Mol Breeding (2012) 30:1325–1335
DOI 10.1007/s11032-012-9719-6

# Discovery and use of single nucleotide polymorphic (SNP) markers in *Jatropha curcas* L.

**Priya Gupta · Asif Idris · Shrikant Mantri · Mehar Hasan Asif · Hemant Kumar Yadav · Joy Kumar Roy · Rakesh Tuli · Chandra Sekhar Mohanty · Samir Vishwanath Sawant**

Received: 8 April 2011 / Accepted: 27 February 2012 / Published online: 24 March 2012


**Abstract** Various programs for genetic improvement in oil yield of the biofuel plant *Jatropha curcas* L. are currently in progress worldwide. In order to develop strategies for genetic improvement, it is important to estimate the degree of diversity at the genetic level among various genotypes of *J. curcas*. High-throughput sequencing of complexity-reduced nuclear genomic DNA of *J. curcas* coupled with computational analysis discovered 2,482 informative single nucleotide polymorphisms (SNPs). Genotyping of selective SNPs among 148 global collections of *J. curcas* lines and further diversity analysis through NTSYS-pc, DARwin and Structure 2.0 software revealed that a narrow level of genetic diversity existed among the indigenous genotypes as compared to the exotic genotypes of *J. curcas*. The level of marker informativeness along with distance-based and Bayesian clustering revealed grouping of the accession from Togo (Africa) with various Indian accessions at $K = 4$ and $K = 5$ values (where $K$ represents the number of populations). The diverse accessions identified in the study will be of further use in genetic improvement of *J. curcas* through quantitative trait loci and association mapping.

**Keywords** *Jatropha curcas* · Molecular marker · Single nucleotide polymorphism · High-throughput sequencing

## Introduction

Biofuels are considered as the preferred fossil-fuel alternatives due to their renewability, non-toxicity, biodegradability and cost-effectiveness (Bondioli et al. 2003, Antolin et al. 2002). The biomass necessary for the production of biofuels can be derived from several sources including oil-producing edible and non-edible crop plants. More than 60 potential biofuel plants have been identified to date for biodiesel production, based on their productivity (El Bassam 1998) and adaptability to growth in adverse soil and harsh climatic conditions (Francis et al. 2005). *Jatropha curcas* L. is considered as the most promising biofuel crop due to its

**Electronic supplementary material** The online version of this article (doi:10.1007/s11032-012-9719-6) contains supplementary material, which is available to authorized users.

P. Gupta · A. Idris · M. H. Asif · H. K. Yadav · C. S. Mohanty (✉) · S. V. Sawant
Plant Molecular Biology and Genetic Engineering Division, National Botanical Research Institute, Rana Pratap Marg, Lucknow 226 001, Uttar Pradesh, India
e-mail: cs.mohanti@nbri.res.in; sekhar_cm2002@rediffmail.com

S. Mantri · J. K. Roy · R. Tuli
National Agri-food Biotechnology Institute, C-127, Industrial Area, SAS Nagar, Phase 8, Mohali 160 071, Punjab, India

reported productivity and adaptability to growth in adverse climatic conditions (Heller 1996).

*Jatropha curcas* is commonly known as purging nut or physic nut or barbados nut. It is a multi-purpose shrub or small tree belonging to the family Euphorbiaceae. This plant is widely distributed throughout arid and semi-arid, tropical and sub-tropical regions of the world (Openshaw 2000). *Jatropha curcas* is a diploid plant species (2n = 2x = 22) (Dehgan 1984) with an estimated genome size (1C) of 416 Mb (Carvalho et al. 2008). Sato et al. (2011) revealed the whole genome of *J. curcas* with a combinatorial conventional Sanger and next-generation multiplex sequencing method and identified 36.6 % repetitive sequences. Comparative analysis of the generated genome sequence data (http://www.kazusa.or.jp/jatropha/) detected a high degree of microsynteny with the genome of castor bean and to a lesser extent with *Glycine max* and *Arabidopsis thaliana*.

