Article

Not peer-reviewed version

Genetic Diversity Analysis and Core Marker Identification of Shanlan Upland Rice Landraces Using Highly Informative InDel Markers

Yin Duan , Ping Gan , Qiuyun Lin , Yujie Zhou , Yuehui Lin , Zhenyu Xie , Xiaoning Wang * , Wei Hu *

Posted Date: 10 November 2025

doi: 10.20944/preprints202511.0565.v1

Keywords: genetic diversity; Shanlan upland rice; DNA fingerprinting; core collection

Article

Genetic Diversity Analysis and Core Marker Identification of Shanlan Upland Rice Landraces Using Highly Informative InDel Markers

Yin Duan [1,2,#], Ping Gan [3,#], Qiuyun Lin [1], Yujie Zhou [1], Yuehui Lin [1], Zhenyu Xie [1], Xiaoning Wang [4,*] and Wei Hu [1,*]

[1] Tropical Crops Genetic Resources Institute, Chinese Academy of Tropical Agricultural Sciences, Haikou 571101, China

[4] College of Tropical Crops, Yunnan Agricultural University, Pu'er 665099, China

[2] National Key Laboratory of Tropical Crop Breeding, Institute of Tropical Bioscience and Biotechnology & Sanya Research Institute, Chinese Academy of Tropical Agricultural Sciences, Sanya 572024, China

[3] Hainan Key Laboratory of Crop Genetics and Breeding, Institute of Food Crops, Hainan Academy of Agricultural Sciences, Haikou 571100, China

* Correspondence: wxning2599@163.com (X.W.); huwei.catas@gmail.com (W.H.)

# Yin Duan and Ping Gan contributed equally to this paper.

Abstract

Upland rice, particularly Shanlan upland rice from the mountainous regions of south China, represents an important genetic resource for improving rice productivity and resilience, especially in the face of climate change and environmental stress. However, understanding its genetic diversity and agronomic potential remains limited. This study systematically analyzed the genetic diversity and agronomic traits of 114 Shanlan upland rice landraces, primarily collected from the mountainous regions of Hainan, China, using both phenotypic evaluation and Insertion/Deletion (InDel) molecular markers. Significant phenotypic variability was observed in key agronomic traits, including plant height, tiller number, panicle length, grain shape, and yield-related traits. The 38 InDel markers used in this study exhibited an average polymorphism information content (PIC) of 0.43, suggesting moderate-to-low genetic diversity, likely due to geographic isolation and localized selection. Pairwise simple matching coefficient (SMC) analysis revealed substantial germplasm redundancy, with 7.9% of comparisons showing high similarity (SMC ≥ 0.85). Phylogenetic and K-means clustering analyses consistently grouped the landraces into three distinct genetic subpopulations. A DNA fingerprinting system was successfully developed using a reduced set of 19 core InDel markers, which was coupled with a QR code database linking molecular profiles with phenotypic data for efficient germplasm management. A network-based strategy identified a core collection of 54 accessions, streamlining the resource for future breeding and conservation efforts. These findings may provide a valuable molecular framework for breeding programs aimed at enhancing yield, lodging resistance, and stress tolerance in Shanlan upland rice.

Keywords: genetic diversity; Shanlan upland rice; DNA fingerprinting; core collection

1. Introduction

Rice (*Oryza sativa* L.) remains the cornerstone of global food security, providing staple nourishment for over half of the world's population. However, the stability and sustainability of rice production are increasingly threatened by global climate change, characterized by erratic precipitation patterns and increasing water scarcity [1]. Besides, the reduction in arable land and the intensification of land marginalization are diminishing its food production potential, thereby threatening long-term food security [2]. In this context, the development and utilization of drought-

 doi:10.20944/preprints202511.0565.v1

tolerant upland rice landraces adapted to rain-fed systems is a critical strategic priority for mitigating the effects of climate change, securing food supply, and fostering sustainable agriculture [3]. This is particularly valuable as upland rice effectively utilizes marginal mountainous and hilly terrains, thereby mitigating the competition for limited freshwater resources typically consumed by irrigated lowland rice.

Shanlan upland rice is a distinct and locally domesticated type of upland rice unique to the mountainous interior of in Haina province, China [4]. These landraces have been traditionally selected and cultivated by the indigenous Li ethnic communities over generations, characterized by heat and drought resistances, resulting in genotypes possessing superior resilience and adaptation to the local agro-ecological conditions. Beyond its ecological adaptability, Shanlan upland rice holds significant cultural memory and economic potential, particularly in the production of specialty foods and traditional wines [5,6], supporting regional economic diversity.

Shanlan upland rice exhibits many traits characteristic of wild rice, such as the presence of awns, lemmas, and strong shattering in many landraces, suggesting that Shanlan upland rice may have a more ancient genetic relationship with wild rice. Based on sequencing five genetic regions from 14 Shanlan upland rice samples in Hainan, compared to Asian cultivated rice and wild rice samples, it was found that Shanlan upland rice has lower genetic diversity than Asian cultivated rice, with about 85% of it being japonica-type, and is more closely related to wild rice from Guangdong and Hunan provinces, suggesting its potential origin from these regions [7]. Additionally, a genetic diversity analysis of 214 upland rice varieties from Southeast Asia and five provinces in southern China using SSR markers further supports this, hypothesizing that the Hainan Shanlan upland rice likely originated from Guangdong province and is genetically distinct from upland rice in Hunan Province [5].

Landraces are typically varieties that have been spontaneously selected by farmers, without undergoing systematic hybridization, inbreeding, or backcrossing processes. landraces are defined by several common characteristics: historical origin, high genetic diversity, local genetic adaptation, distinct identity, absence of formal genetic improvement, and their association with traditional farming systems [8]. Through long-term natural and farmer selection, they have accumulated a rich pool of adaptive alleles, allowing them to withstand diverse environmental pressures [9,10]. However, the Shanlan upland rice resource pool, confined within the limited geography of Hainan Island and subject to traditional, isolated farming practices, faces heightened risks of genetic homogeneity and the irreversible loss of unique genetic information. Furthermore, based on reports from multiple studies, the genetic base of Shanlan upland rice landraces is relatively narrow [7,11]. Despite this, Shanlan upland rice exhibits a broad genetic diversity in starch physicochemical parameters [12], which can be utilized to improve the cooking and eating quality in rice breeding. This apparent paradox—low genome-wide diversity yet high variation in starch-related traits—suggests strong selection pressure on key functional genes, underscoring the value of targeted conservation and utilization.

Systematic genetic diversity analysis constitutes the essential foundational step for effective germplasm identification, genetic improvement, resource conservation, and novel landrace selection [13,14]. Traditional methods of germplasm identification, relying solely on morphology such as plant stature, flower color, or grain shape, are inherently limited. Because morphological traits cannot accurately reveal the underlying genetic variation present within the germplasm collection [15]. Molecular markers, unlike morphological markers, overcome environmental influences and can detect subtle genetic variations that phenotypic evaluation may miss. They offer advantages such as stability, the ability to be detected in all tissues, and independence from factors like cell growth, development, and environmental conditions. These markers provide more reliable and precise genetic identification, making them essential for resource characterization and genetic analysis [16].

Molecular marker technology, which focuses on differences in DNA sequences, provides objective and environment-independent genetic information crucial for germplasm characterization, kinship analysis, and molecular breeding [17,18]. Various marker types have been historically

 doi:10.20944/preprints202511.0565.v1

applied in rice genetics, including simple sequence repeat (SSR), Insertion/Deletion (InDel), and single nucleotide polymorphism (SNP) markers, which are utilized for germplasm identification [5,19,20].

This study addresses the need for genetic improvement and resource conservation in Shanlan upland rice landraces. Despite its ecological adaptability, Shanlan upland rice landraces faces threats due to limited genetic diversity, which could hinder future breeding efforts. The research aimed to develop a comprehensive molecular and phenotypic framework for 114 Shanlan upland rice landraces, evaluating phenotypic variation, assessing genetic diversity using 38 InDel markers, and exploring genetic relationships through phylogenetic clustering. Key achievements include establishing a DNA fingerprinting system with a minimal set of highly discriminatory InDel markers, which facilitates efficient landrace authentication, redundancy control, and germplasm management. By narrowing the genetic pool through core germplasm selection, the study provides a streamlined resource for breeding efforts aimed at improving drought tolerance, yield potential, and culinary quality. The findings offer valuable insights for future breeding programs and provide a reproducible workflow for similar research, advancing the use of molecular markers in crop improvement. Additionally, we have developed a reproducible workflow, and the corresponding scripts are available on GitHub at https://github.com/huweihzau/Rice-DNA-Fingerprint-Analysis for use in similar research studies.

