



Resolving population structure and genetic differentiation associated with RAD-SNP loci under selection in tossa jute (*Corchorus olitorius* L.)

Debabrata Sarkar¹ · Avijit Kundu^{1,2} · Debajeet Das¹ · Avrajit Chakraborty¹ · Nur Alam Mandal¹ · Pratik Satya¹ · Pran Gobinda Karmakar¹ · Chandan Sourav Kar¹ · Jiban Mitra¹ · Nagendra Kumar Singh³

Received: 2 June 2018 / Accepted: 19 December 2018 / Published online: 2 January 2019
© Springer-Verlag GmbH Germany, part of Springer Nature 2019

Abstract

The genetic basis of selection for geographic adaptation and how it has contributed to population structure are unknown in tossa jute (*Corchorus olitorius*), an important bast fibre crop. We performed restriction site-associated DNA (RAD) sequencing-based (1115 RAD-SNPs) population genomic analyses to investigate genetic differentiation and population structure within a collection of 221 fibre-type lines from across nine geographic regions of the world. Indian populations, with relatively higher overall diversity, were significantly differentiated (based on F_{ST} and PCA) from the African and the other Asian populations. There is strong evidence that African *C. olitorius* was first introduced in peninsular India that could perhaps be its secondary centre of origin. However, multiple later introductions have occurred in central, eastern and northern India. Based on four assignment tests with different statistical bases, we infer that two ancestral subpopulations (African and Indian) structure the *C. olitorius* populations, but not in accordance with their geographic origins and patterns of diversity. Our results advocate recent migration of *C. olitorius* through introduction and germplasm exchange across geographical boundaries. We argue that high intraspecific genetic admixture could be associated with increased genetic variance within Indian populations. Employing both subpopulation (F_{ST}/G_{ST} -outlier) and individual-based (PCAdapt) tests, we detected putative RAD-SNP loci under selection and demonstrated that bast fibre production was an artificial, while abiotic and biotic stresses were natural selection pressures in *C. olitorius* adaptation. By reinferring the population structure without outlier loci, we propose ad interim that *C. olitorius* was possibly domesticated as a fibre crop in the Indian subcontinent.

Keywords F_{ST} outlier · Genotyping-by-sequencing · Population genomics · Population structure · RADseq · Single-nucleotide polymorphism

Communicated by Dacheng Tian.

The raw Illumina RADseq reads have been deposited in the NCBI Sequence Read Archive (SRA) under the project SRP064554 vide BioProject PRJNA207496 and BioSamples SAMN03097738 to SAMN03097962, with 225 SRX accessions listed in Online Resource 2. RADseq genotype data and other summary statistics are available as a Figshare (<https://figshare.com/s/05b13e169a8f5ae8e634>) entry doi: <https://doi.org/10.6084/m9.figshare.6339518>.

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s00438-018-1526-2>) contains supplementary material, which is available to authorized users.

✉ Debabrata Sarkar
debabrata_s@yahoo.com; debabrata.sarkar@icar.gov.in

Extended author information available on the last page of the article

Introduction

Corchorus olitorius L. ($2n = 2x = 14$, Malvaceae s. l.), an important bast fibre crop commonly known as tossa or dark jute, originated in Africa (Kundu 1951; Benor et al. 2012; Kundu et al. 2013) and was supposed to have dispersed to Asia via the Mediterranean-Indian trade routes (Benor et al. 2012). Interestingly, however, it was domesticated in Africa as a leafy vegetable crop (Edmonds 1990; Benor et al. 2010). It might have been domesticated in India as a fibre crop through the development of an ennobled type from its African wild type (Kundu et al. 2013). However, not much difference in genetic diversity between the cultivated and wild accessions underpins a rather recent domestication history of this species (Benor et al. 2012). Most of the diversity studies have reported a lack of correspondence between genetic divergence and geographic diversity in *C. olitorius*

(Basu et al. 2004; Roy et al. 2006), suggesting uncertain origins of accessions used or their possible recent migration through introduction across geographical boundaries (Roy et al. 2006).

Ancestry analyses delineated *C. olitorius* populations into varying numbers of genetic groups (Banerjee et al. 2012; Benor et al. 2012; Satya et al. 2014a; Zhang et al. 2015). These results are, however, mostly inconclusive because of uneven geographic representation by a rather small number of accessions sampled, inclusion of artificial mutants, limited number of DNA-based markers used and the use of dominant markers like AFLPs (amplified fragment length polymorphisms) that are known to result in ambiguous genotypes and thus discard potentially valuable information (Foll et al. 2010). In one of the most informative AFLP-based population structure analyses (Benor et al. 2012), only 11 accessions represented the Asian population, without a single one from India that has a long history of tossa jute cultivation (Kundu 1951). Whereas Zhang et al. (2015) did sample only two and three accessions from India and Africa, respectively, to infer the genetic structure. Thus, the correct assignment of *C. olitorius* accessions representing the main gene pool of the species to inferred populations and estimation of the proportion of the genome of an accession that belongs to each inferred one are fundamental to elucidating their ancestries and geographic origins in relation with domestication and/or breeding history.

A reduction in genetic diversity in cultivated tossa jute is associated with domestication bottlenecks (Kundu et al. 2013) that are known to cause selective sweeps in genomic areas containing genes of agronomic importance (Doebley et al. 2006). For most of the *C. olitorius* landraces are yet to be extensively characterized, a lack of knowledge about their biological history and value restricts them to be utilized in breeding programs. It was not until recently that an African origin of the cultivated landraces of *C. olitorius* grown in eastern India has been traced by analyzing the organelle genetic diversity (Basu et al. 2016). Parallel with studies on geographic patterns of genetic variation in *C. olitorius*, understanding the genetic basis of selection for geographic adaptation in relation with population structure is important. Without known phenotypes, identifying loci that have undergone balancing or directional selection would allow us not only to understand how populations have adapted to specific environments and agronomic practices, but to detect novel alleles and haplotypes to improve productivity, adaptation, and quality (Narum and Hess 2011).

Restriction site-associated DNA sequencing (RADseq), a reduced-representation sequencing method, is now the most widely used low-cost genomic approach for discovering single-nucleotide polymorphisms (SNPs) and genotyping in population genomic studies (Andrews et al. 2016). It is similar to genotyping-by-sequencing (GBS), but requires

more demanding sample handling and size-selection of DNA fragments. In addition to improving the precision of demographic inferences, RADseq enables the identification of specific loci and/or genomic regions that have been subject to selection and adaptive evolution (Narum et al. 2013). In an increasing number of plant species, RADseq has been employed to discover RAD-SNP loci that are found useful for population genomic studies, such as geographic population structure and gene flow (Pegadaraju et al. 2013; Xu et al. 2014; Hou et al. 2015; Pan et al. 2016; Valdisser et al. 2016), which often require only several hundred to a few thousand loci to adequately sample the genome (Andrews et al. 2016).

Here, we used individual-based RADseq to discover SNPs across a diverse set of 225 *C. olitorius* accessions and identified a set of 1115 polymorphic RAD-SNP loci, each supporting a single SNP with > 0.05 minor allele frequency (MAF). We determined how genetic diversity was distributed across nine geographically delineated *C. olitorius* populations that represent the main gene pool and how they were structured, with an overreaching objective to trace precisely the ancestry and geographic origin of each accession. We show that the 1115 RAD-SNP panel is effective in clustering the accessions into two ancestral subpopulations (African and Indian) with high genetic resolution but not necessarily in accordance with their geographic origins and patterns of diversity. We applied subpopulation-based outlier tests and an individual-based global approach to detect putative RAD-SNP loci under selection. Finally, we demonstrate that bast fibre production was an artificial, whereas abiotic and biotic stresses were natural selection pressures in *C. olitorius* adaptation. Our results resolve a longstanding caveat underlying the genetic structure of *C. olitorius* and provide an empirical framework for subsequent ecological, evolutionary and population genomic studies of this species.

Materials and methods

Plant materials, sampling, and DNA extraction

A *C. olitorius* association mapping panel comprising 225 accessions was used in this study. They included fibre-type cultivars, improved lines, landraces, and varieties from 15 countries (Mahapatra et al. 2006), and were grouped into nine geographic populations based on their sampled locations (Online Resource 1): AFR1 (Kenya and Sudan), AFR2 (Tanzania), Central India (CI), East India (EI), North India (NI), South India (SI), Nepal and Pakistan (NPPK), East and South-East Asia (ESEA; China, Myanmar, Indonesia and Thailand), and rest of the world (RoW; Australia, Brazil, Germany and Russia). Total genomic DNA was extracted from 10-day-old seedling leaves and purified as described earlier (Kundu et al. 2015).

Restriction site-associated DNA (RAD) sequencing

We used partially methylation sensitive endonuclease *ApeKI* to construct three 75-plex RAD libraries following an optimized protocol as described earlier, with similar adapters, barcodes, and primers (Kundu et al. 2015). Each 75-plex library was sequenced to 100 bp on a single lane of Illumina HiSeq™ 2000 (Illumina, San Diego, CA) based on Illumina's TrueSeq® Version 3.0 single-end sequencing chemistry, and raw sequencing reads were deposited in the NCBI Sequence Read Archive (SRA) under the project SRP064554 vide BioProject PRJNA207496 (Online Resource 2). Non-reference-based universal network-enabled analysis kit (UNEAK; Glaubitz et al. 2014) as implemented in TASSEL 3 (Bradbury et al. 2007) was used to analyze the RADseq data, according to Lu et al. (2013) and Huang et al. (2014). Raw Illumina data were filtered for barcodes and 4-nt *ApeKI* overhang (CWGC), ensuing that there were no adapter–adapter dimers and unknown bases (Ns) across the first 72 bp. Reads were assigned to individual samples according to their barcode sequences and trimmed to 64 bp including the initial 4-nt *ApeKI* restriction-site sequence. A set of unique 64-bp tags was generated by retaining reads with a minimum Q score of 10 and collapsing identical reads into one tag followed by pairwise alignment of tags differing in 1 bp and identification of reciprocal tag pairs to call SNPs. The error tolerance was set at 0.03 to eliminate error tags that were assumed to have resulted from sequencing error (Lu et al. 2013). The tags were collapsed into a tags-by-taxa matrix of presence/absence and genotypes were called for all of the samples.

