight, the area with more markers or haploblocks will dominate the targeting method. Moreover, the chr2 area coincides with the hypervariable region of chr2, making it very challenging to adequately sample all its diversity within only 100 individuals.

Lastly, the genome-wide sampling of haplotype diversity is much lower in the target collection than in the general collection, with a few regions where the target collection performs similarly to the random

collection. This clearly illustrates the trade-off between targeting diversity at specific genomic regions at the cost of covering diversity across the remainder of the genome.

Discussion

A core collection (CC) was originally proposed as minimally repetitive set of accessions that would best represent the genetic diversity of a crop species (Frankel, 1984). Related definitions focusing more on the operational aspects of genebanks have since been proposed (Van Hintum et al., 2000). Regardless of definition, the CC concept has become an essential part of the function and operation of genebank collections across the world for most crops (Diulgheroff et al., 2025). CCs have been usually constructed institutionally to offer a standardized set of diversity but other subset collections are also made *ad hoc* to satisfy experimental needs of specific research projects (Berkner et al., 2024; Bhullar et al., 2009).

Pragmatic subsets of unmanageably large collections are thus necessary, but each subset is also implicitly constructed with a purpose in mind – like representing entire crop diversity or fitting an experimental setup. One of such purposes is to focus on specific trait diversity such as disease resistance or quality traits, which would be target collections if they are defined using target genomic regions. The methods presented and evaluated in this research build target collections as subsets that maximise genetic diversity for specific loci, with the expectation that these loci may affect traits of interest. In essence, these methods imitate the process of generating a CC but focus on a subset of the genome.

Target collection construction is heavily influenced on the choice of genomic region and the genetic diversity of the region(s) of interest. To evaluate diversity and the effectiveness of genetic diversity sampling, we used haplotype diversity and average completeness measures, which help better understand the distribution of diversity within a collection. Construction and usage of such collections, however, may require additional efforts by current genebanks.

Selection of trait-related regions

In this study we show results of a GWAS-based method to identify trait-linked genomic regions. By selecting all markers within a trait-associated region we target the construction of a collection to the diversity of that region, and thus the diversity of the associated trait. Our results support that using such regions achieves the intended goal of increasing the diversity within regions of interest. Moreover, through our random window experiment we also show that any genomic window (or combinations thereof) can be used to build a target collection, regardless of its association to a specific trait.

We foresee that using genes of interest to determine such regions can also be an interesting approach. We initially tested a candidate gene approach using gene ontology (GO) terms but found that it was not feasible to clearly link arbitrary traits to gene functions and did not proceed with this methodology. Nevertheless, if a few genes can be clearly linked to a trait of interest, markers within and around the area of the gene could be used in a similar way as shown in this study to generate gene-target

collections. Especially when target regions are this small, the number of individuals needed to represent the entire collection's diversity would be very small.

The target collection methodology can therefore be applied regardless of trait association and region, being thus a generalized approach to target genetic diversity at regions of interest without the need of a specific trait association study.

Target region diversity determines target collection size

Many processes influence genetic diversity at specific points in the genome including selection, mutation, recombination and demographic forces (Slotte, 2014). We clearly showed in this research that, specifically for the cultivated lettuce material (the *L. sativa* accessions), there are hypervariable regions. Creating target collections for traits encoded in such regions, such as the *Bremia lactucae* resistance which was mapped to chromosome 2, is more challenging. The simple fact that more diversity is available also highlights that a higher number of accessions will be needed to represent the complete genetic diversity present among the entire genebank collection. In other words, when developing target collections, it is essential to evaluate diversity at the genomic loci of interest in order to determine the appropriate target collection size. For the same reason, hypervariable target regions will also require more individuals to be haplotype complete (i.e. all haploblocks at 100% completeness). Similarly, target collections are more likely to decrease haplotype completeness at the remainder of the genome, as was shown by in the *Bremia* example. This highlights that target collections within defined genomic regions can capture all (or most) variation in a specific region, but should not be understood as collections that capture all genomic diversity.

Targeting small genomic regions simplifies the sampling of genetic diversity. Given that the 50kb regions we studied held on average 17.2 confident haplotypes in the entire collection (5.3 in *L. sativa*), around 17 individuals would be sufficient to create a collection that includes all haplotypes for an average 50kb haploblock. Including also those haplotypes that are unique to a certain accession would increase this number to about 25. Adequate assessment of diversity within the region of interest is nevertheless still required, since we found that on some highly diverse haploblocks more than 300 unique haplotypes were detected.

Distribution of haplotype diversity

Haplotype frequencies followed a strongly skewed distribution characterized by a small number of common haplotypes and a large number of rare haplotypes. Such long-tailed frequency distributions are a well-known feature of allele-frequency spectra, and variations of this pattern can be used to estimate evolutionary processes (Ragsdale et al., 2018). Nevertheless, these frequency distributions are further amplified in genebank collections by unequal sampling intensity among crop types, geographic regions, and species groups. In the case of the lettuce accessions held by the CGN genebank, this can be clearly seen: of the over 1000 *L. sativa* accessions, just 50 are of the latin and oilseed types, even though these are the most diverse groups. As a result, genetically optimised core or target collections will inherently

overrepresent rare accessions containing the least frequent haplotypes, leading to a preferential selection of wild relatives, landraces and minor crop types.