Genetic diversity can be assessed using either morphological traits (Kaushik et al. 2007) or molecular markers. The study of genetic diversity based on morphological traits is not reliable as they are highly influenced by environment. Genetic diversity based on molecular markers is independent of environmental factors. There is a surge of interest in identifying a large number of molecular markers for rapid application in the assessment of genetic diversity and the selection of desired genotypes. Among the various molecular markers employed to assess genetic diversity, PCR-based molecular markers such as microsatellites or simple sequence repeats (SSRs) and amplified fragment length polymorphisms (AFLPs) are the choice for many applications.

Intra-specific and inter-specific diversity in *J. curcas* has been studied by employing random amplified polymorphic DNA (RAPD) (Ram et al. 2008), AFLP (Tatikonda et al. 2009), inter-simple sequence repeat (ISSR) (Kumar et al. 2009), SSR and EST-derived simple sequence repeat (eSSR) (Wen et al. 2010) markers. Of the different molecular markers, EST-SSRs are less polymorphic than anonymous SSRs as these are located within genes, and are thus more conserved across species. This has been noticed in a number of taxa, including rice (Cho et al. 2000), pine (Liewlaksaneeyanawin et al. 2004), barley (Chabane et al. 2005) and sunflower (Pashley et al. 2006). However, SNPs are the most abundant type of sequence variations in genome (Batley et al. 2003) and are thus more informative than other markers. SNP-based sequence information can potentially identify casual mutations in the genome. Hence, SNPs provide an approach for developing high-density markers within and near the locus of interest, which is important for map-based gene cloning and haplotype-based association studies (Rafalski 2002).

*Jatropha curcas* is widely distributed throughout the tropical and sub-tropical areas of the world. Understanding the population structure and distribution of *J. curcas* has been a challenge because of its limited genetic variability (Singh et al. 2010). The current investigation presents the results of genetic diversity in terms of SNPs among different genotypes and the underlying population structure of *J. curcas* collected from different parts of the world. The discovery and genotyping of the markers were carried out by using next-generation sequencing technologies. This is the first attempt at this approach to discovering and genotyping SNP markers in *J. curcas*.

## Materials and methods

### Materials

A set of 148 genotypes (Supplementary Table 1) of *J. curcas* collected from different crop-growing regions of India, North America, South America and Africa are now maintained at the garden of the National Botanical Research Institute (NBRI), Lucknow, India. From these, 61 indigenous genotypes (NBRI 001 to NBRI 061) were arbitrarily selected for the extraction of genomic DNA and SNP discovery as these lines were initially available at the time of experimentation. Further validation of these discovered SNPs was carried out on all 148 genotypes including the indigenous as well as the exotic lines which were procured later.

### Extraction and pooling of genomic DNA

The young unexpanded leaves (~5 g) of *J. curcas* plants were ground into fine powder in liquid nitrogen. The powder was used to extract genomic DNA using DNeasy plant DNA kit (Qiagen, Hilden, Germany). The extracted DNA was air-dried, resuspended in 500 µl of sterile ultrapure water and stored at −20 °C for further use. The DNA concentration was estimated

in a spectrophotometer at $OD_{260/280}$ and its quality was checked on 1.0 % agarose gel. Genomic DNA from all 61 accessions was pooled together in equal concentration of 100 ng each for further complexity reduction of the genomic representation. The quantity of DNA from each sample was evenly pooled to avoid losing some markers (Davey et al. 2011).

## Genome complexity reduction

The primary rationale for complexity reduction of the *J. curcas* genome was to maximize the chances of SNP detection. We compared candidate enzymes, namely *Acc*I, *Pst*I, *Pvu*II and *Hin*dIII, for complexity-reduced genomic DNA representation library based on their predicted fragment population, repetitive element content and number of unique fragments (Tassel et al. 2008). We performed in silico digestion of genomic DNA sequences available at the time of experiment in the public database (http://www.ncbi.nlm.nih.gov/) using NEBcutter v2.0 software (tools.neb.com/NEBcutter2/) with these candidate restriction enzymes to identify the number and frequency of the fragments (Table 1). Based on the maximum number of restriction cut-sites, *Hin*dIII was selected as the enzyme of choice. Restriction digestion of *J. curcas* genome with 20 U of each candidate enzyme also confirmed *Hin*dIII as the best choice among the candidate enzymes (Fig. 1).