Population genomic analyses

We used GenoDive v2.27 (Meirmans and Van Tienderen 2004) to estimate the mean number of alleles (A), effective number of alleles (A_e), observed heterozygosity (H_o) and gene diversity (H_s); PowerMarker v3.51 (Liu and Muse 2005) to calculate the average major allele frequency (P) and polymorphism information content (PIC); ADZE v1.0 (Szpiech et al. 2008) to measure rarefied allelic richness (A_r) and private allelic richness (A_p); and PopGenome v2.1.0 (Pfeifer et al. 2014) to estimate MAF and the Ts/Tv ratio. The pairwise estimates of gene flow (N_m) at the population level were calculated as $N_m = (1 - F_{ST})/4F_{ST}$ (Freeland et al. 2011). Statistical significance was estimated by Kruskal–Wallis one-way ANOVA, and associations between all parameters including sample size (n) were tested by Spearman's rank-order correlations using the R Base v3.4.1 (R Core Team 2017).

Population structure analyses

Pairwise F_{ST} values were calculated between all pairs of nine populations based on 80,000 randomization tests followed by strict Bonferroni correction for multiple comparisons (Meirmans and Hedrick 2011). The average pairwise F_{ST} values were used to perform neighbor-joining clustering of populations using Past v3.14 (Hammer et al. 2001). We also inferred the population relationships based on the D_A distance (Nei et al. 1983) using POPTREEW (Takezaki et al. 2014). We tested a higher level of population structure using an AMOVA based on infinite allele model with 10,000 permutations, performed a PCA with an associated permutation test among populations based on 10,000 replicates and assessed the overall population structure using k -means clustering of individuals, with 1,000,000 steps and 100 random starts for several values of k that ranged from 1 to 15 (Meirmans 2012).

Next, we used STRUCTURE v2.3.4 (Pritchard et al. 2000) to infer the population structure based on both admixture and non-admixture ancestries and the correlated allele frequencies, with the burn-in and MCMC (Markov chain Monte Carlo) iterations of 20,000 each and 10 replicates for each of several values of K that ranged from 1 to 15. We also performed STRUCTURE analysis separately on each population using the admixture model, with K that ranged from 1 to 5. The results were collated by STRUCTURE HARVESTER v0.6.97 (Earl and vonHoldt 2012), and the optimal K for each analysis was identified using delta K (Evanno et al. 2005) and delta F_{ST} (Campana et al. 2011) followed by the estimation of K solution's overall stability and the significance of correlation between individual pairs of replicates based on Q matrix correlations using CorrSieve v1.6-8 (Campana et al. 2011). We also used sNMF v1.2 (Frichot et al. 2014) to infer the population structure, without assuming predefined populations, with default run parameters and 10 independent replicates for each of several values of K (1–15). The optimal alignments of clusters were optimized and graphed using CLUMPP v1.1.2 (Jakobsson and Rosenberg 2007) and DISTRUCT v1.1 (Rosenberg 2004), respectively. Accessions were assigned to corresponding Q groups based on their maximum membership probabilities (q) or ancestry coefficients (Remington et al. 2001) and categorized as admixed employing an arbitrary cut-off value of $\leq 80\%$ probability (Vigouroux et al. 2008). The Q -group membership of each accession was evaluated further by performing frequency-based assignment test, with 10,000 permutations for the MCMC test (Paetkau et al. 1995). Finally, we applied discriminant analysis of principal components (DAPC) to analyze the population structure using Adegenet v2.0.0 (Jombart 2008). Accessions were assigned to clusters based on posterior membership probabilities, and the

contribution of individual RAD-SNP alleles to population structure was estimated using the function *loadingplot*.

Estimations of relative kinship (F_{ij}) and linkage disequilibrium (LD)

We estimated pairwise F_{ij} using SPAGeDi v1.5 (Hardy and Vekemans 2002), with an associated permutation test based on 10,000 replicates. All negative F_{ij} estimates between accessions were set to zero (Yu et al. 2006). We used GENEPOP v4.3 (Rousset 2008) to test departures from Hardy–Weinberg equilibrium (HWE) across all loci and samples for heterozygote deficiency, with default MCMC parameters and sequential Bonferroni adjustments over all loci ($P < 0.00004$). We used TASSEL to estimate the LD as r^2 between each RAD-SNP pair, with a corresponding P value < 0.001 based on two-sided Fisher's exact test. With a set of 112 loci with known chromosomal locations, intra-chromosomal LD was calculated, and the decay of r^2 as a function of physical distance was estimated by nonlinear least squares (Marroni et al. 2011) using the *nls* function of the R package Stats v3.4.1.

Detection of RAD-SNP loci under selection

To detect RAD-SNP loci under selection between the two ancestral subpopulations (African and Indian) resolved in this study, we used two population-based F_{ST} -outlier methods as implemented in LOSITAN v1.44 (Antao et al. 2008) and BayeScan v2.1 (Foll and Gaggiotti 2008). For LOSITAN, simulations were run for 100,000 iterations using an infinite allele model at 99% confidence interval and 5% FDR, with forced and neutral mean F_{ST} . Loci that deviated from an expected distribution of neutral expectations outside the 99 and 1% confidence areas were detected as outliers affected by positive and balancing selection, respectively. For BayeScan, we used the default model parameters to run simulations for 100,000 iterations and identified outliers based on the F_{ST} distribution (FDR < 0.05). The probability that a given locus is under selection was estimated by calculating the posterior odds (PO), and inferences were drawn using the Bayes factor (BF) in accordance with Jeffreys' (1961) scale of evidence. We performed ten independent runs for each method and considered only those loci that were consistently identified as outliers across all runs. We further performed a G_{ST} -based EOS (extreme-outlier set) test as implemented in HacDivSel v1.2 (Carvajal-Rodríguez 2017). For each extreme positive outlier (EPO), the LK (Lewontin and Krakauer 1973) test was performed to calculate its P value, which was adjusted for multiple testing for the number of outliers by strict Bonferroni correction. Finally, we used PCAdapt v3.0.2 (Luu et al. 2017) to detect RAD-SNP loci involved in biological adaptation.

The function *pcadapt*, with default parameters, was applied with $K = 20$, and the presence of outliers was confirmed by plotting the histograms of P values and the Mahalanobis (D^2) test statistic, with FDR controlled at < 0.05 using the R package *qvalue* v2.5.2.

Annotations of RAD-SNP loci

We mapped our 1115 RAD-SNP loci to the draft genome of *C. olitorius* cv. JRO-524 (Sarkar et al. 2017a) and its seven chromosome-scale pseudomolecules using BWA v0.7.13 (Li and Durbin 2009) followed by BLASTn v2.7.1+ search (E value $< 10^{-10}$) of the RAD tags against the predicted genes (Sarkar et al. 2017b). They were also mapped to the draft genome of *C. olitorius* cv. O-4 (Islam et al. 2017) and searched against the genes predicted from it. We obtained the gene ontology (GO) annotations of coding loci using Blast2GO Pro v4.0.7 (Conesa et al. 2005) and retrieved their functional classifications using WEGO (Ye et al. 2006). Protein domains associated with candidate genes were annotated against the HMMs profile database Pfam v31 using HMMER v3.1b2 as implemented in DoMosaics (Moore et al. 2014). We used REViGO (Supek et al. 2011) to summarize the GO terms based on their P value-guided semantic similarity measures and retrieved a representative subset of GO terms associated with candidate genes.

Results

Discovery of RAD-SNP loci in *C. olitorius*

RADseq generated 568,217,847 clean reads with 57.98 Gbp of nucleotide sequences from 225 genotypes (Online Resource 2) across three RAD libraries (Online Resource 3). Since we required each locus to be present in all nine geographic populations and in at least 95% of individuals of each one, a set of 1115 informative loci was finally retained across 221 accessions, with a call rate of > 0.95 , MAF of > 0.05 and a Ts/Tv ratio of 1.62 (Online Resource 4a). In total, 1110 loci (99.6%) were mapped to our *C. olitorius* cv. JRO-524 draft genome (Sarkar et al. 2017a; Online Resource 4b), of which 798 (72%) were present in 563 genes (Online Resource 4c). By comparison, 1062 loci (95.2%) were mapped to the *C. olitorius* cv. O-4 draft genome (Islam et al. 2017), of which 766 (69%) were present in 552 genes (BWA and BLASTn mapping data available on Figshare: <https://figshare.com/s/05b13e169a8f5ae8e634>); however, 49% of them were associated with hypothetical proteins based on O-4 genome annotations. In contrast, our manually curated JRO-524 genome annotations (Sarkar et al. 2017b) showed only 3.6% of the loci being associated with hypothetical proteins (Online Resource 4c). The Blast2GO mapping of 563

genes supporting 798 loci resulted in 2292 GO-term annotations (Online Resource 4c) summarized into three main GO categories and 42 sub-categories (Fig. 1). However, only 112 loci were mapped to the seven *C. olitorius* pseudomolecules (Online Resource 4d). This was expected, because though 99.6% of the RAD-SNPs from the *C. olitorius* genetic map (Kundu et al. 2015) were mapped to the JRO-524 draft genome, seven chromosome-scale pseudomolecules covered only a fraction of the genome (Sarkar et al. 2017a).

