



KASP™ based markers reveal a population sub-structure in temperate rice (*Oryza sativa* L.) germplasm and local landraces grown in the Kashmir valley, north-western Himalayas

Shikari, Asif Bashir ; Najeeb, Sofi; Khan, Gazala; Mohidin, Fyaz A.; Shah, Ashaq H.; Nehvi, Firdous A.; Wani, Shafiq A.; Bhat, Nazir A.; Waza, Showkat A.; Subba Roa, L.V. ; Steele, Katherine; Witcombe, John

Genetic Resources and Crop Evolution

DOI:

[10.1007/s10722-020-01025-z](https://doi.org/10.1007/s10722-020-01025-z)

Published: 01/03/2021

Peer reviewed version

[Cyswllt i'r cyhoeddiad / Link to publication](#)

Dyfyniad o'r fersiwn a gyhoeddwyd / Citation for published version (APA):

Shikari, A. B., Najeeb, S., Khan, G., Mohidin, F. A., Shah, A. H., Nehvi, F. A., Wani, S. A., Bhat, N. A., Waza, S. A., Subba Roa, L. V., Steele, K., & Witcombe, J. (2021). KASP™ based markers reveal a population sub-structure in temperate rice (*Oryza sativa* L.) germplasm and local landraces grown in the Kashmir valley, north-western Himalayas. *Genetic Resources and Crop Evolution*, 68(3), 821–834. <https://doi.org/10.1007/s10722-020-01025-z>

Hawliau Cyffredinol / General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

1 **KASP™ based markers reveal a population sub-structure in temperate rice**
2 **germplasm and local landraces grown in the Kashmir valley, north-western**
3 **Himalayas**

4 **¹Asif Bashir Shikari, ¹Sofi Najeeb, ¹Gazala Khan, ¹Fayaz A. Mohidin, ¹Ashaq H.**
5 **Shah, ²Firdous A. Nehvi, ²Shafiq A. Wani, ¹Nazir A. Bhat, ¹Showkat A. Waza, ³L.V.**
6 **Subba Rao, ⁴Katherine A. Steele and ⁴John R. Witcombe**

7 *¹Mountain Research Centre for Field Crops, ²Division of Plant Biotechnology,*
8 *Sher-e-Kashmir University of Agricultural Sciences & Technology of Kashmir, J&K, India, 192*
9 *102*

10 *³ICAR-Indian Institute of Rice Research, Hyderabad, India, 500 030*

11 *⁴Bangor University, U.K.*

12 e-mail: asifshikari@skuastkashmir.ac.in

13 [ORCID: 0000-0002-2911-5536](https://orcid.org/0000-0002-2911-5536)

Name	Affiliation	e-mail
Asif Bashir Shikari (ABS)	<i>Mountain Research Centre for Field Crops, Sher-e-Kashmir University of Agricultural Sciences & Technology of Kashmir, J&K, India, 192 102</i>	asifshikari@skuastkashmir.ac.in
Sofi Najeeb (SN)	<i>Mountain Research Centre for Field Crops, Sher-e-Kashmir University of Agricultural Sciences & Technology of Kashmir, J&K, India, 192 102</i>	najeeb_sofi@rediffmail.com
Gazala Khan (GK)	<i>Mountain Research Centre for Field Crops, Sher-e-Kashmir University of Agricultural Sciences & Technology of Kashmir, J&K, India, 192 102</i>	moazin_khan@yahoo.co.in
Fayaz A. Mohidin (FAM)	<i>Mountain Research Centre for Field Crops, Sher-e-Kashmir University of Agricultural Sciences & Technology of Kashmir, J&K, India, 192 102</i>	famohiddin@rediffmail.com
Ashaq H. Shah (AS)	<i>Mountain Research Centre for Field Crops, Sher-e-Kashmir University of Agricultural Sciences & Technology of Kashmir, J&K, India, 192 102</i>	ahshah71@gmail.com
Firdous A. Nehvi (FAN)	<i>Division of Plant Biotechnology, Sher-e-Kashmir University of Agricultural Sciences & Technology of Kashmir, J&K, India, 192 102</i>	f.nehvi@rediffmail.com
Shafiq A. Wani (SAW)	<i>Division of Plant Biotechnology, Sher-e-Kashmir University of Agricultural Sciences & Technology of Kashmir, J&K, India, 192 102</i>	shafiqawani@gmail.com
Nazir A. Bhat (NAB)	<i>Mountain Research Centre for Field Crops, Sher-e-Kashmir University of Agricultural Sciences & Technology of Kashmir, J&K, India, 192 102</i>	nazirpathology@gmail.com
Showkat A. Waza (SW)	<i>Mountain Research Centre for Field Crops, Sher-e-Kashmir University of Agricultural Sciences & Technology of Kashmir, J&K, India, 192 102</i>	sahmad777@gmail.com
L. V. Subba Rao (LV)	<i>ICAR-Indian Institute of Rice Research, Hyderabad, India, 500 030</i>	lvsubbarao1990@gmail.com
Katherine A. Steele (KAS)	<i>Bangor University, U.K.</i>	k.a.steele@bangor.ac.uk
John R. Witcombe (JRW)	<i>Bangor University, U.K.</i>	j.r.witcombe@bangor.ac.uk

15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40

Abstract

The conservation and utilization of germplasm is contingent on its proper characterization at morphological or molecular levels. The present study aimed to elucidate the population sub-structure of 470 temperate rice germplasm collections of the Kashmir Valley. Analysis was carried out using KASP (Kompetitive Allele Specific PCR) assay on 213 genomic loci. Of these, a restricted set of 114 KASP loci were selected by the elimination of redundant, i.e. tightly linked markers based on map positions. STRUCTURE grouping was carried out to reveal three distinct sub-populations, K1, K2 and K3 comprising of 209, 156 and 105 germplasm accessions, respectively. Population F_{ST} values for K1, K2 and K3 were at 0.60, 0.24, 0.69, respectively, with highest pair-wise F_{ST} obtained between K2-K3 (0.53). Analysis using the restricted set of 114 markers gave a better inferred membership with a low average admixture of 15.1% compared with 22.6% based on the whole marker set. An improved agreement between STRUCTURE grouping and principal coordinate analysis was reached using the restricted marker set. Φ_{ST} values calculated based on nucleotide diversity also suggested three sub-populations: K2, mostly indica germplasm; K1 mostly exotic temperate japonica; and K3, local japonica varieties and landraces. Polymorphic SNPs and haplotypes were discovered which discriminated the three sub-populations. Fifteen KASP markers were most important in discriminating K2 from K1 and K3 and included SNPs associated with domestication within the *Wx*, *Ghd7* and *Ghd8* genes. KASP markers are cheaper than SSR markers. Some of the KASP markers were highly discriminatory, using both model and distance based approaches, and so can be used as a cost-effective tool for efficient maintenance and use of rice genetic resources.