The *Hin*dIII-digested pooled DNA was ethanol-precipitated. *Hin*d-III-specific adaptors were ligated to 100 ng of digested DNA fragments in the ligation buffer (1 × T4 DNA ligase buffer containing ATP) and 0.2 U T4 DNA ligase (New England Biolabs, USA).The ligation mixture was incubated at 16 °C for overnight. The ligated DNA was diluted (1:10) with sterile water before pre-amplification using the primers complimentary to *Hin*dIII adaptors using a PCR

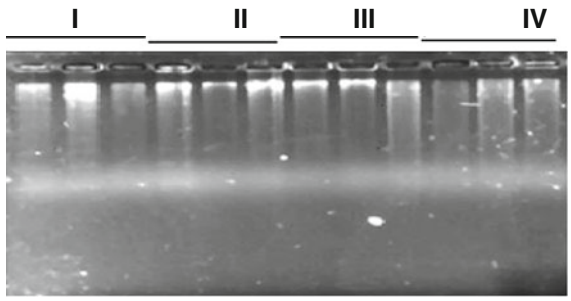


**Fig. 1** Restriction enzyme digestion of the genomic DNA of three genotypes (NBRI 001, NBRI 002 and NBRI 003) of *J. curcas* with four different enzymes, namely *Acc*I, *Pst*I, *Pvu*II and *Hin*dIII

profile of 20 cycles: 94 °C for 30 s, 56 °C for 60 s and 72 °C for 60 s. The amplified product was diluted (1:10) in sterile water for selective amplification Complexity reduction was achieved by taking one selective nucleotide per reaction for extension of the primer (Table 2) using a thermal cycler (MJ Research, USA) with the PCR profile of one cycle of 94 °C for 5 min, 62 °C for 2 min and 72 °C for 2 min, another 35 cycles of 94 °C for 1 min, 62 °C for 1 min and 72 °C for 1 min, and finally three cycles of 72 °C for 5 min. These adaptor and adaptor-based primers were designed as per the Clontech Genome Walker Library (Table 2). The annealing temperature for each selective primer ($S_A$, $S_T$, $S_C$ and $S_G$) was optimized by running gradient PCR. The annealing temperature ranged between 56 and 64 °C. The PCR products were cleaned with a PCR purification kit (Sigma GenElute$^{TM}$ PCR Clean-Up). The sample concentration for each of the four libraries were measured fluorimetrically using Quant-iT$^{TM}$ Picogreen reagent (Invitrogen, Carlsbad, CA, USA) and pooled in equimolar amounts to a final concentration of 4 μg of DNA.

**Table 1** Result of in silico digestion of genomic DNA of *J. curcas* with candidate restriction enzymes

| Sequence characteristics | *Acc*I | *Hin*dIII | *Pst*I | *Pvu*II |
|---|---|---|---|---|
| Total no. of sequences | 33,710 | 35,629 | 18,629 | 18,175 |
| Average length of sequences | 3,932.8 | 3,919.9 | 4,711.3 | 4,690.3 |
| No. of cut sites | 71,248 | **83,108** | 29,061 | 30,593 |
| Average no. of cut sites | 2.11 | 2.33 | 1.55 | 1.68 |

Total number of sequences subjected to in silico digestion was 120586

Bold value indicates maximum number of cut-sites of the genome as compared to the rest of the candidate enzymes

