ISSN: 2641-6379
DOI: 10.33552/WJASS.2025.09.000722
World Journal of Agriculture and Soil Science
Iris Publishers

Research Article

Clustering-based Method for Core Germplasm Collection Constructing via Multiple Correspondence Analysis

Nien-Lun Wu, and Chen-An Tsai*

*Department of Agronomy, National Taiwan University, Taiwan

Corresponding author: Chen-An Tsai, Department of Agronomy, National Taiwan University, Taiwan

Received Date: October 29, 2025

Published Date: November 06, 2025

Abstract

The concept of a core collection, introduced by Frankel in 1984, aims to capture maximum genetic diversity within a minimum number of accessions. While many selection methods exist, the massive datasets generated by Next-Generation Sequencing (NGS) technologies have made some traditional approaches computationally prohibitive. In this study, we propose a new, efficient algorithm for selecting a core set of lines from large genotype datasets based on clustering algorithms. Our goal is to maximize genetic diversity for a user-defined collection size. The method integrates Multiple Correspondence Analysis (MCA) for refined cluster analysis and adapts the successful selection strategy of the Geno Core algorithm as a foundation.

We evaluate the performance of our proposed method against two established algorithms, Geno Core and Core Hunter 3, using four diverse Single Nucleotide Polymorphism (SNP) datasets ranging from 1.5K to 820K SNPs. Quality is assessed using five criteria: coverage rate, Shannon's diversity index, mean modified Roger's value, minimum modified Roger's value, and computational efficiency (time). The experiment results demonstrate the superior performance of our method. While maintaining a high coverage rate (e.g., 99%), our algorithm consistently achieves a higher quality core collection than Geno Core and a higher coverage rate than Core Hunter 3. Critically, our approach successfully processes all four large datasets in a reasonable timeframe, addressing the bottleneck of NGS data analysis.

Keywords: Core collection; next-generation sequencing (NGS); multiple correspondence analysis; cluster analysis; single nucleotide polymorphism (SNP)

Introduction

The conservation of genetic resources for plant and agricultural species is a crucial task, particularly in the face of rapid environmental changes. To address this need, numerous countries have established gene banks to collect and preserve germplasm, thereby safeguarding genetic diversity. However, as the volume and complexity of collected resources grow, more efficient methods for management, classification, and selection are essential. Optimizing the limited storage space requires selection strategies that retain the maximum possible genetic diversity [1]. In response, the concept of the core collection was proposed [2]. He defined a core collection as a small set of accessions that represents the complete genetic diversity of a species, including its wild relatives, with mini-

mal redundancy. Beyond its primary use in gene bank resource conservation, the core collection concept serves as a highly effective tool for researchers to select experimental materials and assess genetic variation. This approach reduces the required sample size for analysis, leading to significant savings in time and resources. Since its inception, the selection of core collections has been extensively studied by numerous scholars [3-5].

These efforts have yielded a variety of selection methods, employing different criteria to achieve the goal of establishing a small-sized core collection that captures the entire genetic diversity of the target germplasm. While there is no rigid limit on the size of a core collection, which can be adjusted based on its intended purpose, published core collections generally comprise between 5% and 20% of the initial accessions [6]. Early core collection selection software, such as MSTRAT [3], Power Core [4], and Core Hunter [5], primarily relied on phenotypic data, often complemented by limited genotypic data, due to the less-developed sequencing technologies of the time. However, the rapid advancement of Next-Generation Sequencing (NGS) and microarray technologies has made it relatively easy for researchers to identify extensive sequence variation among individuals. Given the direct link between genotype and phenotype, current core collection selection is increasingly focused on large-scale genotype data. This shift has introduced a major challenge: the genotype data, often comprised of thousands or even hundreds of thousands of genetic markers (e.g., Single Nucleotide Polymorphisms or SNPs), has become enormously vast. Consequently, many previously effective analysis methods are no longer computationally feasible or efficient for handling such massive datasets.

To address this computational bottleneck, the Geno Core algorithm was first developed specifically for selecting core collections from large categorical datasets (i.e., genotypic data) [7]. Unlike some methods that focus on selecting rare alleles, Geno Core prioritizes the selection of common alleles. The rationale behind this strategy is the authors' view that most traits of current interest are complex traits, influenced by polygenic inheritance and environmental interactions, rather than being controlled by single rare alleles. This selection strategy is designed to effectively increase the genetic coverage of the core collection, thus achieving the goal of high genetic diversity. In this study, we propose a new core collection selection method for large genotype datasets by integrating the selection strategy of Geno Core with cluster analysis. This novel approach is designed to rapidly and efficiently select accessions that are evenly distributed and genetically distant from one another within the initial sample, and it can be implemented efficiently in R. We validate our method using four diverse Single Nucleotide Polymorphism (SNP) datasets: 14K SNP wheat data, 1.5K SNP and 37K SNP rice data, and a high-density 820K SNP wheat dataset. The results are comprehensively compared against those obtained from the state-of-the-art methods, Geno Core [7] and Core Hunter 3 [8].

Materials and Methods

Introduction to the Datasets Used

This study utilizes four genotypic datasets, which can be categorized into four progressively larger sets based on data volume. As noted in the introduction, the information volume requiring processing has increased significantly with the development of next-generation sequencing technology. Both Geno Core and the method proposed in this study were developed in response to this situation. We use these four genotypic datasets to observe whether the core germplasm results obtained using the method proposed in this study maintain the expected screening quality despite the significant differences in the size of the original datasets, and whether the processing time remains within an acceptable range. Furthermore, we compare the results of different methods across these varying data sizes. In this section, we will list the data source, the number of SNP markers, and the number of accessions for each dataset

1.5K Rice SNP Data

This dataset is the smallest in terms of data volume used in this study. Its source is the Rice Diversity website (http://www.ricediversity.org/index.cfm). It contains 1,536 markers and 395 accessions, with no missing values.

14K Wheat SNP Data

This dataset is a sample dataset provided to users on the Geno Core GitHub repository (https://github.com/lovemun/GenoCore/find/master). This sample was reduced from the 35K Axiom® Array data available on Cereals DB (http://www.cerealsdb.uk.net/cerealgenomics/CerealsDB/indexNEW.php). The original 35K data was generated by sequencing 35,143 SNPs across hexaploidy and tetraploid wheat accessions, but the specific criteria for this reduction are not detailed in the sample data. This dataset has 14,099 markers and 556 accessions, with 3.34% missing values.

37K Rice SNP Data

The source of this dataset is the same as the 1.5K Rice SNP Data, specifically the 44K SNP set from Rice Diversity (http://www.ricediversity.org/index.cfm). This data is the result of a genome-wide association study (GWAS) that sequenced 44,100 SNPs in rice accessions collected from 82 countries. It contains 36,901 markers and 413 accessions, and is thus referred to as 37K Rice in this thesis, with 3.77% missing values.

820K Wheat SNP Data

This dataset originates from the 820K Axiom® Array data on Cereals DB (http://www.cerealsdb.uk.net/cerealgenomics/CerealsDB/indexNEW.php). This data was generated from hexaploidy wheat accessions, diploid and tetraploid progenitors, and some closely related wheat species, and was originally used for identifying and tracking introgression in hexaploidy wheat. It contains 819,570 molecular markers and 475 accessions, with 1.59% missing values.

Proposed Core Germplasm Selection Algorithm

In the Methods section, we will describe the core germplasm selection algorithm used in this study, which is primarily divided into four sequential steps:

Step 1: Calculating Relative Coordinates of Accessions

The first step is to calculate the relative coordinates for each ac-

cession based on its similarity to all other accessions. For this study, Multiple Correspondence Analysis (MCA) is selected as the method for calculating these coordinates. It transforms the multi-category genotypic data (where each SNP marker is a variable and each allele is a category) into a set of relative coordinates for each accession. These coordinates essentially capture the genetic relationships and structure within the germplasm. The resulting low-dimensional coordinates are then used as the numerical input for subsequent clustering methods (like Self-Organizing Map and K-means) to partition the accessions based on their genetic similarity.

Step 2: **Determining the Optimal Number of Clusters**

Using the accession coordinates obtained in Step 1, a distance matrix is calculated to establish the clustering criterion. The Average Silhouette score is then used to estimate the optimal number of clusters suitable for the data.

Step 3: Accession Clustering

The third step involves performing the actual clustering based on the results from Step 2. In this study, two distinct clustering methods are used for this step: Self-Organizing Map (SOM) and the K-means algorithm. The differences between the results generated by these two methods will be discussed.