2. Materials and Methods

2.1. Experimental Materials and Field Management

The experimental materials used in this study comprised 114 Shanlan upland rice landraces, predominantly collected from the mountainous areas of Hainan province, China. These materials represent various local types, including black/red-shelled red rice landraces, Shanlan upland rice landraces with red and yellow husks. The experiment was conducted at the base of Chinese Academy of Tropical Agricultural Sciences (Danzhou, China). The seeds were sown on July 31, 2024. After 25 days, the seedlings were transplanted, with 48 plants of each landrace planted per plot. The experiment was conducted with three field replicates. Conventional field management practices for rice cultivation were followed to ensure consistency in environmental conditions, minimizing the impact of environmental factors on phenotypic measurements.

2.2. Phenotypic Data Collection and Evaluation

Throughout the growth period of the plants, phenotypic data were collected on heading date and plant height. For the plant height survey, five plants were randomly selected per landrace for measurement. Upon seed maturity, three plants with consistent growth were selected for harvesting and drying. Subsequently, data such as tiller number, effective tillers, and panicle length were measured. After threshing, seeds were analyzed using a digital seed tester (YTS-5D, China) to determine yield-related traits, including yield per plant, thousand-grain weight, total spikelets, seed setting rate, number of spikelets per panical, grain shape (grain length and width). The phenotypic data from the three field replicates were averaged to obtain the final phenotypic data for each trait.

2.3. Genomic DNA Extraction and Molecular Marker Selection

Genomic DNA from the 114 Shanlan upland rice landraces was extracted using the CTAB method [21]. Thirty-eight InDel molecular markers (Table S1) were selected for the study, based on markers previously published [22]. To ensure the markers represent a wide array of loci across the rice genome, the markers were randomly distributed across the 12 rice chromosomes, with almost three markers selected for each chromosome.

 doi:10.20944/preprints202511.0565.v1

2.4. PCR Amplification System and Procedure

The final volume of the polymerase chain reaction (PCR) reaction system was set to 15 μL. The components included: 20 ng of template DNA, 0.5 μL of primers, 7.5 μL of 2× Rapid Tap Master Mix (P222, Vazyme, China), and the remaining volume made up with deionized water. The PCR program was set as follows: 94°C for 3 minutes (initial denaturation), followed by 35 cycles of 94°C for 30 seconds (denaturation), 58°C for 30 seconds (annealing), and 72°C for 30 seconds (extension). A final extension was performed at 72°C for 2 minutes, followed by a 1-minute hold at 25°C and storage at 4°C.

2.5. Electrophoretic Detection and Polymorphism Analysis

The amplified PCR products were subjected to electrophoresis. Electrophoresis was conducted on a 1.5% agarose gel stained with 3% fluorescent DNA dye (GoldView, Zomanbio, China) under constant conditions of 400 V and 250 mA for 20 to 23 minutes. The gel was observed and photographed using a UV gel imaging system to record differences genotypes of amplified fragments across landraces, thereby determining the polymorphism of the InDel markers. The different genotypes are represented by different Arabic numerals. If a band fails to amplify after more than three attempts, it indicates the locus is absent in the landrace and is recorded as 0, which represents a special genotype.

2.6. Polymorphism Information Content Calculation

The polymorphism information content (PIC) was calculated for each polymorphic marker to assess its informativeness using the following formula:

$$\mathrm{PIC}_i = 1 - \sum_{j=1}^{n} p_{ij}^2,$$

where p_{ij} is the frequency of the j-th genotype of the i-th marker in the sample set, and n is the total number of genotypes [23,24]. A higher PIC value indicates greater polymorphism, broader applicability of the marker, and higher genetic diversity, whereas a value of 0 indicates a monomorphic marker.

2.7. Simple Matching Coefficient and Genetic Distance Calculation

The simple matching coefficient (SMC) is a measure of genetic similarity between two Shanlan upland rice landraces based on the genotype's comparison of molecular markers. The formula is as follows:

$$\mathrm{SMC} = \frac{\sum_{i=1}^{n} x_i}{N},$$

where x_i represents the state of the i-th marker for a pair of Shanlan upland rice landraces (≥ 1 for matching, 0 for non-matching). n is the total number of landraces,

N is the total number of markers being considered. The genetic distance matrix was then calculated by applying the standard transformation: $1 - SMC$ [25].

2.8. Phylogenetic Tree Construction

Subsequently, cluster analysis was performed using the Unweighted Pair-Group Method with Arithmetic Mean (UPGMA) to provide a graphical representation of the genetic relationships among the Shanlan upland rice landraces. The resulting UPGMA dendrogram was visualized using the ggtree package [26].

 doi:10.20944/preprints202511.0565.v1

2.9. DNA Fingerprinting System

To establish a DNA fingerprinting system for Shanlan upland rice landraces, the objective was to identify the minimum number of markers required for effective differentiation among all the landraces. A marker selection process was conducted, resulting in the selection of 19 markers from an initial set of 38. These 19 markers were subsequently utilized to develop a DNA fingerprinting system, which facilitates efficient variety identification, kinship analysis, and intellectual property protection.

Based on both the DNA fingerprinting and phenotypic data, QR codes were generated, encompassing information for 114 Shanlan upland rice landraces. The QR codes were created using the online platform https://cli.im/. Through these QR codes, users can access not only the specific DNA fingerprint data but also relevant agronomic trait information, enhancing the traceability and utility of the rice landrace data.

2.10. Core Germplasm Selection of Shanlan Upland Rice Landraces

To construct a core collection representing the genetic diversity of Shanlan upland rice landraces, we employed a graph-based clustering approach using the SMC matrix derived from genome-wide binary markers. The lower triangular SMC matrix was first symmetrized to generate a complete pairwise similarity matrix. It is generally accepted that an SMC value above 0.75–0.80 indicates similar varieties [19,27]. In this study, we set the threshold for SMC at 0.85, considering landraces with an SMC greater than 0.85 as similar. An undirected similarity network was then constructed by connecting pairs of accessions with SMC ≥ 0.85. Connected components (i.e., maximal subgraphs in which all nodes are reachable from one another) were identified using the igraph package in R [28]. Within each connected component, a single representative accession was selected as the core entry—specifically, the accession with the lexicographically smallest name to ensure reproducibility. Accessions forming singleton components (i.e., with no similarity ≥ 0.85 to any other accession) were retained as genetically unique or "distinctive" germplasm.

2.11. Visualization

Except for some figures and tables created in Excel, all other visualizations were generated using R [29]. The R packages used include adegenet, APE, Dplyr, ggplot2, ggtree, igraph, pheatmap and Poppr [26,28,30–35].

3. Results

3.1. Phenotypic Diversity and Variability in Agronomic Traits

The evaluation of the 114 Shanlan upland rice landraces demonstrated significant phenotypic heterogeneity, confirming the rich genetic resource contained within this landrace collection.

Fourteen representative Shanlan upland rice landraces were observed for plant architecture, panicle type, and grain shape. The observations revealed that the traditional Shanlan upland rice plants were generally tall, although some individuals exhibited shorter stature. Significant differences were observed among the different lines in terms of plant architecture and the number of tillers (Figure 1A–1C). Representative panicles were selected, showing that landraces such as SL79 and SL111 had relatively longer panicles. The glume color varied, including yellow, black, and brown-red glumes, further highlighting the diversity of Shanlan rice germplasm resources (Figure 1D). The grain shape observation exhibited considerable variation. Landraces SL52, SL66, SL111, and SL112 had relatively wider grains, while SL41 and SL92 had the shortest grain lengths. Additionally, seeds from SL52, SL111, and SL112 exhibited lemma, and these morphological differences serve as important reference indicators for variety identification (Figure 1E–1F). Additionally, awns were observed in several landraces, and some Shanlan upland rice landraces exhibited stronger shattering traits, suggesting that Shanlan upland rice may have a more ancient genetic relationship with wild rice.

