Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Genome-wide DArTseq analysis reveals hidden genetic diversity and distinct trait-based lineages in traditional rice from Assam, Northeast India

Thanga Suja Srinivasan

sujasree07@gmail.com

World Vegetable Center South and Central Asia, ICRISAT campus

Demissew Sertse

University of Manitoba

Mahender Anumalla

International Rice Research Institute (IRRI)-South Asia Hub, ICRISAT

Krishnamanikumar Premachandran

Sathyabama Institute of Science and Technology

Neeraj Kumar Tyagi

International Rice Research Institute-South Asia Regional Centre

Swati Nayak

International Rice Research Institute-South Asia Regional Centre

Article

Keywords: Joha rice, Black rice, DArTseq SNP, Genetic diversity, Indigenous, Genotyping

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9351531/v1

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

Additional Declarations: No competing interests reported.

Genome-wide DArTseq analysis reveals hidden genetic diversity and distinct trait-based lineages in traditional rice from Assam, Northeast India

Thanga Suja Srinivasan* [a,b], Demissew Sertse [c], Mahender Anumalla [d], Krishnamanikumar Premachandran [e], Neeraj Kumar Tyagi [b], Swati Nayak [b]

[a] World Vegetable Center South and Central Asia, ICRISAT campus, Hyderabad - 502324, India

[b] International Rice Research Institute (IRRI)-South Asia Regional Centre, Varanasi - 221006, India

[c] University of Manitoba, Winnipeg, Manitoba, Canada

[d] International Rice Research Institute (IRRI)-South Asia Hub, ICRISAT, Patancheru, Hyderabad 502324, India.

[e] Sathyabama Institute of Science and Technology, Chennai – 600119, India

*Corresponding author: Dr. Thanga Suja Srinivasan

ORCID ID: 0000-0001-6748-0166 (Srinivasan T S),

Email: sujasree07@gmail.com; suja.srinivasan@worldveg.org

Mobile No: +91 82480 70534

Abstract

Northeast India is a hotspot of rice diversity, harbouring traditional rice genotypes with distinct grain-quality traits and adaptation to diverse agroecological conditions. In the present study, we investigated the genome-wide diversity, population structure, and trait-associated divergence among indigenous rice genotypes from Assam, Northeast India using 9,714 high-quality DArTseq SNP markers. Genome-wide analyses revealed extensive genetic variation and strong population subdivision, with phylogenetic, Bayesian clustering, and principal component analyses (PCA) consistently identifying three major genetic clusters and six ancestral subpopulations. Genetic structuring was driven predominantly by grain-quality traits, particularly aroma, stickiness (glutinous grain), and pigmentation, rather than geographic origin alone. Trait-specific identity-by-state analyses and chromosome-level investigations further revealed distinct patterns of allele sharing around genomic regions associated with fragrance, grain stickiness, and pigmentation, reflecting the influence of long-term farmer-mediated selection and local adaptation. Analysis of molecular variance showed that 60.49% of total genetic variation was distributed among subpopulations, accompanied by high levels of private allelic diversity, indicating restricted gene flow and independent evolutionary trajectories among rice groups. Aromatic *Joha*, glutinous *Bora*, and pigmented rice lineages emerged as genetically distinct reservoirs of unique variation. These findings provide new insights into the evolutionary diversification of Northeast Indian rice and establish a

comprehensive genomic framework for germplasm conservation, genetic resource management, and the deployment of valuable alleles for grain quality, speciality rice markets, and climate-resilient rice breeding.

Keywords: Joha rice, Black rice, DArTseq SNP, Genetic diversity, Indigenous, Genotyping

1. Introduction

Asian rice (*Oryza sativa*) is one of the world's oldest domesticated crops, with its primary origin traced to the Yangtze River basin in China. Over the years, as rice expanded beyond its centre of origin, the crop encountered diverse ecological niches and farming systems across Asia, leading to extensive genetic diversification. The primary domestication event, followed by subsequent adaptation to heterogeneous environments, led to the formation of distinct subpopulations such as *indica*, *japonica*, and *aus* type, each shaped by region-specific selection pressures and cultural preferences (Molina et al. 2011: Huang et al. 2012). Recent genomic investigations have revealed that rice domestication was considerably more complex than a single-origin model, involving multiple domestication events, contributions from diverse wild progenitors, and extensive introgression among cultivated and wild rice lineages (Civáň and Brown, 2018; He et al. 2021; Zhao et al. 2025). Although genome-wide differentiation among rice groups is well established, several key agronomic and domestication-related genomic regions exhibit signatures of allele sharing and adaptive introgression across subpopulations (Civáň and Brown, 2017, Civáň and Brown. 2018; Zhou et al. 2025). These patterns indicate that both natural and human-mediated selection have facilitated the retention and movement of favourable alleles associated with adaptation, grain quality, and agronomic performance across diverse genetic backgrounds.

Northeast India, part of the Indo-Burma biodiversity hotspot, is widely recognized as the secondary center of rice diversification. Rice occupies nearly 72% of the total cultivated area in the region and it is grown under diverse production systems, including upland, lowland, flood-prone and deep-water ecosystems (Roy et al. 2015). More than 10,000 indigenous rice cultivars have been documented from the region, representing centuries of farmer-led selection, adaptation, and conservation (Hore 2005; Chakraborty and Ray 2019). The remarkable agro-ecological heterogeneity of Northeast India, characterized by diverse topography, climatic variability, and hydrological regimes, has promoted the evolution of genetically distinct rice populations adapted to drought-prone uplands, flood-prone lowlands, deep-water ecosystems, and hill cultivation systems (Choudhury et al. 2014; Durai et al. 2015; Vanalsanga et al. 2019). This ecological distinctness and diversity have made Northeast Indian rice landraces an

invaluable reservoir of genetic variation that can contribute to future breeding efforts targeting climate resilience, nutritional enhancement, grain quality, and adaptation to marginal production environments (Roy et al. 2015; Umakanth et al. 2017; Vanlalsanga et al. 2019; Lazaridi et al. 2024; Chettri et al. 2026).

The uniqueness of Northeast Indian rice diversity is reflected not only in its ecological adaptation but also in the remarkable diversity of grain-quality traits preserved by local farming communities (Roy et al. 2015; Umakanth et al. 2017; Vanlalsanga et al. 2019; Chettri et al. 2026). Traditional landraces of Northeast India encompassing distinct types such as *Joha* (aromatic), *Bora Saul* (glutinous), *Chak-Hao* (black), *Bao-dhan* (red), and *Chokuwa* (semi-glutinous/instant) reflect a wide spectrum of speciality rice types strongly associated with local cultural practices and consumer preferences (Vanlalsanga et al. 2019; Roy et al. 2024). Previous molecular studies have consistently demonstrated high levels of genetic diversity and significant population structuring within these landraces, underscoring the region as a unique reservoir of rice genetic resources (Olsen and Purugganan, 2007; Durai et al. 2015; Roy et al. 2015). Many of the distinctive grain quality attributes found in these landraces are controlled by key functional genes, including *Waxy (Wx)* controlling amylose content and grain stickiness, *BADH2* governing grain fragrance, *Rc* responsible for red pericarp pigmentation, and *Kala4* associated with black grain pigmentation (Brooks et al. 2008; Oikawa et al. 2015; Singh et al. 2022; Lv et al. 2025). The continued cultivation and preference for these nutritionally, culturally important and climate-adaptive traits have likely contributed to the retention of distinct allelic combinations and genomic backgrounds within the specific rice groups of the region (Roy et al. 2024).

Beyond local adaptation and trait-based selection, recent population genomic studies have highlighted the complex evolutionary history of Northeast Indian rice (Choudhury et al. 2014; Roy et al. 2016). Evidence of adaptive introgression and allele exchange among *indica, japonica, and aus* gene pools has suggested that both historical gene flow and selection have played an important role in the emergence of unique regional landrace groups (Chakraborty and Ray, 2019; Zhou et al., 2025). Despite the recognized significance of Northeast Indian rice diversity, the relative contributions of ecological adaptation, farmer-mediated selection, and grain-quality traits to genome-wide population structure remain insufficiently understood. Furthermore, most previous studies have relied on limited marker sets or focused primarily on specific trait groups, leaving a significant knowledge gap regarding the genomic processes

underlying diversification across the broader spectrum of indigenous rice germplasm of the region.

Against this broader background of rice diversity in Northeast India, the present study investigated traditional rice genotypes collected primarily from different agro-climatic regions of Assam, Northeast India encompassing diverse grain-quality characteristics, collection regions and ecosystems (Fig. 1 and Fig. S1). Three additional cultivars- Kalanamak from Uttar Pradesh, and Protezin and CG Dev Bhog from Chhattisgarh were also included to broaden the geographical and genetic representation of the diversity panel. We hypothesized that ecological adaptation, farmer-mediated selection, and historical germplasm exchange have shaped genomic diversity among the sampled genotypes, with grain-quality traits such as aroma, stickiness, and pigmentation potentially contributing to genetic differentiation alongside geographical origin. To test this hypothesis, we used genome-wide DArTseq SNP markers to (i) assess genome-wide genetic diversity, (ii) characterize population structure and genetic differentiation, and (iii) determine the extent to which grain-quality traits and agroecological adaptation influence genomic variation. These findings provide insights into the genomic diversity and grain-quality-associated genetic relationships of the sampled rice genotypes, with implications for the conservation and utilization of traditional rice genetic resources and the improvement of specialty rice varieties from the region.

2. Results

2.1. SNP distribution and genome-wide genetic diversity

After quality filtering, 9,714 high-quality SNPs were retained from the initial 61,957 DArTseq markers across 88 samples, representing 15.68% of the original marker set. The filtered SNP markers were distributed across all 12 chromosomes with an average density of 25.99 SNPs/Mb. Marker density varied moderately among the chromosomes, ranging from 21.9 SNPs/Mb on chromosome 11 to 31.8 SNPs/Mb on chromosome 2 (Fig. 2A). Marker coverage was generally distributed across the genome, although several genomic intervals showed relatively sparse marker representation (Fig. 2C).

Genome-wide genetic diversity analysis revealed moderate genetic variation within the panel. Per marker expected heterozygosity (He) ranged from 0.076 to 0.500, with a mean of 0.341, while polymorphic information content (PIC) values ranged from 0.073 to 0.375, with a mean of 0.275. Observed heterozygosity (Ho) was low (mean 0.039), consistent with the

predominantly self-pollinating nature of rice. Diversity estimates revealed only minor variation among chromosomes, however, chromosome 7 exhibited the highest allelic diversity (He = 0.355) and marker informativeness (PIC = 0.285), whereas chromosome 11 showed the lowest diversity values (Fig. 2B). Collectively, the marker density, genome-wide coverage, and polymorphism levels support the suitability of the SNP dataset for downstream analyses, including population structure and genetic diversity assessments.