Step 4: Core Germplasm Selection

The final step for selecting the core germplasm is adapted from the selection method used in Geno Core [7].

a) Initial Selection: First, from the clustered results obtained in Step 3, one accession with the lowest number of missing values is selected from each cluster and added to the core germplasm.

b) Subsequent Selection: Next, disregarding the clusters, the algorithm iteratively selects the accession from the remaining set that has the greatest genetic dissimilarity to the accessions already present in the current core germplasm.

The detailed operational procedures for this methodology will be explained in the subsequent content.

Multiple Correspondence Analysis (MCA)

MCA is the categorical data equivalent of Principal Component Analysis (PCA) [9]. It transforms the categorical data into a set of coordinates in a low-dimensional space to allow for visualization and analysis of the underlying structure, particularly the associations between the different categories and the similarity between individuals (accessions). The SNP data is initially represented by the indicator matrix **X**, often called the Complete Disjunctive Table (CDT). Each element of X is a binary value (0 or 1) recording the presence of a specific accession within a given category. Each row represents the individual accessions (subjects) and each columns represents the categories (alleles/genotypes) of every variable (SNP marker). A value of 1 in a cell indicates that the accession possesses that specific category (e.g., a specific allele at a certain SNP) and 0 indicates it does not. If you have N accessions and M discrete variables (SNP markers), where the *j*-th variable has K_j categories (e.g., A/A, A/G, G/G for a tri-allelic SNP), the total number of columns in the indicator matrix will be $J = \Sigma_{j=1}^{M} K_j$. As noted, for large genotypic datasets (where the number of SNP markers is high), the indicator matrix **X** can become extremely large, making the direct application of Correspondence Analysis (CA) to **X** computationally intensive and time-consuming. To efficiently overcome the computational burden of a large indicator matrix, the Burt matrix **B** is employed. The Burt matrix is derived as the cross-tabulation of the indicator matrix **X** with its categorical variables, often calculated as

$$B = X^t X. \qquad (1)$$

This matrix summarizes the relationships between all pairs of SNP markers, dramatically reducing the size of the matrix used for the core MCA computation, thus saving significant time and computational resources. The core of MCA involves an Eigenvalue Decomposition (EVD) of the standardized Burt matrix. The central Burt matrix is often defined for simplified computation as:

$$E = D^{-1/2} B D^{-1/2} - 1 \qquad (2)$$

where **D** is the diagonal matrix of the column proportions of the *k*-th SNP. The Eigenvalue Decomposition (EVD) is applied to the transformed matrix **E** to find the principal axes.

$$E = U \Lambda U^t, \qquad (3)$$

where **Λ** is a diagonal matrix containing the eigenvalues λ_i, for $i=1,...,J$, representing the inertia (variance) explained by each dimension; and **U** is a matrix whose columns contain the eigenvectors v_i for $i=1,...,J$. Finally, the principal coordinates **G**, which are the actual coordinates used for the graphical interpretation (the scatter plot), are obtained by scaling the eigenvectors by the square root of the non-trivial eigenvalues:

$$G = D^{-1/2} U \Lambda^{1/2}. \qquad (4)$$

The Inertia (total variance) in the data is equal to the sum of all non-trivial eigenvalues. The proportion of inertia explained by the *i*-th dimension is:

$$f_i = \frac{\lambda_i}{\Sigma_{i=1}^{m} \lambda_i}, \qquad (5)$$

where m is the total number of non-trivial dimensions. This value indicates the amount of variability captured by that dimension. A higher percentage of inertia for a dimension indicates that it represents a greater degree of the association (or underlying structure) among the SNP markers in the dataset. This helps researchers select the most important dimensions for interpretation and visualization.

In our proposed algorithm, MCA serves two primary functions:

a) Coordinate Calculation: It transforms the multi-category genotypic data (where each SNP marker is a variable and each allele is a category) into a set of relative coordinates for each accession. These coordinates essentially capture the genetic

relationships and structure within the germplasm.

b) Input for Clustering: The resulting low-dimensional coordinates are then used as the numerical input for subsequent clustering methods (like Self-Organizing Map and K-means) to partition the accessions based on their genetic similarity.

Clustering Methods

Once the low-dimensional coordinates were obtained, the data was subjected to clustering using two common clustering algorithms: k-means and Self-Organizing Map (SOM). Clustering methods are unsupervised learning techniques used to group data points into subsets (clusters) such that data points in the same cluster are more similar to each other than to those in other clusters. The goal is to discover inherent groupings and patterns in the data without prior knowledge of labels. K-means [10,11] is a popular, simple, and efficient partitioning clustering algorithm. Its goal is to partition N observations into a predefined number of K clusters, minimizing the within-cluster sum of squares (WCSS). The K-means clustering's operational process is as follows:

a) Initialize K centroids randomly.

b) Assignment: Assign each data point to the cluster whose centroid is nearest (usually based on Euclidean distance).

c) Update: Recalculate the new centroid for each cluster as the mean of all data points assigned to that cluster.

d) Repeat steps 2 and 3 until the centroids no longer move significantly or a maximum number of iterations is reached.

The limitations of K-means clustering include that the number of clusters (K) must be specified beforehand, and the result can be sensitive to the initial random placement of the centroids and the presence of outliers.

A Self-Organizing Map (also known as a Kohonen map) [12,13] is an unsupervised neural network used primarily for clustering and dimensionality reduction. It maps high-dimensional data onto a low-dimensional space (typically a 2D grid of neurons) while preserving the original data's topological properties (i.e., data points that are close in the input space are mapped to neighbouring neurons on the grid). The SOM's operational process is as follows:

a) Initialize the weight vectors of all neurons (nodes) on the 2D grid.

b) Competition (Find BMU): For each input vector, find the neuron whose weight vector is closest to it-this is the Best Matching Unit (BMU).

c) Cooperation/Adaptation: Update the weight vector of the BMU and its neighbouring neurons on the grid to move closer to the input vector. The degree of update decreases with distance from the BMU and with time (epochs).

d) Repeat steps 2 and 3 until convergence.

SOMs are effective for visualization and can handle non-linear data relationships better than K-means, as the grid structure explicitly models the data's topology. The neighbourhood update rule is the key difference from K-means (which only updates the winning centroid).

Prior to clustering, the optimal number of clusters is assessed using the Average Silhouette Method [14]. Data is first clustered using the cluster algorithms above, testing a user-defined range of cluster numbers (k). For each data point i, a silhouette value, *s(i)*, is calculated using the following formula:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \qquad (6)$$

where a*(i)* is the average distance between data point *i* and all other data points within its own cluster and b*(i)* is the average distance between data point *i* and all data points in the nearest neighbouring cluster (the cluster to which *i* is not assigned but is closest). From the formula, the silhouette value s*(i)* ranges from -1 to +1. A value close to +1 indicates that the data point is well-clustered: it is close to other points in its own cluster and far from points in other clusters, representing an ideal state of high within-cluster similarity and low cross-cluster similarity. Conversely, a value close to -1 indicates that the data point is poorly clustered: it is dissimilar to points in its own cluster and is closer to points in a neighbouring cluster, suggesting a poor clustering outcome. To evaluate the overall quality of a clustering solution, the Average Silhouette Coefficient is computed by taking the mean of the silhouette values for all data points:

$$\frac{1}{N}\Sigma_{I=1}^{N} s(i), \qquad (7)$$

where N is the total number of data points. The process is repeated for each tested value of k. The value of k that yields the maximum Average Silhouette Coefficient is determined to be the estimated optimal number of clusters for the dataset. In this study, the distance *a(i)* and *b(i)* are calculated using the Euclidean distance between the coordinates of the observation categories derived from the preceding Multiple Correspondence Analysis (MCA).

Method for Core Germplasm Selection

The method for selecting core germplasm in this study is adapted from the Geno Core approach [7]. The core germplasm selecting procedure is as follows:

Step 1: Initial Selection from Clusters. The clustering performed in the preceding steps yields k groups of accessions with similar characteristics. Following the Geno Core approach, the accession with the fewest missing data (NA values) is initially selected from each of these k clusters. This results in an initial set of k selected accessions.

Step 2: Tie-breaking using Coverage Score (CS) If multiple accessions within a cluster have an equal (minimum) number of missing values, their coverage score (CS) is calculated to break the tie. This step introduces a modification to the original Geno Core method: instead of strictly minimizing CS, this study explores two distinct selection strategies for observation: Selecting the accession with the minimum CS or Selecting the accession with the maximum CS.