 doi:10.20944/preprints202511.0565.v1

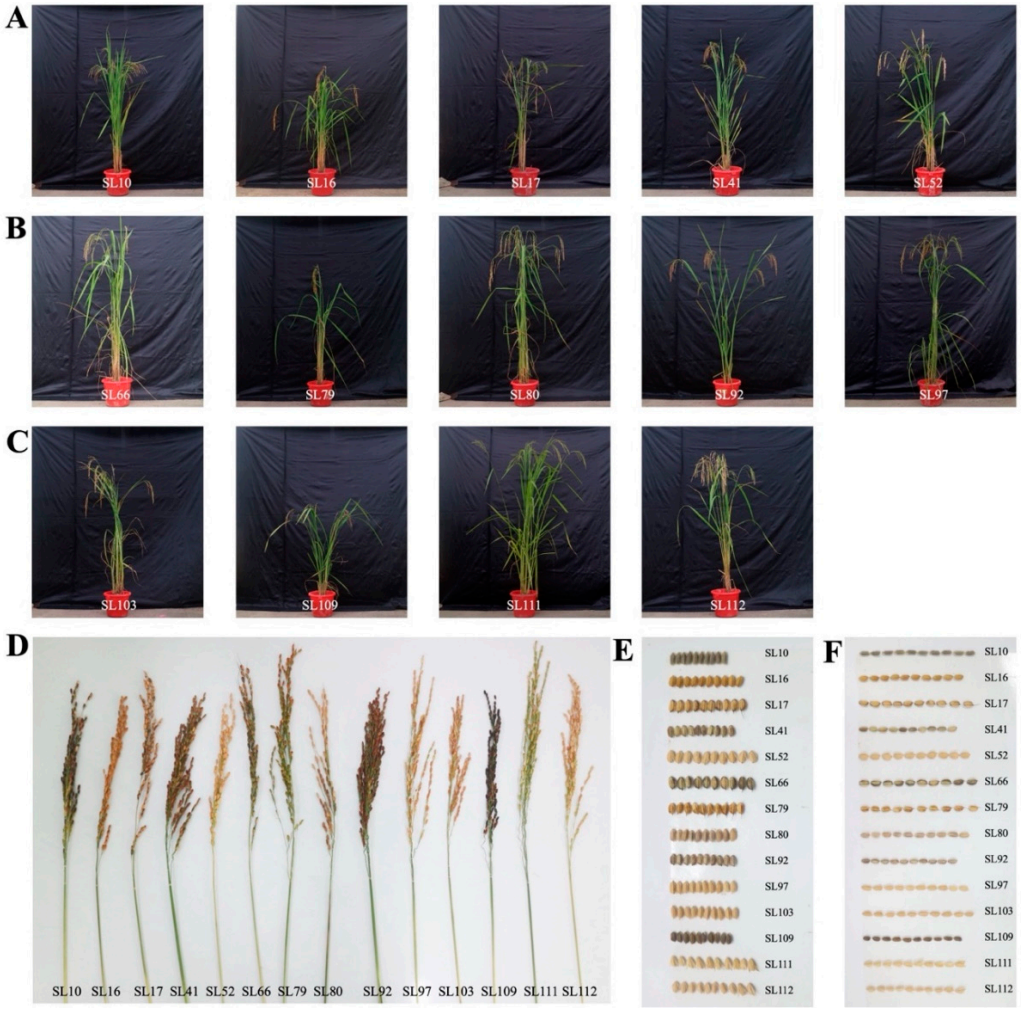


Figure 1. Representative phenotypes of Shanlan upland rice landraces. (**A–C**), plant architecture, (**D**), panicle types, (**E–F**), grain shapes of 14 Shanlan upland rice landraces.

3.2. Variability of Yield Related-Traits in Shanlan Upland Rice Landraces

In 2024, phenotypic traits of 114 Shanlan rice landraces were measured, revealing significant variability across the resource population (Figure S1, Table S2). Traits such as yield per plant, effective tillers, plant height, and seed setting rate showed considerable variation.

The heading date of Shanlan upland rice landraces varied from 70.0 to 96.0 days, with an average of 77.3 days. Although this variation indicates a diverse range of early- and late-maturing varieties, the overall trend leans towards earlier flowering landraces. This preference likely reflects local farmers' selection for early-maturing varieties, which may offer advantages such as reduced susceptibility to lodging and bird predation.

Plant height in the Shanlan upland rice landraces ranged from 88.3 cm to 160.8 cm, with an average of 123.4 cm. The significant variation in plant height suggests that these landraces exhibit diverse growth forms, which may influence traits such as lodging resistance and overall biomass production. While taller plants may accumulate greater biomass, the overall trend toward taller stature could result in lower yield potential and increased susceptibility to lodging, which can negatively affect rice productivity.

Yield per plant ranged from 5.1 g (SL75) to 25.6 g (SL15), with an average of 12.7 g. This trait is closely related to three key yield components: the number of tillers, number of spikelets per panicle, and thousand-grain weight. The interrelationships between number of tillers, spikelets per panicle, and thousand-grain weight significantly impact yield per plant. For example, SL15, with high values for all three yield components, achieved the highest yield per plant (25.6 g). These results highlight

 doi:10.20944/preprints202511.0565.v1

the importance of selecting for optimal combinations of these yield components to enhance rice productivity in breeding programs.

The grain shape in Shanlan upland rice landraces also exhibited significant variation. Grain length ranged from 7.2 mm (SL9) to 9.5 mm (SL82), with a mean of 8.3 mm, and grain width varied from 2.0 mm to 3.6 mm. The length-to-width ratio ranged from 2.3 to 4.3, with an average of 3.0. This suggests that local farmers tend to prefer varieties with shorter or medium-long grains, which likely aligns with local dietary consumption habits.

3.3. Correlation Analysis Among Yield Related-Traits

Correlation analysis was performed to elucidate the intricate relationships between different yield components (Figure 2). It revealed that yield per plant is most strongly associated with the parameters defining reproductive sink capacity. Specifically, Yield per plant showed a robust positive correlation with total spikelets (r = 0.61) and number of spikelets per panicle (r = 0.45). These findings demonstrate that in Shanlan upland rice, the primary mechanism for yield maximization is increasing the number of potential grains (the sink), rather than relying heavily on vegetative characteristics like tillering (number of tillers correlated positively with yield at r = 0.31).

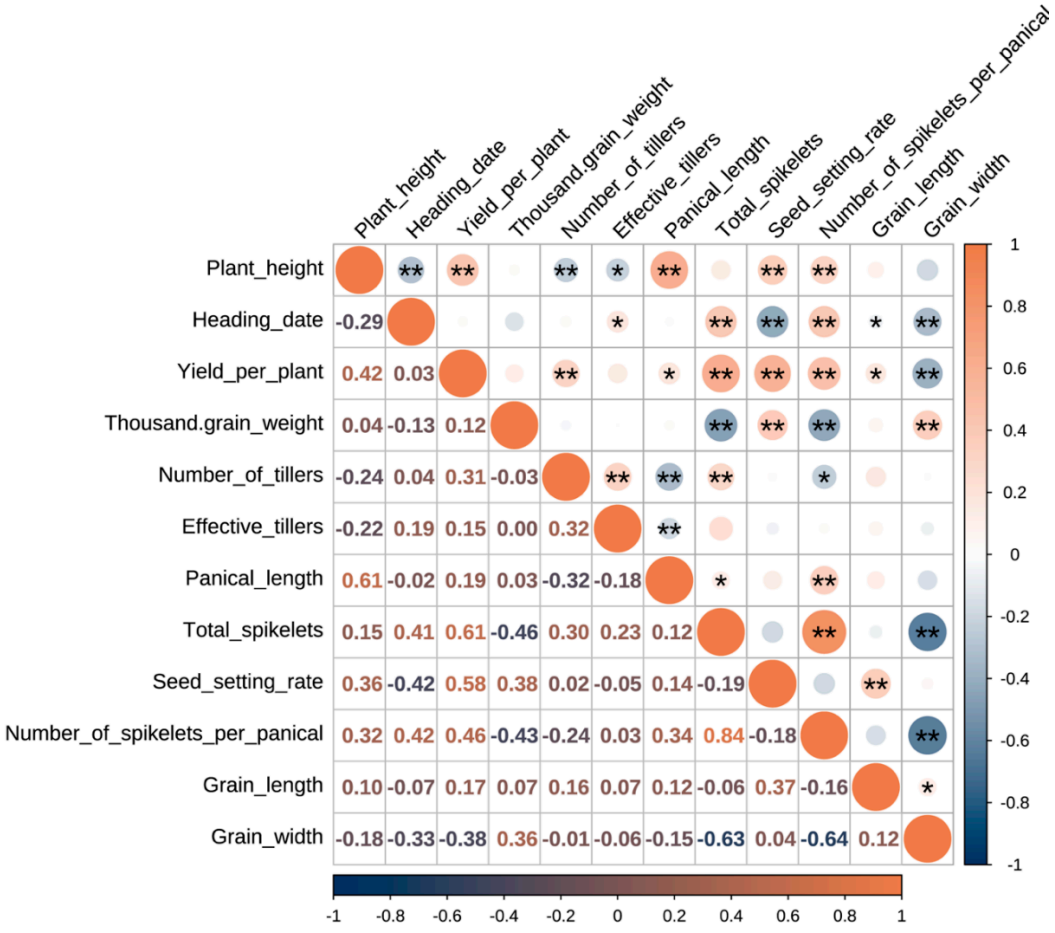


Figure 2. Correlation analysis of major traits in Shanlan upland rice landraces.