This has especially strong implications for multi-species collections. It is well known, and we confirmed it in our analyses of the lettuce collection, that wild species crop relatives are more genetically diverse than domesticated crops. As a result, a target collection built starting from a multi-species collection will likely favour the introduction of many wild species in lieu of their genetic diversity. While this is desired from a haplotype-completeness evaluation, this is likely not useful in many experimental or breeding contexts, where not only diversity but *useful* diversity is of interest. For instance, if one was to construct a target collection focusing on lettuce leaf texture diversity, would it be sensible to build a target collection including mostly wild species with different degrees of hairiness and thorniness? Evidently, a wide range of leaf texture genetics are present in wild species, but they are not *edible* textures, thus they are not particularly *useful* when investigating leaf texture diversity of edible lettuce. This tension between genetic diversity and useful diversity is crucial when developing target collections. The simplest approach to resolving this tension is to select the collection subset that is most relevant to the target of interest. Alternatively, target collections could be developed independently for different species and joined later, keeping a desired proportion of cultivated and wild species. Nevertheless, it is important to remember that if developed on a purely genetic level, the most haplotype-complete target collection will certainly include many rare individuals (wild species, odd crop types or old landraces).

Core and target collections

Subsets of genebank collections are regularly composed to facilitate management, distribution and use of the material stored in a genebank. Researchers that aim to study a specific trait do not have the capacity to study the entire collection, and given the level of redundancy within collections, studying the entire collection would imply unknowingly testing highly similar material multiple times (Anglin et al., 2025; Mascher et al., 2019; Singh et al., 2019). Our analyses have shown that CCs that maximise diversity across the entire genome do not provide the most genetic diversity at every genomic region. Target collections on the other hand, increase that diversity while sacrificing genome-wide diversity.

Core collections remain a useful tool in genebank management, especially when general diversity is of interest. Several approaches have this focus, including genetic studies that focus on multiple traits at once, or that focus on traits whose genetic determinants are unknown, as well as initiatives to introduce novel genetic diversity within breeding programs with the aim of broadening the genepool. But when CCs are used, one must be aware that such collections likely do not represent all diversity at highly variable sites, even if they harbour interesting genetic variation. This trade-off should be accepted when most of the diversity across the entire genome is desired.

In contrast, target collections are most useful when specific genomic regions are of interest. Particularly when focused on relatively small regions of the genome, target collections have the real possibility to offer complete diversity coverage using a very small number of accessions. This can be a very cost-effective way to research and use genetic diversity of genomic regions of interest. On the other hand, if

the target region becomes increasingly large, at some point the target collection becomes equivalent to a CC. In our study, 10Mb target regions still offered improved haplotype coverage. Judging by our limited data, it is possible that up to 100Mb target regions can still offer a noticeable improvement in haplotype completeness. Regardless of the size limit, one can expect that with a constant number of individuals, target collections aimed at larger genomic regions will necessarily sacrifice more diversity of hypervariable regions, similarly to the limitations of CCs.

Towards genomic decision support in genebanks

The introduction of target collections raises methodological questions, but also highlights new opportunities regarding the future of genebank management. Traditionally, CCs have been a practical mechanism to prioritize resource allocation by concentrating demand on a reduced number of accessions. Target collections are likely to select a different subset of accessions, but potentially covering a broader genetic diversity in regions of interest with less individuals. This may lead to a reduced total seed demand while facilitating a broader spread of demand across many accessions. To avoid high demand across the entire collection, future collection-development strategies may consider maximizing overlap between genome-wide and target-specific collections. Rather than replacing CCs, target collections should be viewed as complementary resources tailored to a specific user's interest, broadening the range of services genebanks can provide.

The development of targeted collections is part of a broader transition occurring within genebanks. The value of genomic characterization for genebank management is increasingly recognized, including applications such as duplicate detection, accession classification, allele discovery and phenotype prediction (Anglin et al., 2025; Mascher et al., 2019; McCouch et al., 2012; Singh et al., 2019; Wambugu et al., 2018). As genomic characterization becomes more widespread, opportunities emerge to provide genomics-supported utilization of genetic resources. Target collections represent one example of an approach to translate genomic information into a practical method that addresses user needs.

The implementation of these approaches will require continued investment in genomic resources and computational infrastructure. While large-scale genotyping and whole-genome sequencing remain significant investments for many collections (Díez et al., 2018; Hanson et al., 2025), sequencing costs continue to decrease, and genomic datasets are becoming increasingly available for major crop species. For this reason, computational infrastructure and bioinformatic expertise should progressively become incorporated into genebank operations. This will allow genebanks to expand their role from custodians of biological resources towards stewards of integrated biological and genomic diversity (Mascher et al., 2019).

Ultimately, the future value of genomic characterization will depend not only on the generation of data but also on the development of mechanisms that convert data into useful resources for breeders, researchers, and conservationists. The approaches presented in this study illustrate how genomic information can be transformed into collections that maximize diversity within genomic regions of interest while maintaining practical collection sizes. As genomic datasets become increasingly common,

such user-oriented genomic services may become a natural extension of the traditional mission of genebanks: facilitating access to genetic diversity for research, breeding and conservation for current and future generations of users.

Declarations

Funding

This work was supported by the Fundamental Research Programme ‘Circular and Climate Neutral’ (KB-34-013-001), which is part of the Programme Genetic Resources (WOT-03), funded by the Dutch Ministry of Agriculture, Fisheries, Food Security and Nature. Author ATN has additionally received financial support from the company Rijk Zwaan B.V.

Competing Interests

Author ATN worked for the breeding company Rijk Zwaan B.V during part of this research and at time of publication. Authors EW and TvH were employed by the Center for Genomic Resources Netherlands, the National Dutch Genebank, during this research and at time of publication. The authors do not foresee any of the research activities carried in this study to have direct financial impact for the institutions mentioned.