Step 3: Tie-breaking using Diversity Score (DS) If a tie still exists after calculating the CS (i.e., multiple accessions have the same CS value), the diversity score (DS) is calculated. The accession with the minimum DS is selected. If duplicates still exist, random selection is performed. After this step, a total of k accessions is selected, where k is the number of clusters. The coverage score (C_j) and diversity score (D_j) are defined as follows:

$$C_j = \frac{\Sigma_{i \in N_j} f_{ig_{ij}}}{n(N_j)}; \qquad (8)$$

$$D_j = \frac{\Sigma_{i \in N_j} (f_{ig_{ij}} - C_j)^2}{n(N_j)}, \qquad (9)$$

where f_{igij} is the genotype frequency g_{ij} for the *i*-th marker and *j*-th accession, N_j is the set of non-missing genotype markers in the *j*-th accession, and $n(N_j)$ is the number of elements in N_j.

Step 4: Greedy Selection Pool Initialization. All unselected accessions are aggregated into a single group. The marker types identical to those already present in the initial k core accessions are converted to missing values (NA) within this unselected pool. The accession from this pool with the minimum number of remaining missing values is selected. This process ensures that the next chosen accession contributes the maximum number of novel marker types to the core set.

Step 5: Iterative Core Set Expansion. This step repeats the logic from Step 3: the marker types of the newly selected accession are converted to NA values in the remaining unselected pool. The accession with the fewest NA values is then chosen. In the event of a tie, the selection proceeds in the order of CS and DS (as detailed in Step 3). This iterative expansion continues until the total genetic coverage of the core set reaches a predefined standard of 99%, at which point the loop is terminated.

Core Collection Evaluation Indices

Four performance metrics are used in this study to evaluate and distinguish the quality of the selected core germplasm. Following the approach of Geno Core, the following indices are calculated: Coverage, Shannon's Diversity Index, Mean Modified Rogers Value, and Minimum Modified Rogers Value.

Coverage

Coverage assesses the percentage of distinct marker genotypes present in the original dataset that are successfully captured by the selected core collection. The formula is given by:

$$CV = \frac{1}{m} \Sigma_{i=1}^{m} \frac{G_c(i)}{G_o(i)}, \qquad (10)$$

where m is the total number of SNP markers, $G_c(i)$ is the number of distinct genotypes for the *i*-th marker in the core collection, and $G_0(i)$ is the number of distinct genotypes for the *i*-th marker in the original germplasm dataset. A higher coverage value indicates that the core collection encompasses a greater variety of genotypes, thus more completely preserving the genetic resources of the original collection.

Shannon's Diversity Index (SH)

Shannon's Diversity Index (SH), also known as the Shannon-Wiener Index, is used to estimate the level of genetic diversity within the collection. In this study, each genotype under a molecular marker is considered a distinct "species" or type. The formula is applied for each marker and then averaged across all markers:

$$SH = \sum_{i=1}^{m} \sum_{j=1}^{G_C(i)} p_{ij} \ln(p_{ij}), \qquad (11)$$

where $G_c(i)$ is the number of distinct genotypes for the *i*-th marker in the core collection and p_{ij} is the frequency of the *j*-th genotype under the *i*-th marker. The index considers both the number of types and their relative frequencies (evenness). A larger index indicates higher estimated genetic diversity. The SH value reaches it's maximum when all genotype frequencies within the population are equal, representing the most stable and diverse state.

Mean Modified Rogers Value (MR)

The Modified Rogers Distance d_{xy} is a measure of genetic distance between two accessions, *x* and *y*, based on genotypic data. The formula is:

$$d_{xy} = \frac{1}{\sqrt{2m}} \sqrt{\Sigma_{i=1}^{m} (p_{xi} - p_{yi})^2}, \qquad (12)$$

where p_{xi} and p_{yi} are the frequencies of the *i*-th marker for accession x and accession y, respectively. The Mean Modified Rogers Value (MR) for the core collection is the average of the pair-wise Modified Rogers Distances calculated among all accessions within the core collection.

$$MR = Average(d_{xy}), \qquad (13)$$

for all pairs *i, j* in the core collection. A larger MR value indicates that the selected accessions in the core collection are, on average, more dissimilar from one another, suggesting a more efficient representation of the overall genetic variation.

Minimum Modified Rogers Value (Min.MR)

The Minimum Modified Rogers Value uses the same distance measure as the MR but focuses on the smallest distance found between any pair of accessions within the core collection.

$$Min.MR = Min(d_{xy}), \qquad (14)$$

for all pairs *i, j* in the core collection. A larger Min.MR value guarantees that the accessions are at least separated by a minimum distance. This metric is crucial because it guards against a scenar-

io where the MR is high due to a few widely dispersed accessions, while the minimum distance remains near zero (meaning some accessions are nearly identical), a situation that may be hidden when observing only the average (MR).

Results and Discussion

In this section, a comparative analysis of different selection methods will be performed across the four datasets utilized in this study, namely: Rice with 1.5K SNPs, Wheat with 14K SNPs, Rice with 37K SNPs, and Wheat with 820K SNPs. The comparison proceeds sequentially from the smallest to the largest dataset. Following the methodological steps outlined in the previous section, the results for each dataset will be presented as follows:

a) Multiple Correspondence Analysis (MCA): The results of the MCA, which transforms the data into two-dimensional coordinates for clustering, will be displayed in a 2D scatter plot.

b) Optimal Cluster Number: The optimal number of clusters, as determined by the Average Silhouette Method, will be reported.

c) Clustering Results: The results of the clustering using two different clustering approaches will be presented.

d) Core Collection Selection: The resulting core accessions obtained from the selection procedure applied to the two different clustering results will be listed.

After describing the selection outcomes for this study's method, the results will be compared against two established methods from the literature: Geno Core and Core Hunter 3. For this comparison, the operational settings for the established methods are as follows: 1. Geno Core is configured to stop selection when 99% coverage is achieved. 2. Core Hunter 3 is configured to select the same number of accessions as determined by the Geno Core run, with all other parameters set to the program's default values. The comparison will be based on the four-performance metrics: Coverage, Shannon's Diversity Index (SH), Mean Modified Rogers Value (MR), and Minimum Modified Rogers Value (Min.MR). Furthermore, a key objective of the proposed method is to handle large-scale genotyping data efficiently. Therefore, for each method and dataset, the computational time and the final number of accessions selected for the core collection will also be recorded and compared. These additional factors are crucial for assessing the practical applicability and efficiency of the method when dealing with modern, voluminous genomic data.

1.5K SNP Rice Data

Following the Multiple Correspondence Analysis (MCA) procedure, the resulting dimensional coordinates were used. Next, the Average Silhouette Method was applied, utilizing the coordinates from the top five dimensions (based on proportion of inertia) to find the optimal number of clusters. The Silhouette Coefficient reaches its maximum value at K=9 clusters, with the coefficient exceeding 0.5, suggesting an acceptable clustering quality. Therefore, subsequent calculations for this rice dataset used nine clusters. With K=9 established, the accessions were subjected to both SOM and K-means clustering. The visualization of the clustering results is achieved by projecting the cluster accessions onto the 2D scatter plot using the first two principal dimensions from MCA and colouring the points according to their cluster membership, as shown in (Figures 1&2). The results show notable differences, with the K-means clustering exhibiting a higher degree of intermixing of accessions from different assigned clusters within the same spatial region compared to the SOM clustering.

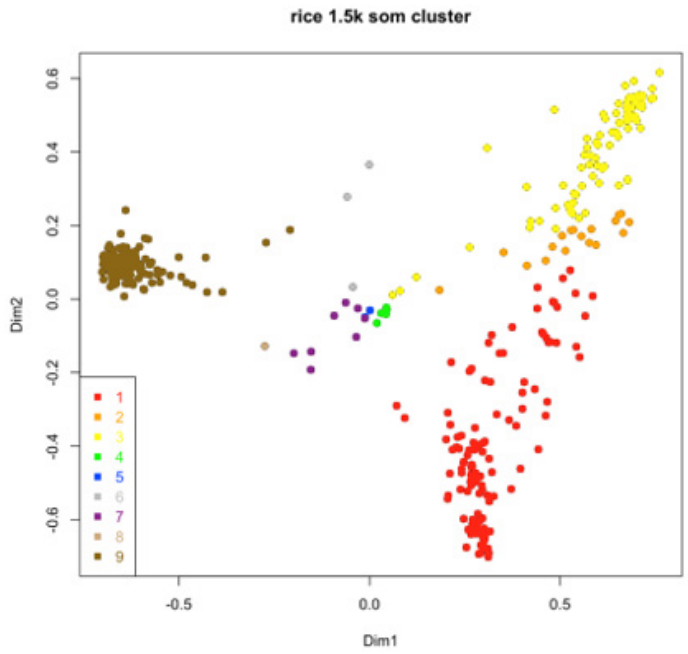

Figure 1: Final clustering results for 1.5K Rice dataset based on the SOM algorithm.