**Table 2** Sequences of oligonucleotide adapter and primers used for complexity reduction

| Name | Sequence |
|---|---|
| Adapter | 5′-GTA ATA CGA CTC ACT ATA GGG CA-3′ |
| | 3′-CAT TAT GCT GAG TAG TAT CCC GTT CGA-5′ |
| Primer (pre-selective) | 5′-GTA ATA CGA CTC ACT ATA GGG C-3′ |
| Primer (selective) $S_A$ | 5′-GAC TCA CTA TAG GGC AAG CTT A-3′ |
| Primer (selective) $S_T$ | 5′-GAC TCA CTA TAG GGC AAG CTT T-3′ |
| Primer (selective) $S_C$ | 5′-GAC TCA CTA TAG GGC AAG CTT C-3′ |
| Primer (selective) $S_G$ | 5′-GAC TCA CTA TAG GGC AAG CTT G-3′ |

### Pyrosequencing of the PCR-amplified reduced representation library

The sample (4 μg) was nebulized by applying pressure of 45 psi liquid nitrogen. Random fragments of size between 300 and 800 bp were quality-checked on a bioanalyzer (Agilent DNA 7500 kit) for the preparation of a single-stranded template (Sst) DNA library. The quality-checked samples were processed for shotgun pyrosequencing on the 454 GS FLX sequencer (Roche, Basel, Switzerland) as per the manufacturer's protocol (Ronaghi et al. 1998). Two sequencing runs were commissioned by using two regions of a four-region gasket from a 454 GS FLX PicoTitre Plate™.

### Detection of single nucleotide polymorphisms

The generated data was further analyzed bioinformatically for the presence of SNP markers. During the analysis, only the quality-filtered reads were used for assembly and identification of SNPs. Assembly of reads was done using the GS assembler at default settings (40 bases overlap and 90 % identity). The generated ACE file was fed into AutoSNP Version 1 Pipeline (Barker et al. 2003) for detection of SNPs. All the aligned sequences in contigs were viewed physically using EagleView software (Huang and Marth 2008).

### Genotyping of SNPs

The discovered SNPs were validated using matrix-assisted laser desorption/ionization time-of-flight mass spectroscopy (MALDI-TOF–MS) assay. The assays were carried out on the Sequenom™ platform. Assays for PCR extension reaction were designed with MassARRAY™ Assay Design™ software (Sequenom™) (Supplementary Table 2). SNPs were genotyped by the iPLEX™ protocol as described by the manufacturer (San Diego, USA). For all the markers tested, the iPLEX™ was found to be robust and required little optimization.

### Statistical analysis

The SNP genotyping data obtained from the Sequenom MassARRAY were used for the estimation of polymorphic information content (PIC) value for each SNP marker. SNP marker diversity in diverse germplasms of *J. curcas* was estimated using the PIC formula proposed by Weir (1996) and implemented in Powermarker software v3.07 (Liu and Muse 2005). The software computes for each marker locus, including allele number, missing proportion, heterozygosity and gene diversity, apart from PIC. This value was estimated from the following formula (Botstein et al. 1980):

$$\mathrm{PIC} = 1 - \sum_{u=1}^{k} p_{lu}^2 - \sum_{u=1}^{k-1} \sum_{v=u+1}^{k} 2 p_{lu}^2 p_{lv}^2$$

where $p_u$ (or $p_{lu}$) represents the population frequency of an allele $A_u$ at the $l$th locus, and $p_v$ (or $p_{lv}$) represents the population frequency of an allele $A_v$ at the $l$th locus.

In PowerMarker, the haplotype estimation is highly optimized for SNP data by taking advantage of the binary feature of these markers (Rohde and Fuerst 2001). Genetic similarity was estimated between pairs of genotypes according to Jaccard's similarity coefficient using NTSYS-pc 2.02 (Rohlf 1993) and DARwin 5.0 (http://darwin.cirad.fr/darwin). A genetic similarity matrix obtained based on Jaccard's similarity coefficient was used to prepare the dendrogram

following the unweighted pair group method with arithmetic average (UPGMA). The discriminatory power of the SNP primers were evaluated by calculating the marker index (Powell et al. 1996) and resolving power (Prevost and Wilkinson 1999):

$$\text{Marker index (MI)} = \text{PIC} \times \text{EMR}$$

where EMR is the "effective multiplex ratio" which is the product of the total number of loci per primer and the fraction of polymorphic loci.

