OPEN

Adaptive fitness of *Sapindus emarginatus* Vahl populations towards future climatic regimes and the limiting factors of its distribution

Ashish Kumar Pal[1,2], Vivek Vaishnav[1], Baleshwar Meena[1], Nalini Pandey[2] & Tikam Singh Rana[1]*

Sapindus emarginatus Vahl (Sapindaceae) also known as 'Indian Soap nut' is significantly important for saponin content in its fruits. However, its current population in India is heavily fragmented due to a lack of sustainable harvesting practices. Moreover, changing climatic regimes may further limit its distribution and possibly compromise the survival of the species in nature. The aim of the present study was to: predict the future distribution range of *S. emarginatus*; identify the bioclimatic variables limiting this distribution and to evaluate its adaptive fitness and genomic resilience towards these variables. To determine future species distribution range and identify limiting bioclimatic variables, we applied two different ecological niche models (ENMs; BioClim and MaxEnt) on real occurrence data (n = 88 locations). The adaptive fitness of the species was evaluated by quantifying the genetic variability with AFLP markers and marker-environmental associations, using AFLP-associated Bayesian statistics. We found 77% overlap between the baseline (2030) and predicted (2100) species distribution ranges, which were primarily determined by maximum temperature (T_{MAX}) and mean annual precipitation (MAP). The T_{MAX} and MAP contributed 43.1% and 27.1%, respectively to ENM model prediction. Furthermore, AFLP loci significantly associated with bioclimatic variables, and T_{MAX} and MAP represent the lowest proportion (6.15%), confirming to the severe response of the species genome towards these variables. Nevertheless, the very low Linkage disequilibrium (LD) in these loci (4.54%) suggests that the current sensitivity to T_{MAX} and MAP is subject to change during recombination. Moreover, a combination of high heterozygosity (0.40–0.43) and high within-population variability (91.63 ± 0.31%) confirmed high adaptive fitness to maintain reproductive success. Therefore, the current populations of *S. emarginatus* have substantial genomic resilience towards future climate change, albeit significant conservation efforts (including mass multiplication) are warranted to avoid future deleterious impacts of inbreeding depression on the fragmented populations.

Sapindus emarginatus Vahl (Sapindaceae) also known as 'Indian 'Soap nut' is a tree species native to Indian sub-continent. In India, it is considered to be originated in the Western-Ghats and later extended up to the West-Central- North Indian biogeographical regions[1]. The species is of substantial importance for global trade in soap and perfumery industries due to the oil and saponin content obtained from its leaves, fruits and seeds. The global trade of saponin from all plant sources is expected to grow up to 970 million US$ in 2023, from 950 million US$ in 2017[2]. India exports approximately 63,368 kg/year of saponin extract obtained from all plant sources to different countries[3]. The exact contribution of *S. emarginatus* to the global saponin market is unclear, but probably substantial as the species is widely used traditionally at Indian homes as detergent to wash precious clothes and ornaments and as shampoo to wash hair. Due to high commercial demand, the species has been enlisted under 'priority species of economic importance' in the country report of state of forest genetic resource in India[4] and the

[1]Plant Diversity, Systematics and Herbarium Division, CSIR-National Botanical Research Institute, Rana Pratap Marg, Lucknow, 226001, India. [2]Plant Nutrition and Stress Physiology Laboratory, Department of Botany, University of Lucknow, Lucknow, 226007, India. *email: ranats@nbri.res.in

1

available natural patches of the species in some places have been restored through reforestation[4]. Lack of sustainable harvesting practices has been causing problems in the natural regeneration, as the fruits are harvested without leaving seeds for its natural regeneration[5,6], which led to fragmentation of the populations from wild and now the plants are found scattered in dry and moist deciduous forests and its periphery. Further, the distribution of the plants is subjected to significant impact due to changing climatic regimes. Therefore, there is a need to evaluate the available genetic resource of *S. emarginatus* and its adaptive fitness in projected bioclimatic regimes to ensure its sustainability.

Species distribution modeling (SDM) based on fundamental ecological niche theory has become an integral tool to provide the biogeographic extent of a species distribution[7]. With the current occurrence data of a species, the ecological niche modeling (ENM) tools such as BioClim (bioclimatic analysis and prediction system) and MaxEnt (maximum entropy), have been employed to predict the suitable niches for species population and their response toward different bioclimatic variables integrating the projected climatic regimes[8,9]. However, the actual climate tolerance of long-lived species is wider than the climatic envelope they currently occupy[10]. Specifically, the forest species are practiced to plant outside of its natural distribution for commercial, breeding and improvement purposes in monoculture. It has also been suggested for covering the commercial forestry trials, botanical gardens, and biodiversity database along with the natural distribution of a tree species to measure its climatic or environmental requirements through SDM/ENM[11]. Nevertheless, these model-based predictions are not practically successful until the species population itself is not adaptively fit[12]. A species needs to be resilient enough for local adaptation to maintain its survivability facing bottleneck selection pressures. In particular, the response of a species may vary toward dynamic extrinsic factors like changing climatic conditions. On the other hand, the existence of a high level of genetic heterogeneity within a population can lead to the increased adaptive fitness and evolutionary potential in longer-term[13–16]. An adaptive fit population is tend to transfer the same genetic characteristic to the future generations[17]. Ultimately, the response of species to the predicted changing climatic condition with high genetic heterogeneity leading to the adaptive fitness is of significant importance for the sustainability of population in limited distribution[18].

The DNA-based marker systems such as dominant markers i.e. amplified fragment length polymorphism (AFLP), inter-simple sequence repeats (ISSR), DAMD (directed amplification of minisatellite DNA) and co-dominant markers i.e. simple sequence repeats (SSR) have become an important tool to evaluate the distribution of heterogeneity and structure of the plant genetic resources. Among the dominant markers, AFLP exhibits higher reproducibility[19], robust informativeness, higher multiplex ratio, wider genome coverage, higher heritability[20] and fewer artifacts compared to ISSR and other markers[21]. In a previous report on *S. emarginatus*, the AFLP markers revealed comparatively higher polymorphism (P%) and polymorphic information content (PIC) values than the ISSR and DAMD markers[22]. Using ISSR markers the genetic diversity and differentiation among *S. emarginatus* populations from three biogeographical regions in India has been evaluated[22]. Assuming the population in Hardy-Weinberg Equilibrium (HWE) and without consideration of spatial and geographical factors, the study resulted into three distinct groups showing affinities towards their biogeographical regions[22]. Unlike the classical genetic algorithm, the Bayesian algorithm developed for AFLP markers[23–25] is advantageous to estimate the genetic diversity and differentiation of populations without considering of HWE and inbreeding coefficient[26]. Moreover, it is also able to handle the biased estimates of genetic diversity measures due to unequal sample sizes.