Key words: Rice, temperate, Structure, Diversity, SNPs, KASP

41 **Introduction**

42 The enormous range of diversity in cultivated rice (*Oryza sativa* L.) is represented by more than 120,000
43 varieties worldwide (Khush 1997, Vasudevan et al. 2014) including cultivars and large number of landraces, with
44 around 50,000 of them present alone in India. Globally, the crop gene banks preserve and maintain a round 250,000
45 rice germplasm accessions which include cultivated types and their wild relatives (Jacob, et al., 2015). However, 95%
46 of these valuable gene repositories have rarely been used in any breeding program me (Peng et al. 2009). Initially, the
47 mapping of large number of microsatellites (SSR) markers (McCouch et al. 2002; Temnykh et al. 2001) and more
48 recent genome saturation through discovery of SNPs (McNally et al. 2009; Singh et al. 2015; Trinh et al. 2017; Trung
49 et al. 2017; Zhou et al. 2011) has helped to delineate genetic diversity and enabled a much better coverage of rice
50 genome and the underlying trait association. Germplasm utilization activity itself depends upon the preliminary
51 characterization and description of germplasm at population and individual level. The natural genetic variation in
52 landraces preserved in gene banks (Diez et al. 2018) or in *ex situ* germplasm repositories (Vanniarajan et al. 2012) has
53 been recently evaluated through the use of molecular markers.

54 The Himalayan tract represents a diversity hot spot of rice and is a home to tens of thousands of landraces.
55 In India, temperate rice is grown in the North-Western Himalayan region (comprising Jammu and Kashmir, Himachal,
56 and Utrakhnad) and North-Eastern hill states. The natural diversity within some areas of this broad region has been
57 recently studied with the help of morphological and molecular markers (Choudhury et al. 2013; Roy et al. 2015;
58 Salgotra et al. 2015; Umakanth et al. 2017). These studies have contributed to the evolutionary classification of rice
59 in mountainous areas and has also helped to quantify the allelic diversity in populations. On the other hand,
60 information on the detailed population structure and classification of high altitude rice from Kashmir is lacking. The
61 valley of Kashmir located at 34°N and 73°E form the northern-most part where rice cultivation extends from an altitude
62 1500 to over 2200 m and is characterised by landraces and varieties having excellent resilience to cold stress (Parray
63 and Shikari 2008). From a breeding point of view, introduced materials with early maturity and a certain degree of
64 cold tolerance have been successfully utilized for trait improvement (Shikari et al. 2018). The present study was
65 designed to assess the population structure of germplasm adapted to high altitude temperate ecology of Kashmir valley
66 with the help of SNP genotyping using KASP assay. KASP is a homogeneous, uniplex, fluorescence-based genotyping
67 technology based on allele-specific oligo extension and fluorescence resonance energy transfer (FRET) for signal

68 generation. KASP has been reported to show improved cost-effectiveness and reliability as compared to some of the
69 contemporary sequence-based markers (Semagn et al. 2014; Steele et al. 2018).

70 **Materials and methods**

71 **Plant materials**

72 We studied 470 rice (*Oryza sativa* L.) germplasm entries of diverse origin which are being maintained at the Mountain
73 Research Centre for Field Crops (MRCFC), Khudwani, SKUAST-Kashmir. The repository comprised of 39 local
74 landraces from Kashmir valley, 4 obsolete cultivars, 27 released varieties, 117 indigenous types (from parts of India
75 other than Kashmir) and 63 exotic collections, as well as 220 advance breeding lines/ derivatives. The landraces and
76 varieties included here mostly were of short round to medium slender grain type and belonged to japonica and indica
77 ecotypes. The accessions designated as ‘exotic’ were those which have been procured / collected from sources other
78 than belonging to Indian sub-continent. Over the years lines have undergone a process of purification, adaptive
79 selection and acclimatization through generations of maintenance and evaluation (Supplementary Table S1).

80 **SNP assay and Genotyping**

81 Leaf samples were collected using a 96 well format ‘Plant Sample Collection Kit’ (Biosearch Technologies,
82 Hoddesdon, Herts., U.K.) and subjected to genotyping. Each germplasm line was genotyped with KASP markers at
83 217 well-distributed genomic loci. Four of the loci with poor genotyping calls were dropped and thus 213 SNPs were
84 pursued for analysis. The KASP markers selected here were designed previously using *indica* varieties and the *indica*
85 reference genome and a large proportion of the loci are located within genes (Steele et al. 2018).

86 **Population structure analysis**

87 A Bayesian model-based clustering approach was implemented using the STRUCTURE v2.3.4 software (Pritchard et
88 al. 2000) in order to define population sub-clustering across 470 germplasm accessions. STRUCTURE performs
89 Bayesian assignment of individuals to a predefined number of K assumed sub-populations. An optimum number of
90 sub-populations were inferred from the software, pre-set at admixture ancestry model with correlated allelic
91 frequencies. SNP data was analysed at three replicate runs per K value, a burn-in period of 50000 and Markov Chain
92 Monte Carlo (MCMC) simulations of 100000. MCMC process randomly assigns individuals to a pre-determined
93 number of K groups followed by estimation of variant frequencies and re-designation of groups. The ideal K value

94 was determined by using adhoc ΔK based on the rate of change in the log probability of data between successive K
95 values (Evanno et al. 2005).