Genetic diversity within *C. olitorius* geographic populations

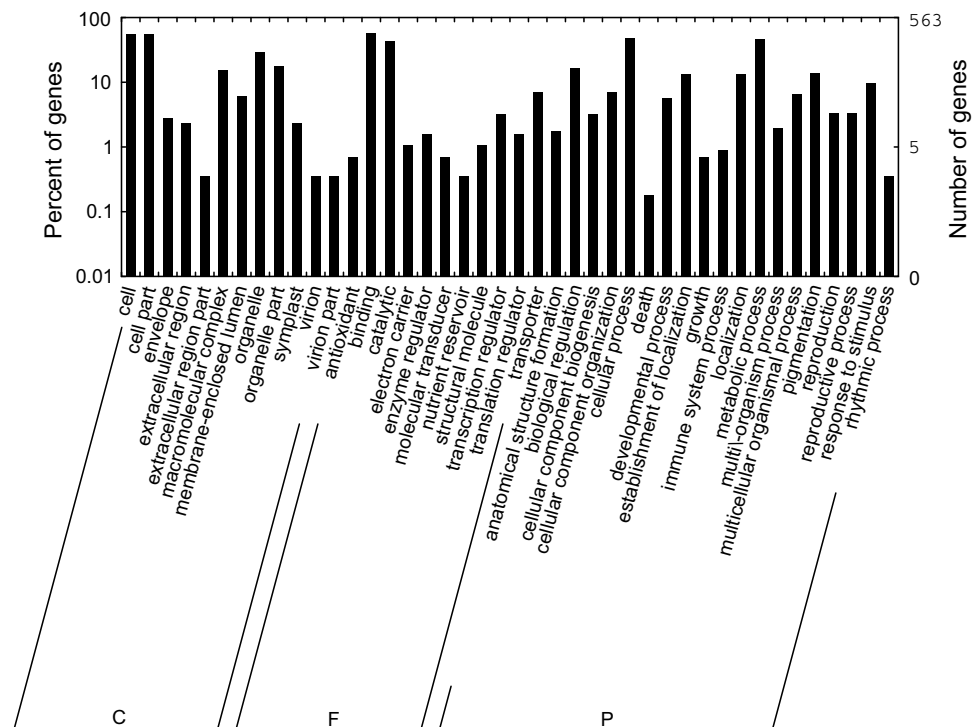
At the population level, there were significant ($P < 0.0001$) differences for all genetic diversity parameters (Online Resource 5a). The four Indian populations showed minimum P but maximum MAF and PIC values. The distribution of P was markedly variable across populations, with a higher number of loci fixed ($P = 1.0$) within each population, except CI, EI and NI (Online Resource 6). The SI and ESEA populations had P values distributed more or less uniformly with those of AFR1 and AFR2, respectively. However, the NPPK and RoW populations had a skewed distribution of P toward 1.0. The H_s estimates (0.192–0.310), with an average of 0.258, were maximum for Indian and minimum for the RoW and NPPK populations. The four Indian populations were characterized by the highest A , A_e and A_r . Both A_r and A_p , when measured across geographic regions, were higher for Asia (A_r $1.75 \pm 0.01/A_p$ 0.220 ± 0.009) than for Africa (A_r

$1.59 \pm 0.01/A_p$ 0.088 ± 0.005). There were strong negative correlations between P and all diversity measures, except A and H_o (Online Resource 5b). However, A_r and A_p had strong positive correlations with PIC and H_s . The population size (n) had a strong positive correlation only with A .

Genetic relatedness of *C. olitorius* geographic populations

The four Indian populations that clustered together (Fig. 2) were significantly ($P < 0.05$) differentiated (F_{ST}) from both the African populations (Table 1). The EI and SI populations showed minimum and maximum genetic divergence from African populations, respectively. The other two Asian populations were, however, genetically related to both the African populations, but significantly differentiated from all Indian populations. The majority of the genetic variation (85.4%) was partitioned across individuals within populations (F_{IS}), whereas only 5.7 and 6.7% of the variations were due to differences among populations nested within (F_{SC}) and across (F_{CT}) regions, respectively (Table 2). Consequently, the overall amount of gene flow among populations was rather high ($N_m = 2.83$). The highest magnitude of N_m was observed between AFR2 and ESEA or RoW (18.78) followed by those between AFR1 and ESEA (8.08), AFR2 and NPPK (6.51), and AFR1 and RoW (5.70) or NPPK (3.17). By comparison, the average estimates of N_m were much lower in the four Indian populations vs. AFR1 (1.37) or AFR2 (1.71), with the minimum and maximum amounts

Fig. 1 Gene ontology (GO) classification of functional RAD-SNP loci of *C. olitorius* obtained with Blast2GO and WEGO. C cellular component, F molecular function, P biological process



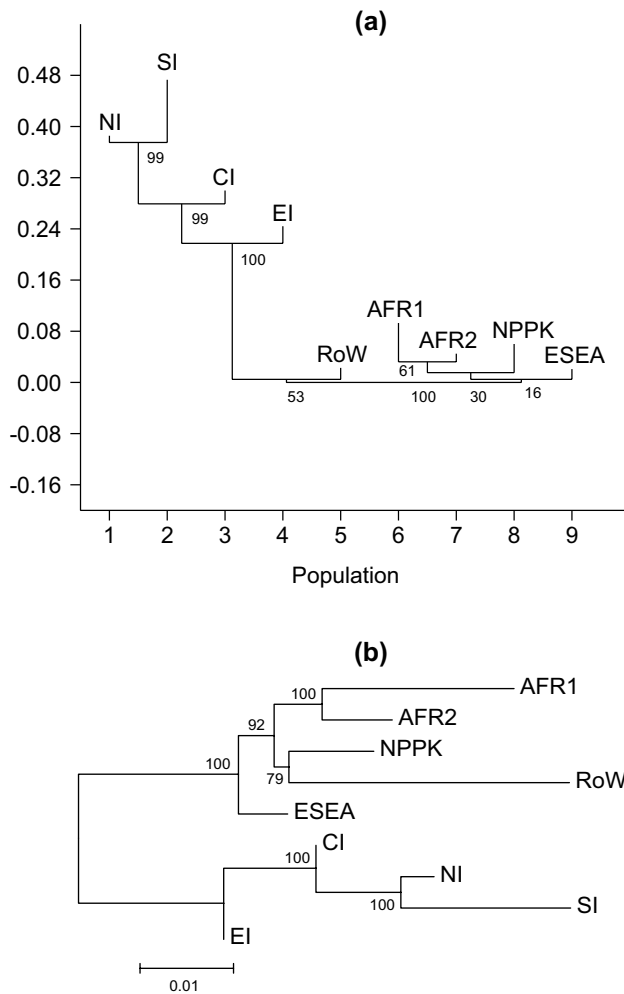


Fig. 2 Neighbor-joining (NJ) tree of clustering of nine geographic *C. olitorius* populations constructed using the average pairwise F_{ST} values (a) and the D_A distance (b), with 1000 bootstrap replicates. For population acronyms, see Table 1

Table 1 Pairwise comparison of genetic distance (F_{ST}) among nine geographic populations of *C. olitorius* based on 1115 RAD-SNP loci

	AFR1 ^a	AFR2	CI	EI	NI	SI	NPPK	ESEA
AFR2	0.017							
CI	0.145	0.119						
EI	0.105	0.082	0.013					
NI	0.189	0.167	0.012	0.031				
SI	0.235	0.209	0.028	0.036	0.017			
NPPK	0.073	0.037	0.123	0.079	0.174	0.231		
ESEA	0.030	0.013	0.102	0.067	0.152	0.195	0.025	
RoW	0.042	0.013	0.083	0.047	0.145	0.190	0.014	0.011

Boldface F_{ST} values are significantly different from 0 at $P < 0.05$, according to 80,000 randomization tests followed by a strict Bonferroni correction for multiple comparisons

AFR1 Africa 1 (Kenya and Sudan), *AFR2* Africa 2 (Tanzania), *CI* Central India, *EI* East India, *NI* North India, *SI* South India, *NPPK* Nepal and Pakistan, *ESEA* East and South-East Asia (China, Myanmar, Indonesia and Thailand), *RoW* rest of the world (Australia, Brazil, Germany and Russia)

^aThe sole Nigerian accession was removed from the *AFR1* population due to > 50% missing RAD-SNP genotypes

detected in the SI (0.88) and EI (2.47) populations, respectively. This is further reflected in our PCA results, where the first PC that was only significant ($P \leq 0.0001$) explained 62.0% of the total genetic variation (Table 3) and differentiated the four Indian populations from African and the other Asian populations (Fig. 3). The second PC that accounts for 11.5% of genetic variation had also some explanatory power.