2.2. Genetic relationship among rice genotypes

Pairwise genetic distance analysis revealed a substantial genome-wide variation among the 88 samples, with a mean bitwise distance of 0.335. The highest genetic divergence was observed between Sital Seni Joha and Kalavati (0.575), indicating highly distinct genetic backgrounds. To avoid bias arising from comparison among the biological samples of the same genotype, within-genotype comparisons were excluded when identifying the most similar genotype pairs. Among distinct genotypes, the lowest genetic distance was observed between Beroi and Padumani (0.025), followed by Chokuwa and Bora Chokuwa (0.035), indicating a high degree of genetic similarity between these genotypes (Table S2).

2.3. Phylogenetic analysis

The phylogenetic distance-based clustering grouped the 88 samples into three major genetic groups (Fig. 2). Group I (green clade) formed a distinct lineage characterized by varying branch lengths, indicating different levels of genetic divergence among the constituent samples. This group encompassed samples from traditional lowland and upland ecosystems and from diverse agro-climatic regions, including the Brahmaputra Valley Zone (BVZ), Upper Brahmaputra Valley Zone (UBVZ), and North Bank Plain Zone (NBPZ) of Assam, Northeast India. Within this group, samples Moynagiri, Defa Bao, and Basudev formed separate sub-branches, indicating relatively greater genetic divergence from the main group I cluster. A second set of upland samples (Derso-Karbianglong, Barza-Karbianglong, and Chingdon-Karbianglon) formed distinct branches with comparatively long terminal lengths, reflecting greater genome-wide divergence from the remaining samples and suggesting the accumulation of unique variants. Towards the distal end of the clade, a dense collection of aromatics, non-sticky, core Joha samples was observed, with short branch lengths indicating high similarity and a shared genomic background. Following these, Lal jeera, and Chudamani from BVZ and Kola Joha from NBPZ occupied intermediate positions between the divergent set and the core Joha group.

Group II (orange) comprised predominantly lowland and deep-water rice samples from NBPZ and UBVZ, encompassing diverse grain-quality characteristics, including sticky and non-sticky types and red-grained genotypes. This group exhibited substantial intra-group diversity, as reflected by the relatively long internal branches. Despite their contrasting grain characteristics, Beroi (sticky) and Padumani (non-sticky), both from NBPZ, clustered closely, suggesting a shared genetic background and possible adaptation to similar agroecological conditions. This clustering pattern was consistent with the pairwise genetic distance analysis (Table S2), which identified Beroi and Padumani as the most genetically similar pair of distinct genotypes. Similarly, Chokuwa and Bora Chokuwa clustered closely within Group II, consistent with their pairwise distance estimates, despite being maintained as distinct landraces. However, given the collection of Chokuwa and Bora Chokuwa from the same geographical region and with similar morphological similarities (Table S1). Chokuwa and Bora Chokuwa may represent the same or closely related genotype maintained under different local names; however, further characterization would be required to confirm this possibility.

Group III (blue clade) formed a compact lineage characterized by short branch lengths and tight clustering, indicating strong genetic relatedness among the samples. This group comprised primarily the sticky (glutinous) rice from different agro-climatic regions of Assam, Northeast India. Within the blue clade, sticky black rice (Upendra black rice 1, Upendra black rice 2, Manipuri black rice and Kalavati) formed a distinct and tight sub-cluster indicating shared allelic states and high overall genomic similarity. In addition to the sticky-black rice group, the blue clade also included a few smaller, distinct sub-lineages of non-glutinous rice (like Protezin and CG Dev Bhog, which are non-sticky rice cultivars from Raichur, Chhattisgarh state, India and also samples from other agro-climatic regions of Assam), suggesting possible partial shared ancestry. Furthermore, the clustering within group III highlights a genetically similar assemblage dominated by glutinous rice types, with a distinct sub-group of sticky/glutinous black rice, within the broader diversity panel. (Fig. 2).

2.4 Population structure analysis

Population structure of the 88 retained samples was inferred using a Bayesian clustering approach, with ancestry coefficients across K values ranging from 1 to 10. Cross-entropy analysis revealed a sharp decline between K=1 and K=2, followed by a gradual decline up to K=7, after which the values plateaued (Fig. 3). Although K = 7 exhibited a marginally lower cross-entropy value, K = 6 provided the most biologically meaningful and interpretable clustering pattern. Therefore, six subpopulations were defined, samples with ancestry

coefficients ≥80% were assigned to a specific subpopulation, whereas those with lower membership proportions were classified as admixed (Table S3, Fig. 3).

At K = 6, the 88 samples (49 genotypes) can be separated into six distinct subpopulations, with varying levels of admixtures (Fig. 3). Subpopulation 1 consisted solely of Moynagiri, indicating its unique genetic background. Subpopulation 2 represented most lowland non-aromatic types from NBPZ and UBVZ, suggesting shared ancestry among ecologically adapted types. Subpopulation 3 grouped mostly the sticky glutinous rice, including the black-rice such as Upendra black rice 1, Upendra black rice 2, Kalavati, and Manipuri black rice, suggesting that grain type and pigmentation contributed strongly to genetic differentiation. Subpopulation 4 comprised the upland-adapted types *(*Derso*,* Barza*,* Chingdon from Karbi Anglong*),* forming a distinct upland genetic cluster consistent with their separation in the phylogenetic analysis (Fig. 2). Subpopulation 5 represented the major aromatic rice group, including Kunkuni Joha*,* Bokul Joha*,* Kon Joha*,* Black Joha*,* Sital Seni Joha*,* and Kalanamak. Despite their diverse geographic origins, the aromatic types clustered together, and the results indicate that aroma- and grain quality-related selection has played a greater role in shaping genetic structure than geographic distribution. Subpopulation 6, comprising Kola Joha, Lal jeera, and Chudamani, formed a separate aromatic lineage, suggesting the existence of a genetically differentiated aromatic rice group within the broader diversity panel.

2.5 Principal component analysis (PCA)

Principal component analysis (PCA) further validated the population structure inferred from phylogenetic and Bayesian clustering analyses. The first two principal components explained 46.65% of total genetic variation, with PC1 and PC2 accounting for 42.63% and 4.02%, respectively (Fig. 4). The substantial contribution of PC1 indicates that a large proportion of genomic variation was captured along a single axis of differentiation. The PC1 component distinctly separated glutinous (sticky) rice samples from diverse agro-climatic regions as a separate group (Fig. 4c), indicating that genomic variation associated with grain quality traits contributed substantially to the observed genetic structure. PC2 further resolved genetic variation associated with ecological adaptation (upland vs. lowland), pigmentation, aroma and geographic distribution (Fig. 4 and Fig. S3). Notably, sticky black rice samples formed a tight cluster distinct from the aromatic Joha group (Fig. 4d), while the upland samples from HZ were clearly differentiated along PC2, reflecting their unique genetic background. Overall, the PCA revealed a strong genetic differentiation based on grain quality and aroma**,** together with ecological adaptation (Fig. 4 and Fig. S3), corroborating the patterns observed in genetic

distance, phylogenetic clustering, and population structure analyses supporting the presence of substantial genetic diversity within the sampled rice from Assam, Northeast India.

2.6 Whole-genome and chromosome-specific IBS analysis

Whole-genome IBS analysis using trait-specific reference sample revealed clear differences in genetic similarity patterns. Using Manipuri black rice as the reference for sticky/glutinous and pigmented type, mean IBS values ranged from 0.44 to 1.00, with black and glutinous rice samples generally exhibiting high allele sharing (IBS ≥0.78), (Table S4, Fig. 5a), including Upendra black rice 1, Upendra black rice 2, Kalavati, Naina, Khamtilahi, and Jaymati showed high levels of similarity to Manipuri black rice. In contrast, aromatic Joha and several non-pigmented rice samples exhibited substantially lower IBS values (<0.55), reflecting marked genetic divergence. These patterns were consistent with the phylogenetic, STRUCTURE, and PCA analyses, which collectively separated black and glutinous (sticky) rice samples from the aromatic Joha group (Fig. 2-5). Using Ketekhi Joha as the aromatic reference, mean values ranged from 0.48 to 1.00 for the Joha group and exhibited high genome-wide allele sharing, whereas most non-aromatic and pigmented rice samples showed substantially lower IBS values (Table S5, Fig. 5b). The concordance between IBS, population structure, phylogenetic, and genetic distance analyses indicates that genetic relationships within the panel are shaped predominantly by grain-quality attributes, including aroma, stickiness, and pigmentation, rather than geographic origin alone (Table S2-S3; Fig. 2 and 5).

Chromosome-specific IBS analysis further reinforced trait-based differentiation. On chromosomes 4 and 8, which harbor the aroma-associated genes *BADH1* and *BADH2*, most Joha samples exhibited high similarity to Ketekhi Joha, whereas non-aromatic and pigmented samples showed substantially lower IBS values (Table S6). Likewise, chromosome 6 analysis using Manipuri black rice as the reference revealed high IBS values among glutinous rice samples, particularly Upendra black rice 1, Aghuni Bora, Kalavati, Gomiri, and Naina, reflecting extensive allele sharing around the *Wx* locus associated with grain stickiness (Table S7). In contrast, aromatic Joha genotypes consistently displayed low IBS values for chromosome 6 (*Wx* locus) (Table S7). Compared with chromosome 6, chromosome 4 for *Kala4* locus exhibited a broader distribution of IBS values, with elevated similarity observed in both black/glutinous and several non-pigmented samples (Table S7). Nevertheless, aromatic Joha samples remained the most divergent group in both the chromosome 6 (*Wx*) and chromosome 4 (*Kala4*) analyses for the sticky/glutinous and pigmented trait. Together, these results indicate stronger genomic differentiation of aromatic and glutinous rice groups than among

pigmentation classes and highlight the contribution of trait-associated genomic regions to the genetic differentiation of rice genotypes from Assam, Northeast India.

2.7 Population divergence and genetic diversity

Analysis of molecular variance (AMOVA) revealed strong hierarchical genetic structure among the rice samples. A total of 60.49% of the genetic variation was attributable to differences among subpopulations, resulting in an AMOVA-derived FST of 0.605 (Table 1). Variation among individuals within subpopulations accounted for 29.86% of total variation, while only 9.65% originated from within-individual variation, consistent with the predominantly self-pollinating nature of rice. The overall SNP-based FST of 0.48 further confirmed strong genomic differentiation among genetic groups. Pairwise FST estimates ranged from 0.08 to 0.81 (Table 2), indicating varying degrees of genetic divergence among subpopulations. The lowest differentiation was observed between subpopulations 2 and 3, suggesting substantial genetic affinity and shared ancestry. In contrast, the aromatic Joha of subpopulation 5 exhibited the highest differentiation from all other groups, highlighting its distinct genetic identity (Table S8). Mean within-subpopulation genetic distances ranged from 0.009 (subpopulation 1) to 0.214 (subpopulation 6) (Table 3), whereas between-subpopulation distances varied from 0.220 to 0.540 (Table S8). Consistently, Nei's genetic distance ranged from 0.071 to 0.643, with value observed between subpopulations 2 and 3 and highest involving subpopulation 5 (Table 2). Subpopulation 6 showed moderate divergence from all other and the highest within-group genetic distance.