Author Contribution

ATN: conceptualized research, performed research, wrote paper. WZ: produced original data, reviewed paper. GT: conceptualized research, performed research, reviewed paper. EW: conceptualized research, reviewed paper, provided funding. TW: produced original data, reviewed paper. TvH: conceptualized research, reviewed paper, provided funding.

Data Availability

The datasets used in this study can be found on the scientific data-sharing platform Zenodo, with the DOI <https://doi.org/10.5281/zenodo.22690557> This includes the following data:

- **Supplementary data 1** (Sup1) contains a full genotype table for all individuals in the collection. This includes >2M marker records for 1229 individuals.
- **Supplementary data 2** (Sup2) contains descriptive data (species classification, origin, common name, date of collection) and aggregated phenotype data (30 traits) for 1229 individuals.
- **Supplementary data 3** (Sup3) contains phased marker data for all individuals in the collection. This is >2M markers, phased in haplotypes 1 and 2 for each individual.

- **Supplementary data 4** (Sup4) contains haplotype-tagged data for all individuals in the collection. Contiguous phased markers, defined by haploblock ranges, were tagged in codes of the type "H1", "H2", "H3", etc. Column definitions are as follows: chrom (chromosome), minpos (left edge of haploblock), maxpos (right edge of haploblock), nmark (number of markers within haploblock), mismatch.exp (expected number of mismatches between variants with the same haplotype label), haplos (number of haplotypes in that block).
- **Supplementary data 5** (Sup5) locations of significant regions identified by GWAS for all traits with positive results.
- **Supplementary data 6** (Sup6) zip file containing all collections (core and target collections) defined in this study.
- **Supplementary data 7** (Sup7) script containing functions with an implementation of a greedy haplotype optimization algorithm to build collections.

References

1. Anglin, N. L., Wenzl, P., Azevedo, V., Lusty, C., Ellis, D., & Gao, D. (2025). Genotyping Genebank Collections: Strategic Approaches and Considerations for Optimal Collection Management. *Plants* 2025, Vol. 14, Page 252, 14(2), 252. <https://doi.org/10.3390/plants14020252>
2. Bari, A., Street, K., Mackay, M., Endresen, D. T. F., de Pauw, E., & Amri, A. (2011). Focused identification of germplasm strategy (FIGS) detects wheat stem rust resistance linked to environmental variables. *Genetic Resources and Crop Evolution* 2011 59:7, 59(7), 1465–1481. <https://doi.org/10.1007/S10722-011-9775-5>
3. Bataillon, T. M., & David, J. L. (1996). Neutral Genetic Markers and Conservation Genetics: Simulated Germplasm Collections. *Genetics*, 144(1), 409–417. <https://doi.org/10.1093/GENETICS/144.1.409>
4. Berkner, M. O., Jiang, Y., Reif, J. C., & Schulthess, A. W. (2024). Trait-customized sampling of core collections from a winter wheat genebank collection supports association studies. *Frontiers in Plant Science*, 15, 1451749. <https://doi.org/10.3389/FPLS.2024.1451749/TEXT>
5. Bhat, J. A., Yu, D., Bohra, A., Ganie, S. A., & Varshney, R. K. (2021). Features and applications of haplotypes in crop breeding. In *Communications Biology* (Vol. 4, Number 1). <https://doi.org/10.1038/s42003-021-02782-y>
6. Bhullar, N. K., Street, K., Mackay, M., Yahiaoui, N., & Keller, B. (2009). Unlocking wheat genetic resources for the molecular identification of previously undescribed functional alleles at the Pm3 resistance locus. *Proceedings of the National Academy of Sciences of the United States of America*, 106(23), 9519–9524. <https://doi.org/10.1073/PNAS.0904152106;PAGE:STRING:ARTICLE/CHAPTER>
7. Brink, M., & van Hintum, T. (2022). Practical consequences of digital sequence information (DSI) definitions and access and benefit-sharing scenarios from a plant genebank's perspective. *PLANTS, PEOPLE, PLANET*, 4(1), 23–32. <https://doi.org/10.1002/ppp3.10201>
8. Brown, A. H. D. (1989). Core collections: a practical approach to genetic resources management. *Genome*, 31(2), 818–824. <https://doi.org/10.1139/g89-144>