Genetic population structure of *C. olitorius*

The k -means identified two clusters that best explain the overall population structure of *C. olitorius* (Online Resource 7). With STRUCTURE, a model with $K=2$ best fitted the complete data for both admixture and non-admixture models (Online Resource 8). However, not a single K solution was stable for the non-admixture model (Online Resource 9), which was further confirmed by analyzing the significance of Q matrix correlations (Online Resource 10). Coding (798) and non-coding (317) RAD-SNPs also inferred two genetic groups (data available on Figshare: <https://figshare.com/s/05b13e169a8f5ae8e634>) under admixture. The plot of posterior probabilities clearly shows these two groupings (Fig. 4), dubbed as Indian (Q_1) and African (Q_2), with mean q of 0.518 and 0.482, respectively (Table 4). We assigned 125 and 96 accessions to Indian and African groups, respectively (Table 4), 96% of which were concordant with that inferred with frequency-based assignment test (Online Resource 1). Indian populations were characterized by rather high levels of admixture (an average of 49.7%), with > 50% of the individuals in CI and EI populations identified as admixed (Table 4). When analyzed separately, each of these nine populations except RoW had two genotypic classes (Online Resource 11).

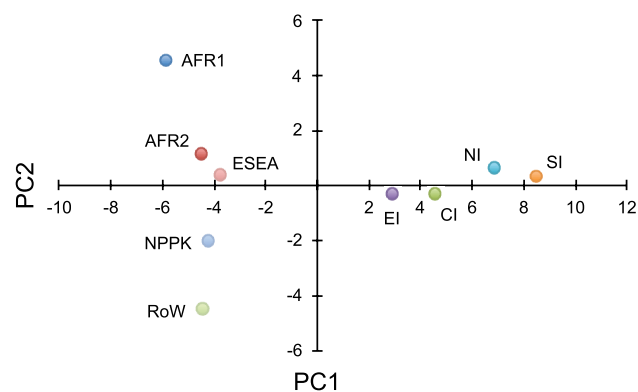
Individual-based sNMF analysis also inferred $K=2$ under admixture and clustered 221 accessions into two ancestral

Table 2 Analysis of molecular variance (AMOVA) partitioning the genetic variation across nine geographic populations of *C. olitorius* over the three regions based on 1115 RAD-SNP loci

Source of variation	Nested in	Variance	F stat	Standard deviation	P value
Within individual	–	0.0227	F_{IT}	0.0005	–
Among individual	Population	0.8537	F_{IS}	0.0005	0.0001
Among population	Region	0.0568	F_{SC}	0.0032	0.0001
Among region	–	0.0668	F_{CT}	0.0041	0.0465

Table 3 Eigenvalues of principal components of genetic variation of 1115 RAD-SNP loci

Axis	Eigenvalue	Variance (%)	Cumulative	P value
1	31.83	61.96	61.96	0.000100
2	5.88	11.45	73.41	1.000000
3	3.66	7.13	80.54	0.999500
4	3.20	6.23	86.76	0.999800
5	2.49	4.85	91.62	0.999700
6	2.15	4.19	95.80	0.999600
7	1.22	2.38	98.18	1.000000
8	0.93	1.82	100.00	1.000000

**Fig. 3** Bivariate plot of average scores along the first two principal axes (PC1 and PC2) of genetic variation for each of nine geographic populations of *C. olitorius*. For population acronyms, see Table 1

groups Indian and African, with mean ancestry coefficients of 0.492 and 0.508, respectively (Online Resource 1). We assigned 118 and 103 accessions to Indian and African groups (Table 4), respectively, with 96.8% concordance with the STRUCTURE groupings. The DAPC also identified two clusters using sequential k -means (Online Resource 12). The correspondence between the geographic populations and k -means-inferred clusters (Online Resource 13) mostly reflected the STRUCTURE groupings (Table 4). For $k=2$, DAPC retained a single discriminant function (Online Resource 14) and assigned 129 and 92 accessions to two clusters, dubbed as $C1$ (Indian) and $C2$ (African), with mean coordinates of -2.333 and 3.272 , respectively (Online Resource 1). We showed that 94.6 and 96.7% of accessions

that were assigned to the STRUCTURE groups $Q1$ and $Q2$, respectively, were correctly assigned to corresponding DAPC clusters (Table 4). There was also high concordance (94.1%) between the sNMF and DAPC groupings. A total of 39 loci were identified as the most contributing to group differentiation (Online Resource 15). Except for CoRAD33, 510, 655 and 946, all loci contributed both alleles. Blast2GO mapping assigned nine of these loci to mostly ‘cell’, ‘cell part’, ‘catalytic’, ‘biological regulation’, ‘cellular process’ and ‘metabolic process’ sub-categories.

Relative kinship (F_{ij}) and LD in the *C. olitorius* diversity panel

The average F_{ij} between two accessions was 0.0582. About 52% of F_{ij} estimates were equal to 0, while 27% ranged from 0 to 0.1 (Fig. 5). The frequency of an F_{ij} estimate of 0.5 was only 0.1%. Significant ($P < 0.00004$) departures from HWE were detected globally across all loci and samples and for both Indian and African subpopulations, suggesting evidence of LD. For 1115 loci (621,055 pairs), we observed that only 23,853 pairs were significantly ($P < 0.001$) in LD. When measured separately, 6484 and 5642 pairs of loci were significantly ($P < 0.001$) in LD for the Indian and African gene pools, respectively. Intrachromosomal LD across 112 chromosomally mapped loci (6216 pairs) was found to extend over a distance of ~ 489 kb (decay_{50}). In all, 409 (6.6%), 271 (4.4%) and 200 (3.2%) pairs of loci were significantly ($P < 0.001$) in intrachromosomal LD for the total, Indian and African gene pools, respectively.

RAD-SNP loci under selection in *C. olitorius*

In the pairwise comparison between African vs. Indian subpopulations, which were significantly ($P < 0.001$) divergent with an average F_{ST} value of 0.147 ($N_m = 1.45$), the majority of loci (48.7%) had F_{ST} values < 0.1 (Fig. 6). About 10% of loci had negative F_{ST} values, implying that they are characterized by more variation within than across subpopulations. Only 2.3% of loci had very high F_{ST} (> 0.5) values. LOSITAN detected the presence of outliers against neutral expectations outside the 99% CI (Online Resource 16). Fourteen outliers (1.3%) were consistently identified through 10 independent iterations (FDR < 0.05 ; Online Resource 17). However, BayeScan through 10 independent iterations

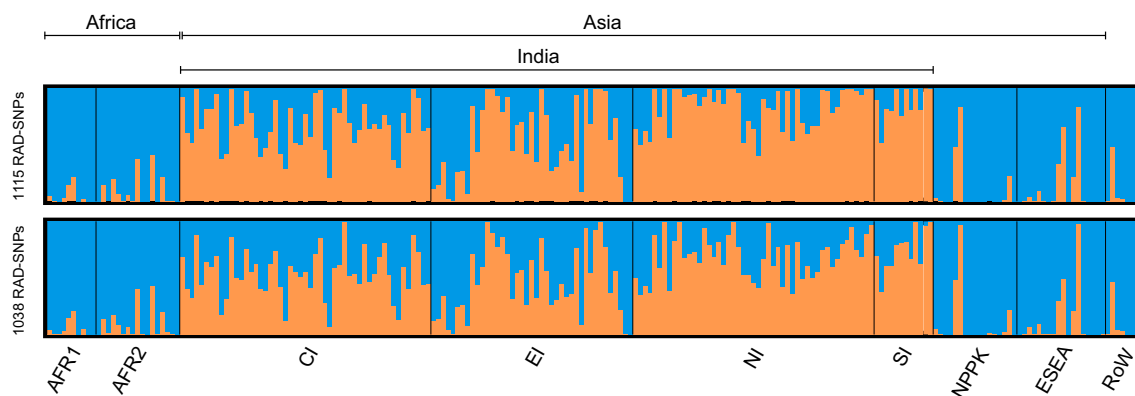


Fig. 4 Plots of posterior probabilities of group membership of each individual into two clusters ($K=2$) across nine geographic populations of *C. olitorius* inferred with STRUCTURE using 1115 RAD-SNP loci, with (top) and without (bottom) the 77 outlier RAD-SNP

loci detected by LOSITAN and PCAdapt. Each vertical bar represents an individual, with its color proportion constituting the posterior probability of assignment to one of two clusters of genetic similarity. For population acronyms, see Table 1

Table 4 Proportions of membership probabilities (q) of two different genotypic classes across nine geographic populations of *C. olitorius* inferred with STRUCTURE (shaded) based on 1115 RAD-SNP loci, with corresponding sNMF group and DAPC cluster assignments

Population	$Q1^a$		$Q2$		sNMF ^a		DAPC ^b		Admixture (%)	
	q	Accessions	q	Accessions	$Q1$	$Q2$	$C1$	$C2$	STRUCTURE	sNMF
AFR1 ^c	0.048	0 (0) ^d	0.952	10 (9)	0 (0) ^d	10 (9)	0	10	10.0 (1) ^d	10.0 (1)
AFR2	0.088	0 (0)	0.912	17 (14)	0 (0)	17 (14)	0	17	17.6 (3)	17.6 (3)
CI	0.679	41 (16)	0.321	10 (0)	38 (9)	13 (0)	45	6	68.6 (35)	82.3 (42)
EI	0.543	21 (12)	0.457	20 (8)	21 (11)	20 (10)	23	18	51.2 (21)	48.8 (20)
NI	0.835	48 (32)	0.165	1 (0)	45 (24)	4 (0)	47	2	34.7 (17)	51.0 (25)
SI	0.865	12 (9)	0.135	0 (0)	11 (6)	1 (0)	11	1	25.0 (3)	50.0 (6)
NPPK	0.095	1 (1)	0.905	16 (14)	1 (1)	16 (15)	1	16	11.8 (2)	5.9 (1)
ESEA	0.122	2 (1)	0.878	16 (14)	2 (1)	16 (14)	2	16	16.7 (3)	16.7 (3)
RoW	0.090	0 (0)	0.910	6 (5)	0 (0)	6 (5)	0	6	16.7 (1)	16.7 (1)
Overall	0.518	125 (71)	0.482	96 (64)	118 (52)	103 (67)	129	92	38.9 (86)	46.1 (102)