The present investigation was performed to address the major questions: How the future climatic regime is going to affect the species population distribution? What is the state of the genomic resilience of different populations of *S. emarginatus* to respond to the projected climatic regime? What are the bioclimatic variables contributing as limiting factor in population distribution? In order to resolve these questions, ENM was performed with real occurrence data of the populations of *S. emarginatus* from three biogeographical regions in India (Fig. 1) with multi-modeling approach with two different models *viz.* BioClim and MaxEnt, comparing the predicted species distribution pattern between baseline (the year 2030) and projected (the year 2100) climatic regimes. The fitness of the populations was evaluated assessing the genetic variability using AFLP markers with Bayesian statistics. The signature of adaptation in genome of the species was evaluated using genome-wide association (GWAS) between marker and bioclimatic variables.

Results

A significant correlation ($p < 0.05$, $r = 0.85$) was found between the predictions by both the models (BioClim and MaxEnt) for the species distribution on its range of occurrence in the baseline (the year 2030) and future (the year 2100) climatic regimes. Therefore, the species distribution prediction (Fig. 2) and output obtained from the MaxEnt was applied for all further estimations. The comparison between the species distribution patterns predicted for the baseline and future climatic regimes resulted into 77% proportion of overlapping range. The niches from Deccan peninsula biogeographic region were predicted to be the most suitable for survival of the species. The niches of species occurrence at Semi-arid biogeographical region showed comparatively less suitability for the survival of species for the year 2100 (Fig. 2). During prediction, MaxEnt resulted no change in area under curve (AUC = 0.99 ± 0.0) for both climatic regimes. The prediction of limiting bioclimatic variables (by BioClim) and their responses to the probability of presence (by MaxEnt) resulted in the relative proportion of maximum temperature ($T_{MAX} = 43.10\%$), and annual precipitation (MAP = 27.10%) with the highest contribution in the model predictions (Fig. 3, Table 1). The species niches were found with sensitive response towards T_{MAX} ($< 10\,°C$), temperature range (T_{RANGE}; $> 10\,°C$) and increased MAP ($< 1000\,mm$).

The AFLP primers amplified 1957 loci (103 ± 39.37 loci/primer) with $94.79 \pm 6.43\%$ polymorphism (1859 polymorphic loci), 0.65 ± 0.02 major allele frequency (MAF) and 0.33 ± 0.01 PIC (Table 2). The 49 loci amplified by the AFLP primers ACG/CTA and ACA/CAG were excluded from the analysis due to low numbers of amplified loci (< 50) and further genetic analysis was performed with the 1908 loci amplified by the 17 AFLP markers (Table 2). The Bayesian model-based analysis performed with Hickory resulted into the lowest deviance
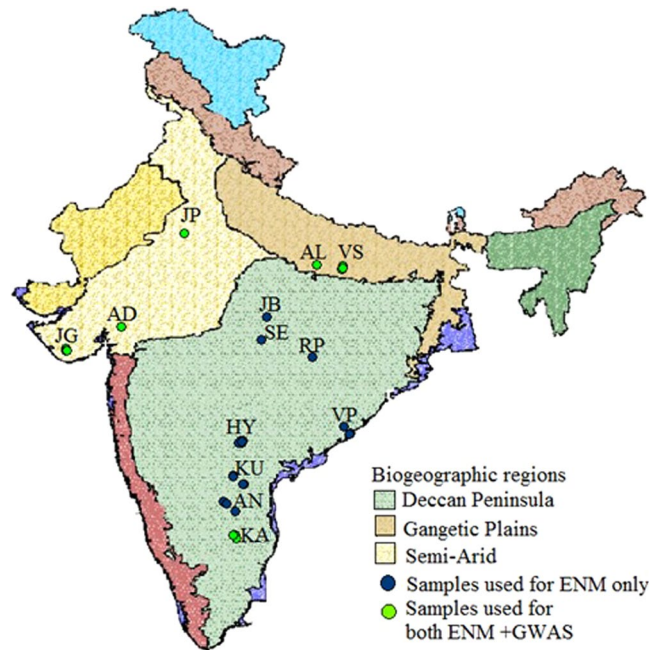
Figure 1. Showing *S. emarginatus* populations at different biogeographic regions sampled for the present investigation. Abbreviated location's code is same as given in Table 4. The map was obtained from the database available in the public domain (http://wiienvis.nic.in/database/htmlpages/biozonemap.htm) and developed in the form of shapefile for modelling after georeferencing through program Q-GIS Desktop (version 3.10; http://qgis.osgeo.org).
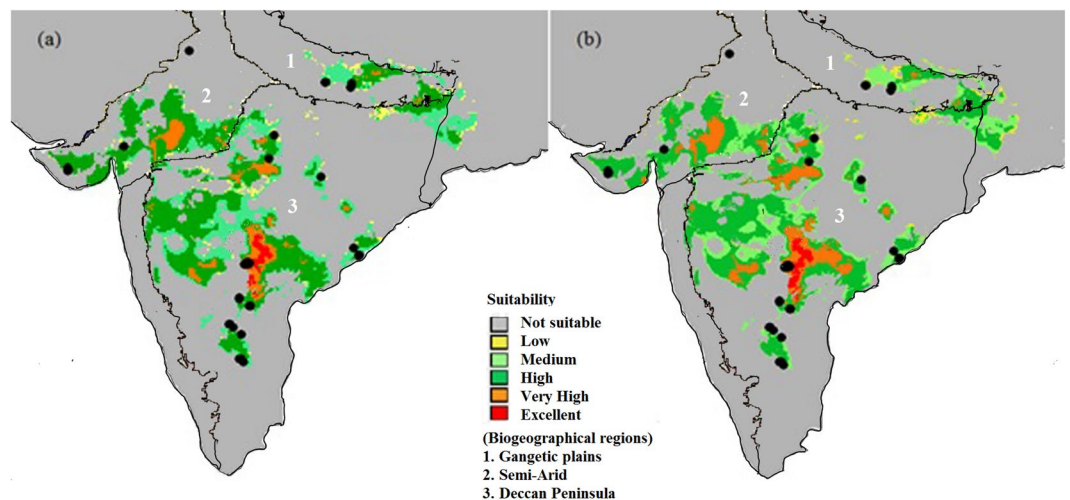

Figure 2. Predicted suitability of ecological niches for the occurrence of *S. emarginatus* in the years (a) 2030 and (b) 2100 in the biogeographical regions 1. Gangetic plain, 2. Semi-Arid, 3. Deccan Peninsula, resulted by MaxEnt algorithm. Black dots represent the sampled locations. The map was obtained from the shapefile of Fig. 1, and the projected output from program MaxEnt[52] is shown with the help of program DIVA-GIS[47].

information criteria (DIC) value (35357.20) supporting the suitability of full-model (inbreeding) among the four models for the data. The average heterozygosity (Hs value) generated for the locations was 0.40 ± 0.0. For the samples as a whole, panmictic heterozygosity (Ht) was 0.43, whereas θ-I and f (F_{IS}) values were 0.18 and 0.98, respectively. The observed heterozygosity (Ho) ranged from 0.42 ± 0.0 to 0.44 ± 0.0 with an average of 0.43 ± 0.01 (Table 3) while F_{ST} value was 0.048. Allahabad (AL) population revealed the highest number of common alleles $<=50\%$, and no rare and common ($<=25\%$) allele was found in the populations analyzed.