The study also documented classic physiological trade-offs inherent in plant architecture. A moderate negative correlation was found between thousand-grain weight (grain size) and number of spikelets per panicle (r = -0.43). This source-sink constraint implies that selection pressures aimed at increasing grain size often result in a corresponding reduction in the number of grains the plant can effectively fill, constraining overall volumetric yield.

 doi:10.20944/preprints202511.0565.v1

Furthermore, a complex adaptive conflict involving heading date was revealed. Heading date correlated positively with total spikelets (r = 0.42), suggesting that a longer growth duration allows more time for photosynthetic accumulation and reproductive development, leading to a larger potential sink size. However, this longer developmental cycle simultaneously resulted in a negative correlation with seed setting rate (r = -0.42). This negative relationship strongly suggests that later-maturing landraces are more likely to encounter environmental stresses, specifically high temperatures common in the late season of tropical Hainan, during the sensitive flowering and fertilization period, resulting in pollen sterility and reduced fertility.

3.4. Polymorphism Analysis and Informativeness of Indel Markers

The PIC values (Table S3) for 38 InDel markers were calculated to assess the informativeness of each marker. The PIC values ranged from 0.12 to 0.64, with an average of 0.43. More than half of the markers had PIC values below 0.5 (Figure S2), indicating that the overall genetic diversity of the Shanlan upland rice landraces is at a moderately low level. This level of genetic diversity suggests that while the population retains sufficient variation for differentiation and effective breeding programs, it may also reflect historical local evolutionary pressures, geographic isolation, and selection bottlenecks typical of landraces. Limited gene flow and long-term selective pressures have resulted in moderate but distinct levels of genetic variation.

Among the markers, 16 showed higher informativeness, with PIC values greater than 0.5. The most effective markers for genotype identification included InD2-136 (PIC = 0.64), InD10-100 (PIC = 0.60), and InD4-75 (PIC = 0.59). These highly polymorphic markers are especially valuable as they have the greatest ability to differentiate closely related landraces.

3.5. Genetic Similarity and Assessment of Germplasm Redundancy

To accurately map the genetic relatedness and redundancy within the Shanlan upland rice landraces, we calculated the pairwise SMC values for all 114 landraces (Table S4, Figure 3). The SMC values ranged widely from 0.18 to 1.00, with an average of 0.54 ± 0.12. In total, there are 6441 pairwise comparisons among 114 landraces. Approximately 72.3% of the comparisons fell between 0.40 and 0.70, confirming that the population shares a common genetic background while exhibiting moderate differentiation. Among these, 506 pairs have an SMC value greater than 0.85, indicating that 506 pairs are highly similar to each other. This accounts for 7.9% of the total comparisons. This aligns with the relatively narrow cultivation range of Shanlan rice, which likely results in lower genetic variation within the population.

 doi:10.20944/preprints202511.0565.v1

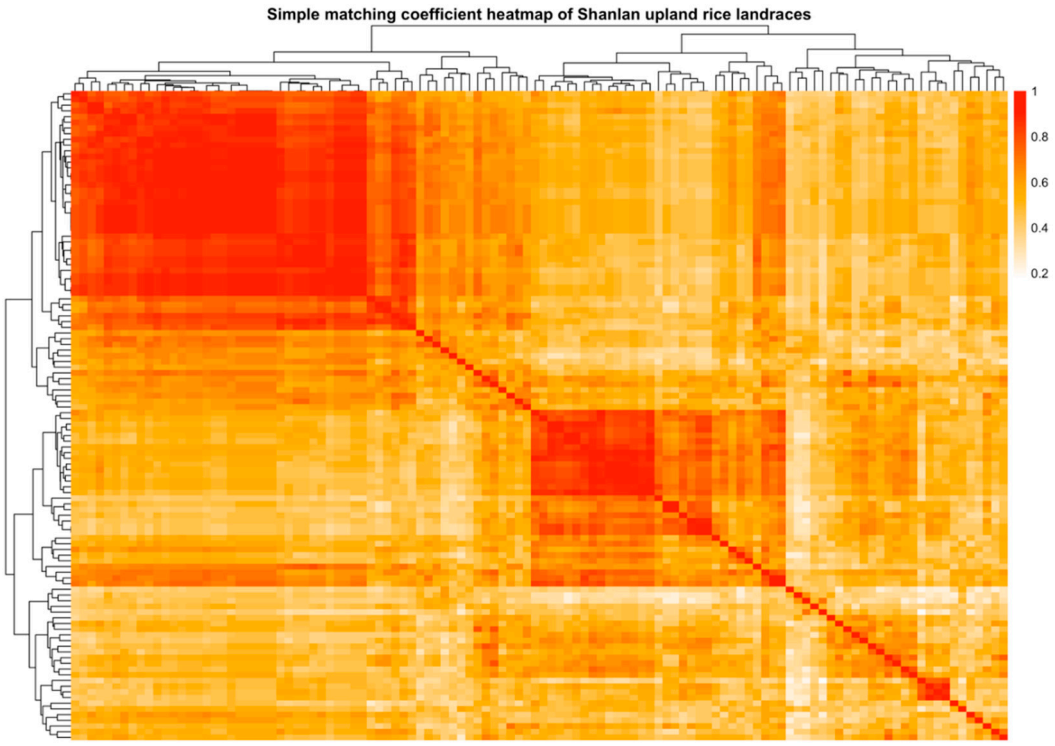


Figure 3. Heatmap of SMC.

The analysis also identified several key examples of genetic redundancy, where certain landrace pairs showed complete genetic identity (SMC = 1.00). This typically indicates the presence of duplicate germplasm due to either sampling repetition or the collection of genetically identical varieties from different locations. Specifically, pairs and groups such as SL27/SL28, SL50/SL51, SL73/SL74, and SL25/SL27/SL28/SL30/SL33 displayed genetic identity or near-identity, suggesting significant redundancy within the population. This also highlights the usefulness of calculating SMC as a rapid method for identifying whether germplasm is genetically identical.

On the other hand, the study successfully identified landraces representing extreme genetic divergence, such as the pair SL92 and SL59 (SMC = 0.18), and SL109 and SL59 (SMC = 0.21). These highly differentiated landraces are crucial for future breeding programs, as they represent genetic outliers that may harbor unique adaptive or performance-enhancing allele combinations.

3.6. Population Structure and Phylogenetic Relationships

Using genetic distances calculated from 38 InDel markers, an UPGMA analysis was performed to explore the potential genetic structure of Shanlan upland rice landraces. (Figure 4). The analysis clearly revealed that the 114 landraces could be grouped into three distinct primary genetic clusters or subpopulations. Notably, varieties SL25, SL27, SL28, SL30, and SL33 clustered together within the same branch, which is consistent with the results obtained from SMC analysis, as the genetic distance is inversely related to the SMC (Genetic distance = 1 - SMC). This clustering reflects the underlying genetic relationships and similarities among the varieties, further corroborating the findings of the SMC-based similarity analysis.