^aAccessions are assigned to STRUCTURE and sNMF groups (Q) based on maximum posterior probabilities and ancestry coefficients, respectively

^bAccessions are assigned to DAPC clusters based on their membership probabilities. $Q1$ ($C1$) and $Q2$ ($C2$) represent Indian and African subpopulations (clusters), respectively. For details, see Online Resource 1

^cThe population codes are the same as expanded in Table 1

^dValues within parentheses represent the number of accessions when they are assigned to corresponding Q groups if their posterior probabilities or ancestry coefficients are higher than an arbitrary cut-off value of 80%, with others classified as admixture

($FDR < 0.05$) consistently detected only one (0.09%) outlier (CoRAD959) at \log_{10} BF of 1.5–2.0 that corresponds to very strong evidence of selection based on Jeffreys' scale of evidence. HacDivSel detected 19 outliers (1.7%), of which 9 represented the K -means EPOs, while 10 were non-EPOs (Online Resource 17). However, none of these EPOs were significant ($P < 0.05$) when assessed by the LK test followed by strict Bonferroni correction. With $K=2$, PCAdapt identified 63 outliers (5.7%) at $FDR < 0.05$ (Online Resource 18), of which 25 (39.7%) and 38 (60.3%) were assigned to PC1 and PC2, respectively (Online Resource 19). None of these outliers were common to those detected by F_{ST} -outlier tests, except the one identified by BayeScan (Fig. 7); however, 26

of them were also identified by DAPC as the most contributing to group discrimination.

Functional classification of outlier RAD-SNP loci

Of the 77 outliers detected by LOSITAN and PCAdapt, 38 had significant (E value $< 10^{-10}$) BLASTn hits, with 20 of them mapped to the best transcriptome (Chakraborty et al. 2015) of *C. capsularis* cv. JRC-212 (Online Resource 20). They were mostly within candidate genes encoding bHLH, FAR1, GRAS and zinc finger RING-type families of transcription factors including SCARECROW 22, TIP-type aquaporins (AQPs), late embryogenesis abundant D-34

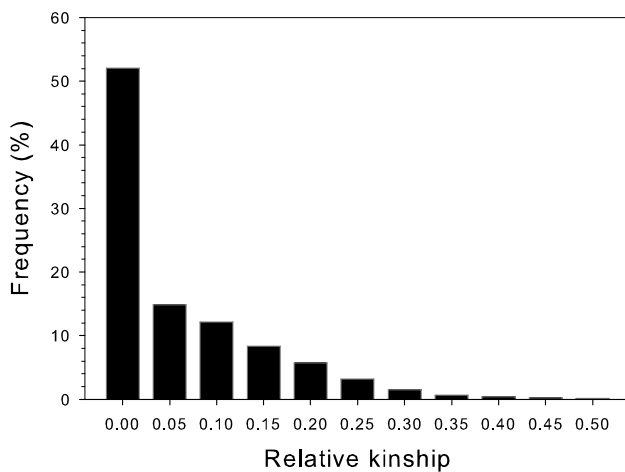


Fig. 5 Distribution of pairwise relative kinship (F_{ij}) estimates between 221 *C. olitorius* accessions across nine geographic populations based on 1115 RAD-SNP loci

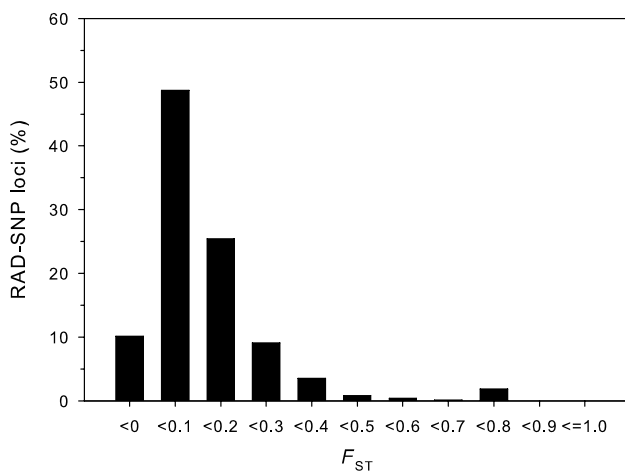


Fig. 6 Distribution of pairwise F_{ST} values between the two STRUCTURE-inferred Indian and African *C. olitorius* subpopulations based on 1115 RAD-SNP loci

(LEA34) and serine carboxypeptidase-like (SCPL) proteins. In all, 42 protein domain families organized in 23 unique arrangements were associated with 24 of these candidate genes (Online Resource 21). Blast2GO mapping assigned 30 of these candidate genes to 88 GO-term annotations across the three main GO categories and 23 sub-categories (Online Resource 22). REViGO semantic analysis identified over-represented subsets of GO functional categories (Online Resource 23).

Inference of population structure without outlier loci

Without 77 outliers, we reinferred two subpopulations with STRUCTURE using ΔK , ΔF_{ST} (Online Resource 24)

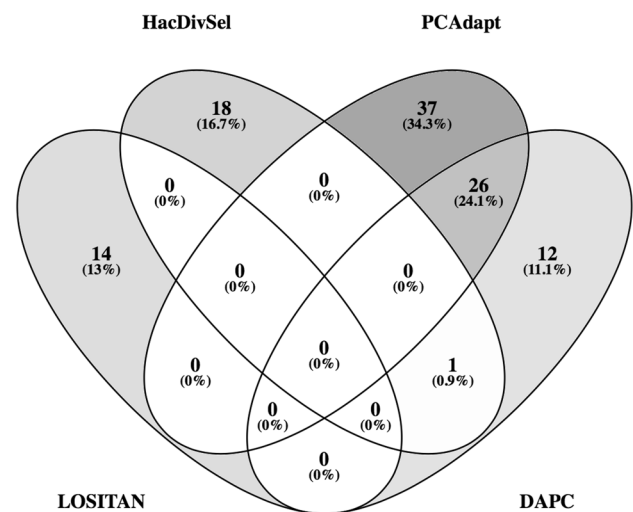


Fig. 7 Outlier RAD-SNP loci detected by population- (HacDivSel and LOSITAN) and individual-based (PCAdapt) approaches, in comparison with those contributing alleles to DAPC-inferred differentiation between African and Indian subpopulations in *C. olitorius*

and the Q matrix correlations (Online Resource 25). With mean q of 0.435 and 0.565, 102 and 119 accessions were assigned to Indian and African subpopulations, respectively (Online Resource 26a; Fig. 4). However, 23 accessions swapped groups from Indian to African in the regrouping retrieved with 1038 loci, of which 22 (95.7%) represent the four Indian populations. The CI accessions (54.5%) were most affected followed by those in NI (31.8%). A 14.5% increase in admixture was due entirely to an increase in the level of admixture (23.0%) in the four Indian populations (Online Resource 26a). There were no changes in the level of admixture within the other geographic populations (Fig. 4). A revised AMOVA with 1038 loci showed that F_{IS} increased by 2.7%, while F_{SC} and F_{CT} decreased by 1.3% and 1.6%, respectively (Online Resource 26b). Without 63 outliers detected by PCAdapt, the STRUCTURE grouping as well as the level of admixture were exactly the same as that retrieved with 1038 loci (STRUCTURE analysis data available on Figshare: <https://figshare.com/s/05b13e169a8f5ae8e634>).

Discussion

We report individual-based RADseq for discovering genome-wide SNPs and their application for inferring the population structure in *C. olitorius* by using a large collection of 225 accessions representing all of the important fibre-type ecotypes (Mahapatra et al. 2006). Most of these accessions have no or very weak kinship, reflecting not only their broad range of collection but also effective exclusion

of duplicates in the study. Despite the recent availability of draft genomes (Islam et al. 2017; Sarkar et al. 2017a), we used the non-reference UNEAK (Glaubitz et al. 2014) to call RAD-SNPs (Lu et al. 2013; Huang et al. 2014). A reference genome representing a single individual is usually incomplete and does not represent the whole genome of a species because of technical difficulties and presence/absence (PAV) variation (Lu et al. 2013). Non-reference RAD-SNP calling not only allowed us to gain access to those missing genomic regions, but ensured a stringent selection of loci with a high call rate across 221 accessions. Differential mapping of these loci to the draft genomes of *C. olitorius* cvs JRO-524 (Sarkar et al. 2017a) and O-4 (Islam et al. 2017) indicates the relative incompleteness of these genomes. This result is not unexpected, because the Bangladeshi cv. O-4 is a pure line selection from a local landrace (Akter et al. 2008), whereas the Indian cv. JRO-524 represents a cross between African and indigenous types (Kar et al. 2010). Non-reference RAD-SNP calling also ensured strict detection of a single SNP per RAD site, because redundant and multiple SNPs are ignored by UNEAK (Huang et al. 2014). In the absence of ascertainment bias (Valdisser et al. 2016), our filtered set of RAD-SNPs afforded more realistic estimates of genetic diversity and genetic differentiation, besides an accurate inference of population structure.