The Jaccard's genetic similarity coefficient among the genotypes ranged from 0.24 to 0.61 with an average of 0.38 ± 0.05. The principal coordinate analysis (PCoA) differentiated the genotypes into three clusters (Fig. 4). Few genotypes from AL, KA, and VS populations (Table 4) were found clustering in distinct clusters but most of them are highly admixed with the genotypes of other locations. Analyzing the results in program STRUCTURE,
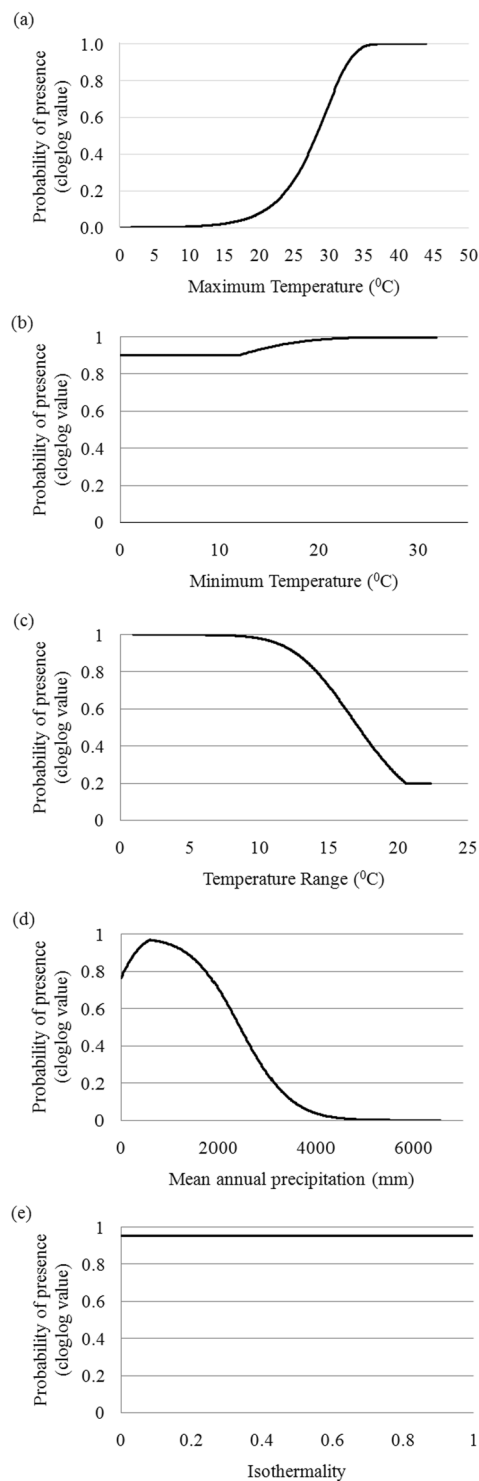
Figure 3. Responses of the climatic variables *viz.* (**a**) maximum temperature (°C), (**b**) minimum temperature (°C), (c) temperature range (°C), (**d**) isothermality, and (**e**) mean annual precipitation (mm) toward the probability of presence of the populations of *S. emarginatus* resulted by MaxEnt[52].

based on the highest Delta-K value, the admixture model with independent allelic frequencies was found the most appropriate for our dataset through the program STRUCTURE HARVESTER. The most suitable cryptic population number was four ($K = 4$, Supplementary Fig. S1). The bar plot showed admixture in populations, and KA population distinctly (Supplementary Fig. S2). The populations VS, GJ, AL, and RJ were admixed, whereas GJ population was found in admixing with KA population. Populations like AL and VS were also found admixed distinctly. The F_{ST} values on these clusters ranged from 0.02 to 0.08. The pair-wise F_{ST} values among

Approach/ Models	MAP	I	T_{RANGE}	T_{MAX}	T_{MIN}
ENM/ MaxEnt	27.10	11.30	2.20	43.10	16.40
GWAS/ MLM	7.55	20.75	39.62	7.55	24.53

Table 1. Estimates of relative contribution (in %) of the climatic variables to the MaxEnt model and of AFLP loci in significant association with the climatic variables. ENM/MaxEnt- ecological niche modeling based on maximum entropy, GWAS/MLM – genome-wide association study based on mixed linear modeling, MAP- mean annual precipitation in mm, I- isothermality, T_{RANGE}- diurnal temperature range in °C, T_{MAX}- maximum temperature °C, T_{MIN}- minimum temperature °C.

Primer Combinations EcoRI/MseI	AL	%P	PIC
AGC/CAT	99	94.95	0.32 ± 0.05
AGC/CTA	112	99.11	0.33 ± 0.03
AGC/CTG	110	100.00	0.34 ± 0.03
ACA/CAT	157	92.99	0.32 ± 0.06
ACA/CAG	27	96.30	0.33 ± 0.04
ACA/CTC	110	100.00	0.33 ± 0.03
AAC/CTG	129	100.00	0.32 ± 0.07
AAG/CAG	126	91.27	0.31 ± 0.06
ACT/CTG	110	95.45	0.33 ± 0.06
ACT/CAT	98	100.00	0.35 ± 0.03
ACT/CAA	135	100.00	0.34 ± 0.02
ACT/CAG	111	85.59	0.33 ± 0.05
ACT/CTT	156	100.00	0.34 ± 0.02
ACC/CAC	136	89.71	0.33 ± 0.06
ACC/CAT	130	90.77	0.33 ± 0.05
AGG/CAT	84	88.10	0.32 ± 0.06
AGG/CTA	53	100.00	0.34 ± 0.02
AGG/CTG	52	76.92	0.33 ± 0.04
ACG/CTA	22	100.00	0.34 ± 0.02
Average	103 ± 39.37	94.79 ± 6.43	0.33 ± 0.01

Table 2. Genetic information revealed by the 19 AFLP primer combinations on 41 genotypes of *S. emarginatus*. AL- the number of amplified loci, %P- the percentage of polymorphism, PIC- polymorphic information content, ±- standard deviation.

the locations ranged between 0.015 to 0.128 and the Nm exchanged among the locations ranged between 3.39 (AL-KA) to 32.39 (GJ-VS, Supplementary Fig. S2). The Mantel's test revealed significant correlation ($p < 0.01$) between genetic distances and geographical distances matrix (Supplementary Fig. S3) and non-significant correlation ($p > 0.05$) was found between the number of migrants (Nm) exchanged among locations to geographic distance matrix (Supplementary Fig. S4) and altitudinal gradient ($r = 0.59$). The variation among locations was $6.46 \pm 1.34\%$ and within location among genotypes was $91.63 \pm 0.31\%$.