96 **Estimation of diversity statistics**

97 Fixation index is a measure of the reduction in heterozygosity or allele sharing at any one level of a population
98 hierarchy relative to another more inclusive level (Weir and Cockerham 1984; Weir and Hill 2002). F-statistics such
99 as inbreeding coefficient (F_{IS}), Fixation index (F_{ST}), and the pairwise F_{ST} , were computed using GenAlEx 6.5. F_{IS}
100 measures the extent of genetic inbreeding within subpopulations and is defined as the mean reduction in heterozygosity
101 of an individual due to non-random mating within a sub-population. F_{IS} can range from -1.0 (all individuals
102 heterozygous) to +1.0 (no observed heterozygotes). F_{ST} measures the extent of genetic differentiation among
103 subpopulations and is defined as the mean reduction in heterozygosity of a subpopulation (relative to the total
104 population) due to genetic drift among subpopulations. F_{ST} can range from 0.0 (no differentiation) to 1.0 (complete
105 differentiation where subpopulations happen to be fixed for different alleles). Further, the parameter $\Phi_{ST} = (\pi_T - \pi_S) / \pi_T$,
106 was calculated and provides an estimate of population differentiation based on nucleotide diversity (Excoffier et al.
107 1992). Here, π_T and π_S are analogous to H_T and H_S , described above, and reflect nucleotide diversity. The Simple
108 matching coefficients (Sokal and Michener 1958) based distance matrix was generated that was utilized for neighbour
109 joining method of clustering (Saitou and Nei 1987) with the help of MEGA X (Kumar et al. 2018). An AMOVA
110 (Analysis of molecular variance) (Peakall et al. 2003) was carried out using the GenAlEx 6.5 software (Peakall and
111 Smouse 2012). It was done with 9999 permutations. The same program was used to carry out principal coordinate
112 analysis across genotypic data. Mean genetic diversity (h) was calculated for each sub-population and was expressed
113 as: $h = [1/m(1 - \sum_{i=1}^n p_i^2)]$, where, m is the number of marker loci, n is the number of individuals in a population, p_i
114 is the allelic frequency. Both h and number of effective alleles ($Ne = 1 / \sum_{i=1}^n p_i^2$; with p_i as the allelic frequency)
115 were worked out using Power Marker V3.0 software.

116 **Restricted marker analysis**

117 After the STRUCTURE analysis was drawn with the help of 213 KASP markers, a sub-set of 114 markers was chosen
118 to repeat the estimation of population parameters. The marker sub-set was chosen after elimination of redundant
119 markers occupying same loci. The purpose was to reverse the overrepresentation of certain chromosomal segments.

120 Secondly, those eliminated were mostly linked to functional genes related to biotic and abiotic stress tolerance and it
121 avoided the clustering arising mainly from variability in such genes.

122 **Results**

123 **Assessment of population sub-structure**

124 A set of 470 rice germplasm accessions were investigated for various population parameters using 213 genome wide
125 SNP markers spotted through KASP technology. STRUCTURE, a program based on Bayesian model was used to
126 define population structure and yielded highest log likelihood estimate and peak ΔK value of 162.79 at $K = 4$, which
127 suggested classification into four sub-populations. The four sub-populations were named as K1, K2, K3 and K4 (after
128 *Khudwani*; location of our Research Centre) and turned out with an allocation of 84 (17.87%), 128 (27.23%), 76
129 (16.17%) and 182 (38.72%) genotypes, respectively (Supplementary Table S2). AMOVA (analysis of molecular
130 variance) revealed that Φ_{PT} , an estimate of population genetic differentiation, was equal to 53% of the total molecular
131 variance confirming a significant population structure. K2 versus K3 and K1 against K2, recorded highest pairwise
132 Φ_{PT} of 0.647 and 0.616, respectively. The variability feature was further explained using three-tiered diversity
133 parameters: H_I (mean observed heterozygosity per individual within subpopulations), H_s (mean expected
134 heterozygosity within subpopulations) and H_T (expected heterozygosity in total population), which were subsequently
135 used in the determination of population F-statistics. An important diversity parameter, F_{ST} was calculated for
136 individual populations and appeared in an order: K3 (0.5947) > K1 (0.4875) > K4 (0.2943) > K2 (0.2213), thereby
137 suggesting strong genetic sub-structure. In line with Φ_{PT} values mentioned above, highest pair-wise F_{ST} values were
138 recorded for K2 – K3 (0.4651), K1 – K2 (0.4184) and K2 – K4 (0.3263) comparisons, therefore, explained discernible
139 population differentiation. Principal Coordinates (PCs) were drawn on the data matrix with PC1 and PC2 explaining
140 53.15 and 5.57% of total variance with corresponding eigen values of 396.1 and 41.5, respectively. In addition to
141 strong signal for admixture as was revealed by STRUCTURE based grouping, further it did not correlate with the
142 pattern depicted by PCoA. Although, K2 plotted separately on negative PC1 axis against K1 and K3 which clustered
143 together within a narrow factor range on positive axis of PC1 with a limited spread of -0.276 to 0.227 on PC2.

144 **Restricted marker analysis:** In an attempt to refine our population estimates, only a sub-set of KASP markers was
145 chosen from a whole set of 213 (see Material and Methods). Analysis using limited marker set lead us to harvest only
146 three sub-populations (instead of four) named K1, K2 and K3 (Table 1, Supplementary Fig. S1) and accommodated

147 209, 156 and 105 germplasm accessions, respectively. F_{ST} values recorded for the three populations stood at 0.60,
148 0.24, 0.69, respectively (Table 2). Highest pair-wise F_{ST} was obtained between K2-K3 (0.531) followed by K1-K2
149 (0.467) and lowest for K1-K3 (0.127) (Table 3). Ancestral relations were deepened through restrictive marker analysis
150 with low average admixture levels of 15.1% compared with 22.6% on whole marker set. Individual sub-populations
151 K1, K2 and K3 had 15.3%, 7.1% and 22.9% individuals with overlapping ancestry (Supplementary Table S3, S4, Fig.
152 1).