In agreement with earlier reports (Basu et al. 2004; Banerjee et al. 2012; Benor et al. 2012; Satya et al. 2014a), we detected a low level of genetic diversity at species level in *C. olitorius*, with most of the variation distributed within populations. Our estimate of RAD-SNP diversity measured in terms of gene diversity (equivalent to nucleotide diversity for bi-allelic SNPs; Pfeifer et al. 2014) was higher than that reported for AFLP diversity (Benor et al. 2012), but comparable to that obtained for peroxidase gene-based (POG) (Satya et al. 2014a) and microsatellite-based (Banerjee et al. 2012) diversity estimates. The overall diversity detected within Indian populations was relatively higher than that within the other populations. Higher sample size of Indian populations is unlikely to account for this result, because most of the diversity estimates were corrected for sample size and had no significant positive correlations with sample size. That all Indian populations are genetically more diverse than African and the other Asian populations are also reflected in the frequencies of major alleles (P). Skewed distributions of P in the NPPK and RoW populations are indicative of young populations, which might have been formed by a few founder individuals and are yet to reach the evolutionary equilibrium through the acquisition of mutations or migrant alleles (Lewontin and Krakauer 1973). Higher allelic diversity in Indian populations implies that a good portion of genetic diversity was captured in the founder lines of our investigated accessions. Hence, Indian populations have not only enhanced adaptive potential to cope up

with future environmental changes, but increased evolutionary potential (Greenbaum et al. 2014).

The F_{ST} , AMOVA, N_m , and PCA all show that Indian populations are significantly differentiated from African and the other two Asian populations. The SI population is the most differentiated from both the African populations. Human Y-chromosome analysis has indicated an African origin of Dravidian agriculture in southern India (Winters 2010). Taken together, it is thus most likely that *C. olitorius* was first introduced in peninsular India together with other African crops, perhaps unconsciously as a contaminant. This region has recently been identified as one of the most frequent habitats of *C. olitorius* (Choudhary et al. 2017). *C. olitorius* flowers and fruits throughout the year in peninsular India, particularly at a geographical location of 11.02° N and 76.96° E (msl 411 m). Here, we provide the first genetic evidence that peninsular India may be the secondary centre of origin of *C. olitorius*. The highest concentration of Indian *Corchorus* spp. in Nilgiri and Palani hills, a biodiversity hotspot, of peninsular India (Choudhary et al. 2017) argues strongly in support of this hypothesis. A peninsular Indian origin has been proposed for roselle (Satya et al. 2014b) and horsegram (Fuller and Murphy 2018). On the other hand, the EI population is the least differentiated from both the African populations. This corresponds to our recent evidence that the *C. olitorius* landraces from eastern India share the common organelle group with accessions from Sudan (AFR1) and Tanzania (AFR2), indicating their common African origin (Basu et al. 2016). Together with the extent of gene flows recorded in the CI and NI populations, these results point to multiple later introductions of African *C. olitorius* to central, eastern and northern India from different source populations. However, the dispersal of peninsular *C. olitorius* to the rest of India ($N_m = 6.7–14.5$), particularly northern India ($N_m = 14.5$), might have played a significant role in population differentiation in the north Indian plains.

Both HWE-assumed (STRUCTURE) and relaxed (DAPC and sNMF) methods consistently identified two distinct geographically isolated Indian and African genetic groups. As compared to previous studies (Banerjee et al. 2012; Benor et al. 2012; Satya et al. 2014a; Zhang et al. 2015), we unequivocally established that the genetic structure of the *C. olitorius* populations is not reflected in the geographical contiguity of individuals. RAD-SNP alleles identified by DAPC as the most contributing to this group discrimination corresponded to genes related to cellular and metabolic process GO categories. All African accessions were correctly assigned to the African group, while nearly all north and south and 80% of central Indian accessions to the Indian one. By contrast, nearly half of the individuals from East Indian, most from the other Asian and all from the RoW populations were assigned to the African group. Intriguingly, the first primitive and modern cultivars Chinsurah

Green and JRO-620, developed in eastern India from local landraces (Kar et al. 2010), were identified as of pure African lineages, with the former reported earlier to be closely related to Sudan Green, an accession from Sudan (Basu et al. 2004). Even though the biological meaning of these two genetic groups appears to be obvious, the fact that they do not always correspond to geographic origin of the accessions advocates recent migration of *C. olitorius* through introduction and germplasm exchange across geographical boundaries (Roy et al. 2006). This is further substantiated by the existence of cryptic population structure within African and the other two Asian populations when examined individually with STRUCTURE. Both the genotypic combinations are present to varying degrees in individuals in each of the four Indian populations, suggesting population structure due to intraspecific genetic admixture. This is consistent with a rather higher level of natural outcrossing (10–17%) in *C. olitorius* (Satya et al. 2014a). With 90% of the admixed individuals belonging to the four Indian populations, a more focused STRUCTURE analysis supported the occurrence of introgressive hybridization between differentiated populations in central, eastern and northern India and potentially in southern India. Intraspecific genetic admixture can be presumed to be associated with increased overall genetic variance within Indian populations (Rius and Darling 2014), resulting in heightened capacity to respond to selection pressures by overcoming the deleterious influences of inbreeding load vis-à-vis founder effects (Verhoeven et al. 2011). Novel allelic combinations appeared through multiple introductions and admixture might have provided the Indian populations with the raw material for local adaptation (Rius and Darling 2014).

We detected a low extension of LD in the whole sample, which may be due to the given gene pool structure and higher level of natural outcrossing. A slightly higher LD in Indian than in African subpopulation may be due to an imposed selection pressure in the former gene pool. The genetic bottleneck associated with the successive breeding efforts allowed the occurrence of increased LD in Mesoamerican as compared to the Andean gene pool of *Phaseolus vulgaris* (Valdisser et al. 2016). Jute was rediscovered as an important fibre crop in British India only around 1800 (Basu et al. 2016). However, systematic breeding efforts could not be traced back earlier than 1900 (Kar et al. 2010). The successive *C. olitorius* breeding efforts in India predominantly used the African gene pool (Kar et al. 2010), but paradoxically many of the indigenous types used in crosses and/or selections proved to be of pure African lineages. Apparently, the repeated use of elite genitors from the same gene pool could have reduced the genetic gains in selection, resulting in relatively narrow genetic base of modern varieties (Roy et al. 2006; Kar et al. 2009). Therefore, the selection process construed to be associated with an increased LD in the

Indian gene pool was extraneous to the systematic breeding efforts, preceding the recorded breeding history of *C. olitorius* as a fibre crop in India.

Of the three population-based outlier tests, LOSITAN was the most efficient but not consistent in identifying loci under positive selection. However, an inadequacy of BayeScan to detect positive selection may be, because its power is severely weakened by the presence of admixed individuals (Luu et al. 2017) that account for as high as 46% in our diversity panel. One possible reason behind the failure of HacDivSel to detect significant ($P \leq 0.05$) outliers may be that it is rather conservative and requires strict independence among loci (Carvajal-Rodríguez 2017), which is an unrealistic assumption while dealing with a large number of markers (Pérez-Figueroa et al. 2010). By contrast, the effectiveness of individual-based PCAdapt in detecting the maximum number of adaptive loci is due to its ability to handle admixed individuals, without the ascertainment bias of F_{ST} -based outlier tests (Luu et al. 2017). For it does not correct for population structure, PCAdapt is likely to detect the set of outliers that will contain more false but also true positives and is thus an efficient method to uncover potential candidate genes (François et al. 2016).

Most of the candidate genes, particularly those encoding TFs, have multiple biological functions. Therefore, instead of BLAST alone, we employed a serial approach based on BLAST, Blast2GO and semantic similarity measures (REViGO) to obtain overrepresented functional GO categories. The two most essential GO biological processes uncovered are related to plant-type cell wall biogenesis and xylem–phloem pattern formation. We have earlier shown that bHLH, FAR1 and GRAS are involved in lignin biosynthesis and secondary cell wall deposition in jute fibres (Chakraborty et al. 2015). Candidate *SCPL* genes control cell elongation (Bienert et al. 2012), a key step in bast fibre formation. Detection of an adaptive gene encoding an ARABIDILLO-1 ortholog, an F-box/ARMADILLO repeat protein that promotes lateral root growth (Mu et al. 2010), could be related to bast fibre production, because a QTL hotspot for bast fibre yield involves a QTL for root growth in *C. olitorius* (Kundu et al. 2015). The wider adaptability of *C. olitorius* in India has been related to deeper and rapidly growing root system (Choudhary et al. 2017). Three adaptive loci associated with the two isoforms of a TIP-type AQP are closely related to LuTIP2-1 and LuTIP2-2 involved in water and solutes transport in flax (Shivaraj et al. 2017). This TIP2 class has long been identified as a shoot AQP involved in vascular development in *Arabidopsis* (Daniels et al. 1996). Together, these results suggest that bast fibre production was an important artificial selection pressure in *C. olitorius*. Overrepresented biological processes related to ion transmembrane transport and oxidation–reduction process are involved in the acquisition of salinity stress tolerance in jute

(Yang et al. 2017). Serine-type carboxypeptidase activity, an overrepresented molecular function GO term, is associated with broad-spectrum disease resistance (Mugford et al. 2009). Manonyl-CoA synthetase activity associated with direct synthesis of manonyl-CoA, a precursor for cauline anthocyanins, may be related to photoprotective acclimation to high light stress (Gould et al. 2010). Therefore, it is possible that abiotic and biotic stresses were natural selection pressures involved in the local adaptation of *C. olerius*. Candidate genes associated with the tolerance of abiotic and biotic stresses were subjected to positive selection during domestication of ramie (Liu et al. 2014).