9. Browning, B. L., Tian, X., Zhou, Y., & Browning, S. R. (2021). Fast two-stage phasing of large-scale sequence data. *The American Journal of Human Genetics*, 108(10), 1880–1890.
<https://doi.org/10.1016/J.AJHG.2021.08.005>
10. Christopoulou, M., McHale, L. K., Kozik, A., Wo, S. R. C., Wroblewski, T., & Michelmore, R. W. (2015). Dissection of Two Complex Clusters of Resistance Genes in Lettuce (*Lactuca sativa*). *Https://Doi.Org/10.1094/MPMI-06-14-0175-R*, 28(7), 751–765. <https://doi.org/10.1094/MPMI-06-14-0175-R>
11. Christopoulou, M., Wo, S. R. C., Kozik, A., McHale, L. K., Truco, M. J., Wroblewski, T., & Michelmore, R. W. (2015). Genome-wide architecture of disease resistance genes in lettuce. *G3: Genes, Genomes, Genetics*, 5(12), 2655–2669. <https://doi.org/10.1534/G3.115.020818/-/DC1>
12. De Beukelaer, H., Davenport, G. F., & Fack, V. (2018). Core Hunter 3: flexible core subset selection. *BMC Bioinformatics* 2018 19:1, 19(1), 203-. <https://doi.org/10.1186/S12859-018-2209-Z>
13. Díez, M. J., De la Rosa, L., Martín, I., Guasch, L., Cartea, M. E., Mallor, C., Casals, J., Simó, J., Rivera, A., Anastasio, G., Prohens, J., Soler, S., Blanca, J., Valcárcel, J. V., & Casañas, F. (2018). Plant genebanks: Present situation and proposals for their improvement. The case of the Spanish network. *Frontiers in Plant Science*, 871, 423954. <https://doi.org/10.3389/FPLS.2018.01794/TEXT>
14. Diulgheroff, S., Furman, B. J., Engels Johannes M M, Ebert, A., Gullotta, G., & Sharock, S. (2025). The state of ex situ conservation. In *The Third Report on The State of the World's Plant Genetic Resources for Food and Agriculture*. FAO. <https://doi.org/10.4060/cd4711en>
15. dos Santos, C. C., de Andrade, L. R. B., do Carmo, C. D., & de Oliveira, E. J. (2023). The Development of Thematic Core Collections in Cassava Based on Yield, Disease Resistance, and Root Quality Traits. *Plants*, 12(19), 3474. <https://doi.org/10.3390/PLANTS12193474/S1>
16. Endresen, D. T. F., Street, K., Mackay, M., Bari, A., Amri, A., de Pauw, E., Nazari, K., & Yahyaoui, A. (2012). Sources of resistance to stem rust (Ug99) in bread wheat and durum wheat identified using focused identification of germplasm strategy. *Crop Science*, 52(2), 764–773.
<https://doi.org/10.2135/CROPSCI2011.08.0427;ISSUE:ISSUE:DOI>
17. Frankel, O. (1984). Genetic perspectives of germplasm conservation. In W. K. Arber, K. Llimensee, W. J. Peacock, & P. Stralinger (Eds.), *Genetic Manipulation: Impact on Man and Society* (pp. 161–170).
18. Hanson, J., Jamora, N., & Bramel, P. (2025). Identifying gaps and challenges for conserving crop diversity in genebanks: Lessons from the reviews of national genebanks. *Plant Genetic Resources*, 1–13. <https://doi.org/10.1017/S1479262125100427>
19. Jansen, J., & Van Hintum, T. (2007). Genetic distance sampling: A novel sampling method for obtaining core collections using genetic distances with an application to cultivated lettuce. *Theoretical and Applied Genetics*, 114(3), 421–428. <https://doi.org/10.1007/s00122-006-0433-9>
20. Jia, G., Huang, X., Zhi, H., Zhao, Y., Zhao, Q., Li, W., Chai, Y., Yang, L., Liu, K., Lu, H., Zhu, C., Lu, Y., Zhou, C., Fan, D., Weng, Q., Guo, Y., Huang, T., Zhang, L., Lu, T., ... Han, B. (2013). A haplotype map of genomic variations and genome-wide association studies of agronomic traits in foxtail millet (*Setaria italica*). *Nature Genetics*, 45(8). <https://doi.org/10.1038/ng.2673>

21. Kim, S., Kim, D. S., Moyle, H., & Heo, S. (2023). ShinyCore: An R/Shiny program for establishing core collection based on single nucleotide polymorphism data. *Plant Methods* 2023 19:1, 19(1), 106-. <https://doi.org/10.1186/S13007-023-01084-0>
22. Mascher, M., Schreiber, M., Scholz, U., Graner, A., Reif, J. C., & Stein, N. (2019). Genebank genomics bridges the gap between the conservation of crop diversity and plant breeding. In *Nature Genetics* (Vol. 51, Number 7, pp. 1076–1081). Nature Research. <https://doi.org/10.1038/s41588-019-0443-6>
23. McCouch, S. R., McNally, K. L., Wang, W., & Hamilton, R. S. (2012). Genomics of gene banks: A case study in rice. *American Journal of Botany*, 99(2), 407–423. <https://doi.org/10.3732/ajb.1100385>
24. Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4). <https://doi.org/10.1093/genetics/157.4.1819>
25. Odong, T. L., Jansen, J., van Eeuwijk, F. A., & van Hintum, T. J. L. (2012). Quality of core collections for effective utilisation of genetic resources review, discussion and interpretation. *Theoretical and Applied Genetics* 2012 126:2, 126(2), 289–305. <https://doi.org/10.1007/S00122-012-1971-Y>
26. Pessoa-Filho, M., Rangel, P. H. N., & Ferreira, M. E. (2010). Extracting samples of high diversity from thematic collections of large gene banks using a genetic-distance based approach. *BMC Plant Biology* 2010 10:1, 10(1), 127-. <https://doi.org/10.1186/1471-2229-10-127>
27. Ragsdale, A. P., Moreau, C., & Gravel, S. (2018). Genomic inference using diffusion models and the allele frequency spectrum. *Current Opinion in Genetics & Development*, 53, 140–147. <https://doi.org/10.1016/J.GDE.2018.10.001>
28. Reyes-Chin-Wo, S., Wang, Z., Yang, X., Kozik, A., Arikat, S., Song, C., Xia, L., Froenicke, L., Lavelle, D. O., Truco, M. J., Xia, R., Zhu, S., Xu, C., Xu, H., Xu, X., Cox, K., Korf, I., Meyers, B. C., & Michelmore, R. W. (2017). Genome assembly with in vitro proximity ligation data and whole-genome triplication in lettuce. *Nature Communications*, 8. <https://doi.org/10.1038/NCOMMS14953>
29. Singh, N., Wu, S., Raupp, W. J., Sehgal, S., Arora, S., Tiwari, V., Vikram, P., Singh, S., Chhuneja, P., Gill, B. S., & Poland, J. (2019). Efficient curation of genebanks using next generation sequencing reveals substantial duplication of germplasm accessions. *Scientific Reports* 2019 9:1, 9(1), 650-. <https://doi.org/10.1038/s41598-018-37269-0>
30. Sivabharathi, R. C., Rajagopalan, V. R., Suresh, R., Sudha, M., Karthikeyan, G., Jayakanthan, M., & Raveendran, M. (2024). Haplotype-based breeding: A new insight in crop improvement. In *Plant Science* (Vol. 346). <https://doi.org/10.1016/j.plantsci.2024.112129>
31. Slotte, T. (2014). The impact of linked selection on plant genomic variation. *Briefings in Functional Genomics*, 13(4), 268. <https://doi.org/10.1093/BFGP/ELU009>
32. Tewhey, R., Bansal, V., Torkamani, A., Topol, E. J., & Schork, N. J. (2011). The importance of phase information for human genomics. In *Nature Reviews Genetics* (Vol. 12, Number 3). <https://doi.org/10.1038/nrg2950>
33. Upadhyaya, H. D., Ravishankar, C. R., Narasimhudu, Y., Sarma, N. D. R. K., Singh, S. K., Varshney, S. K., Reddy, V. G., Singh, S., Parzies, H. K., Dwivedi, S. L., Nadaf, H. L., Sahrawat, K. L., & Gowda, C. L. L. (2011). Identification of trait-specific germplasm and developing a mini core collection for efficient