In order to detect the outlier loci with a signature of adaptation, the only locus E-AGC/M-CTG-71 ($F_{ST} = 0.23$) was found with the posterior odd ($\log_{10}PO$) > 0.5 (Fig. 5). For the rest of the loci, their corresponding F_{ST} values were found real positive. Linkage disequilibrium (LD) was confirmed to avoid false-discovery in genome-wide association study (GWAS). Among all loci pairs, only 1.99% were found in significant LD ($p < 0.01$). Major LD decay was observed within a distance of 100 base pair (bp) and the length of LD block extended up to 1450 bp (Supplementary Fig. S5). The mixed linear model (MLM) based analysis detected a significant ($p < 0.001$) association of 65 loci (3.59% out of 1810 polymorphic loci) from 15 AFLP markers with bioclimatic variables and altitude (Supplementary Table S1). Among these 65 loci, the highest proportion (32.30%) was found associated with T_{RANGE} and the lowest proportion (6.15%) was with MAP and T_{MAX} (Table 1). Among 1651 possible combinations of these 65 loci, only 4.54% were found in significant ($p < 0.001$) LD.

Discussion

The occurrence of *S. emarginatus* was found at the ecological niches exhibiting varied altitudinal ranges (CV = 74.40%), and MAP distribution (CV = 22.51%) but with a narrow temperature gradient (CV = 3.90%; Table 4). Sun *et al.* studied the natural distribution of another species of the same genus (*S. mukorossi*) which showed comparatively wider range of MAP (CV = 26.84%) and mean annual temperature (CV = 17.84%)[27]. The comparison of the CV and MAP data revealed that *S. emarginatus* might have been facing intrinsic incompetency

Regions	Location	N	CA	He	Hs	Ho
Gangetic Plain	AL	15	8	0.35 ± 0.01	0.41 ± 0	0.43 ± 0
	VS	7	2	0.30 ± 0.06	0.40 ± 0	0.43 ± 0
Semi-arid	RJ	5	0	0.28 ± 0.02	0.40 ± 0	0.44 ± 0
	GJ	9	3	0.33 ± 0.02	0.40 ± 0	0.44 ± 0
Deccan peninsula	KA	5	3	0.26 ± 0.02	0.40 ± 0	0.42 ± 0
Average (±SD)				0.30 ± 0.04	0.40 ± 0	0.43 ± 0.01

Table 3. Genetic diversity of *S. emarginatus* based on AFLP markers. N- the number of genotypes investigated, ±- standard deviation, CA- common alleles (<=50%), He- expected heterozygosity, Hs- panmictic heterozygosity resulted from Hickory, Ho- observed heterozygosity resulted from AFLP-Surv.

Biogeographic Regions	States	Locations	N	Alt (m)	T (°C)	MAP (mm)
Gangetic Plain	Uttar Pradesh	Allahabad (AL)	15[a,b]	88 ± 0	27.80 ± 7.35	1090 ± 0
		Varanasi (VS)	7[a,b]	78.14 ± 1.34	28.14 ± 0.05	1095 ± 21.47
Semi-arid	Rajasthan	Jaipur (JP)	5[a,b]	416 ± 0	27.10 ± 0	668 ± 0
	Gujarat	Aanand (AD)	1[a,b]	105 ± 30.41	27.70 ± 0.53	1331.67 ± 74.10
		Junagarh (JG)	8[a,b]	106.25 ± 27.50	27.60 ± 0.20	1339.25 ± 55.50
Deccan peninsula	Karnataka	Kolar (KA)	5[a,b]	875 ± 34.66	25.32 ± 0.21	849.80 ± 27.13
	Telangana	Hyderabad (HY)	17[b]	547.20 ± 24.14	27.80 ± 0.2	1117 ± 33.78
	Andhra Pradesh	Vizag (VP)	10[b]	108.33 ± 91.90	28.73 ± 0.55	1189.67 ± 135.50
		Kurnool (KU)	7[b]	271 ± 41.61	29.50 ± 0.26	919 ± 30.47
		Anantpur (AN)	7[b]	424 ± 110	28.46 ± 0.68	723 ± 56.92
	Madhya Pradesh	Jabalpur (JB)	1[b]	361 ± 0	26.80 ± 0	1301 ± 0
		Seoni (SE)	1[b]	608 ± 0	26.50 ± 0	1288 ± 0
	Chhattisgarh	Raipur (RP)	4[b]	281 ± 0	28.5 ± 0	1452 ± 0
		CV%		74.40	3.90	22.51

Table 4. Geo-climatic variables at locations of *S. emarginatus* populations in different biogeographical regions of India. N- the number of genotypes, Alt (m)- altitude in meters, T (°C)- the temperature in °C, MAP (mm)- mean annual precipitation in millimeter, [a]-used for DNA profiling, [b]-used for species distribution modeling, ± - standard deviation, CV- coefficient of variation.

to shift beyond the specified range of temperature gradient due to bionomic characteristics. Since the species distribution thrives on the fragmented populations in varied biogeographical regions, it may be significantly influenced by the extrinsic barriers like local conditions and selection pressures as well. Hence the adaptive fitness of the species led by the population genetic characteristics was determined to ensure the sustainability and genomic resilience of the population through dominant AFLP markers. The dominant markers have already been applied to confirm the adaptive fitness of trees[28,29] and other plant species[30–32]. In the present study also, a high MAF (0.65 ± 0.02) of the AFLP markers and a high LD decay (1.99%) in loci-combinations supported their suitability for high-resolution GWAS[33] confirming their frequent coverage of the species genome.

Influence of future climatic regimes on the species population distribution. In both; BioClim and MaxEnt models, no change in AUC on projected climatic regimes confirmed a stable distribution pattern of the species population. However, our projection has confirmed the increased suitability of the niches of the species in Deccan peninsula and none of the locations of species occurrence is going to be vulnerable by the future changes, except one (JP) from the Semi-arid biogeographic region (Fig. 1), which has been found to be comparatively less suitable. This might be possible because the northern regions of India have been predicted to be more influenced by the warmer climate than other regions[34]. It can be inferred also from the results of ENM establishing the T_{MAX} as one of the limiting factors to the species distribution with the highest (43.10%) relative contribution to the model prediction (Table 1).