No marked changes in population structure inferences in the absence of outlier loci may be because relatively larger number of RAD-SNP loci might have buffered the potential bias caused by outlier loci. With increasing numbers of gSSR loci, a buffering effect on the flawed genetic structure results was observed in flax (Soto-Cerda and Cloutier 2013). However, correction of our population structure results in terms of individual group membership and admixture level following the removal of outliers implies that these loci or others in LD with them were potentially affected by artificial and/or natural selection pressures. Since geographically sampled Indian accessions were mostly affected by the flawed group membership, it is evident that these selective loci could be involved in the differentiation of Indian populations. The exclusion of the outlier loci resulted in 23–24% reduction in genetic differentiation among geographic populations concomitant with 21.1% reduction in genetic divergence (F_{ST}) between the two ancestral subpopulations. Given that bast fibre production was a potential artificial selection pressure, we would suggest that *C. olerius* was possibly domesticated as a fibre crop most likely in central and northern India. In the future, this hypothesis will be tested further by studying the genetic association between these candidate RAD-SNP loci and domestication-related bast fibre yield and yield components.

Acknowledgements This work was funded by National Agricultural Science Fund (NASF), Indian Council of Agricultural Research (ICAR), New Delhi (Grant ID: GB-2018) and ICAR-Network Project on Transgenics in Crops (Grant ID: ICAR-NPTC-3070). We thank NxGenBio Life Sciences, New Delhi for assistance in RADseq library preparation, Illumina HiSeq™ 2000 sequencing and raw data processing. We also thank Dr. Subhojit Datta for helpful feedback on tossa jute aquaporins. The manuscript was reviewed and approved by the institute. Comments and suggestions on the manuscript from the Editor and two anonymous reviewers are gratefully acknowledged.

Compliance with ethical standards

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

Ethical approval All laboratory experiments complied with appropriate ethical standards, according to existing rules and regulations of the

Indian Council of Agricultural Research (ICAR), Government of India. The article does not pertain to any laboratory experiments involving human participants or animals.

References

- Akter J, Islam MS, Sajib AA, Ashraf N, Haque S, Khan H (2008) Microsatellite markers for determining genetic identities and genetic diversity among jute cultivars. *Austral J Crop Sci* 1:97–107
- Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA (2016) Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat Rev Genet* 17:81–92
- Antao T, Lopes A, Lopes R, Beja-Pereira A, Luikart G (2008) LOSITAN: a workbench to detect molecular adaptation based on a F_{st} -outlier method. *BMC Bioinform* 9:323
- Banerjee S, Das M, Mir RR, Kundu A, Topdar N, Sarkar D et al (2012) Assessment of genetic diversity and population structure in a selected germplasm collection of 292 jute genotypes by microsatellite (SSR) markers. *Mol Plant Breed* 3:11–25
- Basu A, Ghosh M, Meyer R, Powell W, Basak SL, Sen SK (2004) Analysis of genetic diversity in cultivated jute determined by means of SSR markers and AFLP profiling. *Crop Sci* 44:678–685
- Basu T, Satya P, Sarkar D, Kar CS, Mitra J, Karmakar PG (2016) Organellar genetic diversity in a global collection of jute (*Corchorus capsularis* and *C. olerius*, Malvaceae). *S Afr J Bot* 103:54–60
- Benor S, Blattner FR, Demissew S, Hammer K (2010) Collection and ethnobotanical investigation of *Corchorus* species in Ethiopia: potential leafy vegetables for dry regions. *Genet Resour Crop Evol* 57:293–306
- Benor S, Demissew S, Hammer K, Blattner FR (2012) Genetic diversity and relationships in *Corchorus olerius* (Malvaceae s. l.) inferred from molecular and morphological data. *Genet Resour Crop Evol* 59:1125–1146
- Bienert MD, Delannoy M, Navarre C, Boutry M (2012) NtSCP1 from tobacco is an extracellular serine carboxypeptidase III that has an impact on cell elongation. *Plant Physiol* 158:1220–1229
- Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007) TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23:2633–2635
- Campana MG, Hunt HV, Jones H, White J (2011) *CorrSieve*: software for summarizing and evaluating Structure output. *Mol Ecol Resour* 11:349–352
- Carvajal-Rodríguez A (2017) HacDivSel: two new methods (haplotype-based and outlier-based) for the detection of divergent selection in pairs of populations. *PLoS One* 12:e0175944
- Chakraborty A, Sarkar D, Satya P, Karmakar PG, Singh NK (2015) Pathways associated with lignin biosynthesis in lignomaniac jute fibres. *Mol Genet Genom* 290:1523–1542
- Choudhary SB, Sharma HK, Anil Kumar A, Maruthi RT, Karmakar PG (2017) The genus *Corchorus* L. (Malvaceae) in India: species distribution and ethnobotany. *Genet Resour Crop Evol* 64:1675–1686
- Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21:3674–3676
- Daniels MJ, Chaumont F, Mirkov TE, Chrispeels MJ (1996) Characterization of a new vacuolar membrane aquaporin sensitive to mercury at a unique site. *Plant Cell* 8:587–599
- Doebley JF, Gaut BS, Smith BD (2006) The molecular genetics of crop domestication. *Cell* 127:1309–1321
- Earl DA, vonHoldt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output

- and implementing the Evanno method. *Conserv Genet Resour* 4:359–361
- Edmonds JM (1990) Herbarium survey of African *Corchorus* L. species. International Board for Plant Genetic Resources, Rome
- Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol* 14:2611–2620
- Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* 180:977–993
- Foll M, Fischer MC, Heckel G, Excoffier L (2010) Estimating population structure from AFLP amplification intensity. *Mol Ecol* 19:4638–4647
- François O, Martins H, Caye K, Schoville SD (2016) Controlling false discoveries in genome scans for selection. *Mol Ecol* 25:454–469
- Freeland JR, Kirk H, Petersen S (2011) *Molecular ecology*, 2 edn. Wiley-Blackwell, Chichester
- Frichot E, Mathieu F, Trouillon T, Bouchard G, François O (2014) Fast and efficient estimation of individual ancestry coefficients. *Genetics* 196:973–983
- Fuller DQ, Murphy C (2018) The origins and early dispersal of horsegram (*Macrotyloma uniflorum*), a major crop of ancient India. *Genet Resour Crop Evol* 65:285–305
- Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q et al (2014) TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PLoS One* 9:e90346
- Gould KS, Dudle DA, Neufeld HS (2010) Why some stems are red: cauline anthocyanins shield photosystem II against high light stress. *J Exp Bot* 61:2707–2717
- Greenbaum G, Templeton AR, Zarmi Y, Bar-David S (2014) Allelic richness following population founding events- a stochastic modeling framework incorporating gene flow and genetic drift. *PLoS One* 9:e115203
- Hammer Ø, Harper DAT, Ryan PD (2001) Past: paleontological statistics software package for education and data analysis. *Palaeontol Electron* 4:1–9
- Hardy OJ, Vekemans X (2002) SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Mol Ecol Notes* 2:618–620
- Hou Y, Nowak MD, MirrÈ V, BJORÅ CS, Brochmann C, Popp M (2015) Thousands of RAD-seq loci fully resolve the phylogeny of the highly disjunct Arctic-Alpine genus *Diapensia* (Diapensiaceae). *PLoS One* 10:e0140175
- Huang Y-F, Poland JA, Wight CP, Jackson EW, Tinker NA (2014) Using genotyping-by-sequencing (GBS) for genomic discovery in cultivated oat. *PLoS One* 9:e102448
- Islam M, Saito JA, Emdad E, Ahmed B, Islam M, Halim A et al (2017) Comparative genomics of two jute species and insight into fibre biogenesis. *Nat Plants* 3:16223
- Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* 23:1801–1806
- Jeffreys H (1961) *Theory of probability*. Oxford University Press, Oxford
- Jombart T (2008) *adegen*: a R package for the multivariate analysis of genetic markers. *Bioinformatics* 24:1403–1405
- Kar CS, Kundu A, Sarkar D, Sinha MK, Mahapatra BS (2009) Genetic diversity in jute (*Corchorus* spp.) and its utilization: a review. *Indian J Agric Sci* 79:575–586
- Kar CS, Satya P, Mitra J, Sarkar D, Sinha MK, Kundu A et al (2010) Varietal development of jute and allied fibres in India. *Indian Farm* 60:5–9
- Kundu BC (1951) Origin of jute. *Indian J Genet Plant Breed* 11:95–99
- Kundu A, Topdar N, Sarkar D, Sinha MK, Ghosh A, Banerjee S et al (2013) Origins of white (*Corchorus capsularis* L.) and dark (*C. olitorius* L.) jute: a reevaluation based on nuclear and chloroplast microsatellites. *J Plant Biochem Biotechnol* 22:372–381
- Kundu A, Chakraborty A, Mandal NA, Das D, Karmakar PG, Singh NK et al (2015) A restriction-site-associated DNA (RAD) linkage map, comparative genomics and identification of QTL for histological fibre content coincident with those for retted bast fibre yield and its major components in jute (*Corchorus olitorius* L., Malvaceae s. l.). *Mol Breed* 35:19
- Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of theory of selective neutrality of polymorphisms. *Genetics* 74:175–195
- Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760
- Liu K, Muse SV (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics* 21:2128–2129
- Liu T, Tang S, Zhu S, Tang Q, Zheng X (2014) Transcriptome comparison reveals the patterns of selection in domesticated and wild ramie (*Boehmeria nivea* L. Gaud). *Plant Mol Biol* 86:85–92
- Lu F, Lipka AE, Glaubitz J, Elshire R, Cherney JH, Casler MD et al (2013) Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. *PLoS Genet* 9:e1003215
- Luu K, Bazin E, Blum MGB (2017) *pcadapt*: an R package to perform genome scans for selection based on principal component analysis. *Mol Ecol Resour* 17:67–77
- Mahapatra AK, Saha A, Gupta D (2006) Catalogue on evaluation of tossa jute germplasm (*Corchorus olitorius* L.). Central Research Institute for Jute and Allied Fibres, Kolkata
- Marroni F, Pinosio S, Zaina G, Fogolari F, Felice N, Cattonaro F et al (2011) Nucleotide diversity and linkage disequilibrium in *Populus nigra* cinnamyl alcohol dehydrogenase (*CAD4*) gene. *Tree Genet Genomes* 7:1011–1023
- Meirmans PG (2012) AMOVA-based clustering of population genetic data. *J Hered* 103:744–750
- Meirmans PG, Hedrick PW (2011) Assessing population structure: F_{ST} and related measures. *Mol Ecol Resour* 11:5–18
- Meirmans PG, Van Tienderen PH (2004) GENOTYPE and GENODIVE: two programs for the analysis of genetic diversity of asexual organisms. *Mol Ecol Notes* 4:792–794
- Moore AD, Held A, Terrapon N, Weiner J, Bornberg-Bauer E (2014) DoMosaics: software for domain arrangement visualization and domain-centric analysis of proteins. *Bioinformatics* 30:282–283
- Mu C, Chen N, Li X, Jia P, Wang Z, Liu H (2010) F-box protein Arabidillo-1 promotes lateral root development by depressing the functioning of GA_3 in *Arabidopsis*. *J Plant Biol* 53:374–380
- Mugford ST, Qi X, Bakht S, Hill L, Wegel E, Hughes RK et al (2009) A serine carboxypeptidase-like acyltransferase is required for synthesis of antimicrobial compounds and disease resistance in oats. *Plant Cell* 21:2473–2484
- Narum SR, Hess JE (2011) Comparison of F_{ST} outlier tests for SNP loci under selection. *Mol Ecol Resour* 11:184–194
- Narum SR, Buerkle AC, Davey JW, Miller MR, Hohenlohe PA (2013) Genotyping-by-sequencing in ecological and conservation genomics. *Mol Ecol* 22:2841–2847
- Nei M, Tajima F, Tateno Y (1983) Accuracy of estimated phylogenetic trees from molecular data. II. Gene frequency data. *J Mol Evol* 19:153–170
- Paetkau D, Calvert W, Stirling I, Strobeck C (1995) Microsatellite analysis of population structure in Canadian polar bears. *Mol Ecol* 4:347–354
- Pan Y, Wang X, Sun G, Li F, Gong X (2016) Application of RAD sequencing for evaluating the genetic diversity of domesticated *Panax notoginseng* (Araliaceae). *PLoS One* 11:e0166419
- Pegadaraju V, Nipper R, Hulke B, Qi L, Schultz Q (2013) *De novo* sequencing of sunflower genome for SNP discovery using RAD (Restriction site Associated DNA) approach. *BMC Genom* 14:1–9