The model-based clustering method Structure 2.0 (Pritchard et al. 2000) was used for detecting the population structure of the different genotypes and the multilocus data. Population structure is an important factor in association mapping studies (Breseghello and Sorrells 2006). To estimate the number of sub-populations, the SNP marker datasets were analyzed separately (Pritchard et al. 2000; Falush et al. 2003). A burn-in period of 100,000 and Monte Carlo Markov Chain replications of 200,000 were used for the analyses. The "admixture" and "correlated" options were used for the ancestry and allele frequency models, respectively. Structure was run to test the hypothesis of two to ten ($K = 2$–10) sub-populations.

For each dataset, five independent runs were carried out for each possible number of clusters ($K$) from $K = 2$ to $K = 5$ in order to quantify the variation in the likelihood of the data for a given $K$. The range of tested $K$ was set according to the true number of the simulated population. Each data set took between 5 and 30 h to run, depending on the number of markers and individuals simulated in the dataset.

## Result and discussion

### SNP discovery in *J. curcas*

The restriction enzyme *Hin*dIII was used for digestion of genomic DNA. The digestion product produced a continuous smear of DNA (Fig. 1). Complexity-reduced genomic DNA was pooled and further sequenced on the 454 GS FLX sequencer. The generated data provided substantially more genome coverage than achieved with the previous analysis based on AFLP (Tatikonda et al. 2009) and SSR molecular markers. The generated sequence data was assembled using the GS Assembler. The entire data was assembled into 871 contigs containing 26,940 sequences (Table 3). The assembled contigs of the sequencing run are available in the web-link (http://www.nbri.res.in/downloads/jatropha.txt/). A total of 2,482 candidate SNPs were identified from the total contigs at an average frequency of one SNP per 100 bp and one insertion/deletion (InDel) per 500 bp according to the analysis done by AutoSNP program. When compared with the genome sequence data of Sato et al. (2011) (http://www.kazusa.or.jp/jatropha/) and queried using BLASTn (E-value $\leq$ 1e−5), a significant match was found between the contig sequence data concluding that, of 871 contigs, 817 contigs gave a significant match with a sequence similarity of 80–100 % (Supplementary Table 3).

The genome analysis of *J. curcas* indicated that a transition bias existed in the analyzed sequence of the *J. curcas* genome. This may be due to methylated cytosines in CpG dinucleotides changed into thymines during the genesis of the SNPs (Tsaftaris and Polidoros 1999). A similar abundance of SNPs has been reported in maize (Batley et al. 2003). The present study revealed the presence of approximately one SNP per 100 bp of *J. curcas* nuclear DNA. However, in *Picea rubens* and *Picea mariana* five SNPs per 1 kbp of nuclear DNA (Germano and Klein 1999), one SNP

**Table 3** De novo assembly of *J. curcas*

| | |
|---|---|
| Total bases (bp) | 4,174,913 |
| Total reads | 26,940 |
| No. of contigs (bp) | 871 |
| Av. contig size (bp) | 353 |
| Singletons | 869 |
| N50 contig size[a] (bp) | 757 |
| Q40 plus bases[b] (%) | 93.52 |

[a] N50 corresponds to the contigs larger than those that have 50 % of the base assembly

[b] Percentage of bases called that have a quality score of 40 or above

**Table 4** Summary of *J. curcas* sequence variant analysis based on 454 pyrosequencing

| Type of SNP | Number of SNPs |
|---|---|
| Transition (C/T or G/A) (%) | 1,751 (70.54 %) |
| Transversion (C/T, A/G, C/A, or T/G) (%) | 731 (29.54 %) |
| InDels | 757 |
| Frequency of InDel | 1/500 bp |

in approximately 200-bp sequence in *Glycine max* (Coryell et al. 1999) and one SNP per every 48 bp and every 130 bp in 3′-untranslated regions and coding regions, respectively, were reported in maize (Tenaillon et al. 2002). Compared to SNPs, the frequency of InDels was much lower in *J. curcas* (Table 4). In *A. thaliana*, the levels of InDels (one every 6.1 kbp) and SNPs (one every 3.3 kbp) between the genomes of Columbia and Landsberg erecta differ by a factor of only two (Drenkard et al. 2000). These results suggest that, for fine mapping of a gene in *J. curcas*, the use of SNPs rather than InDels as a marker appear to be a good choice.