153 **Principal Coordinate Analysis**

154 The first two principal coordinates marked eigen values of 207.9 and 21.3 and explained cumulative variance of 58.7%
155 (Supplementary Table S5). The grouping based on STRUCTURE and Principal Coordinate Analysis (PCoA) was
156 observed to follow a similar pattern under restricted marker analysis. Individuals in K2 clustered on negative PC-1 in
157 contrast to K1 and K3 those appeared in proximity along PC-1 with positive loadings (Fig. 2). The PCoA grouping
158 corresponded well with pair-wise F_{ST} values among the three sub-populations.

159 **Gene diversity**

160 A statistic, Φ_{ST} is a measure of population differentiation based on nucleotide diversity and was equal to 0.6795 (K1),
161 0.2915 (K2) and 0.7665 (K3) bearing a similar trend as that for Φ_{ST} . Sub-populations K1 to K3 recorded unbiased
162 mean diversity (u_h) estimates of 0.13 (K1), 0.25 (K2) and 0.10 (K3). The number of alleles per locus for a bi-allelic
163 SNP marker has to be two in every case and as such N_e (number of effective alleles) were 1.14 (K1), 1.33 (K2) and
164 1.11 (K3) (Table 2). As regards the nature of marker polymorphism, the information on frequency of transversions
165 was notably found to discriminate the sub-populations with respective values of 16.39% (K2), 7.58% (K3) and 7.04%
166 (K1) across populations. Coefficients of Nei's Genetic identity among populations were highest (0.663) between K1
167 – K3 and lowest (0.000) between K2 – K3. These values corresponded to the relationship explained by pairwise Φ_{ST}
168 coefficients and the results of the PCoA.

169 The Neighbour Joining method based on Simple Match Coefficients was applied to estimate the pattern of
170 genetic divergence and clearly defined two major clusters at a molecular distance of around 0.50 (Fig. 3). Most of the
171 japonica grouped into cluster-I and those of indica represented cluster-II. Out of a total of 470, Cluster-I and Cluster-
172 II included 313 and 157 accessions, respectively. Cluster-I was further partitioned into two sub-clusters, Cluster-Ia
173 and Cluster-Ib comprising of 166 and 147 accessions, respectively. Overall the individuals were categorized at an

174 average divergence coefficient of 0.32. The highest distance coefficients were recorded between genotype *GS-592*
175 against *Pusa Sugandh 3* (0.78) and HPR-2373 (0.77). Thirty nine local landraces originating from altitudinal range of
176 1500 to 2300 msl grouped closely within cluster-Ia (33) and cluster-Ib (4). Two other landraces *Yemberzul* and GS-
177 23 appeared in cluster-II. The landraces with red pericarp namely, *Tangdhar Zag* and *Karnah Zag* and popular
178 aromatic landraces, *Kamad* and *Mushk Budji* occupied similar clusters. Other temperate exotic and indigenous
179 collections with japonica background occupied cluster-I. Cluster-II featured with almost all the indica and derivative
180 lines. The 12 released and locally adapted varieties, spread across the tree circumference. Of these, high altitude
181 japonica varieties *K-332* and *Shalimar Rice-5* grouped into Cluster-Ia and *Barkat* and *Kohsar* in Cluster-Ib. All the
182 eight indica varieties (*China-988*, *China-1007*, *K-39*, *Chenab*, *Jhelum*, *Shalimar Rice-1*, *Shalimar Rice-2* and
183 *Shalimar Rice-3*) grouped in cluster-I. Fine grained *Pusa Sugandh 3* and Basmati variety *Pusa Basmati 1509* clustered
184 mid-way indica and japonica with proximity to the accessions adapted to North-western Himalayan region on one side
185 and japonica group at the other. Germplasm accessions representing different clusters are given in Fig. 4.

186 **Allelic polymorphism and distribution**

187 STRUCTURE grouping into sub-populations K1, K2 and K3 revealed a pattern in terms with distribution of
188 germplasm into indica and japonica. K2 mostly comprised of indica germplasm, K3 of local japonica (landraces) while
189 K1 included other japonica collections. Graphical genotypes over 114 SNP loci depicted the discriminatory alleles
190 (Fig. 5). At least fourteen SNPs discriminated K2 from K1 and K3 including two SNPs *Waxy* and *Amy_W2_R_1*,
191 that are both associated with the *Wx* locus on chromosome 6. Likewise, *Ghd7_05_SNP_ff_1* and *Ghd8_SNP_ff_2*
192 showed A/T and A/G polymorphism, respectively between K2 versus both K1-K3. SNP *SSII_1_SNP_ff_1* on
193 chromosome 10 is associated with Starch synthase II and produced A/G polymorphism between K2 / K1-K3
194 populations. In addition, 9 other SNP markers discriminated K2 from rest two populations, K1 and K3. A GAG
195 haplotype on chromosome 7 differentiated K2 from other two sub-populations which carried CGA at corresponding
196 sites. Likewise, K2 had the haplotype CC at two loci, on chromosome 9 against TT for K1 and K3. Markers
197 *RM171_SNP_nn_1*, *RM147_SNP_nn_3* and *RM590_SNP_ff_1* on chromosome 10 amplified a haplotype GAG in
198 K1 and K3 against ACT in K2. The unique locus *OsR498G0713985600_SNP_ff_1* differentiated between K1 and K3
199 with C/T polymorphism. (Fig. 5; Supplementary Table S6; Supplementary Fig. S2).

200 Near about 90% of accessions in K3 were local landraces and 83% of K1 were exotic (japonica) germplasm.
201 Of the 220 advanced breeding lines, more than 90% were grouped into K1 and K2 which indicates that varietal
202 breeding programmes have largely been carried through utilization of exotic and indigenous germplasm while
203 landraces have been promoted in their original form (Table 4). Pertinently, six KASP loci largely differentiated local
204 landraces from temperate exotic germplasm and included RM9B_SNP_nn_2; OsR498G0510120000_SNP_ff_3;
205 ALK_SNP_ff_1; RM51_SNP_nn_2; OsR498G0713985600_SNP_ff_1 and CRG4_SNP_nn_1 (Supplementary Fig.
206 S3).