 doi:10.20944/preprints202511.0565.v1

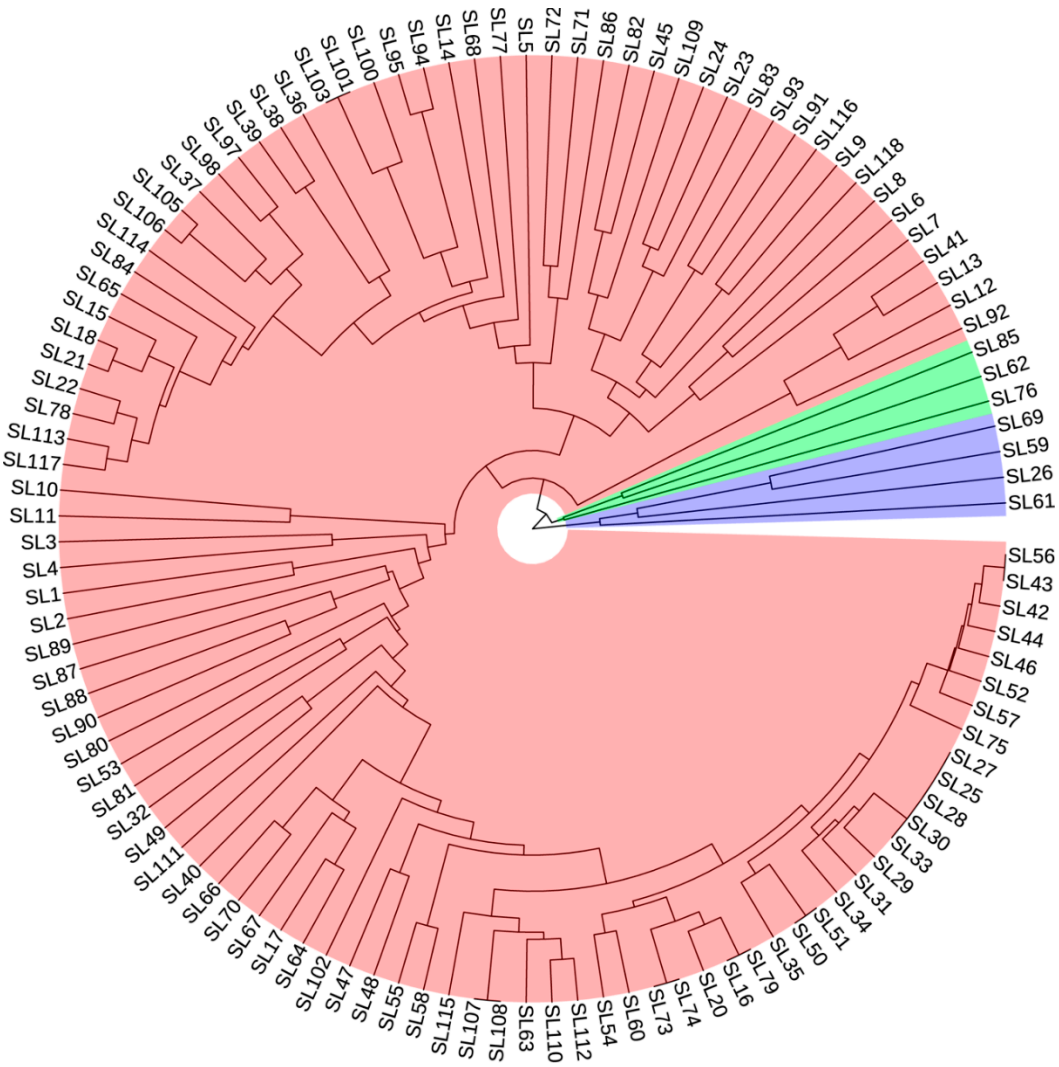


Figure 4. Clustering results of 114 Shanlan upland rice landraces based on 38 InDel marker pairs.

Within each identified cluster, the landraces displayed high genetic similarity, confirming close kinship and a shared history of localized selection. However, the genetic distances between these three major clusters were significantly larger, establishing a robust and clear population structure.

Using K-means clustering analysis with a 75% confidence interval, the majority of the varieties were grouped into three distinct subgroups (Figure S3). This result aligns well with the findings from the evolutionary tree analysis, further supporting the consistency and reliability of the genetic structure observed in the Shanlan upland rice landraces. The clustering analysis corroborates the pattern observed in the Neighbor-Joining tree, where the same three primary genetic clusters were identified.

3.7. Construction and Validation of the Minimum DNA Fingerprinting Marker Set

Based on the genotypic data from 38 markers, we constructed a heatmap to visualize the relationship between the varieties and the markers (Figure 5). The heatmap clearly highlights the genotypic variations across different landraces at various loci, providing an insightful overview of the genetic diversity within the Shanlan upland rice landraces.

 doi:10.20944/preprints202511.0565.v1

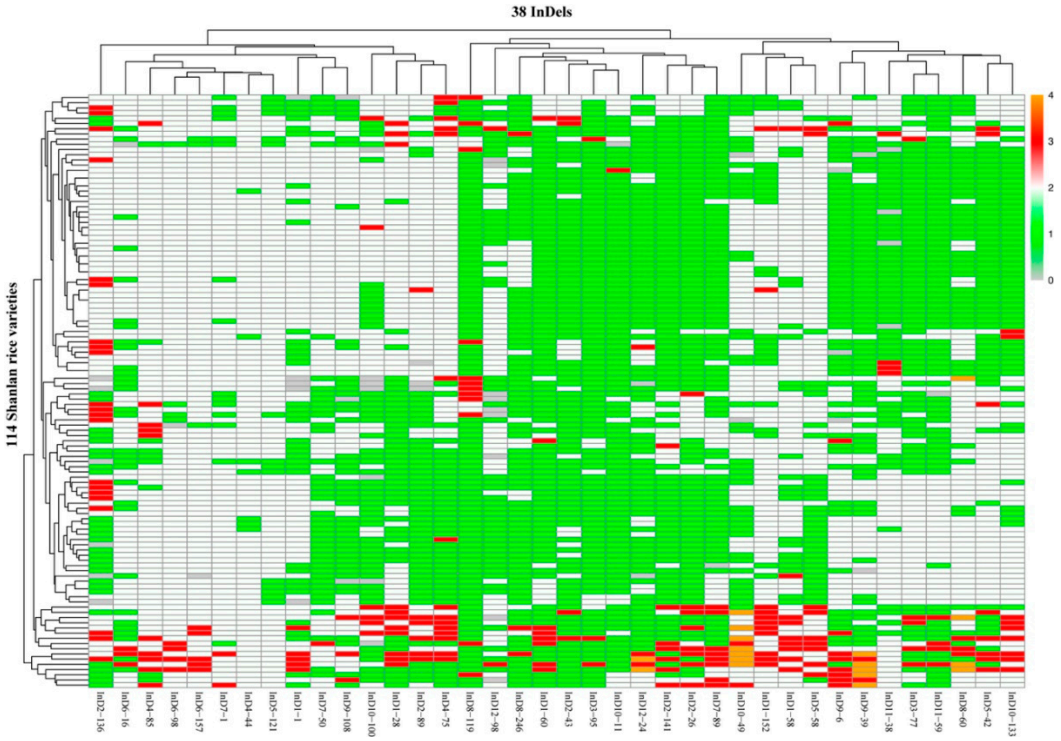
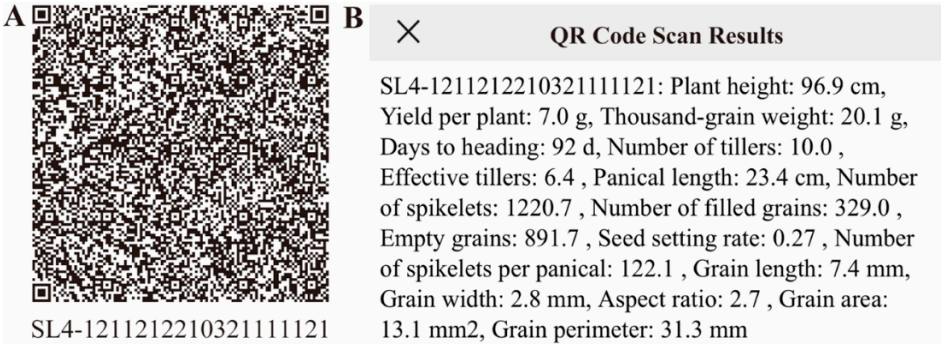


Figure 5. Heatmap illustrating the genotypic variation of 114 Shanlan upland rice landraces across 38 markers.

In the identification of germplasm resources, it is essential to establish an optimized minimal InDel marker set that can uniquely identify each landrace. Based on a comprehensive evaluation of PIC values and even distribution across the 12 rice chromosomes, a total of 19 core markers were selected from the original 38 InDel markers. These markers include: InD1-1, InD1-28, InD1-58, InD1-60, InD1-152, InD2-26, InD2-43, InD2-89, InD2-136, InD2-141, InD6-16, InD8-60, InD8-246, InD9-6, InD9-39, InD10-49, InD10-100, InD11-38, and InD12-98.