Population diversity analysis revealed substantial variation among subpopulations (Table 3). Subpopulations 2, 3, and 5 exhibited the highest genotypic diversity, as reflected by their high Shannon-Wiener diversity indices and Simpson’s diversity indices and expected heterozygosity. In contrast, diversity estimates were lower for subpopulations 1, 4, and 6, largely due to their smaller sample size. Analysis of private alleles further highlighted genetic differentiation, with subpopulation 3 possessing the highest number of unique alleles, followed by subpopulation 5, 2, and 4 and this corroborates with the strong population structure identified by STRUCTURE, AMOVA, and genetic distance analyses, and demonstrating the presence of distinct reservoirs of unique genetic variation within specific rice groups from Assam, Northeast India.

3 Discussion

The present study provides a comprehensive assessment of genetic diversity and population structure within a representative panel of rice genotypes from Assam, Northeast India using

genome-wide SNP markers. The high-density marker dataset provided broad genome coverage and revealed substantial genetic variation, enabling robust characterization of relationships among traditional rice landraces differing in grain quality, cultivation ecology, and geographic origin. The relatively uniform distribution of SNPs across all 12 chromosomes, coupled with moderate levels of heterozygosity and marker informativeness, confirms the suitability of the marker set for investigating population structure and trait-associated genetic differentiation.

3.1 Genome-wide diversity and strong population subdivision

The genome-wide SNP dataset revealed substantial genetic variation together with pronounced population structure among traditional rice samples from Assam, Northeast India (Fig. 2). Despite broad genomic coverage, observed heterozygosity was consistently lower than expected heterozygosity across chromosomes, reflecting rice's predominantly self-pollinating nature and the consequent reduction in heterozygous loci. Similar patterns have been reported in rice landraces and global rice collections, where prolonged selfing and local selection reduce within-population heterozygosity while maintaining high divergence among populations (Huang et al. 2012).

The strong genetic subdivision detected in the present study (Fig. 2-4; Table 1-3 and Table S3) was higher compared with large Asian rice panels, where most variation lies between the major varietal groups (*indica, japonica, aus, aromatic*) (Garris et al. 2005 and Wang et al. 2018). It is interesting to note that the panel included in the study falls within the regional gene pool of Northeast India yet shows stronger differentiation. This reflects the unique evolutionary history of Northeast Indian rice, where cultivation across contrasting ecosystems, including upland, lowland, flood-prone, and deep-water environments, has facilitated the maintenance of distinct allele pools. Similar patterns of localized genetic differentiation have been documented in traditional rice landraces maintained through farmer-led seed systems and region-specific selection practices (Pusadee et al. 2019; Cui et al. 2019). Together, these findings indicate that Northeast Indian rice genotypes represents a highly structured gene pool shaped by both ecological adaptation and long-term on-farm conservation. However, these estimates should nevertheless be interpreted with caution. As subpopulations defined by STRUCTURE were used for AMOVA and FST, similarly sample size within certain groups, marker ascertainment, and the analyses statistical power may overstate true differentiation. For example, subpopulation 1 contains a single genotype and cannot support population-level inference. We therefore treat our estimates as relative contrasts among groups rather than absolute measures of gene-flow.

3.2 Grain-quality traits as major drivers of population structure

Phylogenetic, STRUCTURE, and PCA analyses consistently revealed genetic differentiation among rice samples associated with grain-quality characteristics, particularly aroma, stickiness, and pigmentation. Although Assam, encompasses diverse agro-climatic regions, genotypes sharing similar grain-quality attributes clustered together irrespective of their geographical regions (Fig. 2 and 3), suggesting that farmer-mediated selection and the maintenance of preferred grain characteristics may have contributed to the observed population structure.

Aromatic Joha types formed a highly cohesive genetic group characterized by limited admixture and short phylogenetic branch lengths, indicating a shared genetic background maintained through continued selection for aroma and cooking quality. Notably, Kalanamak clustered closely with the Joha group despite its distinct geographical origin, indicating a close genetic affinity among aromatic rice groups from geographically separated regions. This pattern is consistent with previous reports of genetic affinity between Joha rice and aromatic landraces from the Himalayan foothill region (Gayacharan et al. 2018). The phylogenetic and STRUCTURE based clustering of glutinous and pigmented rice genotypes (Fig. 2 and Table S3) further suggests an association between grain-quality characteristics and genetic differentiation. Sticky/glutinous black rice genotypes clustered with other glutinous genotypes despite differences in pigmentation (Fig. 2-3 and Table S3), highlighting genetic affinity among rice types sharing glutinous grain texture. This finding complements previous reports of genetic differentiation among Northeast Indian rice landraces in relation to grain and cooking-quality characteristics (Roy et al. 2015). The presence of non-glutinous cultivars Protezin and CG Dev Bhog, originating from geographically distant locations, within the same major clade (Fig. 2) indicates that this genetic group encompasses variation beyond grain texture and geographical origin. Nevertheless, the clustering of glutinous and pigmented genotypes underscores the relevance of grain-quality characteristics in interpreting the genetic structure of Northeast Indian rice.

3.3 Chromosome level allele sharing among aromatic and glutinous rice

Trait-based IBS analysis provided additional insights into genomic similarities among rice samples grouped by grain quality characteristics. Glutinous rice samples showed extensive allele sharing across chromosome 6 (Table S7), which harbors the *Waxy* (*Wx*) gene, a major determinant of amylose content and grain stickiness (Shao et al. 2013; Roy et al. 2015; Zhang et al. 2019). The aromatic samples also showed high chromosome-level IBS across chromosomes 4 and 8; and aromatic rice genotypes are strongly associated with loss-of-function variants of *BADH1* and *BADH2* genes on chromosome 4 and 8 respectively (Kovach

et al. 2009). Similar genetic differentiation among rice groups has been reported in Asian rice, where strong selection around loci controlling starch composition and grain quality promotes distinct genetic groups (Glaszmann et al. 2010).

In contrast, allele sharing among black rice samples was less pronounced than that observed for the glutinous and aromatic groups. Black pericarp pigmentation has been linked to regulatory changes involving *Kala4* (Oikawa et al. 2015), while rice pigmentation is also reported to be influenced by other genetic factors. Thus, genotypes sharing pigmentation characteristics may not exhibit the same degree of broader genomic similarity as those grouped by aroma or grain texture. However, the present analysis assesses allele sharing rather than directly resolving haplotypes or causal variants at *Wx*, *BADH2*, or *Kala4*. These chromosome-level patterns therefore should be interpreted as evidence of genomic similarity rather than direct confirmation of trait-associated genetic variation.

3.4 Ecological contribution to diversification within groups:

Upland rice samples from the HV region formed a relatively distinct subpopulation, whereas lowland non-aromatic samples exhibited greater internal genetic divergence and formed a separate subpopulation (Fig. 3; Table S3). This pattern suggests that ecological differentiation may contribute to genetic structuring within broader quality- or trait-defined groups. Similar patterns have been reported in rice, where dispersal into diverse environments further differentiated local populations after the establishment of major varietal groups (Gutaker et al. 2020). Thus, the distinct clustering of the upland and lowland genotypes may reflect genetic differentiation associated with their geographic and ecological background.

Similarly, lowland non-aromatic genotypes exhibited greater internal divergence (orange clade Fig. 2), forming multiple genetic lineages likely shaped by adaptation to distinct production environments and farmer selection for agronomic traits. The coexistence of upland and deep-water ecotypes (subpopulation 3) within related genetic backgrounds (Fig. 2 and 3) further suggests that adaptive divergence may have occurred through selection at a relatively limited number of genomic regions associated with drought tolerance, water-use efficiency, and flooding response (Xia et al. 2019). Thus, ecological adaptation has contributed to diversification within broader genetic groups, whereas grain-quality traits have largely determined the major population subdivisions.

3.5 Conservation and breeding implications

The strong agreement among various analytical approaches indicates that the observed population structure reflects a combination of historical divergence, ecological adaptation, and persistent farmer-mediated selection. The aromatic and glutinous groups carried the highest

numbers of private alleles and diversity indices (Table 3), which makes them priority targets for conservation and potential donors of grain-quality alleles. Such enrichment of unique allelic variants suggests long-term independent evolutionary trajectories and limited genetic exchange with other rice groups.

The distinct genetic differentiation observed among aromatic, sticky/glutinous, pigmented, and ecologically adapted rice groups has implications for both germplasm conservation and rice improvement. The enrichment of private alleles in the aromatic Joha and sticky-pigmented groups identifies these landraces as potentially valuable reservoirs of genetic variation for grain quality and adaptation-related traits. Their distinct genomic backgrounds may also provide useful diversity for parental selection and targeted introgression, particularly where specific grain-quality or adaptive traits are desired. However, functional validation of candidate genomic regions and associated alleles will be required to establish their direct contributions to these phenotypes.

Overall, the present study demonstrates that rice genotypes from Assam, Northeast India harbors highly structured and evolutionarily distinct genetic groups shaped by both quality traits and ecological adaptation. The differentiation of aromatic Joha and glutinous-pigmented rice lineages, together with their rich reservoir of private alleles, underscores their significance for germplasm conservation and future rice improvement. These findings provide important insights into the evolutionary dynamics of Northeast Indian rice and establish a valuable genomic framework for the discovery and utilization of adaptive and grain-quality-associated alleles in breeding programs.

4. Methods

4.1 Plant materials

The study comprised a panel of 49 rice genotypes, including 46 traditional landraces collected from diverse agroclimatic regions of Assam, Northeast India (Fig. 1) and three reference genotypes from other rice-growing regions of India. The reference set included Protezin, a released-high protein rice variety, and CG Dev Bhog, an aromatic cultivar from Raichur, Chhattisgarh, India (Krishna et al. 2025), together with Kalanamak, a renowned aromatic landrace originating from the Himalayan foothills of Northern India (Singh et al. 2021) (Fig. 1). Inclusion of these reference genotypes provided a broader comparative framework for evaluating the genetic distinctiveness and population structure of rice from Assam, Northeast India. Details of genotype origin, agroecological adaptation, and key phenotypic characteristics are presented in Figure 1, Table S1 and Supplementary Figure 1.

To minimize genetic heterogeneity within the sampled materials and ensure representative sampling for genome-wide genotyping, all genotypes were established under uniform field conditions at the KVK research farm, Nagaon, Assam. Each genotype was grown in five rows of 10 plants at a spacing of 20 × 20 cm. At the active tillering stage, ten phenotypically uniform plants were selected from each genotype and subjected to preliminary quality-control (QC) genotyping using the IRRI Indica QC 10-SNP panel (https://isl.irri.org/services/genotyping/qc-snp-panels#h.69rgvnkyf10f; Ndjiondjop et al. 2018). This screening enabled the identification of genetically uniform plants and the detection of residual heterogeneity within farmer-maintained landraces. We then selected one or two representative plants with identical QC-SNP profiles from each genotype for genotyping. This strategy ensured robust representation of genotype-specific genetic diversity while reducing artefacts associated with intra-accession variation. A total of 94 leaf samples each from an individual QC controlled plant were collected, dried in silica gel, and stored until DNA extraction.