207 Discussion

208 The restricted markers analysis procedure helped us to estimate population sub-structure among a set of 470
209 germplasm lines. The markers were selected to give a more uniform genome representation by elimination of
210 redundant, i.e., tightly linked loci that were mostly trait-based markers. An average of 50% (0.51) of variability was
211 explained by population sub-structure across 470 germplasm lines. The strong pair-wise F_{ST} values obtained between
212 K2 and other two populations were in line with the evolutionary expectations, since K2 was mostly comprised of
213 indica accessions and the japonica accessions were concentrated in K1 and K3. Inbreeding coefficient (F_{IS}) for all the
214 three populations was high (> 0.88) as expected for self-fertilizing species. The close proximity of K1-K3, as evident
215 from low pair-wise F_{ST} estimates (0.127), suggests high allele sharing between these two sub-populations. The average
216 F_{ST} of progenitor *Oryza rufipogon* measures 0.18 against 0.55 for domesticated *O. sativa* (Huang et al. 2010). Indica
217 are believed to have descended from Or-I (*O. rufipogon*-I) group with preservation of 75% of total genetic diversity
218 and $F_{ST} = 0.17$. On the other hand japonica and aromatic rice have descended from Or-III with strong bottleneck with
219 representation of 33% divergence and high level of population differentiation ($F_{ST} = 0.36$) (Huang et al. 2012). These
220 theories are supportive of the population differentiation levels for japonica (K1 and K3) and indica (K2) in our
221 materials. Between K1 and K3, the latter contained most of the landraces and recorded a higher F_{ST} than the former.
222 Landraces symbolize an intermediate stage of evolution between wild and cultivated germplasm. 'Inferred ancestry'
223 on individual basis was calculated from Q-Q plots based on percentage admixed individuals in a sub-population, i.e.,
224 where an admixture population was defined as having a greater than 15% probability of belonging to another
225 subpopulation. The estimate for admixture was 15.1% for 114 SNPs, while for the whole marker analysis (213 SNPs)
226 it was 22.6%, thereby validating the usefulness of elimination of redundant markers. The highest value for admixture

227 was in K3 (22.9%) followed by K1 (15.3%) and K2 (7.1%). High admixture levels in K3 were because of considerable
228 genome sharing with K1 as both populations mostly represent japonica. However, local landraces of Kashmir in sub-
229 population K3 were highly differentiated from those in K1, probably reflecting their distinct ancestry

230 STRUCTURE operates on assigning membership coefficients of individual samples towards sub-populations
231 (Pritchard et al. 2009), while PCoA aligns samples along meaningful coordinates (Mohammadi and Prasanna, 2003).
232 In our case, the two approaches showed a similar clustering pattern although the fine difference between K1-K2 was
233 dissipated in PoCA but was clearly resolved through STRUCTURE. The population defined through STRUCTURE
234 analysis under whole marker set differed to that produced from more uniformly distributed markers that proved to
235 reflect a more reliable grouping. The over-representation of certain parts of genome may lead to false conclusions
236 when estimating genetic diversity or population sub-structure. This statement is supported by the close agreement of
237 the results from STRUCTURE grouping and the PCoA on the restricted set of markers. Secondly, the restricted marker
238 analysis gave lower admixture levels compared with the analysis using all of the markers.

239 The more useful nucleotide diversity coefficients, Φ_{IS} , Φ_{ST} and Φ_{IT} were computed from the SNP data. The
240 coefficients are analogous to F-coefficients but are not dependent on heterozygosity. The Φ_{ST} coefficients are based
241 on the nature of SNP polymorphisms and, thereby, substantiate the presence of the population structure determined
242 by other methods. There was a varying proportion of transversions among the sub-populations, and this pattern
243 explains the differing Φ_{ST} of the sub-populations. While F_{ST} has been regarded as the outcome of recent sharing of
244 alleles, Φ_{ST} is an outcome of a long evolutionary history and, therefore, possess higher values (Excoffier et al. 1992).

245 **Estimates of genetic diversity:** The high genetic diversity of sub-population K2 is an outcome of greater allele sharing
246 in indica as compared to japonica. The level of genetic diversity happens to be low in japonica compared to indica
247 which is in agreement with the findings of Choudhary et al. (2013). Further, the individuals in sub-population K3
248 mostly belong to higher hills and have a lower diversity compared with K1 and K2 which originate from plains. The
249 diversity gradient across altitude has been mentioned by Roy et al. (2016). Since the SNPs are bi-allelic markers, N_e
250 (number of effective alleles) was less than two in all three sub-populations. Coefficients of Nei's Genetic identity
251 among populations were highest (0.66) between K1 – K3 and lowest (0.00) between K2 – K3. These values
252 corresponded to the relationship explained by pairwise F_{ST} coefficients. The genetic distance (GD) was computed
253 based on Simple Match coefficients followed by grouping through Neighbour joining principle. Two major cluster