- Pérez-Figueroa A, García-Pereira MJ, Saura M, Rolán-Alvarez E, Caballero A (2010) Comparing three different methods to detect selective loci using dominant markers. *J Evol Biol* 23:2267–2276
- Pfeifer B, Wittelsbürger U, Ramos-Onsins SE, Lercher MJ (2014) PopGenome: an efficient Swiss army knife for population genomic analyses in R. *Mol Biol Evol* 31:1929–1936
- Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155:945–959
- R Core Team (2017) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna
- Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J et al (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc Natl Acad Sci USA* 98:11479–11484
- Rius M, Darling JA (2014) How important is intraspecific genetic admixture to the success of colonising populations? *Trends Ecol Evol* 29:233–242
- Rosenberg NA (2004) DISTRUCT: a program for the graphical display of population structure. *Mol Ecol Notes* 4:137–138
- Rousset F (2008) GENEPOP'007: a complete re-implementation of the GENEPOP software for Windows and Linux. *Mol Ecol Resour* 8:103–106
- Roy A, Bandyopadhyay A, Mahapatra AK, Ghosh SK, Singh NK, Bansal KC et al (2006) Evaluation of genetic diversity in jute (*Corchorus* species) using STMS, ISSR and RAPD markers. *Plant Breed* 125:292–297
- Sarkar D, Mahato AK, Satya P, Kundu A, Singh S, Jayaswal PK et al (2017a) The draft genome of *Corchorus olitorius* cv. JRO-524 (Navin). *Genom Data* 12:151–154
- Sarkar D, Menor C, Singh NK (2017b) Does RNA-seq-based Eukaryotic GeneFinding of Blast2GO require repeat-masking the whole genome shotgun (WGS) sequence? A case study in jute (*Corchorus olitorius* L., Malvaceae s. l.). *BioBam Bioinformatics S. L.* <https://www.blast2go.com/support/blog/22-blast2goblog/178-is-repeat-masking-necessary-in-blast2go>. Accessed 24 Jul 2017
- Satya P, Banerjee R, Biswas C, Karan M, Ghosh S, Ali N (2014a) Genetic analysis of population structure using peroxidase gene and phenylalanine ammonia-lyase gene-based DNA markers: a case study in jute (*Corchorus* spp.). *Crop Sci* 54:1609–1620
- Satya P, Karan M, Chakraborty K, Biswas C, Karmakar PG (2014b) Comparative analysis of diversification and population structure of kenaf (*Hibiscus cannabinus* L.) and roselle (*H. sabdariffa* L.) using SSR and RGA (resistance gene analogue) markers. *Plant Syst Evol* 300:1209–1218
- Shivraj SM, Deshmukh RK, Rai R, Bélanger R, Agrawal PK, Dash PK (2017) Genome-wide identification, characterization, and expression profile of aquaporin gene family in flax (*Linum usitatissimum*). *Sci Rep* 7:46137
- Soto-Cerda BJ, Cloutier S (2013) Outlier loci and selection signatures of simple sequence repeats (SSRs) in flax (*Linum usitatissimum* L.). *Plant Mol Biol Rep* 31:978–990
- Supek F, Bošnjak M, Škunca N, Šmuc T (2011) REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* 6:e21800
- Szpiech ZA, Jakobsson M, Rosenberg NA (2008) ADZE: a rarefaction approach for counting alleles private to combinations of populations. *Bioinformatics* 24:2498–2504
- Takezaki N, Nei M, Tamura K (2014) POPTREEW: web version of POPTREE for constructing population trees from allele frequency data and computing some other quantities. *Mol Biol Evol* 31:1622–1624
- Valdisser PAMR, Pappas GJ, Menezes IPP, Müller BSF, Pereira WJ, Narciso MG et al (2016) SNP discovery in common bean by restriction-associated DNA (RAD) sequencing for genetic diversity and population structure analysis. *Mol Genet Genom* 291:1277–1291
- Verhoeven KJF, Macel M, Wolfe LM, Biere A (2011) Population admixture, biological invasions and the balance between local adaptation and inbreeding depression. *Proc R Soc Lond B Biol Sci* 278:2–8
- Vigouroux Y, Glaubitz JC, Matsuoka Y, Goodman MM, Sánchez GJ, Doebley J (2008) Population structure and genetic diversity of New World maize races assessed by DNA microsatellites. *Am J Bot* 95:1240–1253
- Winters CA (2010) Y-chromosome evidence of an African origin of Dravidian agriculture. *Int J Genet Mol Biol* 2:30–33
- Xu P, Xu S, Wu X, Tao Y, Wang B, Wang S et al (2014) Population genomic analyses from low-coverage RAD-Seq data: a case study on the non-model cucurbit bottle gourd. *Plant J* 77:430–442
- Yang Z, Yan A, Lu R, Dai Z, Tang Q, Cheng C et al (2017) De novo transcriptome sequencing of two cultivated jute species under salinity stress. *PLoS One* 12:e0185863
- Ye J, Fang L, Zheng H, Zhang Y, Chen J, Zhang Z et al (2006) WEGO: a web tool for plotting GO annotations. *Nucleic Acids Res* 34:W293–W297
- Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF et al (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* 38:203–208
- Zhang L, Yuan M, Tao A, Xu J, Lin L, Fang P et al (2015) Genetic structure and relationship analysis of an association population in jute (*Corchorus* spp.) evaluated by SSR markers. *PLoS One* 10:e0128195

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

Affiliations

Debabrata Sarkar¹  · Avijit Kundu^{1,2} · Debajeet Das¹ · Avrajit Chakraborty¹ · Nur Alam Mandal¹ · Pratik Satya¹ · Pran Gobinda Karmakar¹ · Chandan Sourav Kar¹ · Jiban Mitra¹ · Nagendra Kumar Singh³

¹ Biotechnology Unit, Division of Crop Improvement, ICAR-Central Research Institute for Jute and Allied Fibres (CRIJAF), Nilganj, Barrackpore, Kolkata, West Bengal 700 120, India

² Department of Genetics and Plant Breeding, North Bengal Agricultural University, Pundibari, Cooch Behar, West Bengal 736 165, India

³ Rice Genome Lab, ICAR-National Research Centre on Plant Biotechnology (NRCPB), Pusa, New Delhi 110 012, India