4.2 DNA extraction and DArTseq genotyping

Genomic DNA was extracted from dried leaf tissue using the CTAB protocol of Doyle and Doyle (1987). DNA quantity and quality were assessed using a NanoDrop spectrophotometer (Thermo Fisher Scientific, USA), and only high-quality DNA samples were advanced for genotyping. Genome-wide SNP genotyping was performed by Diversity Arrays Technology Pty. Ltd. (DArT, Bruce, Australia) using the medium density Rice DArTseq™ 1.0 platform (https://www.diversityarrays.com). This genotyping-by-sequencing approach enables simultaneous SNP discovery and genotyping across the rice genome, providing dense marker coverage suitable for population genetic and diversity analyses.

4.3 SNP filtering and quality control

Raw DArTseq data obtained were processed using the proprietary DArT analytical pipeline (https://www.diversityarrays.com/services/dartseq/). SNP-containing sequences were aligned against the *Oryza sativa* reference genome (Rice RGAP v7) using BLAST with stringent filtering criteria, including an E-value threshold of 5×10^{-7} and a minimum sequence identity of 70%, to ensure high-confidence genomic assignment of marker loci. Initial genotyping generated 61,957 SNP markers across the 94 samples. Subsequent quality filtering was performed in R version 4.4.2 (R Core Team, 2024) to retain high-confidence markers for downstream analyses. SNPs with more than 10% missing data, minor allele frequency (MAF) below 0.05, reproducibility less than 0.95, or monomorphic loci were also removed. Following quality filtering, six samples representing biological replicates of six genotypes were excluded. The remaining 88 samples represented all 49 genotypes, as the exclusion of individual replicate

samples did not result in the loss of any genotype. The final filtered dataset comprised 9,714 high-quality SNP markers, representing 15.68% of the 61,957 SNP markers. All retained markers exhibited a minimum call rate of 90% and were subsequently used for analyses of genome-wide diversity, population structure, genetic differentiation, and trait-associated genomic relationships.

4.4 Data analysis

4.4.1 Genetic diversity and population structure

Genetic diversity parameters, including observed heterozygosity (Ho), expected heterozygosity (He), minor allele frequency (MAF), and polymorphic information content (PIC), were computed using the filtered SNP dataset with *adegenet package* in R (Jombart 2008). Pairwise genetic distances were estimated using the bitwise distance implemented in poppr (Kamvar et al. 2014) where biological samples of the same genotype were excluded in identifying the most genetically similar genotype pairs. These analyses provided an overview of genetic variation and relatedness among the rice genotypes. Population structure was inferred using the *LEA* package in R, evaluating K values from 1 to 10 based on the cross-entropy criterion (Frichot et al. 2014). Samples with ancestry coefficients ≥80% were assigned to a specific subpopulation, whereas others were classified as admixed. To further quantify genetic differentiation at the subpopulation level, Nei's genetic distance was used to assess genetic divergence within and among inferred subpopulations.

To complement model-based clustering, phylogenetic relationships among samples were reconstructed using the Neighbor-Joining method based on the bitwise genetic distance matrix. Tree construction and visualization were performed using the *ape* package in R. Genetic relationships were further explored through Principal Component Analysis (PCA) using TASSEL 5. PCA was conducted on the filtered SNP dataset (MAF ≥ 0.05; missing data ≤10%), and eigenvectors were exported to R for visualization. Clustering patterns were examined in relation to inferred genetic groups, grain-quality traits, grain pigmentation, agroecological adaptation, and geographical origin.

4.4.2 Identity-by-State (IBS) analysis of trait-associated genomic relationships

To investigate genome-wide and trait-specific allele-sharing patterns, Identity-by-State (IBS) analysis was performed using the *SNPRelate package* in R (Zheng et al. 2012). Two representative landraces from the study were selected as reference based on their distinctive grain-quality characteristics (i.e. Ketekhi Joha, for the aromatic rice group, and Manipuri black rice for the pigmented and glutinous rice group). Genome-wide IBS values were calculated using non-overlapping sliding windows of 50 SNPs, generating approximately 200 windows

across the genome. Window-wise IBS values for each sample were calculated relative to the reference selected and summarized to obtain overall genome-wide similarity profiles. Integration of IBS estimates with physical marker positions enabled visualization of genomic regions exhibiting elevated allele sharing or divergence.

To determine whether population differentiation was associated with major grain-quality loci, chromosome-specific IBS analyses were performed for chromosomes harbouring key candidate genes. For the aromatic trait, analyses focused on chromosome 8, which contains the fragrance-associated *BADH2* gene, and chromosome 4, which harbours the putative aroma-associated *BADH1* region (Shao et al. 2013; Roy et al. 2015). For grain stickiness and pigmentation, analyses focused on chromosome 6, containing the *Waxy* (*Wx*) locus involved in amylose synthesis and grain texture, and chromosome 4, harbouring the *Kala4* locus associated with black pericarp pigmentation (Oikawa et al. 2015; Zhang et al. 2019; Lv et al. 2025). IBS estimates were calculated across the entire length of each target chromosome rather than being restricted to the specific candidate-gene regions because the number of SNP markers available within some of these regions was insufficient for reliable locus-specific comparisons. The chromosome-wide analysis therefore provided a broader assessment of genetic similarity and allele sharing among rice samples on chromosomes containing these candidate genes, and the observed patterns were interpreted as chromosome-level evidence

4.4.3 Analysis of molecular variance and population differentiation

The distribution of genetic variation within and among inferred subpopulations was quantified using Analysis of Molecular Variance (AMOVA) implemented in the *pegas* package in R (Paradis 2010). Pairwise genetic differentiation among subpopulations was estimated using fixation index (FST) statistics. To characterize genetic diversity within each inferred subpopulation, several population genetic parameters were calculated using the *poppr* package in R, including the number of multilocus genotypes (MLG), expected multilocus genotypes (eMLG), Shannon-Wiener diversity index (H), Simpson's genotypic diversity index (λ), expected heterozygosity (Hexp), index of association (Ia), and standardized index of association ($\bar{r}D$). In addition, private alleles were identified for each subpopulation to identify unique genetic variation and assess their potential value for conservation and breeding. This integrated analytical framework enabled comprehensive evaluation of genome-wide diversity, population structure, genetic differentiation, and trait-associated genomic relationships among rice samples from Assam, Northeast India.

Conclusion

This study demonstrates that traditional rice genotypes from Assam, Northeast India harbour substantial genome-wide diversity, pronounced population structure, and clear trait-associated differentiation shaped by long-term farmer selection and ecological adaptation. Genome-wide SNP analysis identified distinct genetic groups associated with aromatic Joha, sticky/glutinous, pigmented, upland, and lowland rice types, highlighting the contribution of both grain-quality traits and agro-ecological specialization to population divergence. The high levels of private allelic variation observed in several groups, particularly aromatic and sticky-pigmented rice, underscore the unique genetic resources preserved within these traditional germplasm pools. These findings provide a valuable genomic framework for the conservation and utilization of Northeast Indian rice diversity and support their potential use in breeding programmes targeting grain quality, specialty traits, and climate resilience. Future studies integrating whole-genome resequencing, haplotype-based analyses, and functional characterization of unique adaptive and trait-associated loci will further enhance the identification and deployment of beneficial alleles for rice improvement while strengthening strategies for the long-term conservation of these important genetic resources.

References

Brooks, S.A., Yan, W., Jackson, A.K., Deren, C.W., 2008. A natural mutation in rc reverts white-rice-pericarp to red and results in a new, dominant, wild-type allele: Rc-g. Theor. Appl. Genet. 117, 575–580. https://doi.org/10.1007/s00122-008-0801-8

Chakraborty, D. and Ray, A. 2019. Population genetics analyses of North-East Indian indigenous rice landraces revealed divergent history and alternate origin of aroma in aus group. Plant Genet. Resour. 17, 437–447. https://doi.org/10.1017/S1479262119000273

Chettri, D., Lepcha, P., Sathyanarayana, N., 2026. Phenotypic and nutritional diversity of indigenous rice (*Oryza sativa* L.) cultivars from Sikkim, India. *Discov Agric* 4, 75. https://doi.org/10.1007/s44279-026-00533-3

Choudhury, B.I., Khan, M.L., Dayanandan, S., 2014. Patterns of nucleotide diversity and phenotypes of two domestication-related genes (OsC1 and Wx) in indigenous rice varieties in Northeast India. BMC Genet. 15, 71.

Civáň, P. and Brown, T. A. 2017. Origin of rice (Oryza sativa L.) domestication genes. Genet. Resour. Crop Evol. 64, 1125–1132. https://doi.org/10.1007/s10722-017-0518-0

Civáň, P. and Brown, T. A. 2018. Role of genetic introgression during the evolution of cultivated rice (Oryza sativa L.). BMC Evol. Biol. 18, 57 https://doi.org/10.1186/s12862-018-1180-7

Cui, D., Lu, H., Tang, C., Li, J., A, X., Yu, T., Ma, X., Zhang, E., Wang, Y., Cao, G., Xu, F., Qiao, Y., Dai, L., Li, R., Tian, S., Koh, H.J., Han, L., 2019. Genomic analyses reveal selection footprints in rice landraces grown under on-farm conservation conditions during a short-term period of domestication. Evol. Appl. 13, 290–302. https://doi.org/10.1111/eva.12866

Doyle, J.J., Doyle, J.L., 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem. Bull. 19, 11–15.

Durai, A., Tomar, J. M. S., Devi, P., Arunachalam, A. & Mehta, H. 2015. Rice diversity – the genetic resource grid of North-East India. Indian J. Plant Genet. Resour. 28, 205. https://doi.org/10.5958/0976-1926.2015.00024.8

Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G., and François, O. 2014. Fast and efficient estimation of individual ancestry coefficients. Genetics *196*, 973–983.

Garris, A. J., Tai, T. H., Coburn, J., Kresovich, S. and McCouch, S. 2005. Genetic structure and diversity in *Oryza sativa* L. *Genetics* **169**, 1631–1638

Gayacharan, Bisht, I.S., Pandey, A., Yadav, M.C., Singh, A.K., Pandravada, S.R., Rana, J.C., 2018. Population structure of some indigenous aromatic rice (Oryza sativa L.) landraces of India. Indian J. Biotechnol. 17, 110–117.