use of foxtail millet genetic resources in crop improvement. *Field Crops Research*, 124(3), 459–467. <https://doi.org/10.1016/J.FCR.2011.08.004>

34. Van Hintum, T. J. L., Brown, A. H. D., Spillane, C., & Hodgkin, T. (2000). *Core collections of plant genetic resources*. Core collections of plant genetic resources. IPGRI Technical Bulletin No.3. International Plant Genetic Resources Institute, Rome, Italy. 48p.
35. van Rossum, B.-J., & Kruijer, W. (2026). *An R package for fast GWAS • statgenGWAS*. <https://biometris.github.io/statgenGWAS/index.html>
36. Wambugu, P. W., Ndjiondjop, M. N., & Henry, R. J. (2018). Role of genomics in promoting the utilization of plant genetic resources in genebanks. *Briefings in Functional Genomics*, 17(3), 198–206. <https://doi.org/10.1093/bfpg/ely014>
37. Zhang, L., Su, W., Tao, R., Zhang, W., Chen, J., Wu, P., Yan, C., Jia, Y., Larkin, R. M., Lavelle, D., Truco, M. J., Chin-Wo, S. R., Michelmores, R. W., & Kuang, H. (2017). RNA sequencing provides insights into the evolution of lettuce and the regulation of flavonoid biosynthesis. *Nature Communications*, 8(1), 2264. <https://doi.org/10.1038/S41467-017-02445-9>