254 were identified at inter-cluster distance of ~0.52. Cluster-I comprised of japonica and Cluster-II with most of indica
255 germplasm lines. Basmati and other derived accessions grouped mid-way. In spite of the correlations that were found
256 of GD with geographical origin, the division into indica, japonica and derivative class dominated over the clustering
257 pattern based on geographical area. For example, the varieties bred and released for same geographical area occupied
258 separate clusters: Shalimar Rice-1, -2, -3 (all indica from SKUAST-Kashmir) fell in Cluster-II, whereas K-78, K-332,
259 Shalimar Rice-5 grouped into Cluster Ia (all japonica from SKUAST-Kashmir). Earlier we performed the principal
260 component analysis on a set of 150 germplasm lines using 31 agro-morphological traits (Shikari et al., 2009) The
261 study delineated the accessions into two major clusters with some accessions falling mid-way. Our results are in close
262 conformity with the classification based on morphological markers. Although morphological markers are governed
263 by genic loci they may, in many cases, be different from molecular markers which also originate from non-genic
264 regions. Recently, Gaur et al. (2019) performed analysis based on kernel dimensions and found that the local landraces
265 plotted across two clusters. In the present study, we classified them in two sub-clusters, namely Ia and Ib. Some of the
266 landraces which are grown under mid-mountains like *Mushk Budji* and *Kamad* and belong to japonica, clustered
267 together. Similarly, the red pericarp landraces, *Tangdhar Zag* and *Karnah Zag*, which belong to same region were
268 hardly differentiated. Recently we carried out the studies on expression of quality related genes where these two
269 showed similar expression levels for quality (Hussain et al. 2020). On the other hand, even though some accessions
270 were placed within the same cluster within low genetic distance, they belonged to different ecologies.

271 Among the SNPs which differentiated the sub-population K2 from K1-K3, *Waxy_SNP*, *Amy_W2_R_1*
272 associated with *Wx* locus, *SNP SSII_1_SNP_ff_1* related to endosperm starch synthesis, *Ghd7_05_SNP_ff_1* and
273 *Ghd8_SNP_ff_2* linked to heading date were most prominent. Among the six KASP loci which differentiated local
274 landraces from exotic temperate germplasm, the marker, *ALK_SNP_ff_1* at the *ALK* locus determines kernel starch
275 properties. The *ALK* gene is linked to amylopectin chain-length in rice endosperm, and it co-segregates with starch
276 synthase II enzyme that determines gelatinization temperature in rice (Gao et al. 2011). Strong selection under
277 domestication has been reported for several important genes and include *Wx* for amylose (Wang et al. 1995), *qSH1*
278 for seed shattering (Konishi et al. 2006), *Rc* for pericarp colour (Sweeney et al. 2007), and *Ghd 7* related to heading
279 date (Huang et al. 2012). InDel (Sahu et al. 2017) and SSR (Vanniarajan et al. 2012) markers have also been reported
280 to differentiate indica and japonica populations. Such markers or genes reveal high degree of polymorphism between
281 indica and japonica and have possibly evolved before the divergence of the two ecotypes from a common progenitor.

282 The present level of genetic divergence points towards possible useful variability for traits of economic importance
283 and grain quality. Few of the germplasm accessions studied here were previously evaluated for cold tolerance
284 (Sanghera et al. 2011), apart from the work on characterization of landraces for stress resistance (Umakanth et al.
285 2017). We recently identified certain specific alleles for blast resistance (Shikari et al. 2014) and also revealed
286 differential expression for γ -amino butyric acid among rice landraces (Hussain et al. 2020). Besides, the
287 characterization and genetic improvement of landraces was carried out for resistance towards rice blast (Khan et al.
288 2018). The process of germplasm characterization helps in the documentation and the long-term conservation of
289 germplasm which, in turn, may help in the better utilization of genetic resources for the development of improved rice
290 varieties. Further, the genotyping process can help define a core germplasm set and may also help in the selection of
291 a population for mapping useful alleles linked to traits of economic importance.

292 KASP markers are more cost effective than SSR markers (Steele et al., 2018) that have previously been
293 commonly used to characterise germplasm in rice (Yang et al., 2019), wheat (Roncallo et al., 2019) and Brassica (Li
294 et al. (2019). The markers effectively divided a population of germplasm of the temperate region of the Kashmir valley
295 into sub-populations with the greatest distinction between indica and japonica groupings. A small number of KASP
296 markers were highly discriminatory and were usually associated with domestication traits. KASP markers, and
297 specifically highly discriminatory markers, can be used as a cost-effective tool for the more efficient maintenance and
298 use of rice genetic resources.

299 **Compliance with Ethical Standard**

300 **Conflict of interest: The authors declare that they have no conflict of interest.**

301 **Author Contributions:** ABS carried out the field work for the study; ABS, KAS, and JRW were
302 involved in the KASP genotyping process; KAS identified the KASP markers for the study; ABS,
303 GK performed statistical analysis; SN, NAB, LV facilitated the access to and maintenance of
304 germplasm; FAN, SAW supported for critical inputs; SW helped in editing the analyses; ABS and
305 GK wrote the article with assistance from JRW

306 **KASP™ based markers reveal a population sub-structure in temperate rice**
307 **germplasm and local landraces grown in the Kashmir valley, north-western**
308 **Himalayas**

309 **¹Asif Bashir Shikari, ¹S. Najeeb, ¹Gazala Khan, ¹F. A. Mohidin, ¹Ashaq H. Shah**
310 **¹Showkat A. Waza, ²F. A. Nehvi, ²Shafiq A. Wani, ¹N. A. Bhat, ³L.V.S. Rao, ⁴K. A.**
311 **Steele and ⁴J. R. Witcombe**

312 *¹Mountain Research Centre for Field Crops, ²Division of Plant Biotechnology,*
313 *Sher-e-Kashmir University of Agricultural Sciences & Technology of Kashmir, J&K, India, 192*
314 *102*

315 *³ICAR-Indian Institute of Rice Research, Hyderabad, India, 500 030*

316 *⁴Bangor University, U.K.*

317 e-mail: asifshikari@skuastkashmir.ac.in

318 ORCID: 0000-0002-2911-5536

319

320 **Acknowledgement:**

321 Authors duly acknowledge the funding and technical support facilitated by Innovate U.K. and
322 Biosearch Technologies, Hoddesdon, Herts, U. K. for the conduct of this study