Glaszmann, J.C., Kilian, B., Upadhyaya, H.D., Varshney, R.K., 2010. Accessing genetic diversity for crop improvement. Curr. Opin. Plant Biol. 13, 167–173. https://doi.org/10.1016/j.pbi.2010.01.004

610 Gutaker, R.M., Groen, S.C., Bellis, E.S., Choi, J.Y., Pires, I.S., Bocinsky, R.K., Slayton, E.R.,
611 Wilkins, O., Castillo, C.C., Negrão, S., Oliveira, M.M., Fuller, D.Q., Guedes, J.A. d'Alpoim,
612 Lasky, J.R. and Purugganan, M.D. 2020. Genomic history and ecology of the geographic
613 spread of rice. *Nat. Plants* **6**, 492–502
614
615 He, W., Chen, C., Xiang, K., Wang, J., Zheng, P., Tembrock, L.R., Jim, D., Wu, Z., 2021. The
616 history and diversity of rice domestication as resolved from 1464 complete plastid genomes.
617 Front. Plant Sci. 12. https://doi.org/10.3389/fpls.2021.781793
618
619 Hore, D.K., 2005. Rice diversity collection, conservation and management in North-eastern
620 India. Genet. Resour. Crop Evol. 52, 1129–1140.
621
622 Huang, X., Kurata, N., Wei, X., Wang, Z.X., Wang, A., Zhao, Q., Zhao, Y., Liu, K., Lu, H.,
623 Li, W., Guo, Y., Lu, Y., Zhou, C., Fan, D., Weng, Q., Zhu, C., Huang, T., Zhang, L., Wang,
624 Y., Feng, L., Furuumi, H., Kubo, T., Miyabayashi, T., Yuan, X., Xu, Q., Dong, G., Zhan, Q.,
625 Li, C., Fujiyama, A., Toyoda, A., Lu, T., Feng, Q., Qian, Q., Li, J., Han, B., 2012. A map of
626 rice genome variation reveals the origin of cultivated rice. Nature 490, 497–501.
627 https://doi.org/10.1038/nature11532
628
629 Jombart, T. 2008. adegenet: a R package for the multivariate analysis of genetic markers.
630 Bioinformatics 24, 1403–1405.
631
632 Kamvar, Z. N., Tabima, J. F. & Grünwald, N. J. 2014. Poppr: an R package for genetic analysis
633 of populations with clonal, partially clonal, and/or sexual reproduction. PeerJ 2, e281.
634 https://doi.org/10.7717/peerj.281
635
636 Kovach, M. J., Calingacion, M. N., Fitzgerald, M. A. and McCouch, S. R. 2009. The origin
637 and evolution of fragrance in rice (*Oryza sativa* L.). *Proc. Natl Acad. Sci. USA* **106**, 14444–
638 14449
639
640 Krishna, Bhambri, M.C., Kumar, S., 2025. Studies on growth and yield behavior of short grain
641 aromatic traditional, improved scented and fortified quality rice varieties under organic
642 production system. J. Adv. Biol. Biotechnol. 28, 1–11.
643

Lazaridi, E., Kapazoglou, A., Gerakari, M., Kleftogianni, K., Passa, K., Sarri, E., Papasotiropoulos, V., Tani, E., Bebeli, P.J., 2024. Crop landraces and indigenous varieties: A valuable source of genes for plant breeding. Plants 13, 758. https://doi.org/10.3390/plants13060758

Lv, Y., Deng, W., Zuo, X., Lan, D., Tan, J., Zhang, J., Dong, Y., Xu, Y., Zhang, J., Zhang, X., Tu, J., Kui, L., Gu, A., Shen, X., Li, X., 2025. Allelic variations of the Waxy gene and their associations with indica–japonica differentiation and amylose content in Yunnan local rice germplasm. Genes 16, 1198. https://doi.org/10.3390/genes16101198

Molina, J., Sikora, M., Garud, N., Flowers, J.M., Rubinstein, S., Reynolds, A., Huang, P., Jackson, S., Schaal, B.A., Bustamante, C.D., Boyko, A.R. and Purugganan, M.D. 2011. Molecular evidence for a single evolutionary origin of domesticated rice. Proc. Natl. Acad. Sci. U.S.A. 108, 8351-8356. https://doi.org/10.1073/pnas.1104686108

Ndjiondjop, M.N., Semagn, K., Zhang, J., Gouda, A.C., Kpeki, S.B., Goungoulou, A., Wambugu, P., Dramé, K.N., Bimpong, I.K., Zhao, D., 2018. Development of species diagnostic SNP markers for quality control genotyping in four rice (*Oryza* L.) species. Mol Breed. 38(11): 131. doi: 10.1007/s11032-018-0885-z.

Oikawa, T., Maeda, H., Oguchi, T., Yamaguchi, T., Tanabe, N., Ebana, K., Yano, M., Ebitani, T., Izawa, T., 2015. The birth of a black rice gene and its local spread by introgression. Plant Cell 27(9), 2401-2414. https://doi.org/10.1105/tpc.15.00310.

Olsen, K.M., Purugganan, M.D., 2007. Molecular evidence on the origin and evolution of glutinous rice. Genetics 162, 941–950.

Paradis, E. 2010. pegas: an R package for population genetics with an integrated-modular approach. Bioinformatics **26**, 419–420

Pusadee, T., Wongtamee, A., Rerkasem, B., Olsen, K.M., Jamjod, S., 2019. Farmers drive genetic diversity of Thai purple rice (Oryza sativa L.) landraces. Econ. Bot. 73, 76–85.

R Core Team, 2024. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Roy, S., Banerjee, A., Mawkhlieng, B., Misra, A.K., Pattanayak, A., Harish, G.D., Singh, S.K., Ngachan, S.V., Bansal, K.C., 2015. Genetic diversity and population structure in aromatic and quality rice (Oryza sativa L.) landraces from North-Eastern India. PLOS ONE 10, e0141405. https://doi.org/10.1371/journal.pone.0141405

Roy, S., Marndi, B.C., Mawkhlieng, B., Banerjee, A., Yadav, R.M., Misra, A.K., and Bansal, K.C., 2016. Genetic diversity and structure in hill rice (Oryza sativa L.) landraces from the North- Eastern Himalayas of India. BMC Genetics, 17:107. DOI 10.1186/s12863-016-0414-1

Roy, S., Patra, B.C., Kumar, J., Sar, P., Jogi, U.S., Konyak, Z., Banerjee, A., Basak, N., Mandal, N.P., Bansal, K.C., 2024. Ethnolinguistic associations and genetic diversity of rice landraces in Nagaland, India. Plants, People, Planet. https://doi.org/10.1002/ppp3.10454

Shao, G., Tang, S., Chen, M., Wei, X., He, J., Luo, J., Jiao, G., Hu, Y., Xie, L., Hu, P., 2013. Haplotype variation at Badh2, the gene determining fragrance in rice. Genomics 157–162. https://doi.org/10.1016/j.ygeno.2012.11.010

Singh, D.P., Chandra, V., Tiwari, T., 2021. Evaluation of different Kalanamak rice genotypes for yield and yield-related traits of eastern Uttar Pradesh. Int. J. Chem. Stud. 9, 97–101. https://doi.org/10.22271/chemi.2021.v9.i1b.11449

Singh, G., Gopala Krishnan, S., Kumar, A., Vinod, K.K., Bollinedi, H., Ellur, R.K., Nagarajan, M., Bhowmick, P.K., Madhav, S.M., Singh, K., Singh, A.K., 2022. Molecular profiling of BADH2 locus reveals distinct functional allelic polymorphism associated with fragrance variation in Indian aromatic rice germplasm. Physiol. Mol. Biol. Plants 28, 1013–1027. https://doi.org/10.1007/s12298-022-01181-9

Umakanth, B., Vishalakshi, B., Sathish Kumar, P., Rama Devi, S.J.S., Bhadana, V.P., Senguttuvel, P., Kumar, S., Sharma, S.K., Sharma, P.K., Prasad, M.S., and Madhav, M.S., 2017. Diverse Rice Landraces of North-East India Enables the Identification of Novel Genetic Resources for Magnaporthe Resistance. Front. Plant Sci. 8:1500. doi: 10.3389/fpls.2017.01500

Vanlalsanga, Singh, S.P., Singh, Y.T., 2019. Rice of Northeast India harbor rich genetic diversity as measured by SSR markers and Zn/Fe content. BMC Genet. 20(1): 79. doi: 10.1186/s12863-019-0780-6. PMID: 31646978; PMCID: PMC6806518.

Wang, W., Mauleon, R., Hu, Z., Chebotarov, D., Tai, S., et al. 2018. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature 557, 43–49

Xia, H., Luo, Z., Xiong, J., Ma, X., Lou, Q., Wei, H., Qiu, J., Yang, H., Liu, G., Fan, L., Chen, L., Luo, L., 2019. Bi-directional selection in upland rice leads to its adaptive differentiation from lowland rice in drought resistance and productivity. Plant Cell Rep. 12, 170–184.

Zhang, C., Zhu, J., Chen, S., Fan, X., Li, Q., Lu, Y., Wang, M., Yu, H., Yi, C., Tang, S., Gu, M., Liu, Q., 2019. Wx^{lv}, the ancestral allele of rice Waxy gene. Mol Plant. 12(8), 1157-1166. doi: 10.1016/j.molp.2019.05.011.

Zhao, Y., Li, T., Liu, D., Yin, H., Wang, L., Lu, S., Yu, H., Sun, X., Zhang, T., Zhao, Q., 2025. The origin and evolution of cultivated rice and genomic signatures of heterosis for yield traits in super-hybrid rice. BMC Biol. 23, 153. https://doi.org/10.1186/s12915-025-02255-2

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., and Weir, B. S. 2012. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics 28(24), 3326–3328.

Zhou, J., Li, J., Zhang, Y., Yang, Y., Lv, Y., Pu, Q., Deng, X., Tao, D. 2025. Introgression among subgroups is an important driving force for genetic improvement and evolution of the Asian cultivated rice (Oryza sativa L.). Front. Plant Sci. 16, 1535880. https://doi.org/10.3389/fpls.2025.1535880

Acknowledgments

The authors thankfully acknowledge the support received from APART project, Assam team and IRRI, Varanasi.

Funding Information

The authors declare that this study was conducted without any financial support.

Author Contributions

Conceptualization—T.S.S., N.K.T., D.S., S.N.; Data collection: T.S.S.; Data curation—T.S.S., D.S.; Formal analysis—D.S., K.P., and T.S.S.; Funding acquisition—T.S.S. and N.K.T.; Writing original draft—T.S.S.; Writing, review and editing—T.S.S., N.K.T., M.A., K.P., D.S., S.N. All authors have read and agreed to the manuscript.

Data Availability Statement

The raw sequencing reads and genome-wide DArTseq SNP variant datasets generated during this study have been deposited in the European Nucleotide Archive (ENA) and European Variation Archive (EVA) at EMBL-EBI under project accession number PRJEB123558 (variant analysis accession: ERZ29802000). All raw sequencing data (runs ERR17742746–ERR17742833) and SNP variant calls are accessible at https://www.ebi.ac.uk/ena/browser/view/PRJEB123558.