This reduced marker set provides sufficient resolution to generate unique digital fingerprints for all 114 Shanlan upland rice landraces. This critical achievement offers a robust and scientifically verifiable molecular identification system. The reduction in marker quantity from 38 to 19 significantly enhances the cost-effectiveness and efficiency of future germplasm screening, facilitating the rapid identification and validation of germplasm in the resource bank.

The utility of this minimal marker set is summarized in the digital fingerprint code (Table S5). At the same time, a QR code database have been constructed for the DNA fingerprint profiles and phenotypic data of the 114 Shanlan upland rice landraces. By scanning the QR codes, one can access the DNA fingerprint of a landraces, as well as its corresponding phenotypic traits. For example, scanning the QR code in Figure 6A will yield the results shown in Figure 6B. The first part contains the landrace number and DNA fingerprint profile, followed by phenotypic traits such as plant height, yield per plant, etc. This approach provides a convenient way for breeders to directly access and query the data in the future.



 doi:10.20944/preprints202511.0565.v1

Figure 6. Schematic diagram of DNA fingerprint QR codes (**A**) and their scanning results (**B**).

3.8. A Network-Based Core Germplasm of Shanlan Upland Rice Landraces

Using the network-based strategy, the core germplasm including 54 Shanlan upland rice landraces were identified from the original 114 landraces (Figure 7, Table S6), effectively reducing redundancy while preserving genetic representation. Among these, 7 landraces were classified as genetically distinctive, as they formed singleton connected components—indicating no close similarity (SMC ≥ 0.85) to any other landrace in the dataset. The remaining 47 core landraces each represent a distinct similarity cluster, collectively capturing the major genetic groups within the Shanlan upland rice germplasm. This core set provides a streamlined yet comprehensive resource for future phenotypic evaluation, genomic analysis, and breeding utilization.

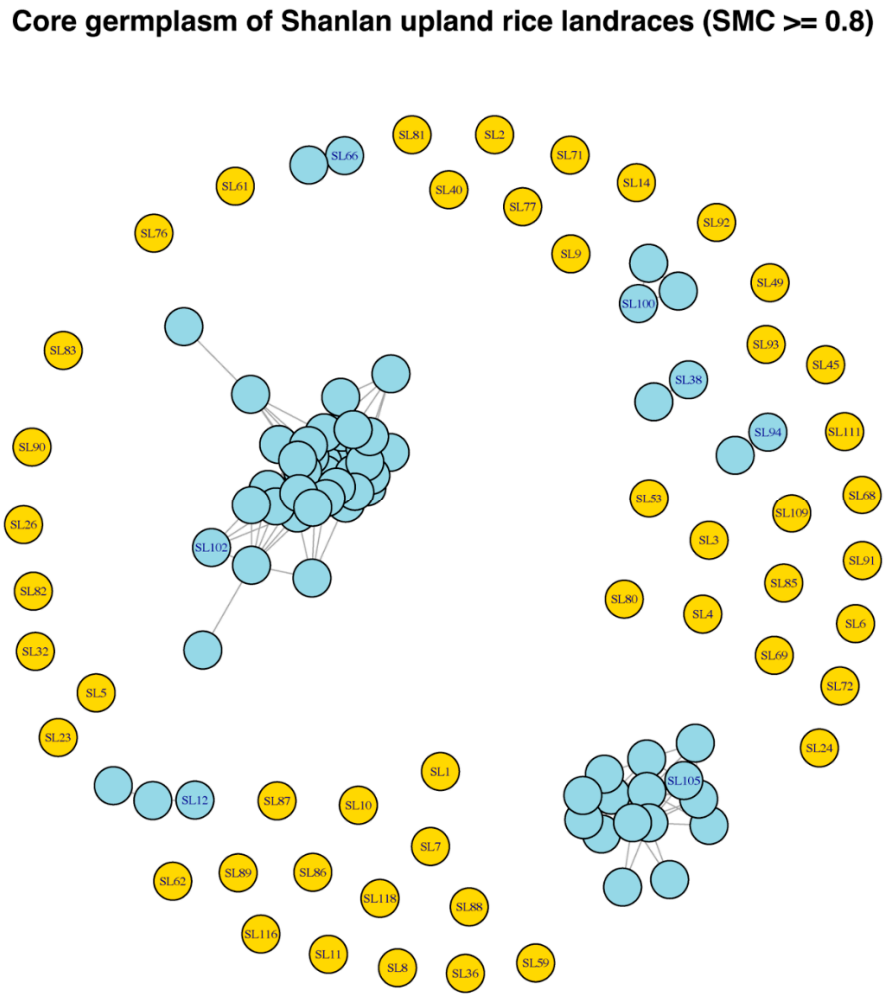


Figure 7. Network visualization of genetic similarity among Shanlan upland rice landraces. Nodes represent landraces, and edges connect pairs with SMC ≥ 0.85. Node colors indicate connected components (genetic groups), and core accessions are labeled.

4. Discussion

This study aimed to assess the genetic diversity and agronomic traits of 114 Shanlan upland rice landraces, focusing on their potential for breeding and conservation. Our findings highlighted significant phenotypic variation in traits such as plant height, tiller number, panicle length, and grain shape, which are crucial for future breeding strategies aimed at improving rice productivity and resilience. The moderate genetic diversity observed, reflected by a PIC of 0.43, suggests a narrow genetic base for the Shanlan upland rice landraces. This is consistent with previous studies indicating

 doi:10.20944/preprints202511.0565.v1

the limited genetic diversity of landraces from isolated farming areas, which could impede their adaptation to new environmental pressures and modern breeding needs [7,11].

4.1. Genetic Diversity and Its Implications for Breeding

Despite the narrow genetic base, the broad variation in starch physicochemical properties observed in this study is significant for breeding programs focusing on rice quality, particularly cooking and eating traits [12]. This paradox of low genetic diversity at the genomic level and high variation in agronomic traits underscores the strong selection pressure on specific genes related to starch composition. This finding supports the hypothesis that targeted conservation of these traits could enhance the culinary quality of Shanlan upland rice while preserving the genetic adaptability necessary for its ecological niche [10].

Molecular markers, particularly InDel markers, have proven to be valuable tools for identifying and tracking genetic diversity in rice germplasm [20]. The development of a minimal set of 19 core markers in this study enables more efficient identification, tracking, and management of Shanlan upland rice landraces. This marker set, combined with phenotypic data linked via QR codes, offers a practical solution for improving the traceability and commercial value of these landraces, making it an invaluable resource for both breeders and conservationists [19].

4.2. Challenges in Improving Agronomic Traits

The moderate-to-low genetic diversity, particularly within the three identified genetic clusters, indicates that breeding for higher yield potential and broader adaptability may require the introduction of new genetic material. The strategy of introgressing beneficial genes from conventional rice varieties into the Shanlan upland rice gene pool could enhance traits such as disease resistance, improved panicle architecture, and semi-dwarfism, which are critical for increasing yield and improving lodging resistance [3]. However, most Shanlan upland rice landraces are still traditional farmer varieties that have been propagated mainly through seed exchange and local circulation. Currently, the local government is actively promoting the breeding and utilization of Shanlan upland rice, aiming to develop regionally characteristic varieties and derivative products, such as Shanlan rice wine, to enhance the economic and cultural value of this unique genetic resource.

Plant height was identified as a major determinant of lodging resistance, with taller varieties showing a higher susceptibility to lodging and reduced yield. The findings of significant phenotypic variation in plant height (123.4 cm on average) highlight the necessity for breeding efforts that focus on achieving optimal plant height. Indeed, upland rice varieties are generally characterized by relatively tall plant stature. For instance, a report indicated that the average plant height of upland rice is around 125.3 cm [36]. The *sd1* gene has been widely used in rice breeding to reduce plant height and improve lodging resistance [37], which is crucial for increasing yield potential and enhancing overall crop stability. By incorporating this gene into the high-plant rice landraces, breeding programs could help balance plant stature, reduce susceptibility to lodging [38], and ultimately improve yield in these traditional landraces, all while maintaining their ecological and stress-resilient traits. Future breeding should prioritize these dual goals of improved lodging resistance and maintained drought tolerance for Shanlan upland rice.