The adaptive fitness of the species population. In the present investigation, a higher value (0.40–0.43) for the measures of heterozygosity (Table 3) and higher within-population variability (91.63 ± 0.31%) have been resulted. This could be possible that the species prefers xenogamy over geitonogamy due to asynchrony in flowering[35]. Moreover, the sweet incense of its flowers attracts the pollinators, like Indian honeybees[36] that may help to admix the population through pollination between the trees scattered away from each other[35] maintaining within population variability. The Bayesian algorithm-based analysis of genetic structure has found the admixture model with independent allele frequency as the fittest and assumed four cryptic populations for the samples. Nevertheless, a weak or negligible genetic structure (Delta K < 300, Fig. S1) confirmed high admixture of the *S. emarginatus* populations exchanging genes to limited distance. The same can be inferred from the Mantel's test
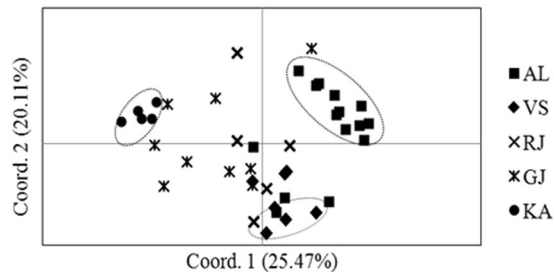
Figure 4. The PCoA differentiating 41 genotypes into three clusters. Few genotypes from AL, KA, and VS were found representing distinct clusters but most of them are highly admixed with the genotypes of other locations.
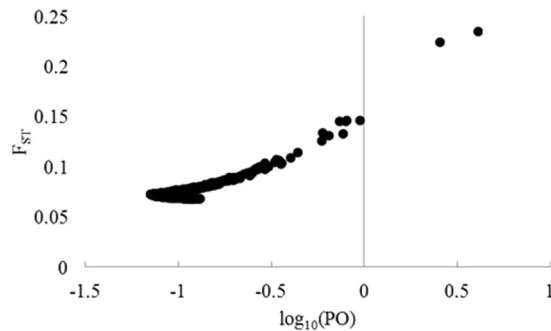

Figure 5. BayeScan analysis resulting F_{ST} values on Bayes Factor, i.e. Log_{10} (PO) for 1859 AFLP loci, confirms only one locus namely E-AGC/M-CTG-71 as an outlier (Log_{10} (PO) > 0.5) depicting substantial evidence to be in balancing selection.

resulting no significant effect of geographical distance on a number of migrants (Nm) exchanged among the sampled genotypes (Fig. S4) leading to higher within-population variability.

The high heterozygosity of *S. emarginatus* populations in light of high admixing within-population strongly supports its adaptive fitness to maintain reproductive success. Only one locus from the AFLP markers has been found with signature of substantial selection, which confirms no possibilities of local adaptation by the populations. Nevertheless, the Bayesian model-based program HICKORY suggested the 'full-model' as the fittest for the data with a high value for the inbreeding coefficient ($f = 0.98$) supporting the sign of inbreeding depression on the species population. High heterozygosity with inbreeding depression in sub-divided populations represents the state of overdominance caused by heterozygote superiority[37]. The current status of the *S. emarginatus* populations in severe fragmentation along with the absence of natural regeneration through seeds, act as barrier for inter-population or long-distance out-crossing and seems to be the reason for inbreeding[38]. Inbreeding depression, causing loss of genetic variability in small or fragmented populations, may result in the reduced fitness of the offspring[39,40]. Despite inbreeding, the higher heterozygosity of the populations indicated that the population might not have crossed many generations with the inbreeding depression[41]. Therefore, being a long-lived forest tree, the reproductive success of the highly heterozygous population of *S. emarginatus* through mass multiplication leading to restoration may mask the deleterious impact of inbreeding[42].

Limiting factors of the species population distribution. A multi-model climate change projection for the Intergovernmental Panel on Climate Change (IPCC, Assessment Report-V) has predicted a 4% increase in warming and 14% increase in precipitation by the year 2080[43]. This increment will have some influence on the climatic envelope of fundamental ecological niches leading their shifting to suitable conditions. Temperature and rainfall have been referred to as prime contributing factors in most of the projections for future climatic regimes[44]. In our investigation also, we found that T_{MAX} and MAP with maximum contribution to the model prediction. The ENM establishes T_{MAX} and MAP as extrinsic limiting factors of the species population distribution. Since the population distribution of the species has been observed within a narrow range of temperature gradient, it may have an influence of the predicted warm climate[45]. The GWAS approach found the most of the genomic proportion covered by the AFLP markers significantly associated with T_{RANGE} and T_{MIN}. It revealed the genomic resilience of the species to survive within varied T_{RANGE} and T_{MIN}. It is obvious, as the species is known to be frost hardy with occurrence in tropical deciduous forests in India[46]. The least proportion of the AFLP markers was found associated with T_{MAX} and MAP establishing them as limiting factors for the genomic resilience of the species. In MLM, 3.45% of the species genome was found in significant association with the bioclimatic variables. The least proportion (6.15%) of the loci linked with T_{MAX} and MAP has confirmed the severe response of the species genome toward these variables. Nevertheless, these associations are highly dependent on high LD to transfer to future generations. We observed only 4.54% LD among the loci-combinations found in significant association

with the bioclimatic variables with high LD decay (within 100 bp). Since *S. emarginatus* maintains high heterozygosity preferring xenogamy, the current status of susceptibility of the species genome responding to T_{MAX} and MAP is subject to change during recombination. In addition, the mating system preferred by *S. emarginatus* may help to avoid the deleterious effect of inbreeding on the species population. Therefore, it can be assumed that the *S. emarginatus* populations may have substantial genomic resilience toward the bioclimatic variables.

Conclusions

To the best of our knowledge, the present study seems to be a maiden attempt to study the adaptive fitness of *S. emarginatus* populations using molecular marker technique with SDM/ENM approach. The ENM-based predictions have supported a stable distribution pattern of the species population towards the projected future climatic regime and confirmed maximum temperature and annual precipitation as the major limiting factors. However, the AFLP markers have detected the signature of adaptation on the genome of *S. emarginatus*, which supports that the species is resilient enough to survive on the threat of rising temperature and precipitation. The Deccan peninsular biogeographic region was found more suitable for the species, therefore the niches in the region can be preferred to conserve as strict nature reserves to maintain the existing genetic variability of the species populations. In light of the threat and vulnerability of ecological niches of Semi-arid region, sincere efforts can be made for screening of resilient genotypes of the species. Further, initiatives for the conservation and mass multiplication of the species are recommended to avoid the future deleterious impact of inbreeding depression on the fragmented populations. There is a scarcity of the genomic information related to *S. emarginatus*. Therefore, the 15 markers found significantly associated with the bioclimatic variables and altitudinal gradients, which can be further employed to find out their association with the functional adaptive traits.