Conflicts of Interest

There are no conflicts of interest to disclose.

Generative AI statement

The authors declare that no Gen AI was used in the creation of this manuscript.

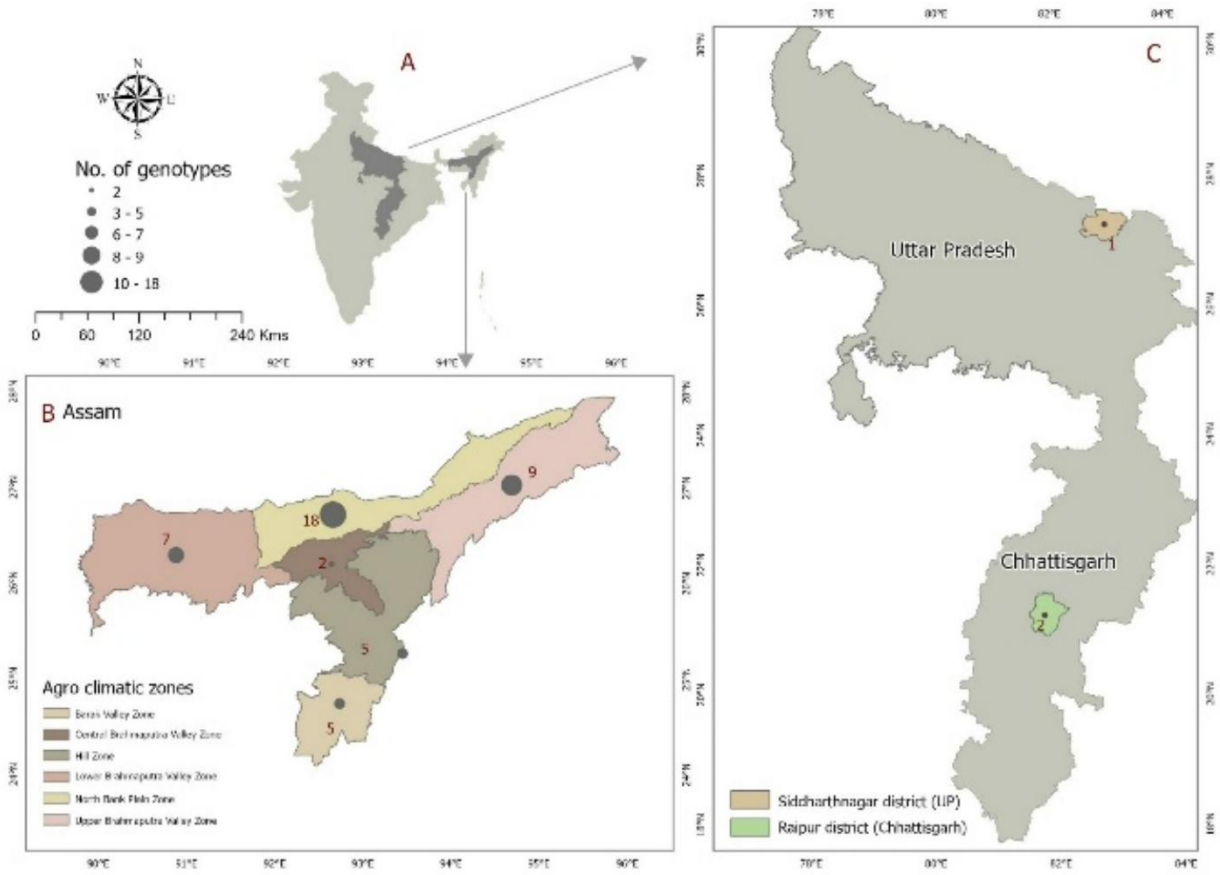


Figure 1. Geographic distribution and collection sites of the rice genotypes included in the study. **(A)** Map of India showing the geographic regions represented in the study, with Assam, Uttar Pradesh, and Chhattisgarh states of India highlighted. **(B)** Enlarged map of Assam showing the agro-climatic regions and the number of unique rice genotypes collected from each agro-climatic region. **(C)** Geographic locations of the reference genotypes from Uttar Pradesh and Chhattisgarh. The 49 unique rice genotypes included in the study were represented as 88 biological samples in the study.

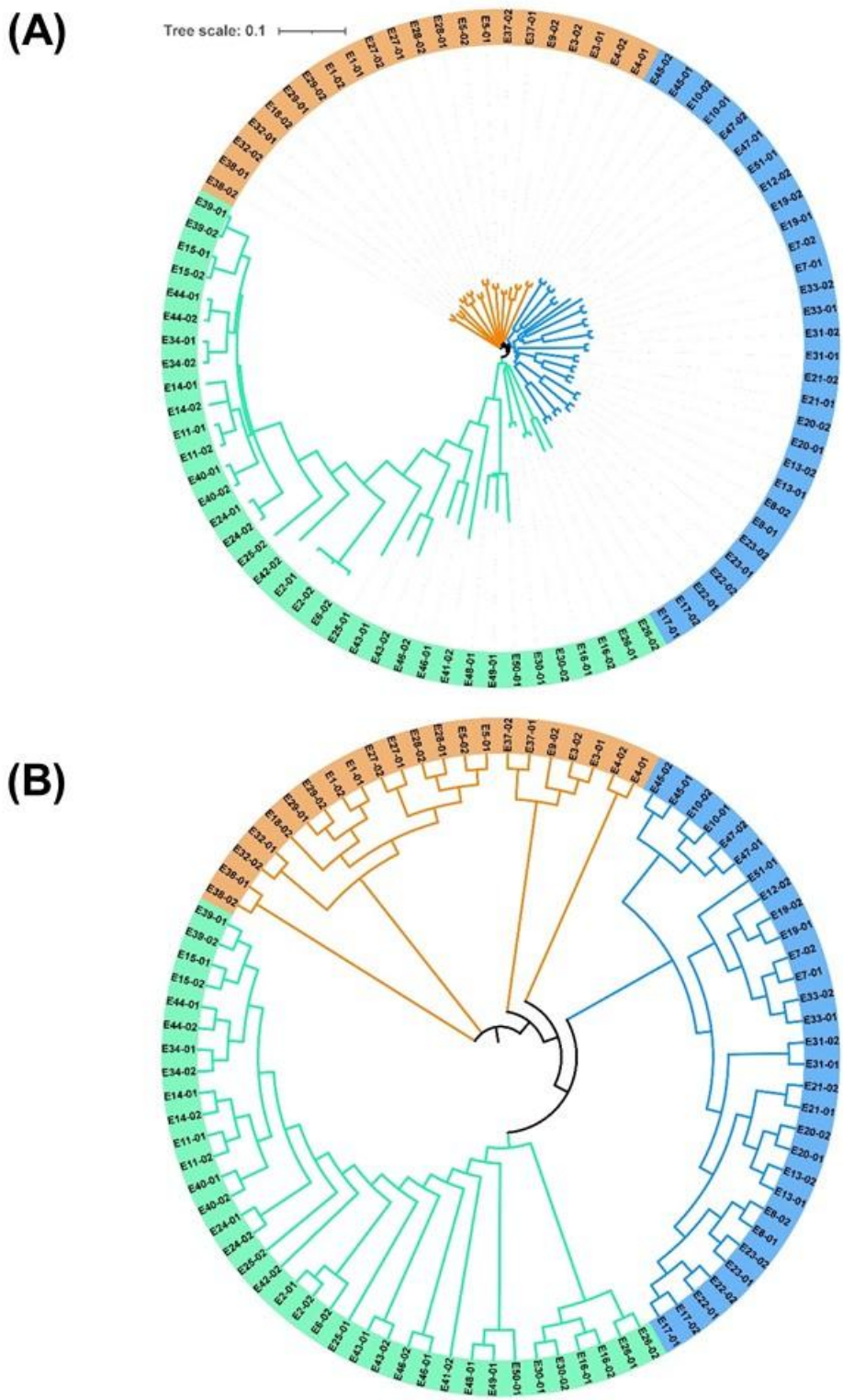


Figure 2. Phylogenetic relationships among rice samples from Assam, Northeast India inferred from 9,714 genome-wide DArTseq SNP markers. (A) Radial phylogenetic tree showing overall clustering patterns and major genetic groups. (B) Expanded circular phylogram of the same dataset highlighting relative branch-length variation within and between clusters.

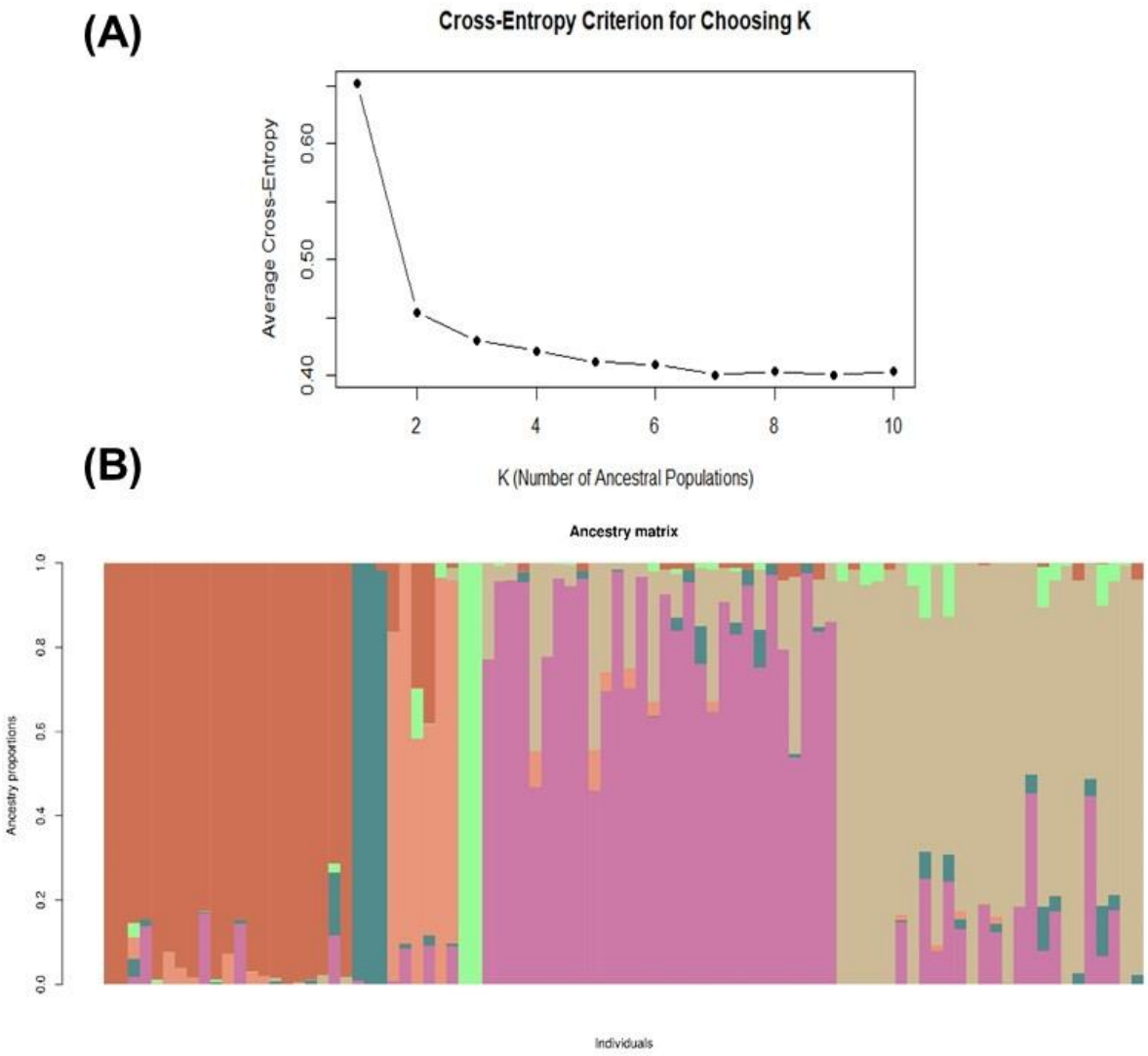


Figure 3. Population structure analysis of rice samples from Assam, Northeast India. (A) Cross-entropy-based estimation of the optimal number of ancestral populations (K). (B) Population structure of the rice samples is based on inferred ancestry at the optimal K. Samples with a membership probability ≥80% were assigned to the corresponding inferred subgroup, while those with membership probabilities below 80% were categorized as admixed.