4.3. Adapting to Environmental Stressors

The negative correlation between heading date and seed setting rate observed in this study ($r = -0.42$) points to a key physiological constraint—high-temperature stress during the late reproductive phase. In recent years, extreme climate events, especially frequent high temperatures, have severely impacted crop yields [39]. Late-maturing varieties exposed to high temperatures in tropical regions often experience reduced fertility due to pollen sterility [40]. Thus, breeding for early-maturing varieties is a priority to mitigate these stress effects. Early-maturing varieties would escape the high-temperature stress period, completing their reproductive cycle before the peak heat of the season.

 doi:10.20944/preprints202511.0565.v1

However, it is crucial to balance the shortening of the growth period with maintaining a high sink capacity for maximum yield [41,42].

Another strategy is the identification and utilization of genotypes that can maintain high seed setting rates despite late-heading, possibly through the introgression of quantitative trait loci (QTLs) that confer heat tolerance [43,44]. Such genotypes would allow for sustained productivity under high-temperature stress, a critical consideration for tropical upland rice production.

4.4. Practical Applications of DNA Fingerprinting and Core Germplasm

The establishment of the DNA fingerprinting system using 19 InDel markers represents a significant advancement in the characterization of Shanlan upland rice. The high-resolution capability of this system provides a reliable method for distinguishing between varieties, thus aiding in the identification and protection of intellectual property associated with rice landraces. In China, the current standard for new rice variety identification requires the use of 48 SSR markers [45]. In contrast, our system employs far fewer markers, which greatly reduces the experimental workload while maintaining sufficient discriminatory power. Moreover, the agarose gel–based InDel markers used in this study are highly practical, as they can be easily applied by most research institutions for self-assessment prior to official variety registration or evaluation. The QR code database developed in this study enhances the accessibility of both genetic and phenotypic data, promoting greater transparency and traceability in breeding and conservation efforts [46].

The identification of a 54-accession core germplasm collection, selected based on genetic similarity and redundancy, is particularly important for streamlining breeding efforts. By reducing genetic redundancy, the core collection retains the maximum amount of genetic diversity while minimizing the costs and resources required for future breeding projects [47,48]. Furthermore, the inclusion of unique accessions in the core set opens opportunities for genome-wide association studies (GWAS) and fine mapping of important traits such as drought tolerance and yield potential.

5. Conclusions

This research successfully characterized the genetic resources of 114 Shanlan upland rice landraces using an integrated phenotypic and molecular approach. We confirmed the significant phenotypic variation and, importantly, established a reliable, cost-effective DNA fingerprinting system using a minimal set of 19 core InDel markers for germplasm authentication and management. The identification of a 54-accession core germplasm landraces reduces redundancy while preserving the population's genetic breadth. The findings highlight critical breeding objectives for Shanlan upland rice, specifically the need to reduce plant height for lodging resistance, improve seed setting rate under high-temperature stress, and strategically broaden the genetic base by incorporating exotic germplasm. The developed QR code database provides an efficient molecular and phenotypic data query system, enhancing the traceability and utility of this vital resource for future breeding efforts.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org, Figure S1. Analysis of phenotypic variation in major traits of Shanlan upland rice landraces; Figure S2. Polymorphism statistics of 38 InDel molecular markers; Figure S3. Principal component analysis (PCA) of 114 Shanlan upland rice landraces based on InDel markers; Table S1. InDel markers used in this study; Table S2. Agronomic trait measurement data for Shanlan upload rice landraces; Table S3. Genotypes and PIC values of the 38 InDel markers identified; Table S4. Simple matching coefficient among 114 Shanlan upland rice landraces; Table S5. Minimum DNA fingerprint set for Shanlan upland rice landraces based on 19 InDel markers; Table S6. Core germplasm of Shanlan upland rice landraces.

Author Contributions: Conceptualization, W.H, P.G. and X.W.; methodology, P.G. and Q.L.; software, W.H. and P.G.; validation, Y.D., Q.L., and Y.Z.; formal analysis, W.H.; investigation, Y.D., Y.L., and Z.X; resources, Q.L. and Z.X.; data curation, W.H. and P.G.; writing—original draft preparation, Y.D., P.G., W.H. and X.W.; writing—review and editing, W.H. and X.W.; visualization, W.H. and P.G.; supervision, W.H.; project

 doi:10.20944/preprints202511.0565.v1

administration, W.H. and X.W.; funding acquisition, W.H. and X.W. Y.D. and P.G. contributed equally to this paper. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Key Project of Regional Joint Fund of National Natural Science Foundation (U22A20476), the Central Public-interest Scientific Institution Basal Research Fund for Chinese Academy of Tropical Agricultural Sciences (1630032021015), and the Earmarked Fund for Hainan Agriculture Research System (HNARS-04-G02).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The data presented in this study are available in the main text and supplementary materials. Additional data can be requested from the corresponding author. The code and example data for this study are available at https://github.com/huweihzau/Rice-DNA-Fingerprint-Analysis.

Acknowledgments: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

InDel	Insertion/Deletion
PIC	polymorphism information content
SMC	simple matching coefficient
SSR	simple sequence repeat
SNP	single nucleotide polymorphism
PCR	polymerase chain reaction
UPGMA	Unweighted Pair-Group Method with Arithmetic Mean
QTLs	quantitative trait loci
GWAS	genome-wide association studies

References

1. Habib-ur-Rahman, M.; Ahmad, A.; Raza, A.; Hasnain, M.U.; Alharby, H.F.; Alzahrani, Y.M.; Bamagoos, A.A.; Hakeem, K.R.; Ahmad, S.; Nasim, W. Impact of climate change on agricultural production; Issues, challenges, and opportunities in Asia. *Front. Plant Sci.* **2022**, *13*, 925548.
2. Lu, D.; Wang, Z.; Su, K.; Zhou, Y.; Li, X.; Lin, A. Understanding the impact of cultivated land-use changes on China's grain production potential and policy implications: A perspective of non-agriculturalization, non-grainization, and marginalization. *Journal of Cleaner Production* **2024**, *436*, 140647.
3. Bernier, J.; Atlin, G.N.; Serraj, R.; Kumar, A.; Spaner, D. Breeding upland rice for drought resistance. *J. Sci. Food Agric.* **2008**, *88*, 927-939.
4. Yang, X.; Liu, C.; Niu, X.; Wang, L.; Li, L.; Yuan, Q.; Pei, X. Research on lncRNA related to drought resistance of Shanlan upland rice. *BMC Genomics* **2022**, *23*, 336.
5. Li, R.; Huang, Y.; Yang, X.; Su, M.; Xiong, H.; Dai, Y.; Wu, W.; Pei, X.; Yuan, Q. Genetic diversity and relationship of Shanlan upland rice were revealed based on 214 upland rice SSR markers. *Plants* **2023**, *12*, 2876.
6. Hao, Y.; Li, J.; Zhao, Z.; Xu, W.; Wang, L.; Lin, X.; Hu, X.; Li, C. Flavor characteristics of Shanlan rice wines fermented for different time based on HS-SPME-GC-MS-O, HS-GC-IMS, and electronic sensory analyses. *Food Chem.* **2024**, *432*, 137150.
7. Yuan, N.; Wei, X.; Xue, D.; Yang, Q. The origin and evolution of upland Rice in Li ethnic communities in Hainan Province, China. *Journal of Plant Genetic Resources* **2013**, *14*, 202-207.
8. Villa, T.C.C.; Maxted, N.; Scholten, M.; Ford-Lloyd, B. Defining and identifying crop landraces. *Plant genetic resources* **2005**, *3*, 373-384.
9. Marone, D.; Russo, M.A.; Mores, A.; Ficco, D.B.; Laidò, G.; Mastrangelo, A.M.; Borrelli, G.M. Importance of landraces in cereal breeding for stress tolerance. *Plants* **2021**, *10*, 1267.