323

324

325

326

327

328

329

330

331

332

333

334

335 **References**

- 336 Choudhary G et al. (2013) Molecular Genetic Diversity of Major Indian Rice Cultivars over Decadal Periods PloS
337 one 8:e66197 doi:10.1371/journal.pone.0066197
- 338 Choudhury B, Khan ML, Dayanandan S (2013) Genetic structure and diversity of indigenous rice (*Oryza sativa*)
339 varieties in the Eastern Himalayan region of Northeast India SpringerPlus 2:228 doi:10.1186/2193-1801-2-
340 228
- 341 Diez MJ et al. (2018) Plant Genebanks: Present Situation and Proposals for Their Improvement. the Case of the
342 Spanish Network Frontiers in plant science 9:1794 doi:10.3389/fpls.2018.01794
- 343 Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software
344 STRUCTURE: a simulation study Molecular ecology 14:2611-2620 doi:10.1111/j.1365-
345 294X.2005.02553.x
- 346 Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among
347 DNA haplotypes: application to human mitochondrial DNA restriction data Genetics 131:479-491
- 348 Gao Z, Zeng D, Cheng F, Tian Z, Guo L, Su Y, Yan M, Jiang H, Dong G, Huang Y, Han B (2011) ALK, the Key
349 Gene for Gelatinization Temperature, is a Modifier Gene for Gel Consistency in Rice Journal of integrative
350 plant biology, 53(9), pp.756-765
- 351 Gaur A, Pararay GA, Shikari AB and Najeeb S (2019). Capturing the Genetic Diversity for Grain Quality Attributes in
352 a Set of Temperate Rice (*Oryza sativa* L.) Germplasm by Cluster Analysis and the Assessment of Wx gene
353 Polymorphism Int. J. Pure Applied Biosciences, 7(3): 67-73
- 354 Huang X et al. (2010) Genome-wide association studies of 14 agronomic traits in rice landraces Nature genetics
355 42:961
- 356 Huang X et al. (2012) A map of rice genome variation reveals the origin of cultivated rice Nature 490:497-501
- 357 Hussain SZ, Jabeen R, Naseer B, Shikari AB (2020) Effect of soaking and germination conditions on γ -
358 aminobutyric acid and gene expression in germinated brown rice Food Biotechnology 34:132-150
359 doi:10.1080/08905436.2020.1744448
- 360 Iqbal AM, Chrungoo N, Shikari A, Najeeb S, Gayle A, Mujtaba A (2017) Molecular Diversity of Rice Germplasm
361 Grown Under High Altitude Conditions Plant Cell Biotechnology And Molecular Biology: 481-488

362 Jacob SR, Tyagi V, Agrawal A, Chakrabarty SK, Tyagi RK (2015) Indian plant germplasm on the global platter: an
363 analysis. *PloS one*, 10(5), p.e0126634.

364 Khan GH et al. (2018) Marker-assisted introgression of three dominant blast resistance genes into an aromatic rice
365 cultivar Mushk Budji *Scientific reports* 8:4091 doi:10.1038/s41598-018-22246-4

366 Khush GS (1997) Origin, dispersal, cultivation and variation of rice *Plant molecular biology* 35:25-34

367 Konishi S, Iza wa T, Lin SY, Ebana K, Fukuta Y, Sasaki T, Yano M (2006) An SNP caused loss of seed shattering
368 during rice domestication *Science* 312:1392-1396

369 Kumar S, Stecher G, Li M, Knyaz C, Tamura K (2018) MEGA X: Molecular Evolutionary Genetics Analysis across
370 Computing Platforms *Molecular biology and evolution* 35:1547-1549 doi:10.1093/molbev/msy096

371 Li, P. et al. (2019) Development of a core set of KASP markers for assaying genetic diversity in *Brassica rapa*
372 subsp. *chinensis* Makino *Plant Breeding* 138.3: 309-324 doi.org/10.1111/pbr.12686

373 McCouch SR et al. (2002) Development and mapping of 2240 new SSR markers for rice (*Oryza sativa* L.) *DNA*
374 *research* 9:199-207

375 McNally KL et al. (2009) Genomewide SNP variation reveals relationships among landraces and modern varieties
376 of rice *Proceedings of the National Academy of Sciences of the United States of America* 106:12273-12278
377 doi:10.1073/pnas.0900992106

378 Mohammadi SA and Prasanna BM (2003) Analysis of genetic diversity in crop plants—salient statistical tools and
379 considerations. *Crop science*, 43(4), pp.1235-1248

380 Najeeb S, Parry GA, Shikari AB, Zaffar G, Kashyap SC, Ganie MA, Shah A (2017) Farmers participatory selection
381 of new rice varieties to boost production under temperate agro-ecosystems *Journal of Integrative*
382 *Agriculture* (17)61810 doi:10.1016/s2095-3119

383 Parry G, Shikari AB (2008) Conservation and characterization of indigenous rice germplasm adapted to
384 temperate/cooler environments of Kashmir valley *ORYZA-An International Journal on Rice* 45:198-201

385 Peakall R, Ruibal M, Lindenmayer DB (2003) Spatial autocorrelation analysis offers new insights into gene flow in
386 the Australian bush rat, *Rattus fuscipes* *Evolution; international journal of organic evolution* 57:1182-1195
387 doi:10.1111/j.0014-3820.2003.tb00327.x

388 Peakall R, Smouse PE (2012) GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and
389 research—an update *Bioinformatics* 28:2537-2539 doi:10.1093/bioinformatics/bts460

390 Peng ZY et al. (2009) Characterization of the genome expression trends in the heading-stage panicle of six rice
391 lineages *Genomics* 93:169-178 doi:10.1016/j.ygeno.2008.10.005

392 Porras-Hurtado L, Ruiz Y, Santos C, Phillips C, Carracedo Á, Lareu M (2013) An overview of STRUCTURE:
393 applications, parameter settings, and supporting software *Frontiers in genetics* 4:98

394 Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data
395 *Genetics* 155:945-959

396 Pritchard V, Metcalf J, Jones K, Martin A, Cowley D (2009) Population structure and genetic management of Rio
397 Grande cutthroat trout (*Oncorhynchus clarkii virginalis*) *Conservation Genetics* 10:1209

398 Roncallo PF et al. (2019) Genetic diversity and linkage disequilibrium using SNP (KASP) and AFLP markers in a
399 worldwide durum wheat (*Triticum turgidum* L. var *durum*) collection *PloS one* 14.6
400 doi:10.1371/journal.pone.0218562