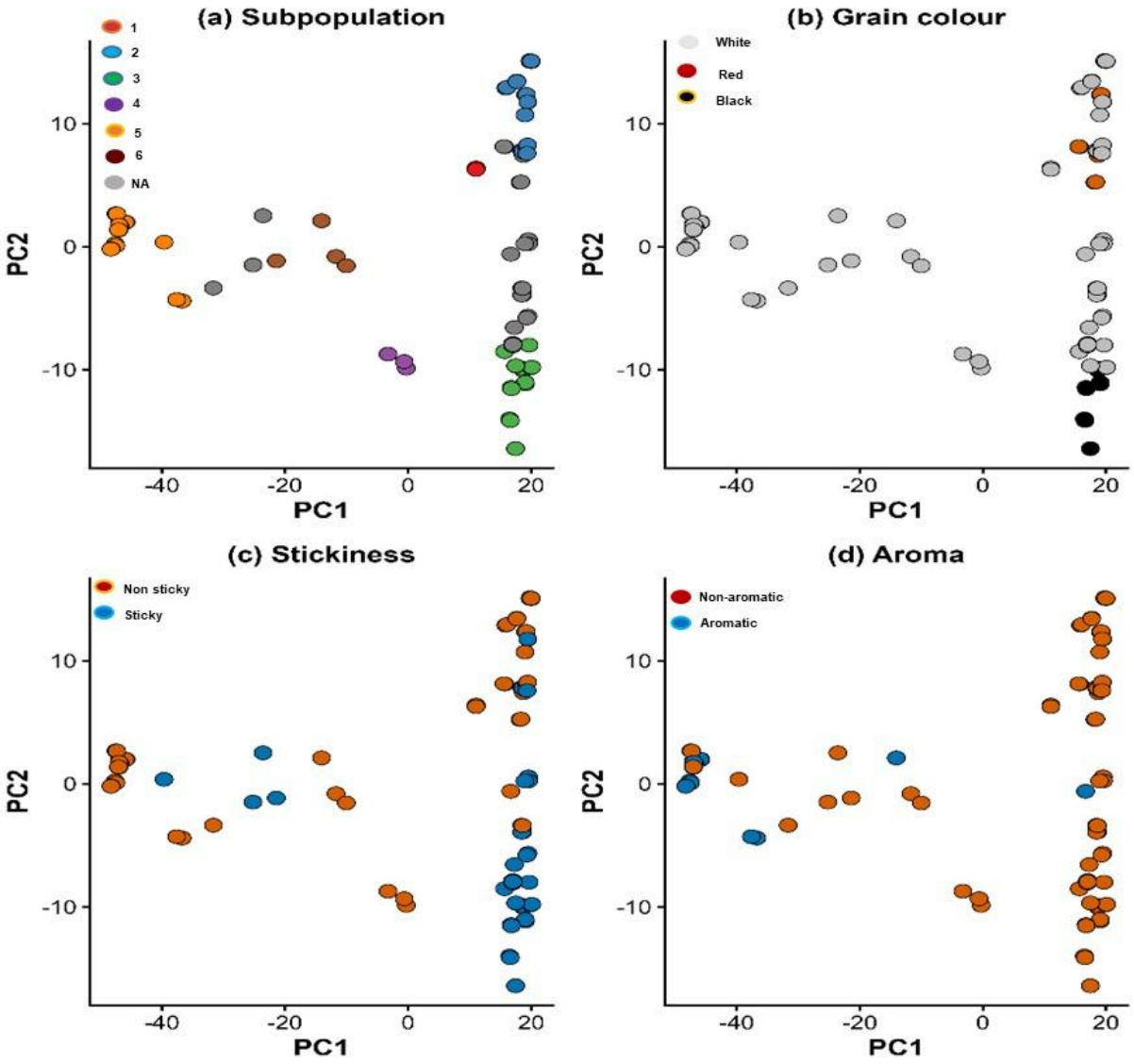


Figure 4. Principal component analysis (PCA) of rice samples from Assam, Northeast India genotyped using 9,714 genome-wide SNP markers. Samples are coloured according to (a) STRUCTURE-defined subpopulations, (b) grain colour, (c) stickiness phenotype, and (d) aroma phenotype. PC1 and PC2 explain the major axes of genetic variation and reveal clustering associated with population structure and grain-quality traits.

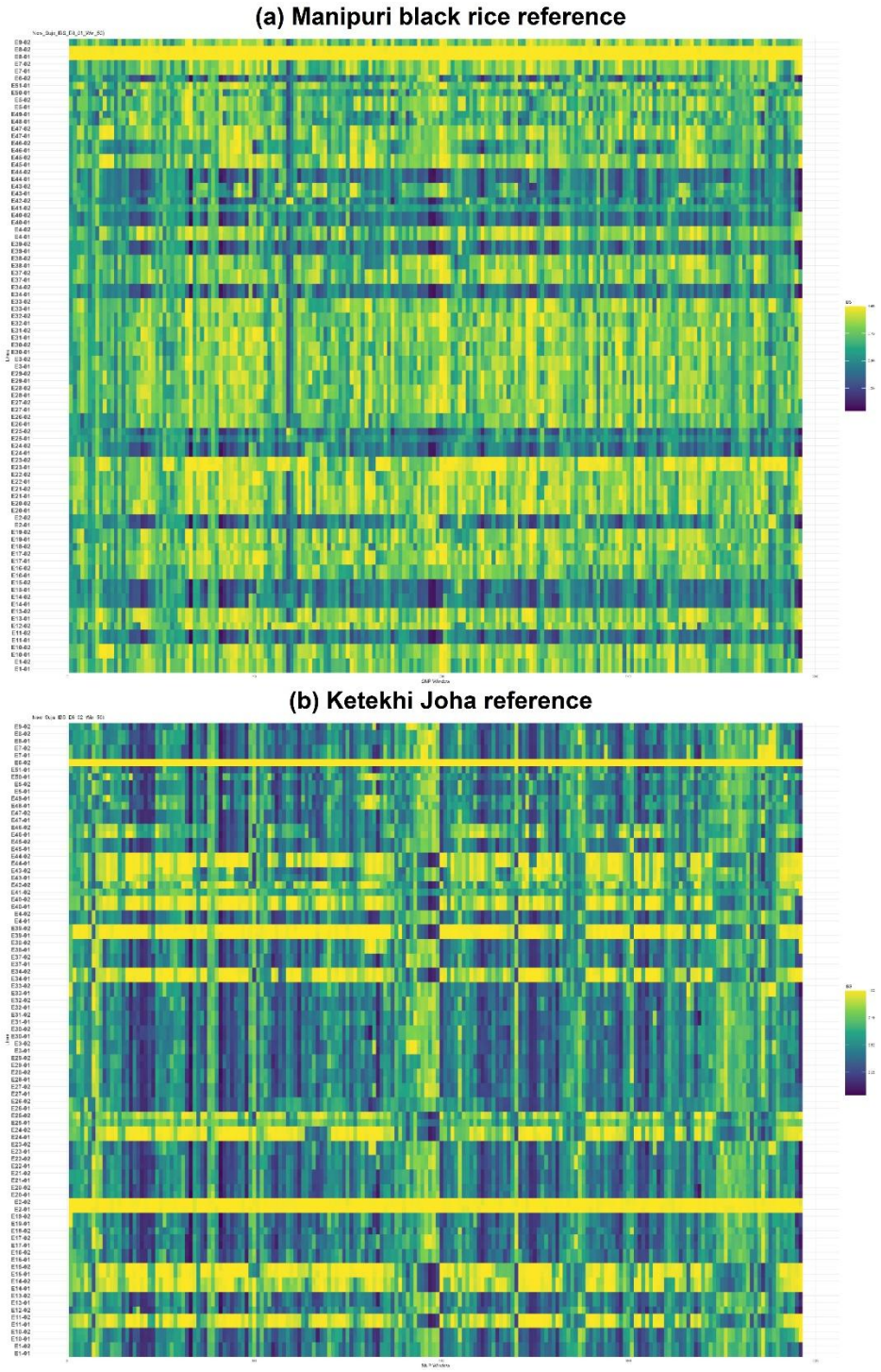
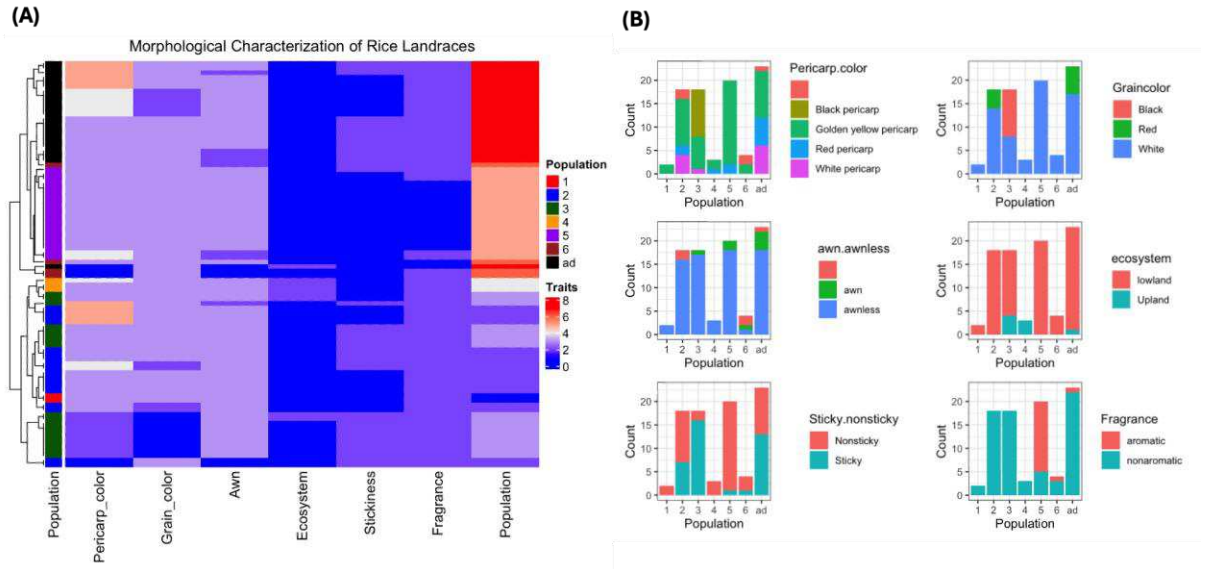
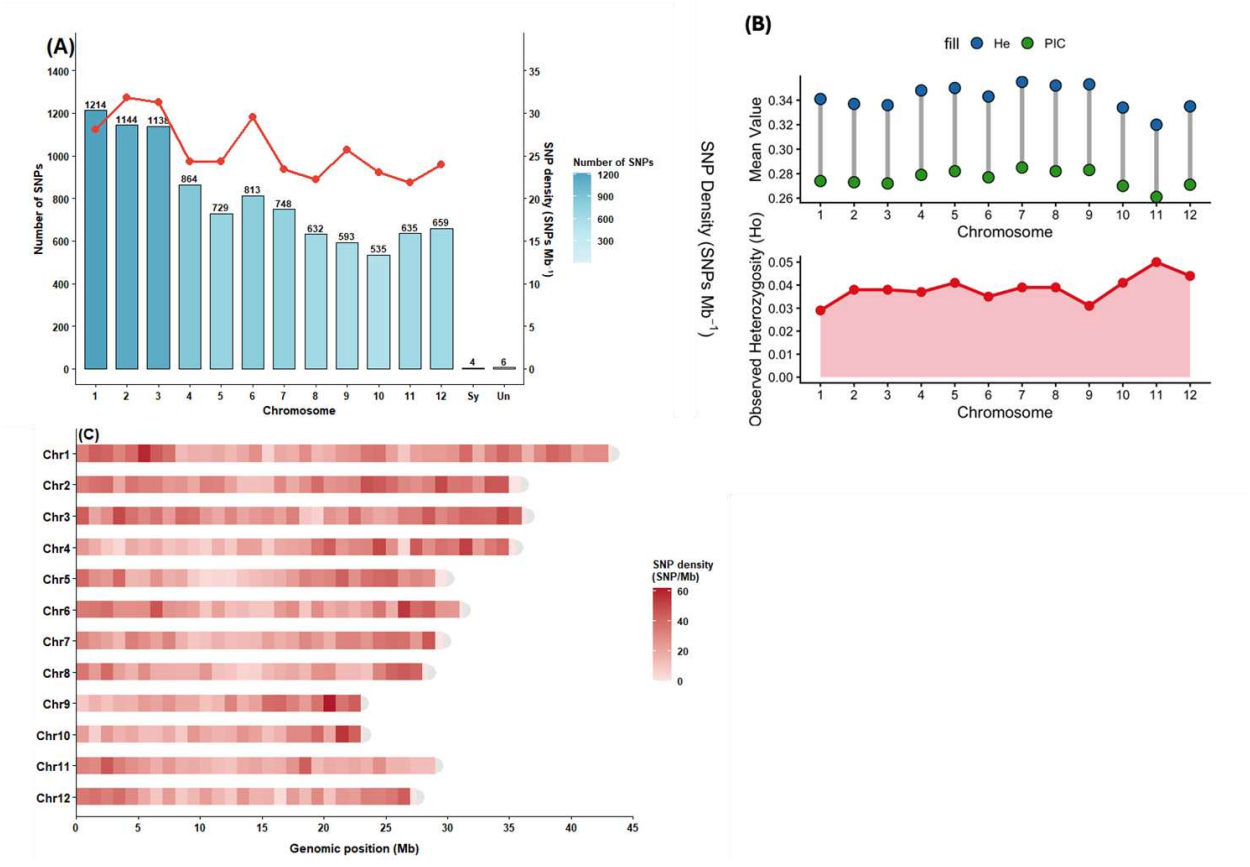