Methods

Population occurrence data and sampling. The information on occurrence of the species was obtained from the global plant database like GBIF (https://www.gbif.org/occurrence/search?taxon_key=8086757), European herbaria; KEW (https://www.kew.org/search?textsearch=sapindus) and the herbarium records available at CSIR- National Botanical Research Institute, Lucknow (LWG), Central National Herbarium, Kolkata (CAL), Botanical Survey of India (Central Circle), Allahabad (BSA), Botanical Survey of India, Hyderabad (BSID) and Forest Research Institute, Dehradun (DD). The 157 GPS coordinates obtained were subjected to spatial filtration to avoid the sampling biases and equal richness based grid sampling applied through the program DIVA-GIS v7.5[47]. Remaining GPS coordinates were selected for ground-truthing to obtain the real occurrence data of *S. emarginatus* populations. The locations were visited and finally 88 coordinates were recorded at the locations with verified abundance of the species in eight Indian states representing three biogeographical regions *viz*. Gangetic plain, Semi-arid and Deccan peninsula (Table 4, Fig. 1) for ENM. For DNA profiling leaf samples were collected from 41 genotypes of five localities (Table 4, Fig. 1) representing the above biogeographical regions.

Bioclimatic data. The data of 35 bioclimatic variables[48] for baseline (year 2000–2030) and projected (the year 2070–2100) regimes was obtained from the website http://www.climond.org (Climond model CSIRO- MK 3.0) and extracted for the GPS coordinates of the locations using DIVA-GIS v7.5[47] (available at http://www.diva-gis.org/) in the resolution of 2.5 m.

Prediction of the distribution pattern. The GPS coordinates were employed for ENM to predict the potential distribution of populations determining suitable niches for the species. For confirmation of the result both, BioClim and MaxEnt based models were applied as the assembling and comparing the results from different models can give robust projection[49,50]. The BioClim algorithm was applied using the envelope method as implemented in program DIVA-GIS v7.5[47] for baseline and predicted bioclimatic variables. MaxEnt model was applied through program MaxEnt v3.4.1[50–52] to evaluate the response of niches to the bioclimatic variables for baseline and predicted climatic regimes. For training with threshold-depended sensitivity model (threshold = 0.5), default parameters in program setting were preferred with 50 replications and 500 iterations doing jackknife and to measure variable importance. The results obtained for two different models for two climatic regimes were correlated and the proportion of overlapping niches was determined through program ENMtools v1.8[53] applying relative rank statistics[54]. Based on the significant correlation among the bioclimatic variables obtained through program ENMtools v1.8, finally climatic variables *viz*. minimum temperature (T_{MIN}), maximum temperature (T_{MAX}), temperature range (T_{RANGE}), isothermality (I), and mean annual precipitation (MAP) were tested to find out the limiting factor for the species distribution.

DNA isolation and AFLP amplification. Genomic DNA was isolated from the collected leaf samples following a modified CTAB method[22]. The conventional AFLP protocol[19] was used with minor modifications[55] to develop an amplification profile for the 41 genotypes on 19 AFLP primer combinations (Table 2).

Assessment of genetic diversity. DNA profiling of 41 genotypes with 19 combinations of AFLP primers was developed based on scoring the presence (1) and absence (0) of the bands with help of program Genemapper® v4.0 (Applied Biosystems, Foster City, CA, USA). The program eliminates the band profiles with short sizer peaks (<50) to avoid the error on scoring. Further, the AFLP markers with <50 amplified loci were excluded for further analysis. The profile was analyzed applying both band-based and allele frequency-based approach[26]. The mean allele frequency (MAF), polymorphism percentage (P %) and polymorphic information content (PIC) of the markers were calculated by program POWERMARKER v3.25[56]. The genetic diversity of the sampled populations was estimated using Bayesian models. We preferred the θ-statistics (*viz*. θ-I, θ-II, θ-III and G_{ST}-B) implemented in program HICKORY v1.1[24] that allows direct estimates of genetic differentiation measure (F_{ST}) from dominant markers without prior assumption of inbreeding and Hardy -Weinberg Equilibrium (HWE) within population

even with small sample size. Location-wise panmictic heterozygosity (Hs) and total panmictic heterozygosity (Ht) were also estimated. All these estimations were performed with 50,000 steps of burn-in, 500,000 replicates, 'thinning' = 20 following the best-suited model for the data amongst all the four models ('full', '$f = 0$', '$\theta = 0$' and 'f-free' models) implemented in the program based on their corresponding DIC (deviance information criterion) value[25]. The inbreeding coefficient ($f = F_{IS}$) resulted from the best-fitted model in HICKORY was incorporated in another Bayesian program AFLP-Surv v1.0[57] with the non-uniform prior distribution of allele frequencies at 10000 permutations to estimate the observed heterozygosity (Ho). Rare (private) and common alleles ($<25\%$) among the populations were calculated through program GenAlEx v6.5[58]. Jaccard's genetic similarity coefficient among the genotypes was calculated by program DARWIN v5.0[59]. A principal coordinate analysis (PCoA) was performed to cluster the genotypes in different axes applying program GenAlEx v6.5. To determine the genetic structure of the sampled population with an appropriate number of cryptic population (K) assigned by the genotypes, Bayesian statistics-based program STRUCTURE v2.3.1[60] was used. The project was run applying four prior-assumption with combinations of admixture/no-admixture models with correlated/independent allele frequencies on 100000 burn-in and 1000000 MCMC repeats with three-run of K from 1 to 9 for $2 \le K \le 8$. The most suitable model and the best-suited 'K' for the data set were determined based on the highest Delta-K value resulted from an online program STRUCTURE HARVESTER[61]. The ancestry coefficient (Q) values of the genotypes were further used for mixed linear modeling (MLM) based association mapping. Pair-wise F_{ST} among populations were also calculated by program AFLP-Surv. The gene flow among populations was determined by calculating the number of migrants (Nm) among the populations[62]. To determine the effect of geographical distance and altitude on gene flow among locations, the Mantel's tests were performed among pair-wise Nm, pair-wise geographical distance and pair-wise Euclidean distance based on altitude through program GenAlEx. Analysis of molecular variance (AMOVA) was also calculated.

Detection of signature of local adaptation. To detect the signature of local adaptation on AFLP loci, program BAYESCAN v2.1[63] was applied that identifies outlier loci with a Bayesian test based on a logistic regression model that decomposes F_{ST} values into a locus-specific component (α, selection effect) and a population-specific component (β, demographic effect). Twenty pilot runs of 5,000 iterations were used to estimate the distribution of α-parameters, followed by 50,000 iterations for sampling. Outlier loci with substantial signature of adaptation were identified using a 5% false discovery rate with posterior odds of >10 (log10 (PO) > 0.5). Linkage disequilibrium among loci-combinations was performed by program TASSEL v3.0[64] to confirm the influence of demographic variables on the population genetic structure inferring LD decay. A genome-wide association study (GWAS) was performed between climatic variables (i.e. T_{MIN}, T_{MAX}, T_{RANGE}, I, and MAP), altitude and AFLP marker profile by program TASSEL v3.0 applying MLM to control the false-discovery due to the structured population[65,66]. The MLM considers the markers applied to the study as a fixed–effect factor and the population structure information of the sampled genotypes are considered as random effect factors. The required kinship (K) information was obtained from program TASSEL and the ancestry coefficient (Q) value was obtained from the program STRUCTURE output to implement the MLM + K + Q method. The loci significantly ($p \le 0.01$) associated with one or more climatic variables were taken into consideration. To find out the limiting factor, the proportions of AFLP markers linked with these climatic variables and altitude were calculated based on their number of loci in significant association with the climatic variables.