10. Dwivedi, S.L.; Ceccarelli, S.; Blair, M.W.; Upadhyaya, H.D.; Are, A.K.; Ortiz, R. Landrace germplasm for improving yield and abiotic stress adaptation. *Trends Plant Sci.* **2016**, *21*, 31-42.
11. Yang, G.; Yang, Y.; Guan, Y.; Xu, Z.; Wang, J.; Yun, Y.; Yan, X.; Tang, Q. Genetic diversity of Shanlan upland rice (Oryza sativa L.) and association analysis of SSR markers linked to agronomic traits. *BioMed Research International* **2021**, *2021*, 7588652.
12. Zhang, L.; Deng, B.; Peng, Y.; Gao, Y.; Hu, Y.; Bao, J. Population structure and genetic diversity of Shanlan landrace rice for GWAS of cooking and eating quality traits. *Int. J. Mol. Sci.* **2024**, *25*, 3469.
13. Cheng, S.; Feng, C.; Wingen, L.U.; Cheng, H.; Riche, A.B.; Jiang, M.; Leverington-Waite, M.; Huang, Z.; Collier, S.; Orford, S. Harnessing landrace diversity empowers wheat breeding. *Nature* **2024**, *632*, 823-831.
14. Salgotra, R.K.; Chauhan, B.S. Genetic diversity, conservation, and utilization of plant genetic resources. *Genes* **2023**, *14*, 174.
15. Jansky, S.H.; Dawson, J.; Spooner, D.M. How do we address the disconnect between genetic and morphological diversity in germplasm collections? **2015**, doi:10.3732/ajb.1500203.
16. Jacob, S.R.; Singh, N.; Srinivasan, K.; Gupta, V.; Radhamani, J.; Kak, A.; Pandey, C.; Pandey, S.; Aravind, J.; Bisht, I.; et al. Molecular Characterization of Plant Genetic Resources. *National Bureau of Plant Genetic Resources* **2015**, 323.
17. Bunjkar, A.; Walia, P.; Sandal, S.S. Unlocking genetic diversity and germplasm characterization with molecular markers: strategies for crop improvement. *Journal of Advances in Biology & Biotechnology* **2024**, *27*, 160-173.
18. Nadeem, M.A.; Nawaz, M.A.; Shahid, M.Q.; Doğan, Y.; Comertpay, G.; Yıldız, M.; Hatipoğlu, R.; Ahmad, F.; Alsaleh, A.; Labhane, N. DNA molecular markers in plant breeding: current status and recent advancements in genomic selection and genome editing. *Biotechnol. Biotechnol. Equip.* **2018**, *32*, 261-285.
19. Salgotra, R.; Gupta, B.; Bhat, J.A.; Sharma, S. Genetic diversity and population structure of Basmati rice (Oryza sativa L.) germplasm collected from North Western Himalayas using trait linked SSR markers. *PLoS ONE* **2015**, *10*, e0131858.
20. Sahu, P.K.; Mondal, S.; Sharma, D.; Vishwakarma, G.; Kumar, V.; Das, B.K. InDel marker based genetic differentiation and genetic diversity in traditional rice (Oryza sativa L.) landraces of Chhattisgarh, India. *PLoS ONE* **2017**, *12*, e0188864.
21. Murray, M.; Thompson, W. Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Res.* **1980**, *8*, 4321-4326.
22. Hu, W.; Zhou, T.; Wang, P.; Wang, B.; Song, J.; Han, Z.; Chen, L.; Liu, K.; Xing, Y. Development of Whole-Genome Agarose-Resolvable LInDel Markers in Rice. *Rice* **2020**, *13*, 1–11.
23. Anderson, J.A.; Churchill, G.; Autrique, J.; Tanksley, S.; Sorrells, M. Optimizing parental selection for genetic linkage maps. *Genome* **1993**, *36*, 181-186.
24. Liu, J.; Li, J.; Qu, J.; Yan, S. Development of Genome-Wide Insertion and Deletion Polymorphism Markers from Next-Generation Sequencing Data in Rice. *Rice* **2015**, *8*, 27.
25. Le, D.; Nguyen, C.M.; Mann, R.K.; Yerkes, C.N.; Kumar, B.V. Genetic diversity and herbicide resistance of 15 Echinochloa crus-galli populations to quinclorac in Mekong Delta of Vietnam and Arkansas of United States. *J. Plant Biotechnol.* **2017**, *44*, 472-477.
26. Yu, G.; Smith, D.K.; Zhu, H.; Guan, Y.; Lam, T.T.Y. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution* **2017**, *8*, 28-36.
27. Mazumder, S.R.; Hoque, H.; Sinha, B.; Chowdhury, W.R.; Hasan, M.N.; Prodhan, S.H. Genetic variability analysis of partially salt tolerant local and inbred rice (Oryza sativa L.) through molecular markers. *Heliyon* **2020**, *6*.
28. Csardi, M.G. Package 'igraph'. *Last accessed* **2013**, *3*, 2013.
29. Ihaka, R.; Gentleman, R. R: a language for data analysis and graphics. *J. Comput. Graphical Stat.* **1996**, *5*, 299-314.
30. Jombart, T. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **2008**, *24*, 1403-1405.

 doi:10.20944/preprints202511.0565.v1

31. Kamvar, Z.N.; Tabima, J.F.; Grünwald, N.J. Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ* **2014**, *2*, e281.
32. Paradis, E.; Claude, J.; Strimmer, K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **2004**, *20*, 289-290.
33. Wickham, H. ggplot2. *WIREs Comp. Stats.* **2011**, *3*, 180–185.
34. Yarberry, W. Dplyr. In *CRAN recipes: DPLYR, stringr, lubridate, and regex in R*; Springer: 2021; pp. 1-58.
35. Kolde, R.; Kolde, M.R. Package 'pheatmap'. *R package* **2015**, *1*, 790.
36. Luo, Z.; Xia, H.; Bao, Z.; Wang, L.; Feng, Y.; Zhang, T.; Xiong, J.; Chen, L.; Luo, L. Integrated phenotypic, phylogenomic, and evolutionary analyses indicate the earlier domestication of Geng upland rice in China. *Mol. Plant* **2022**, *15*, 1506-1509.
37. Cho, Y.; Eun, M.; McCouch, S.; Chae, Y. The semidwarf gene, sd-1, of rice (Oryza sativa L.). II. Molecular mapping and marker-assisted selection. *Theor. Appl. Genet.* **1994**, *89*, 54-59.
38. Niu, Y.; Chen, T.; Zhao, C.; Zhou, M. Improving crop lodging resistance by adjusting plant height and stem strength. *Agronomy* **2021**, *11*, 2421.
39. Zhao, C.; Liu, B.; Piao, S.; Wang, X.; Lobell, D.B.; Huang, Y.; Huang, M.; Yao, Y.; Bassu, S.; Ciais, P. Temperature increase reduces global yields of major crops in four independent estimates. *Proceedings of the National Academy of sciences* **2017**, *114*, 9326-9331.
40. Rang, Z.; Jagadish, S.; Zhou, Q.; Craufurd, P.; Heuer, S. Effect of high temperature and water stress on pollen germination and spikelet fertility in rice. *Environ. Exp. Bot.* **2011**, *70*, 58-65.
41. Cheng, F.; Bin, S.; Iqbal, A.; He, L.; Wei, S.; Zheng, H.; Yuan, P.; Liang, H.; Ali, I.; Xie, D. High sink capacity improves rice grain yield by promoting nitrogen and dry matter accumulation. *Agronomy* **2022**, *12*, 1688.
42. White, A.C.; Rogers, A.; Rees, M.; Osborne, C.P. How can we make plants grow faster? A source–sink perspective on growth rate. *J. Exp. Bot.* **2016**, *67*, 31-45.
43. Ye, C.; Ishimaru, T.; Lambio, L.; Li, L.; Long, Y.; He, Z.; Htun, T.M.; Tang, S.; Su, Z. Marker-assisted pyramiding of QTLs for heat tolerance and escape upgrades heat resilience in rice (Oryza sativa L.). *Theor. Appl. Genet.* **2022**, *135*, 1345-1354.
44. Vanitha, J.; Mahendran, R.; Raveendran, M.; Jegadeeswaran, M. Marker assisted backcross analysis for high temperature tolerance in rice. *Vegetos* **2024**, *37*, 731-737.
45. Ministry of Agriculture of the PRC. *NY/T1433-2014, Protocol for identification of rice varieties-SSR marker method*; China Agriculture Press: Beijing, 2024.
46. Volk, G.M.; Byrne, P.F.; Coyne, C.J.; Flint-Garcia, S.; Reeves, P.A.; Richards, C. Integrating genomic and phenomic approaches to support plant genetic resources conservation and use. *Plants* **2021**, *10*, 2260.
47. Gu, R.; Fan, S.; Wei, S.; Li, J.; Zheng, S.; Liu, G. Developments on Core collections of plant genetic resources: do we know enough? *Forests* **2023**, *14*, 926.
48. Zhang, H.; Zhang, D.; Wang, M.; Sun, J.; Qi, Y.; Li, J.; Wei, X.; Han, L.; Qiu, Z.; Tang, S. A core collection and mini core collection of Oryza sativa L. in China. *Theor. Appl. Genet.* **2011**, *122*, 49-61.