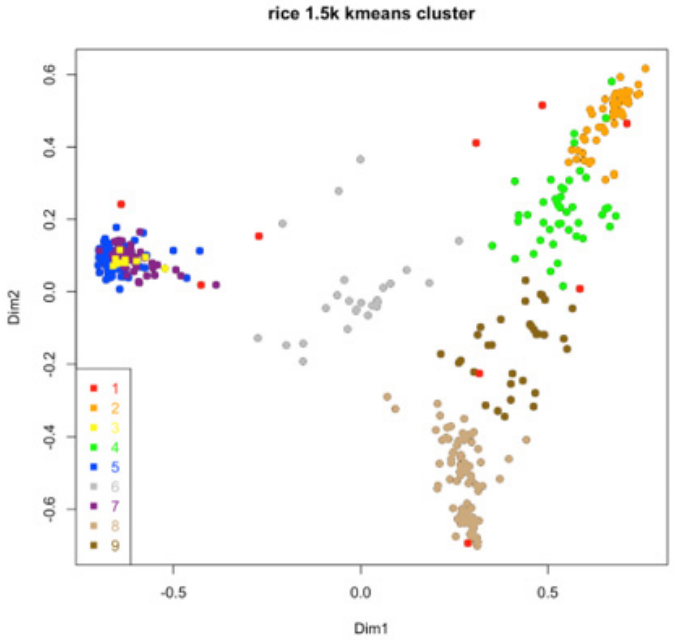

Figure 2: Final clustering results for 1.5K Rice dataset based on the k-means algorithm.

The SOM topology was set to a rectangular shape; testing showed that using a rectangular versus a hexagonal topology had no significant impact on the final core germplasm selection results for this study. The core germplasm selection method proposed in this study was applied to the resulting clusters. (Table 1) summarizes the results, comparing our method against Geno Core and Core Hunter 3. It is observed that our method, when compared to Geno Core (which stopped at 99% coverage), generally selects between 1 and 5 more accessions than Geno Core. Specifically, the K-means-based selection yielded 60 accessions, while the SOM-based selections yielded 63 and 64 accessions. The remaining performance metrics are comparable across our method and Geno Core. Core Hunter 3, constrained to select the same number of accessions (59), performed slightly worse than our proposed method only on the Minimum Modified Rogers Value (Min.MR).

Table 1: Core Germplasm Selection Results of Different Methods for the 1.5K Rice Dataset.

Method	size	MR	Min.MR	SH	CV	Time (sec)
Geno Core	**59**	0.2363	0.1192	7.9910	99.0147	35.5429
corehunter3	59	0.2395	**0.1464**	7.9791	97.0125	48.0344
K means	60	0.2411	0.1154	7.9911	**99.0338**	**28.5933**
K means(max)	60	0.2303	0.1271	**7.9965**	99.0148	29.1795
Som	63	0.2429	0.0859	7.9837	99.0211	32.7130
Som (max)	64	**0.2493**	0.0866	7.9866	**99.0338**	33.4937

This difference is expected, as Core Hunter 3 explicitly uses the Min.MR as a primary optimization criterion, whereas our method primarily focuses on maximizing coverage. Computation time for this small dataset was negligible for all methods (in the order of seconds), and the minor differences in speed are likely within the margin of error, making them practically indistinguishable (Figure 3). Illustrates the coverage growth efficiency for each method, showing the accumulated coverage as accessions are sequentially added to the core set. While the growth efficiency of our method is slightly lower than that of Geno Core, it remains stable and rapid, eventually achieving near 100% coverage. Core Hunter 3 shows the slowest growth curve and the lowest final coverage among the three. Although the figure plots all four variants of our method (K-means/SOM × max/min CS), only the max CS curves for K-means and SOM are distinctly visible. This is because the corresponding min CS curves are virtually overlaid, demonstrating that the choice between maximizing or minimizing CS has a small impact on coverage growth efficiency.

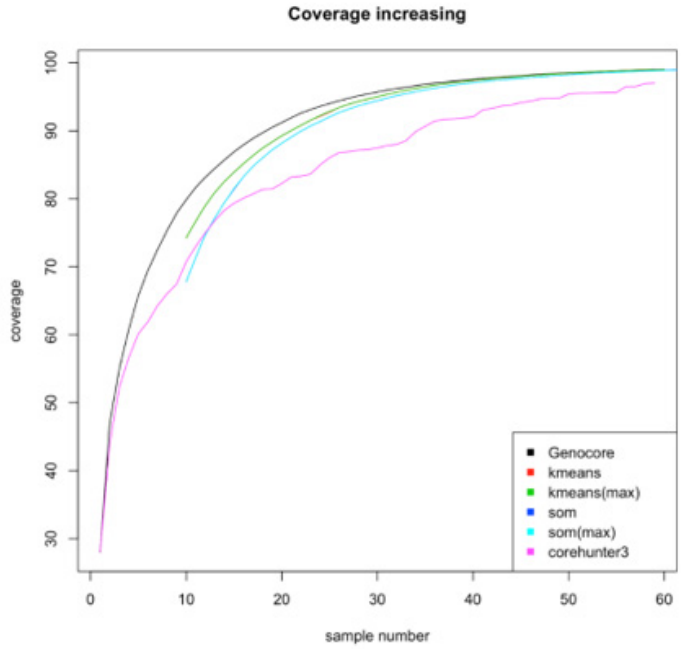

Figure 3: Coverage values versus the number of selected samples for each software in the 1.5K Rice dataset.

The difference in initial clustering method (K-means vs. SOM) likely has a greater influence on the initial core accessions selected. The Venn Diagram quantifies the overlap in selected accessions (Figure 4). For simplicity and given the minimal observed difference, only the max CS selection variant of our method is included. The diagram confirms that Core Hunter 3 has the most distinct selection, contributing 18 unique accessions. In contrast, our method and Geno Core share over 86% of their core accessions, with each having only a small number (3-6) of unique accessions.

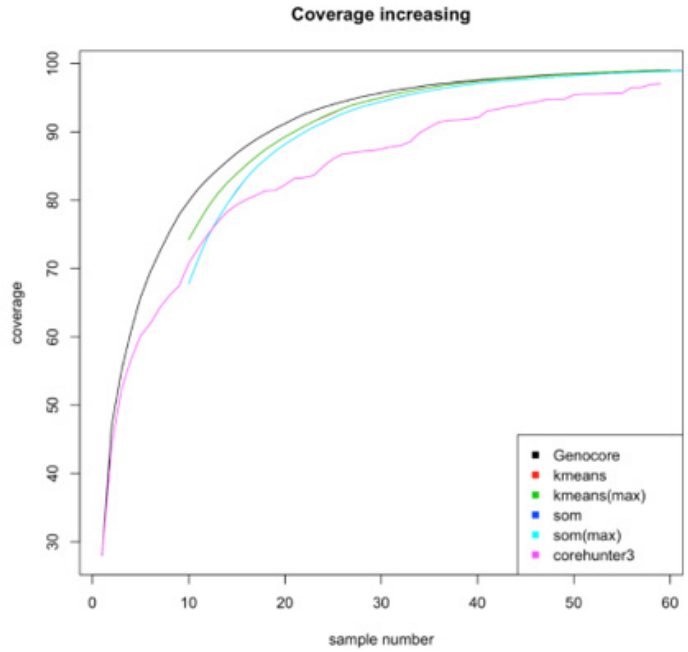

Figure 3: Coverage values versus the number of selected samples for each software in the 1.5K Rice dataset.