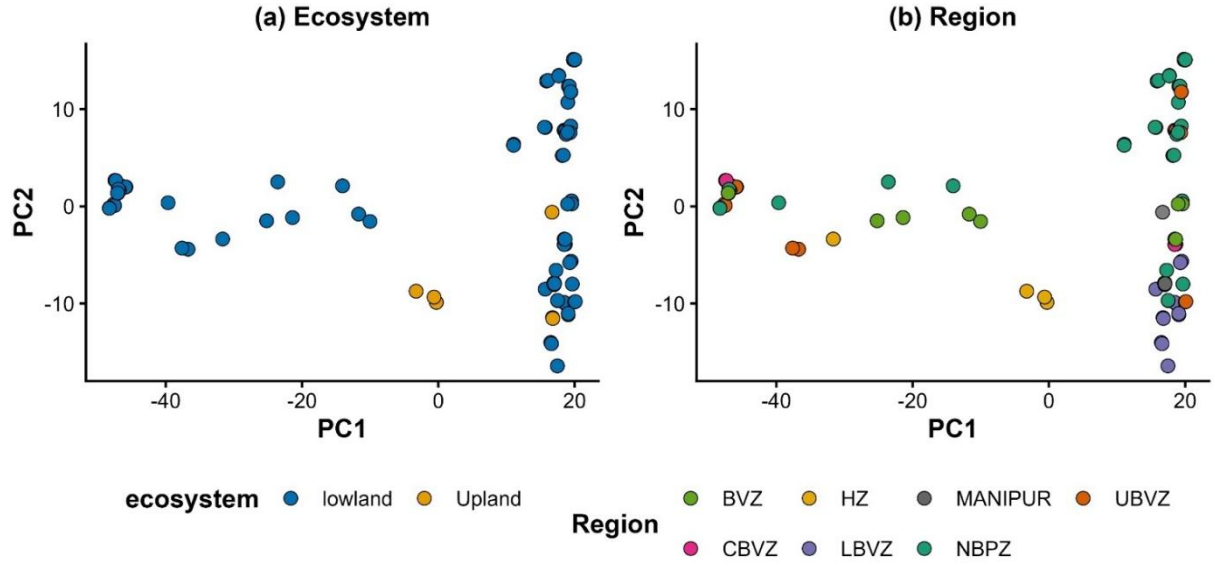
Figure 5. Identity-by-state (IBS) analysis of rice samples from Assam, Northeast India using trait-specific reference. (a) Genome-wide IBS values of rice samples relative to sticky- pigmented rice- Manipuri black rice as reference. (b) Genome-wide IBS values of rice samples relative to aromatic rice- Ketekhi Joha as reference. Each row represents a sample and each column represents genome-wide IBS estimates. Warmer colours indicate higher allele sharing with the reference.



Supplement Figure 1. Morphological diversity and trait distribution across structure-based subpopulation groups of rice samples. (A) Heatmap showing variation in pericarp color, grain color, awn status, ecosystem, grain stickiness, and fragrance among the rice samples from Assam, with structure-based population assignments (1–6 and admixed) indicated alongside. (B) Stacked bar plots showing the frequency of trait categories within each structure-based population group, highlighting phenotypic variation associated with genetic structure.



Supplement Figure 2. Chromosome-wise distribution of SNP markers and genetic diversity parameters. **(A)** Number of SNPs (bars) and SNP density (line; SNPs/Mb) across the 12 rice chromosomes. **(B)** Mean expected heterozygosity (He), polymorphic information content (PIC), and observed heterozygosity (Ho) estimated from chromosome-specific SNP datasets. **(C)** Chromosome-wise distribution of SNP density across the 12 rice chromosomes, shown as a genomic-position heatmap. Together, the panels illustrate genome-wide marker coverage and the distribution of genetic diversity across the rice genome.



Supplement Figure 3: Principal component analysis (PCA) of rice samples from Assam, Northeast India coloured according to (a) ecosystem (upland and lowland) and (b) geographic region of origin. These plots illustrate the contribution of ecological adaptation and regional distribution to the observed genomic variation

Table 1

Analysis of molecular variance (AMOVA) of the six STRUCTURE defined subpopulations

Source of Variation	Df	Sum of Squares	Mean Square	Variance Components	% Variation
Among populations	5	1,38,082.92	27,616.58	1,345.90	60.4
Among individuals within populations	59	91,055.28	1,543.31	664.35	29.86
Within individuals	65	13,950.09	214.62	214.62	9.65
Total	129	2,43,088.28	—	2,224.87	100

Note: Population assignments were based on STRUCTURE analysis (K = 6). AMOVA was conducted using the 65 non-admixed samples assigned to the six inferred subpopulations, whereas admixed samples (n = 23) were excluded from variance partitioning. AMOVA-derived FST = 0.605.

Table 2

Pairwise FST (below diagonal) and Nei's genetic distance (above diagonal) among STRUCTURE-defined subpopulations

Subpopulation	Subpop1	Subpop2	Subpop3	Subpop4	Subpop5	Subpop6
1	-	0.157	0.173	0.356	0.609	0.31
2	0.190	-	0.071	0.296	0.643	0.251
3	0.216	0.079	-	0.278	0.632	0.249
4	0.354	0.293	0.267	-	0.493	0.32
5	0.777	0.814	0.774	0.593	-	0.211
6	0.389	0.375	0.353	0.373	0.366	-

Table 3: Genotypic diversity metrics and mean within-subpopulation genetic distance of STRUCTURE-defined subpopulations

Sub population	N	MLG	eMLG	H	λ	Hexp	Ia	rbarD	Mean within subpopulation distance	No. of private alleles
1	2	2	2	0.693	0.5	0.78	NA	NA	0.009	23
2	18	18	10	2.89	0.944	0.665	225.4	0.0472	0.160	122
3	18	18	10	2.89	0.944	0.675	210.6	0.0394	0.186	359
4	3	3	3	1.099	0.667	0.761	26.4	0.0162	0.107	67
5	20	20	10	2.996	0.95	0.662	382.6	0.0878	0.121	214
6	4	4	4	1.386	0.75	0.776	1807.4	0.2863	0.214	22
Total	65	65	10	4.174	0.985	0.687	1384.9	0.1456	-	-

N: Number of individuals; MLG: Number of unique multilocus genotypes; eMLG: Expected MLG at smallest population size; H: Shannon–Wiener genotypic diversity; λ: Simpson's index of genotypic diversity; Hexp: Expected heterozygosity; Ia: Index of association; rbarD: Standardized index of association

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryfileS1andS22.xlsx