Figures

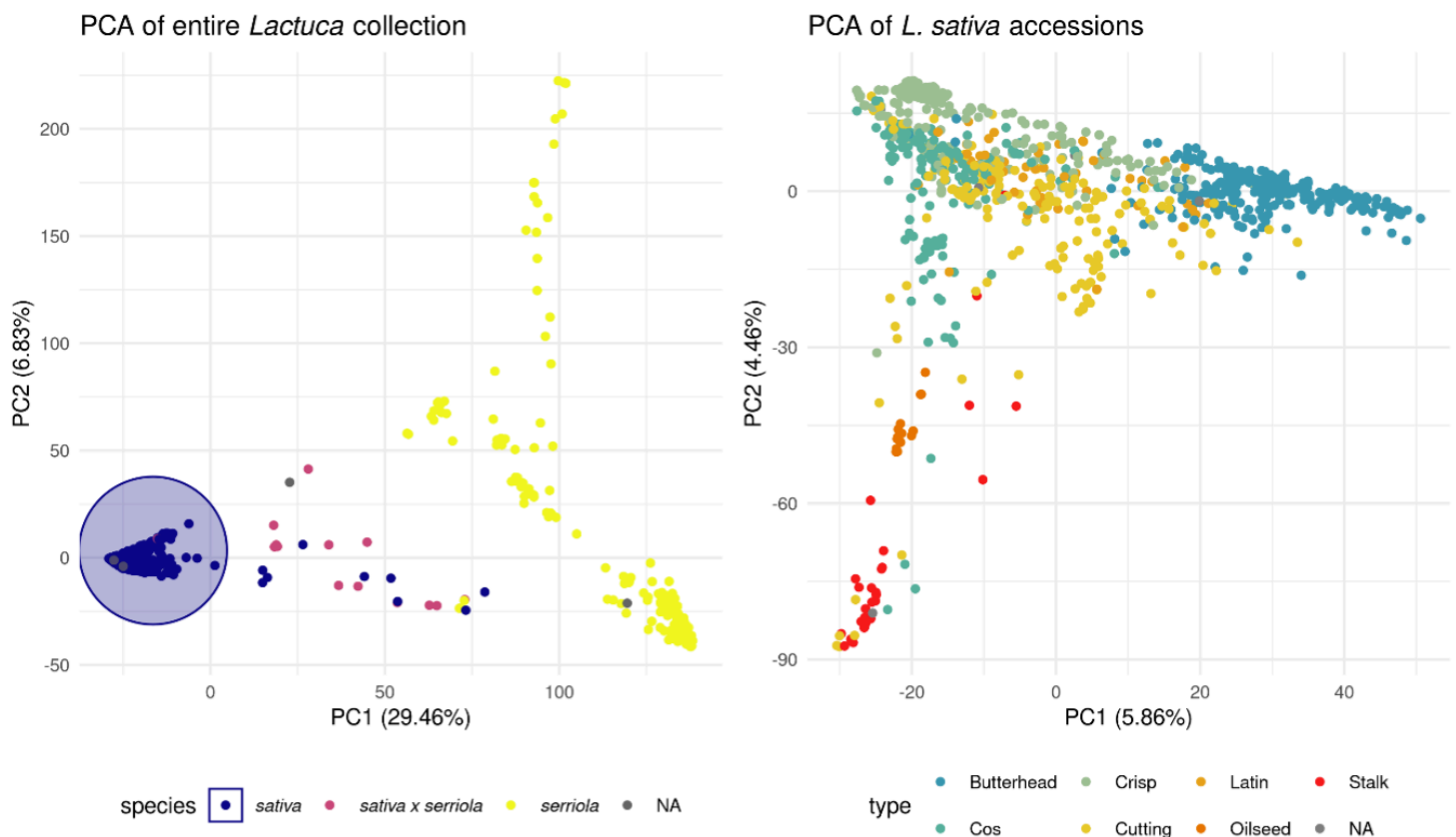


Figure 1

Genetic diversity of the *Lactuca* material used in this study. PCA of total and subset collections used in this study. Percentage of explained variance can be found on x and y axis labels. **Left**, total population, including wild species. Individuals in the *L. sativa* subset population are within the blue circle. Several individuals labelled as “*L. sativa*” were excluded due to high genetic differentiation compared to the rest of *L. sativa*. **Right**, *L. sativa* subset, coloured by crop type.

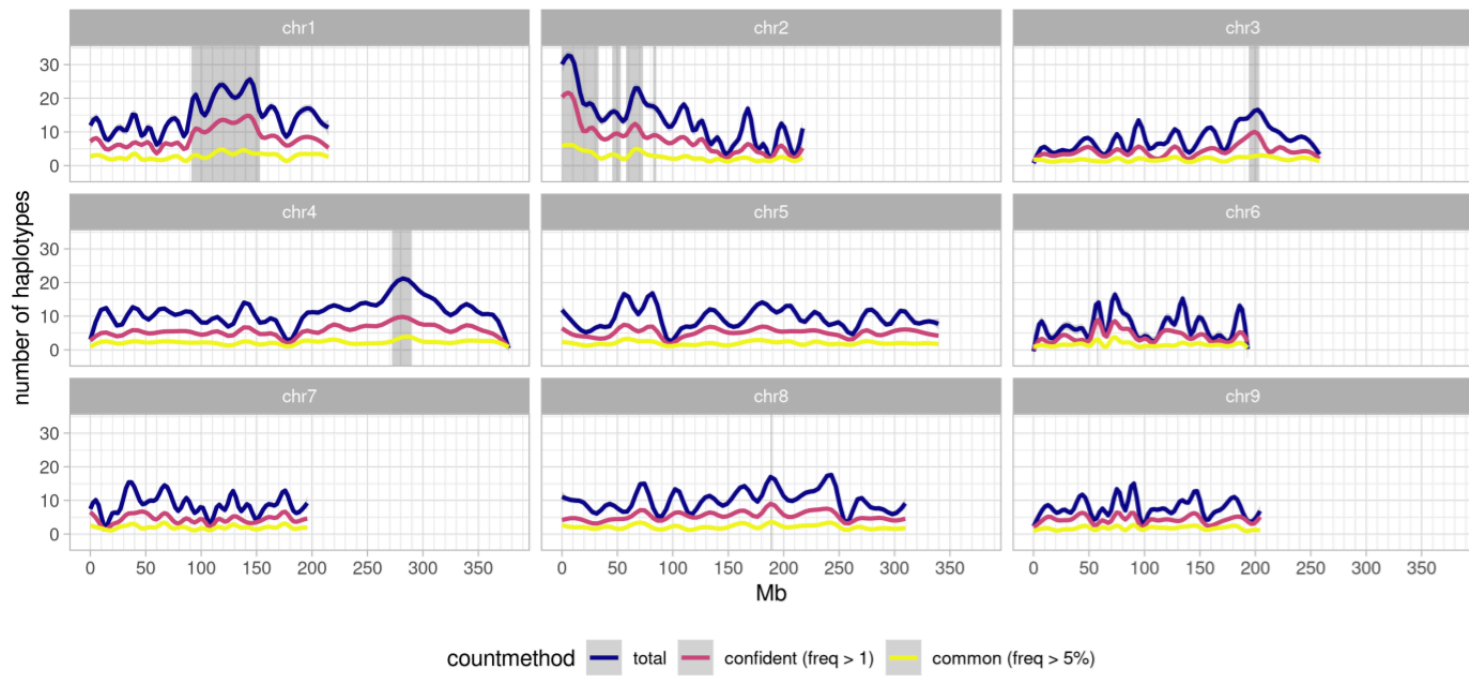


Figure 2

Haplotype diversity throughout the genome. Smoothed averages are shown for haplotype counts along each of the 9 lettuce chromosomes. Three counts are shown, either total number of detected haplotypes (dark blue, total); number of haplotypes detected more than once (magenta, confident); and number of haplotypes with frequency >5% (yellow, common). High haplotype diversity regions, where more than 9 confident haplotypes were detected are highlighted in grey.