14K SNP Wheat Data

The second dataset analysed is the 14K SNP Wheat data, which represents a significantly larger volume of information compared to the previous 1.5K SNP data. Following the established procedure, MCA was first performed to obtain the dimensional coordinates for each accession. The optimal number of clusters, k=5, maximizes the average silhouette score, and the 556 accessions were then partitioned into five clusters. Clustering was subsequently performed using both SOM and K-means (Figures 5&6). Similar to the 1.5K Rice results, K-means clustering showed a greater degree of intermixing among different cluster assignments (Table 2). Shows that Geno Core selected 89 core accessions. Our method selected slightly more: the SOM (max CS) variant selected 92 accessions, while the others, K-means and SOM (min CS), selected 91 accessions. Regarding the evaluation metrics, the K-means (max CS) variant demonstrated superior performance over Geno Core across all metrics. However, our method's Min.MR and SH values were slightly lower than those achieved by Core Hunter 3.

Table 2: Core Germplasm Selection Results of Different Methods for the 14K Wheat Dataset.

Method	size	MR	Min.MR	SH	CV	Time (min)
Geno Core	89	0.1830	0.0774	9.5460	99.0082	7.8771
corehunter3	89	0.1837	**0.1367**	**9.7994**	86.5428	**0.6921**
km0065ans	91	0.1822	0.0776	9.5421	99.0106	13.6298
K means(max)	91	**0.1847**	0.0777	9.5575	**99.0153**	12.6858
Som	91	0.1819	0.0772	9.5370	99.0047	13.4388
Som (max)	92	0.1843	0.0773	9.5502	99.0129	12.8916

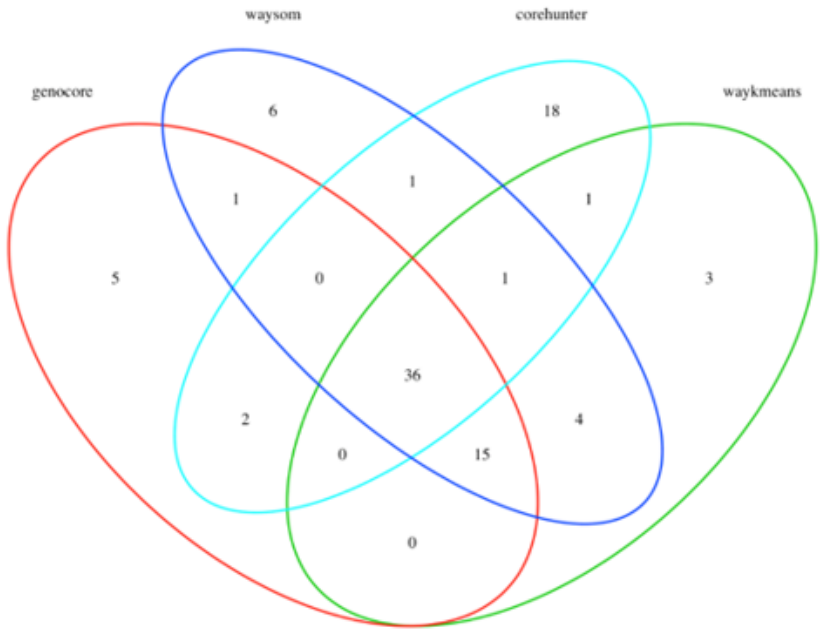

Figure 4: Venn diagram for the 1.5K Rice dataset.

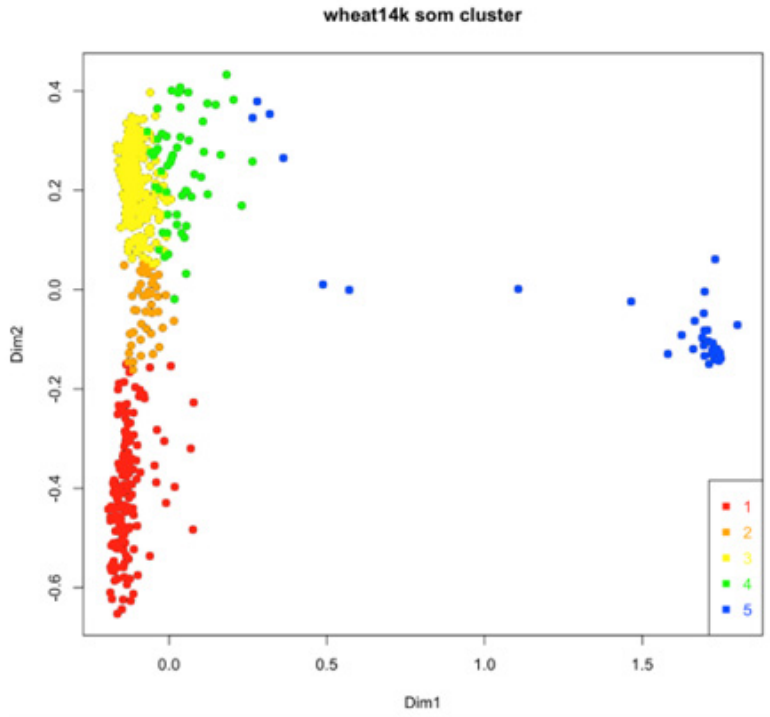

Figure 5: Final clustering results for 14K Wheat dataset based on the SOM algorithm.

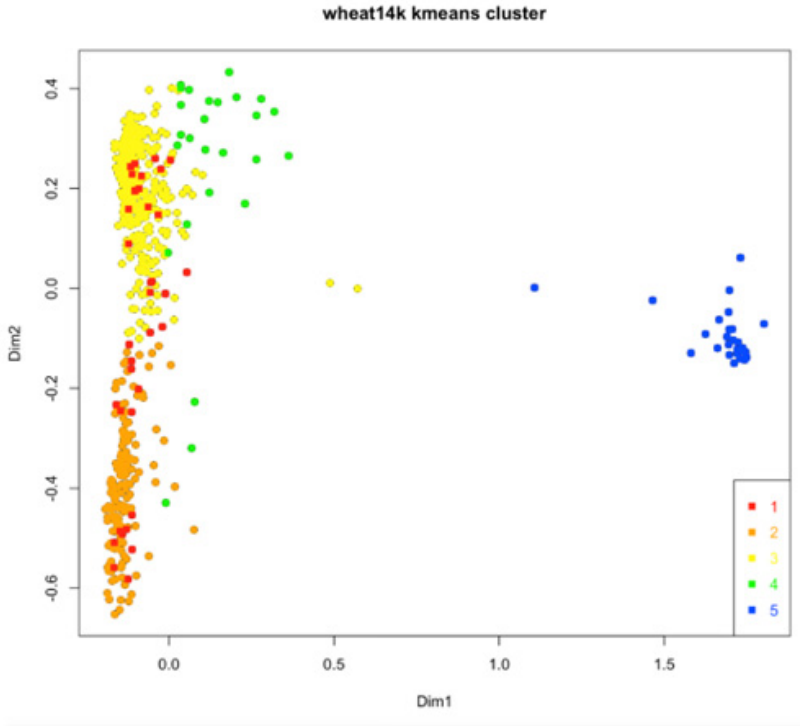

Figure 6: Final clustering results for 14K Wheat dataset based on the k-means algorithm.

Furthermore, Core Hunter 3 was the fastest method, requiring only 0.69 minutes, compared to Geno Core's 7.88 minutes and our method's 12–13 minutes. This difference in performance and speed stems from the distinction in optimization criteria: Core Hunter 3 uses the Min.MR as a major selection standard, while our method and Geno Core prioritize maximizing Coverage. The time discrep-

ancy is likely due to Core Hunter 3's use of a stochastic (randomized) selection process, which is more flexible and unstable in its stopping time (Figure 7). Illustrates the coverage growth efficiency. Similar to the 1.5K Rice data, our method and Geno Core show comparable, rapid, and stable growth curves, whereas Core Hunter 3 exhibits a noticeably slower rate of coverage accumulation. Additionally, the selection curves for the same clustering method are almost perfectly overlapped, further confirming that the choice between maximizing or minimizing the CS value has little effect on the final coverage efficiency. The Venn Diagram (Figure 8) shows that Core Hunter 3 again has the largest difference, with 69 unique accessions. The other two methods (Geno Core and our method) show high agreement, sharing over 80% of the accessions, with only 7-8 unique accessions each.