401 Roy S et al. (2015) Genetic Diversity and Population Structure in Aromatic and Quality Rice (*Oryza sativa* L.)
402 Landraces from North-Eastern India *PloS one* 10:e0129607 doi:10.1371/journal.pone.0129607

403 Roy S, Mamdi BC, Mawkhlieng B, Banerjee A, Yadav RM, Misra AK, Bansal KC (2016) Genetic diversity and
404 structure in hill rice (*Oryza sativa* L.) landraces from the North-Eastern Himalayas of India *BMC genetics*
405 17:107 doi:10.1186/s12863-016-0414-1

406 Sahu PK, Mondal S, Sharma D, Vishwakarma G, Kumar V, Das BK (2017) InDel marker based genetic
407 differentiation and genetic diversity in traditional rice (*Oryza sativa* L.) landraces of Chhattisgarh, India
408 *PloS one* 12:e0188864 doi:10.1371/journal.pone.0188864

409 Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees
410 *Molecular biology and evolution* 4:406-425

411 Salgotra RK, Gupta BB, Bhat JA, Sharma S (2015) Genetic Diversity and Population Structure of Basmati Rice
412 (*Oryza sativa* L.) Germplasm Collected from North Western Himalayas Using Trait Linked SSR Markers
413 *PloS one* 10:e0131858 doi:10.1371/journal.pone.0131858

414 Sanghera G, Hussaini A, Anwer A, Kashyap S (2011) Evaluation of some IRCTN rice genotypes for cold tolerance
415 and leaf blast disease under temperate Kashmir conditions *Journal of Hill Agriculture* 2:28-32

416 Semagn K, Babu R, Hearne S, Olsen M (2014) Single nucleotide polymorphism genotyping using Kompetitive
417 Allele Specific PCR (KASP): overview of the technology and its application in crop improvement
418 Molecular breeding 33:1-14

419 Shikari AB, Parray GA, Rather AG and Sheikh, FA (2009). Principal component analysis for evaluation of rice
420 (*Oryza sativa* L.) germplasm. Journal of Rice Research, 2(1): 16-22

421 Shikari AB, Najeeb S, Khan GH, Ali G, Parray G, Zargar S, Sheikh F (2018) DNA fingerprinting of rice (*Oryza*
422 *sativa* L.) varieties cultivated in Kashmir SKUAST Journal of Research 20:32-36

423 Shikari AB et al. (2014) Identification and validation of rice blast resistance genes in Indian rice germplasm Indian
424 Journal of Genetics and Plant Breeding (The) 74:286-299

425 Singh N et al. (2015) Single-copy gene based 50 K SNP chip for genetic studies and molecular breeding in rice
426 Scientific reports 5:11600 doi:10.1038/srep11600

427 Sokal R, Michener C (1958) A statistical method for evaluating systematic relationships. iniv. kansas sci. bull., 38:
428 1409–1438 Prim Product Ecol Factors Lake Maggiore 127

429 Steele KA et al. (2018) Accelerating public sector rice breeding with high-density KASP markers derived from
430 whole genome sequencing of indica rice Molecular breeding : new strategies in plant improvement 38:38
431 doi:10.1007/s11032-018-0777-2

432 Sweeney MT, Thomson MJ, Cho YG, Park YJ, Williamson SH, Bustamante CD, McCouch SR (2007) Global
433 dissemination of a single mutation conferring white pericarp in rice PLoS genetics 3

434 Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch S (2001) Computational and
435 experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon
436 associations, and genetic marker potential Genome research 11:1441-1452 doi:10.1101/gr.184001

437 Trinh H et al. (2017) Whole-Genome Characteristics and Polymorphic Analysis of Vietnamese Rice Landraces as a
438 Comprehensive Information Resource for Marker-Assisted Selection International journal of genomics
439 2017:9272363 doi:10.1155/2017/9272363

440 Trung KH, Nguyen TK, Khuat HBT, Nguyen TD, Khanh TD, Xuan TD, Nguyen XH (2017) Whole Genome
441 Sequencing Reveals the Islands of Novel Polymorphisms in Two Native Aromatic Japonica Rice Landraces
442 from Vietnam Genome biology and evolution 9:1816-1820 doi:10.1093/gbe/evx135

443 Umakanth B et al. (2017) Diverse Rice Landraces of North-East India Enables the Identification of Novel Genetic
444 Resources for Magnaporthe Resistance Frontiers in plant science 8:1500 doi:10.3389/fpls.2017.01500

445 Vanniarajan C, Vinod KK, Pereira A (2012) Molecular evaluation of genetic diversity and association studies in rice
446 (*Oryza sativa* L.) Journal of genetics 91:9-19 doi:10.1007/s12041-012-0146-6

447 Vasudevan K, Vera Cruz CM, Grisse W, Bhullar NK (2014) Large scale germplasm screening for identification
448 of novel rice blast resistance sources. Frontiers in plant science, 5, p.505 doi: 10.3389/fpls.2014.00505

449 Wang ZY et al. (1995) The amylose content in rice endosperm is related to the post-transcriptional regulation of the
450 waxy gene The Plant Journal 7:613-622

451 Weir BS, Cockerham CC (1984) Estimating F-Statistics for the Analysis of Population Structure Evolution;
452 international journal of organic evolution 38:1358-1370 doi:10.1111/j.1558-5646.1984.tb05657.x

453 Weir BS, Hill WG (2002) Estimating F-statistics Annual review of genetics 36:721-750
454 doi:10.1146/annurev.genet.36.050802.093940

455 Yang G et al. (2019) Development of a core SNP arrays based on the KASP method for molecular breeding of
456 rice Rice 12: 21 doi.org/10.1186/s12284-019-0272-3

457 Zhou M et al. (2011) Genome-wide analysis of clustering patterns and flanking characteristics for plant microRNA
458 genes The FEBS journal 278:929-940 doi:10.1111/j.1742-4658.2011.08008.x