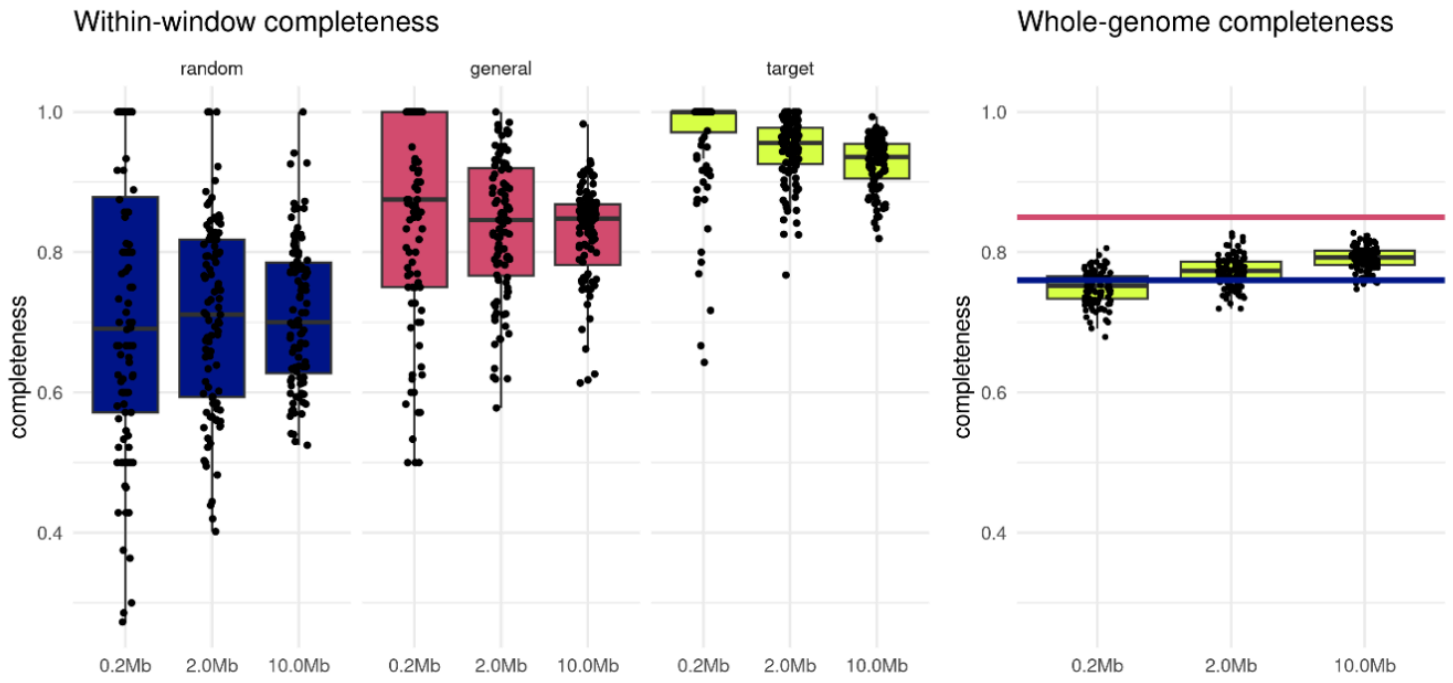


Figure 3

Comparison of collection selection methods based on haplotype completeness of target windows.

Target windows of varying size were defined (0.2, 2 and 10 Mb), and haplotype completeness was computed within those windows, and for the whole genome, for collections generated using a random sample of individuals (blue), a CC using the 2M SNP set (magenta), or our target collection method (yellow). **Left**, within window completeness for varying window sizes using the three collection methods. **Right**, whole-genome haplotype completeness for the random collection (blue line), the general CC (magenta line), or the target collections for different window sizes (yellow boxplots).