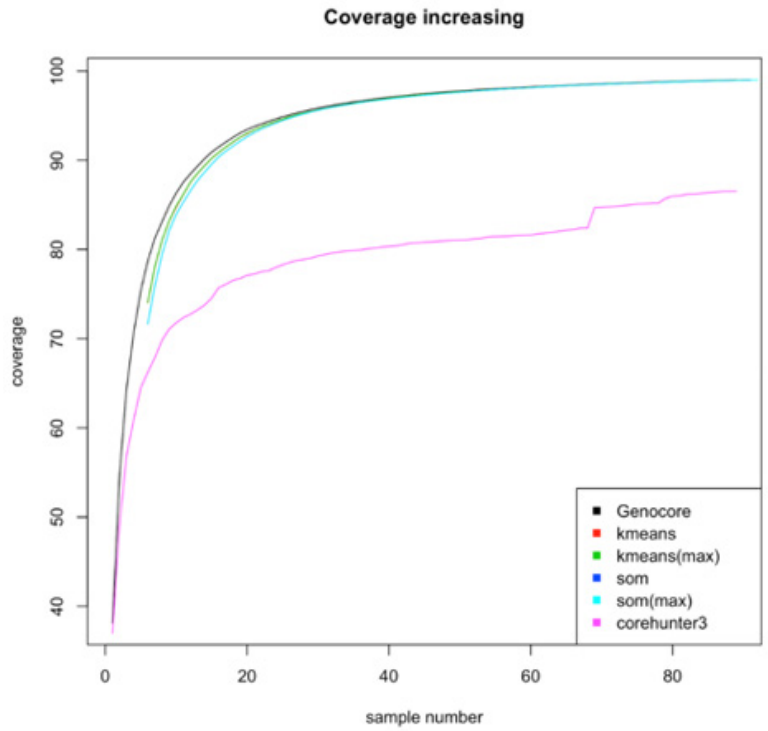

Figure 7: Coverage values versus the number of selected samples for each software in the 14K Wheat dataset.

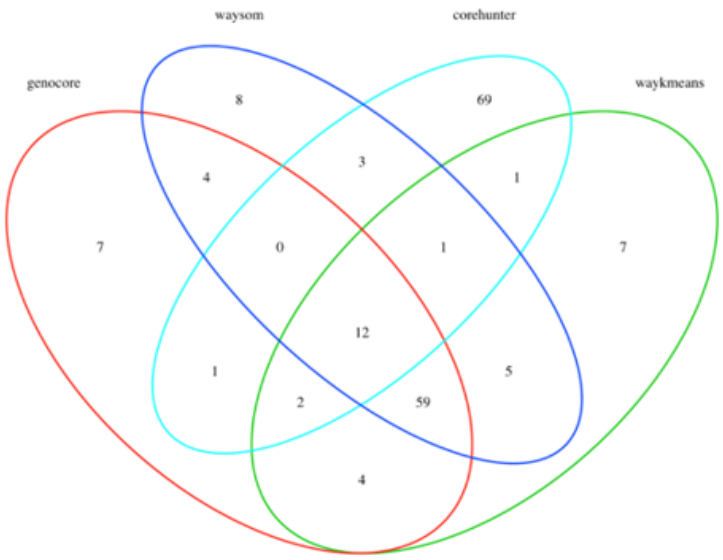

Figure 8: Venn diagram for the 14K Wheat dataset.

37K SNP Rice Data

The 37K SNP Rice data contains more than twice the amount of information compared to the previous Wheat dataset. Calculating the optimal number of clusters via the average Silhouette method showed that the Silhouette coefficient reached its maximum value at K=7, with a value exceeding 0.6, which is deemed acceptable for clustering. Following this result, the 413 accessions were partitioned into seven clusters using both SOM and K-means (Figures 9&10). As shown in the Figures, the distribution of the 413 accessions generated from the MCA coordinates exhibits a triangular distribution, but the tendency for accessions to aggregate at the three corner vertices appears even more pronounced than the 1.5K SNP Rice data. Similar to the 1.5K Rice data, one spatial region shows particularly high diversity, leading to more complex cluster boundaries. After clustering, core germplasm was selected, and the quality of the resulting core sets from all methods was compared (Table 3). Geno Core selected 96 accessions and our method selected one more accession across all variants, resulting in 97 accessions.

Table 3: Core Germplasm Selection Results of Different Methods for the 37K Rice Dataset.

Method	size	MR	Min.MR	SH	CV	Time (min)
Geno Core	**96**	0.1572	0.0157	10.5323	99.0068	22.7655
corehunter3	96	**0.1697**	**0.0906**	**10.6179**	94.5026	**18.5933**
K means	97	0.1571	0.0158	10.5386	**99.0136**	34.0492
K means(max)	97	0.1573	0.0158	10.5419	99.0104	39.1785
Som	97	0.1572	0.0158	10.5421	99.0113	38.4628
Som (max)	97	0.1572	0.0158	10.5427	99.0109	38.7124

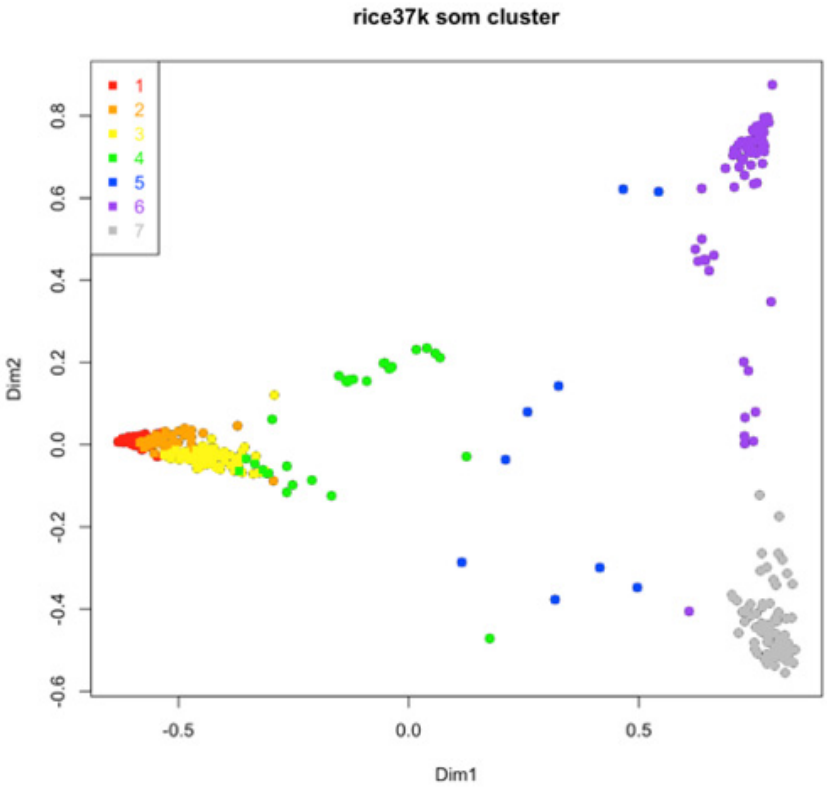

Figure 9: Final clustering results for 37K Rice dataset based on the SOM algorithm.

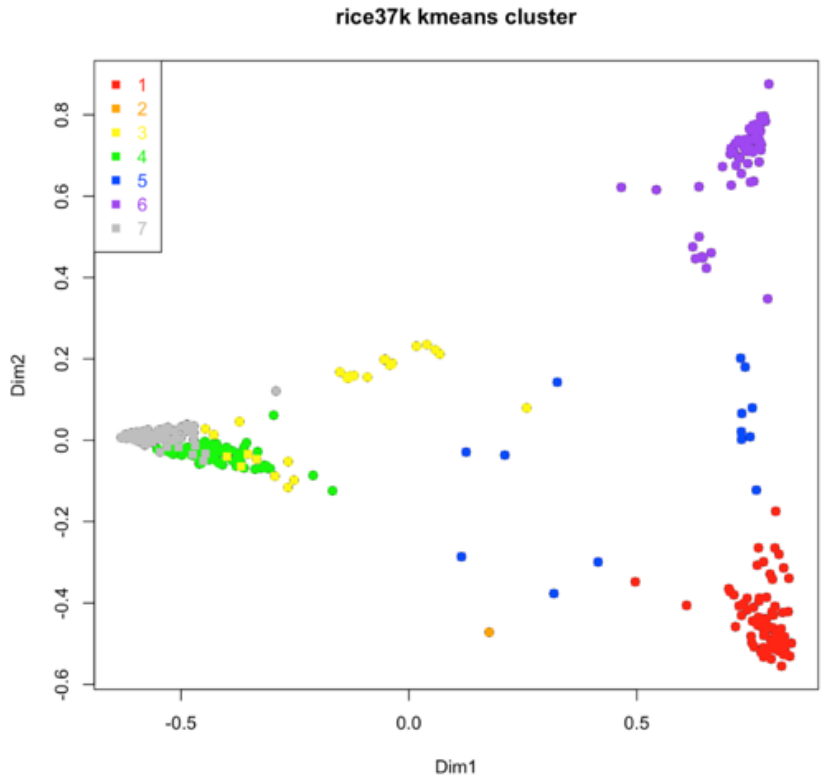

Figure 10: Final clustering results for 37K Rice dataset based on the k-means algorithm.