Data availability
The datasets generated and analyzed during the present study will be available from the corresponding author on reasonable request.

References
1. Champion, H. G. & Seth, S. K. *A Revised Survey of Forest Types of India* (Government of India Press New Delhi, 1968).
2. Anonymous. Global saponin market by manufacturers, regions, type and application, forecast to 2023, https://www.planetmarketreports.com/reports/global-saponin-market-5797. (2019).
3. Anonymous. Saponin export data of India, https://www.exportgenius.in/export-data/india/saponin.php. (2019).
4. Anonymous. State of Forest Genetic Resources in India: A Country Report, http://www.fao.org/3/i3825e/i3825e32.pdf (Institute of Forest Genetics and Tree Breeding Coimbatotre, Indian Council of Forestry Research and Education, 2012).
5. Murali, K. S., Shankar, U., Shaanker, R. U., Ganeshaiah, K. N. & Bawa, K. S. Extraction of non-timber forest products of Biligiri Rangan Hills, India. 2. Impact of NFTP extraction on regeneration, population structure, and species composition. *Econ. Bot.* **50**, 252–269 (1996).
6. Murthy, K., Bhat, P. R., Ravindranath, N. H. & Sukumar, R. Financial valuation of non-timber forest product flows in Uttara Kannada district, Western Ghats, Karnataka. *Curr. Sci.* **88**, 1573–1579 (2005).
7. Pereira, I. M. & Groppo, M. Ecological Niche Modeling: Using Satellite Imagery and New Field Data to Support Ecological Theory and its Applicability in the Brazilian Cerrado. *J. Ecosyst. Ecography*. **2**, https://doi.org/10.4172/2157-7625.1000111 (2012).
8. Nagaraju, S. K. *et al.* Do ecological niche model predictions reflect the adaptive landscape of species?: a test using Myristica malabarica Lam., an endemic tree in the Western Ghats, India. *PLoS One* **8**, https://doi.org/10.1371/journal.pone.0082066 (2013).
9. Guisan, A. *et al.* Predicting species distributions for conservation decisions. *Ecol. Lett.* **16**, 1424–35 (2013).
10. Hughes, L., Cawsey, E. M. & Westoby, M. Climatic Range Sizes of Eucalyptus Species in Relation to Future Climate Change. *Glob. Ecol. Biogeogr.* **5**(1), 23–29 (1996).
11. Booth, T. H. Assessing species climatic requirements beyond the realized niche: some lessons mainly from tree species distribution modelling. *Clim. Change.* **145**, 259–271 (2017).
12. Kremer, A., Potts, B. M., Delzon, S. & Bailey, J. Genetic divergence in forest trees: understanding the consequences of climate change. *Funct. Ecol.* **28**, 22–36 (2014).

13. Fischer, M., Hock, M. & Paschke, M. Low genetic variation reduces cross-compatibility and offspring fitness in populations of a narrow endemic plant with a self-incompatibility system. *Conserv. Genet.* **4**, 325–336 (2003).

14. Szulkin, M., Bierne, N. & David, P. Heterozygosity-fitness correlations: a time for reappraisal. *Evolution* **64**, 1202–1217 (2010).

15. Reed, D. H. & Frankham, R. Correlation between fitness and genetic diversity. *Conserv. Biol.* **17**, 230–237 (2003).

16. Spielman, D., Brook, B. W. & Frankham, R. Most species are not driven to extinction before genetic factors impact them. *Proc. Natl. Acad. Sci. USA* **101**(42), 15261–15264 (2004).

17. West, S. A. & Gardner, A. Adaptation and Inclusive Fitness. *Curr. Biol.* **23**, 577–584 (2013).

18. Thode, V. A. *et al.* Genetic diversity and ecological niche modelling of the restricted Recordia reitzii (Verbenaceae) from southern Brazilian Atlantic forest. *Bot. J. Linn. Soc.* **176**, 332–348 (2014).

19. Vos, P. *et al.* AFLP- a new technique for DNA fingerprinting. *Nucleic. Acids. Res.* **23**, 4407–4414 (1995).

20. Lerceteau, E. & Szmidt, A. E. Properties of AFLP markers in inheritance and genetic diversity studies of *Pinus sylvestris* L. *Heredity* **82**, 252–260 (1999).

21. Meudt, H. M. & Clarke, A. C. Almost forgotten or latest practice? AFLP applications, analyses, and advances. *Trends. Plant. Sci.* **12**(3), 106–117 (2007).

22. Mahar, K. S., Rana, T. S., Ranade, S. A. & Meena, B. Genetic variability and population structure in *Sapindus emarginatus* Vahl from India. *Gene* **485**, 32–39 (2011).

23. Holsinger, K. E., Lewis, P. O. & Dey, D. K. A Bayesian approach to inferring population structure from dominant markers. *Mol. Ecol.* **11**, 1157–1164 (2002).

24. Holsinger, K. E. & Lewis, P. O. Hickory: A package for analysis of population genetic data, version 1.1. Computer program and documentation, Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, Connecticut, USA (2007)

25. Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & Vanderlinde, A. Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Series B Stat. Methodol.* **64**, 583–689 (2002).

26. Bonin, A., Ehrich, D. & Manel, S. Statistical analysis of amplified fragment length polymorphism data: a toolbox for molecular ecologists and economists. *Mol. Ecol.* **16**, 3737–3758 (2007).

27. Sun, C. *et al.* Genetic structure and biogeographic divergence among Sapindus species: An inter-simple sequence repeat-based study of germplasms in China. *Ind. Crops. Prod.* **118**, 1–10 (2018).

28. Wang, T., Wang, Z., Xia, F. & Su, Y. Local adaptation to temperature and precipitation in naturally fragmented populations of *Cephalotaxus oliveri*, an endangered conifer endemic to China. *Sci. Rep.* **6**, 25–31 (2016).

29. Vaishnav, V. & Ansari, S. A. Genetic differentiation and adaptation in teak (*Tectona grandis* L.f.) metapopulations. *Plant Mol. Biol. Report.* https://doi.org/10.1007/s11105-018-1101-3 (2018).

30. Wang, T., Chen, G., Zan, Q., Wang, C. & Su, Y. J. AFLP genome scan to detect genetic structure and candidate loci under selection for local adaptation of the invasive weed *Mikania micrantha*. *PLoS One.* **7**, e41310 (2012).