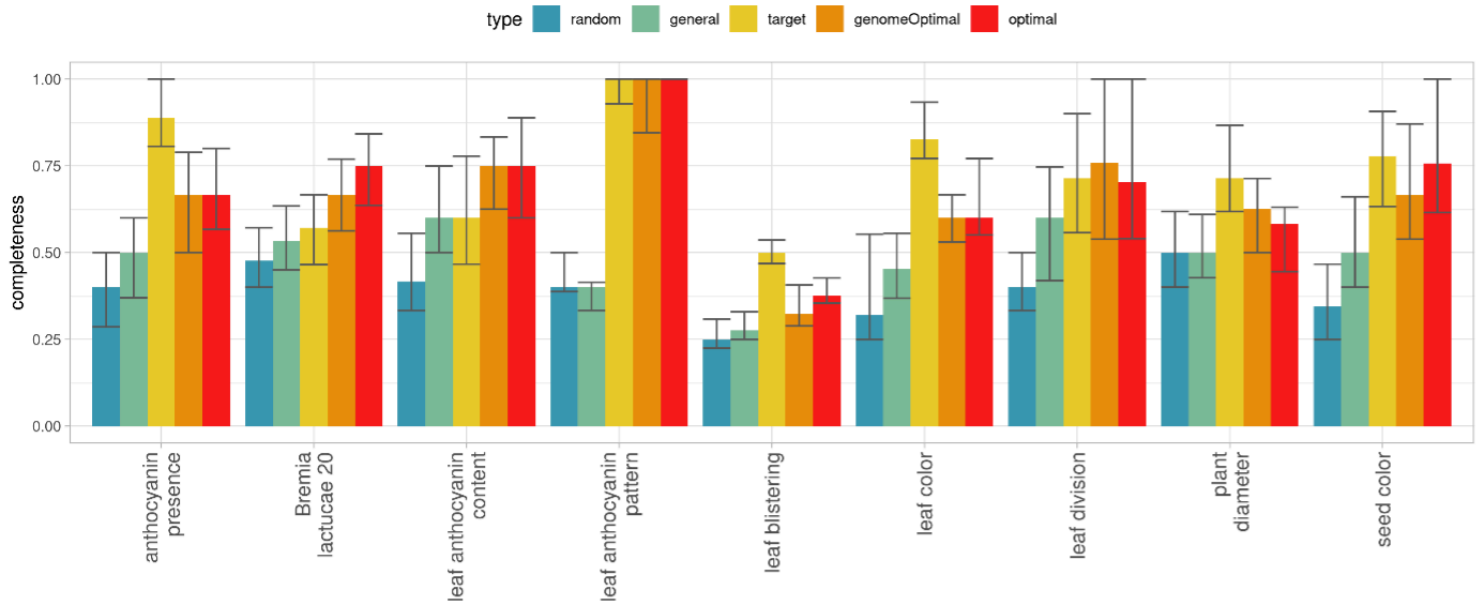


Figure 4

Haplotype completeness within target regions of different traits. Haplotype completeness within each trait-associated genomic region was assessed for the five types of collection generated. Each bar shows the distribution of haplotype completeness for haploblocks within the target regions. The random (blue), general (green) and genome optimal (orange) collections were optimized using entire genome information; while the target (yellow) and optimal (red) collections were optimized for the target regions.

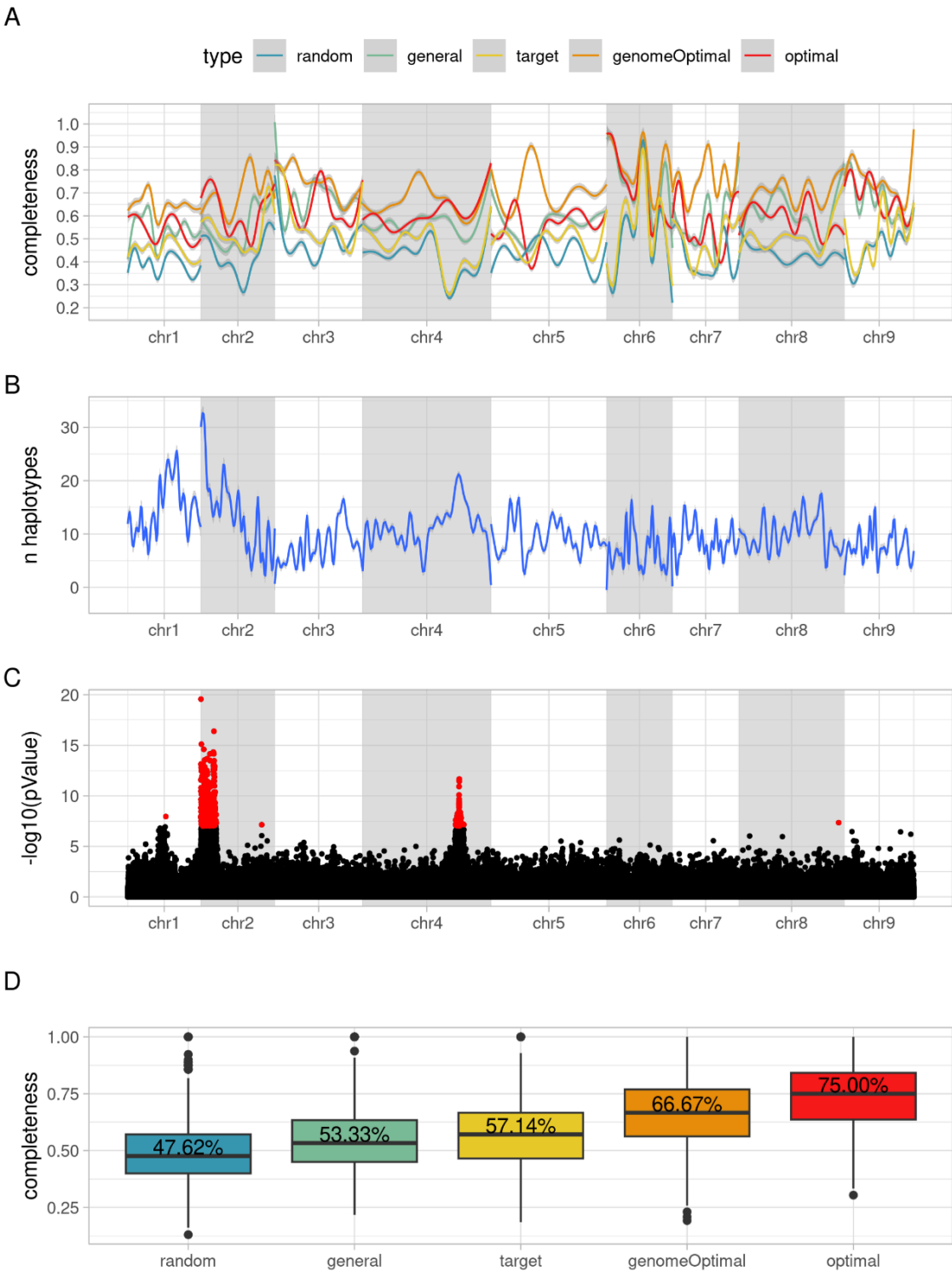


Figure 5

Haplotype completeness of collections obtained using genomic regions associated to *Bremia*

resistance. A) Haplotype completeness (smoothed line) of non-target collections and target collections across the entire genome. **B)** Number of haplotypes across the genome (smoothed line). **C)** GWAS Manhattan plot for *Bremia* resistance. **D)** Completeness distribution within the target windows in the five collections assessed. Whiskers indicate 25% and 75% quantiles. Above the black line the median completeness is shown.