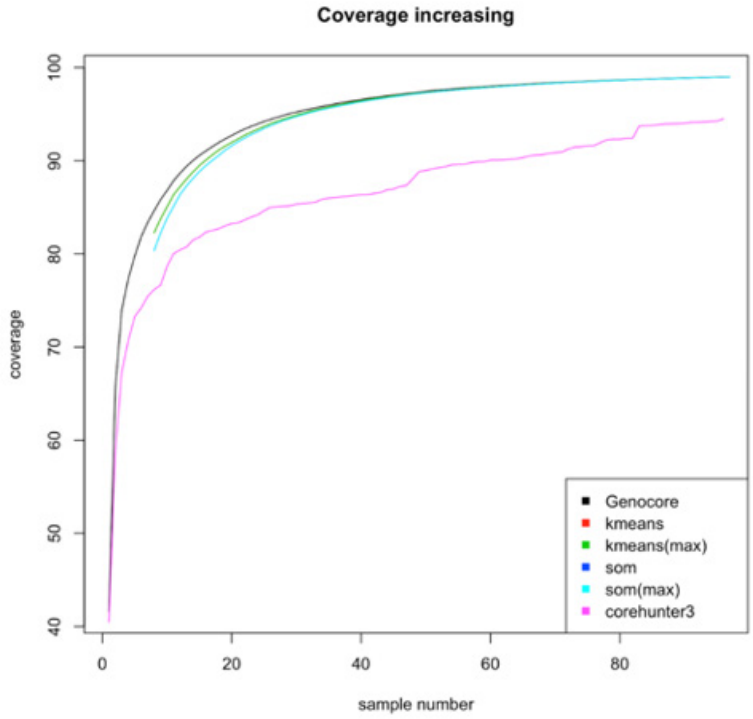

Figure 11: Coverage values versus the number of selected samples for each software in the 37K Rice dataset.

Regarding the other metrics, the results demonstrate that our method generally outperforms Geno Core. However, as the data volume increases, the computational time for our method also increases, requiring 12 to 17 minutes more than Geno Core. Core Hunter 3, despite being the fastest (only 18 minutes, 4 minutes less than Geno Core), clearly shows superior performance in all metrics except Coverage. As discussed with the 14K SNP Wheat data, Core Hunter 3's stopping criterion is likely the cause of its low coverage. In this dataset, surprisingly, despite the SNP count being more than double the Wheat data, Core Hunter 3's coverage deficit is smaller (94.5% coverage compared to 86.5% for the Wheat data, while the others reached 99%). This may be related to the more concentrated distribution of accessions in the MCA plot compared to the other two datasets. Since most data points are located near the three vertices, it might have been easier for Core Hunter 3, with its stochastic selection process, to randomly find accessions that achieve a certain level of coverage more quickly (Figure 11). Shows the Coverage Growth Efficiency, which yields similar results to the previous two datasets: our method and Geno Core show comparable, fast, and stable growth, while Core Hunter 3's growth is notably slower.

This phenomenon is likely due to the specific optimization criteria of Core Hunter 3 used in this simulation, which focuses on both the minimum distance between core accessions and the distance between unselected accessions and the core set. This may lead it to Favor accessions located at a certain optimal intermediate distance from others, rather than those at the extreme edges, which is visually supported by a slight tendency to avoid peripheral accessions. The Venn Diagram (Figure 12), reinforces these findings: Core Hunter 3 selected 54 accessions that were unique to its core set. The other two methods (Geno Core and our method) showed a high degree of overlap, sharing over 95.8% of their accessions, with only a 2 or 3 accession difference.

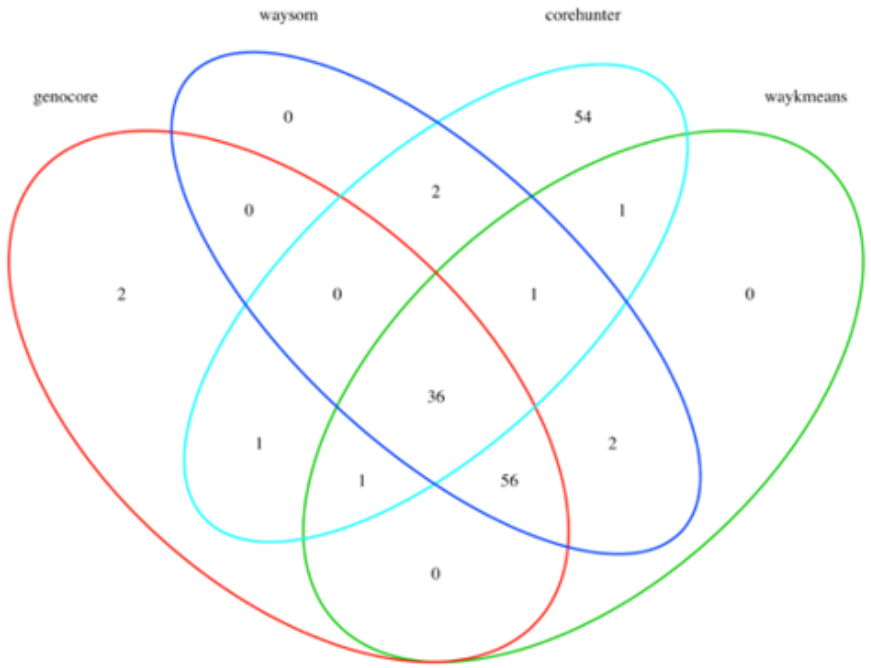

Figure 12: Venn diagram for the 37K Rice dataset.

820K SNP Wheat Data

Finally, we analyse the high-density 820K SNP Wheat dataset. Based on the comparative results from the previous three datasets, only the relatively best-performing variant of our method, K-means (max CS), is used as the representative for this analysis. Due to the sheer size of this dataset, the Multiple Correspondence Analysis (MCA) procedure was slightly modified: the calculation of dimensional coordinates employed the Burt Matrix method instead of the indicator matrix. While both methods yield similar relative positions, the coordinate values calculated using the Burt Matrix method have a smaller numerical range. Although the optimal number of clusters was found to be K=2 using the average Silhouette method, K=3, which yielded the second-highest silhouette value, was also included for analysis. The K-means algorithm was applied to partition the data into K=2 and K=3 clusters (Figure 13). The three-cluster result separates the accessions on the right into an additional distinct group (Table 4). Shows the performance metrics for core germplasm selection results.

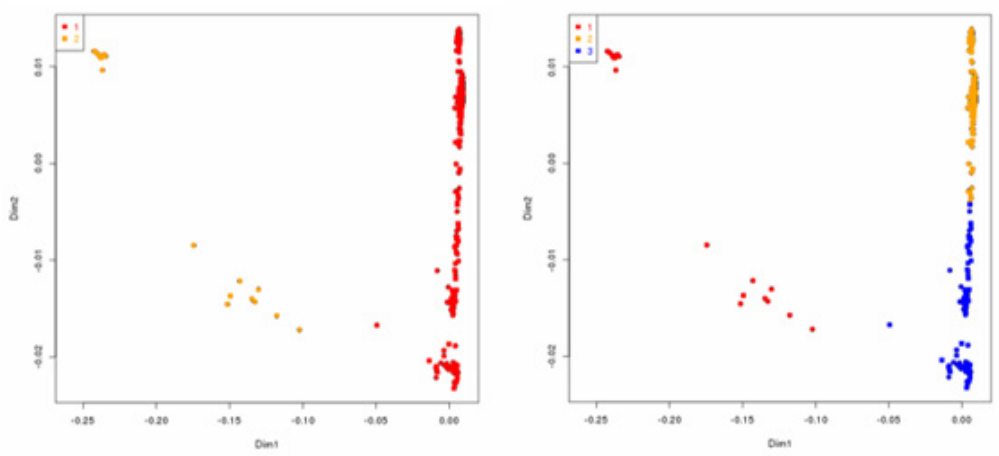

Figure 13: Final clustering results for 820K Wheat dataset based on the k-means algorithm. The left panel corresponds to the clusters of 2, the right panel to the clusters of 3.

Table 4: Core Germplasm Selection Results of Different Methods for the 820K Wheat Dataset.