31. Krauss, S. L., Sinclair, E. A., Bussell, J. D. & Hobbs, R. J. An ecological genetic delineation of local seed-source provenance for ecological restoration. *Ecol. Evol.* **321**, 38–49 (2013).

32. Leempoel, K., Parisod, C., Geiser, C. & Joost, S. Multiscale landscape genomic models to detect signatures of selection in the alpine plant *Biscutella laevigata*. *Ecol. Evol.* **8**, 1794–1806 (2018).

33. Kim, S. Y. *et al.* Estimation of allele frequency and association mapping using next-generation sequencing data. *BMC Bioinformatics* **12**, 231 (2011).

34. Revadekar, J. V., Kothawale, D. R., Patwardhan, S. K., Pant, G. B. & Rupa, K. K. About the observed and future changes in temperature extremes over India. *Natural Hazards* **60**, 1133–1155 (2012).

35. Reddi, C. S., Reddi, E. U. B., Reddi, N. S. & Reddi, P. S. Reproductive ecology of Sapindus emarginatus Vahl (Sapindaceae). *Proc. Ind. Nat. Sci. Acad.* **49B**(1), 57–72 (1983).

36. Rao, A. S. Preliminary studies on the seasonal occurrence of insect pests a soap nut (Sapindus sp). *Ind. Forester.* **118**(6), 432–437 (1992).

37. Charlesworth, D. & Charlesworth, B. Inbreeding depression and its evolutionary consequences. *Annu. Rev. Ecol. Syst.* **18**, 237–268 (1987).

38. Severns, P. M., Liston, A. & Wilson, M. V. Habitat fragmentation, genetic diversity, and inbreeding depression in a threatened grassland legume: is genetic rescue necessary? *Conserv. Genet.* **12**, 881–893 (2011).

39. Saccheri, I. *et al.* Inbreeding and extinction in a butterfly metapopulation. *Nature* **392**, 491–494 (1998).

40. Walisch, T. J., Matthies, D., Hermant, S. & Colling, G. Genetic structure of Saxifraga rosacea subsp. sponhemica, a rare endemic rock plant of Central Europe. *Plant. Syst. Evol.* **301**, 251–263 (2015).

41. Charlesworth, B. & Charlesworth, D. The genetic basis of inbreeding depression. *Genet. Res.* **74**, 329–340 (1999).

42. Taylor, H. R. *et al.* Cryptic inbreeding depression in a growing population of a long-lived species. *Mol. Ecol.* **26**, 799–813 (2017).

43. Chaturvedi, R. K., Joshi, J., Jayaraman, M., Bala, G. & Ravindranath, N. H. Multi-model climate change projections for India under representative concentration pathways. *Curr. Sci.* **103**(7), 791–802 (2012).

44. Song, X. & Zeng, X. Evaluating the responses of forest ecosystems to climate change and CO_2 using dynamic global vegetation models. *Ecol. Evol.* **7**, 997–1008 (2017).

45. Gao, Y., Gao, X. & Zhang, X. The 2 °C Global Temperature Target and the Evolution of the Long-Term Goal of Addressing Climate Change-From the United Nations Framework Convention on Climate Change to the Paris Agreement. *Engineering* **3**, 272–278 (2017).

46. Troup, R. S. The Silviculture of Indian Trees: Volume I. 239–240 (Oxford University Press 1921).

47. Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G. & Jarvis, A. Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* **25**, 1965–1978 (2005).

48. Kriticos, D. J. *et al.* CliMond: global high-resolution historical and future scenario climate surfaces for bioclimatic modelling. *Methods Ecol. Evol.* **3**(1), 53–64 (2012).

49. Shabani, F., Kumar, L. & Ahmadi, M. A comparison of absolute performance of different correlative and mechanistic species distribution models in an independent area. *Ecol. Evol.* **6**(16), 5973–5986 (2016).

50. Phillips, S. J., Anderson, R. P. & Schapire, R. E. Maximum entropy modeling of species geographic distributions. *Ecol. Modell.* **190**(3–4), 231–259 (2006).

51. Elith, J. *et al.* A statistical explanation of MaxEnt for ecologists. *Divers. Distrib.* **17**, 43–57 (2011).

52. Phillips, S. J., Dudik, M. & Schapire, R. E. [Internet] Maxent software for modeling species niche and distribution (Version 3.4.1). Available from url: http://biodiversityinformatics.amnh.org/open_source/maxent/.

53. Warren, D. L., Glor, R. E. & Turelli, M. Environmental niche equivalency versus conservatism: quantitative approaches to niche evolution. *Evol.* **62**, 2868–2883 (2008).

54. Warren, D. L. & Seifert, N. Ecological niche modeling in MaxEnt: the importance of model complexity and the performance of model selection criteria. *Ecol. Application.* **21**(2), 335–342 (2011).

55. Tripathi, P. K., Jena, S. N., Rana, T. S. & Sathyanarayana, N. High levels of gene flow constraints population structure in *Mucuna pruriens* L. (DC.) of northeast India. *Plant. Gene.* **15**, 6–14 (2018).

56. Liu, K. & Muse, S. V. Powermarker: Integrated analysis environment for genetic marker data. *Bioinformatics.* **21**(9), 2128–2129 (2005).

57. Zhivotovsky, L. A. Estimating population structure in diploids with multilocus dominant DNA markers. *Mol. Ecol.* **8**, 907–913 (1999).
58. Peakall, R. & Smouse, P. E. GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research-an update. *Bioinformatics.* **28**, 2537–2539 (2012).
59. Perrier, X. & Jacquemoud-Collet, J. P. DARwin software: Dissimilarity Analysis and Representation for Windows http://darwin.cirad.fr/darwin (2006).
60. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genet.* **155**, 945–959 (2000).
61. Evanno, G., Regnaut, S. & Goudet, J. Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Mol. Ecol.* **14**, 2611–2621 (2005).
62. McDermott, J. M. & McDonald, B. A. Gene flow in plant pathosystems. *Annu. Rev. Phytopathol.* **31**, 353–373 (1993).
63. Foll, M. & Gaggiotti, O. E. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective. *Genet.* **180**(2), 977–993 (2008).
64. Bradbury, P. J. *et al*. TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics.* **23**(19), 2633–2635 (2007).
65. Pritchard, J. K., Stephens, M., Rosenberg, N. A. & Donnelly, P. Association mapping in structured populations. *Am. J. Hum. Genet.* **67**, 170–181 (2000).
66. Yu, J. *et al*. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **38**, 203–208 (2006).

Acknowledgements

Author contributions

A.K.P., N.P. and T.S.R. conceived the research hypothesis and experiment design; A.K.P., B.M., sampled the species from different localities and recorded the data. A.K.P. performed the laboratory works; V.V. & A.K.P. analyzed the data and drafted the manuscript; all authors contributed equally in the final version after critical reviews.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-60219-8.

Correspondence and requests for materials should be addressed to T.S.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.