Method	size	MR	Min. MR	SH	CV	Time (hr)
Geno Core	39	0.1458	0.0495	13.1281	99.0185	6.5517
corehunter3	39	**0.1523**	**0.0539**	**13.3315**	97.0793	**0.0689**
K means(max) 2 clusters	39	0.1457	0.0493	13.1233	99.0109	4.4917
K means(max) 3 clusters	40	0.1435	0.0492	13.1317	**99.0223**	4.5067

For this dataset, Core Hunter 3 again outperformed Geno Core and our method on all metrics except Coverage. Intriguingly, Core Hunter 3's coverage reached 97.07%, a performance that is atypically high compared to its previous results. As with the 37K Rice data, this phenomenon is likely due to the concentrated distribution of accessions and the low inherent diversity of the markers. A detailed examination of the 819,570 SNP markers showed a high frequency of markers with few genotype categories: 144,824 markers had only one category, 513,709 had only two, and only 161,037 had three categories. This implies that, excluding missing values, selecting a single accession instantly achieves 55.56% coverage, compared to only 38.11% for the 14K Wheat data. This suggests that achieving a high level of coverage is relatively easy in this dataset. The underlying cause of this marker distribution is likely high-density SNP analysis, where a strong Linkage Disequilibrium (LD) exists between tightly linked SNPs, meaning little or no recombination has occurred.

In addition, the results of our method with K=3 clusters showed an increase in both Shannon's Diversity Index and Coverage. In terms of Computational Time, the results here differ from the previous two datasets. For the 14K Wheat and 37K Rice data, our method was slower than Geno Core. However, using the Burt Matrix for MCA significantly reduced the time required for this 820K dataset, making our method approximately two hours faster than Geno Core. Core Hunter 3 was remarkably fast, requiring only 4.134 minutes, suggesting that algorithms incorporating stochastic selection may hold a decisive advantage when dealing with very large datasets exhibiting low marker diversity. The Coverage Growth Efficiency plot is omitted as it did not show significant differences from previous results. The Venn Diagram (Figure 14) shows that Core Hunter 3 selected 23 unique accessions, while the other two methods shared over 84.6% of their core accessions.

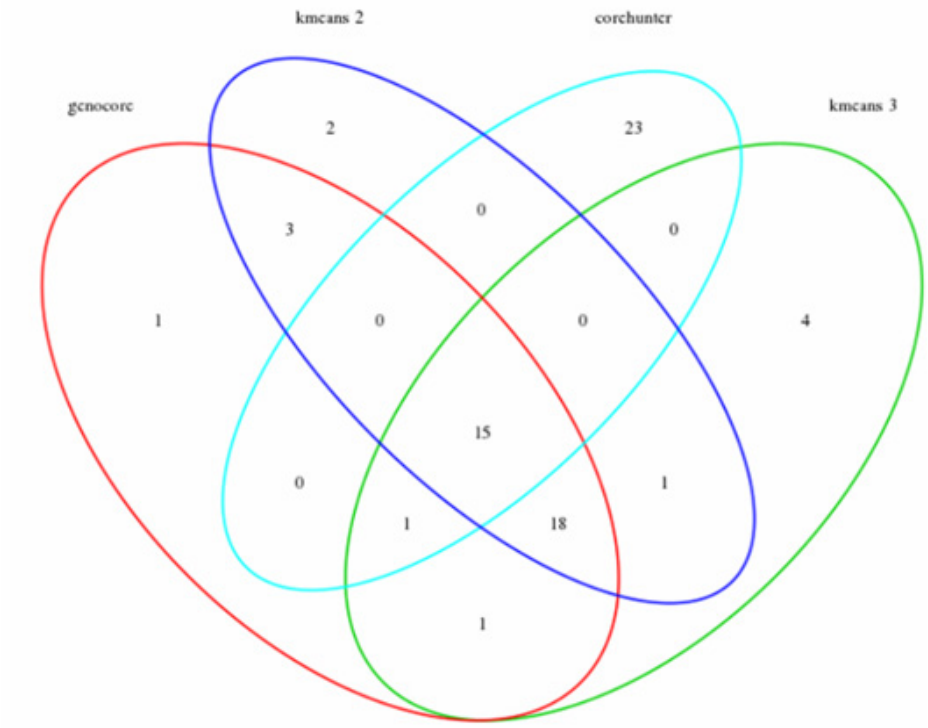

Figure 14: Venn diagram for the 820K Wheat dataset.

Conclusion

The results obtained across the four datasets demonstrate that while the differences between the selection methods were not especially significant for the smallest 1.5K SNP Rice data, the distinctions became more pronounced as the data volume increased. Overall, our proposed method consistently improves upon Geno Core's performance on three key metrics, Shannon's Diversity Index, Mean Modified Rogers Value, and Minimum Modified Rogers Value, while maintaining the standard of 99% coverage. However, our method typically requires slightly more computational time than Geno Core, increasing by approximately 0.5 times depending on the dataset size. For the current researches, the total execution time remains within an acceptable range. Notably, when the Burt Matrix method is used for coordinate calculation (as with the massive 820K SNP Wheat data), the computational burden is significantly reduced, resulting in a processing time approximately two hours faster than Geno Core.

The primary selection criterion of our proposed method is maximizing Coverage. Consequently, it generally does not match the performance of Core Hunter 3 on the other three diversity and distance metrics. Core Hunter 3 explicitly optimizes for these metrics by considering the distance between core accessions and the distance between unselected and core accessions. However, in the fundamental goal of core collection creation-that is, capturing the maximum amount of genetic information from the original population-our proposed method significantly outperforms Core Hunter 3, which frequently failed to reach the 99% coverage standard due to its stochastic selection process and stopping criteria. Future improvements will focus on incorporating more precise cluster analysis and refining selection standards by considering measures of inter-accession distance to further maximize the genetic distance within the core collection.

Acknowledgement

This work was supported by a grant from Taiwan's National Science and Technology Council (NSC 113-2118-M-002-009-).

Conflict of Interest

No conflict of interest.

References

1. Odong TL, Jansen J, van Eeuwijk FA, van Hintum TJ (2013) Quality of core collections for effective utilisation of genetic resources review, discussion and interpretation. Theor Appl Genet 126(2): 289-305.
2. Frankel OH (1984) Genetic Perspectives of Germplasm Conservation. In: Arber W, Illmensee K, Peacock WJ, Starlinger P, (Eds.,). Genetic Manipulation: Impact on Man and Society, Cambridge University Press, Cambridge, England 161-170.
3. Gouesnard B, Bataillon TM, Decoux G, Rozale C, Schoen DJ, et al. (2001) MSTRAT: An Algorithm for Building Germ Plasm Core Collections by Maximizing Allelic or Phenotypic Richness. J Hered 92(1): 93-94.
4. Kim KW, Chung HK, Cho GT, Ma KH, Chandrabalan D, et al. (2007) Power Core: a program applying the advanced M strategy with a heuristic search for establishing core sets. Bioinformatics 23(16): 2155-2162.

5. Thachuk C, Crossa J, Franco J, Dreisigacker S, Warburton M, et al. (2009) Core Hunter: an algorithm for sampling genetic resources based on multiple genetic measures. BMC Bioinformatics 10: 243.

6. Vargas AM, de Andrés MT, Ibáñez J (2016) Maximization of minority classes in core collections designed for association studies. Tree Genetics & Genomes 12: 28.

7. Jeong S, Kim JY, Jeong SC, Kang ST, Moon JK, et al. (2017) Geno Core: A simple and fast algorithm for core subset selection from large genotype datasets. PLoS ONE 12(7): e0181420.

8. H De Beukelaer, Davenport GF, Fack V (2018) Core Hunter 3: flexible core subset selection. BMC Bioinformatics 19(1): 203.

9. Greenacre M (2006) Multiple Correspondence Analysis in Practice 3rd edition, Florida: CRC Press.

10. Lloyd SP (1957) Least squares quantization in PCM. Technical Report RR-5497, Bell Lab.

11. MacQueen JB (1967) Some methods for classification and analysis of multivariate observations. In L. M. Le Cam, J. Neyman, (Eds.,). Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Vol. 1, University of California Press. pp. 281-297.

12. Kohonen T (1982) Self-organized formation of topologically correct feature maps. Biol. Cybern 43: 59-69.

13. Kohonen T (1984) Self-organization and associative memory. Heidelberg: Springer-Verlag, Berlin.

14. Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Computat Appl Mathemat 20: 53-65.