The PIC value along with the marker index of several crop plants have been determined by AFLP primer combinations to be soybean (PIC = 0.32, MI = 6.14) (Powell et al. 1996), wheat (PIC = 0.32, MI = 3.41) (Bohn et al. 1999) and cornsalad (PIC = 0.25, MI = 4.47) (Muminovic et al. 2004). Here the average SNP–locus combinations of the entire analysis and their corresponding PIC and MI values for *J. curcas* showed a PIC value of 0.102 and a MI value of 0.102. Out of the total of 60 validated SNPs, 13 SNPs were monomorphic and the remaining 47 SNPs were polymorphic in nature. The PIC value of polymorphic SNPs ranged between 0.01 (SNP locus 50) and 0.37 (SNP locus 17). The average PIC value was 0.06 ± 0.1, indicating a low level of informativeness in SNPs. The major allele frequency (MAF) of the polymorphic SNP loci varied between 0.99 and 0.5, with an average of 0.94 ± 0.1. The heterozygosity level of the polymorphic locus also varied between 0.98 and 0.01 with an average of 0.1 ± 0.2. The multiplex effective ratio (MER) is 1.0 per SNP marker (Varshney et al. 2008). The MI values for these markers therefore ranged between 0.014 and 0.104 with a mean value of 0.068 (Supplementary Table 4). Estimation of genetic diversity through SSR markers by Sato et al. (2011) employed on 12 *J. curcas* lines also concluded a narrow level of genetic diversity with a mean PIC value of 0.06.

## Distribution and characterization of the putative SNPs

The candidate SNPs were categorized according to nucleotide substitution as either transitions (C/T or A/G) or transversions (A/C, C/G, A/T, G/T). There was a relative increase in the proportion of transitions (70.54 %) over transversions (29.54 %). A relative increase in the frequency of transitions was observed over transversions. A total of 757 InDels were observed (Table 3). Since there are four types of transitions ($T_s$) and eight types of transversions ($T_v$), the expected ratio of transition to transversion ($T_s/T_v$) is 0.5. The $T_s/T_v$ ratio for *J. curcas* was calculated to be 2.3.

## Validation of SNPs

A total of 103 putative SNPs flanked by ~100 bp on each side were submitted to the primer design software (MassARRAY Assay Design 3.0) of Sequenom. It was possible to design primers for 78 SNP loci for 34, 27, 22, 16 and 13 multiplex assays. The majority of these multiplex assays provided results consistent with 60 SNPs. The assay was carried out using the manufacturer's guidelines. The genotyping data was acquired using the Sequenom MassARRAY and processed using Sequenom Typer 3.4 software™.

## Genetic relationships between *J. curcas* genotypes

The genotyping data was used to estimate pair-wise Jaccard's genetic similarity among the 148 accessions. The Jaccard's genetic similarity coefficient ranged from 0.629 to 0.984. The maximum similarity value of 0.984 was observed between the genotypes NBRI 096 (Uttar Pradesh), NBRI 086 (Andhra Pradesh) and NBRI 097 (Uttaranchal) with NBRI 118 (Uttar Pradesh). These results indicated that the genotypes collected from different locations of Uttar Pradesh, Uttaranchal and Andhra Pradesh have a close genetic similarity with each other. The maximum dissimilarity of 0.65 was observed between genotypes NBJC 186 (Togo) and NBRI 048 (Himachal Pradesh). Of the Indian genotypes, NBRI 079 (Balasore, Orissa) showed maximum dissimilarity (0.629) with the genotype NBRI 054 (Coimbatore, Tamil Nadu). To understand the genetic relationships between the genotypes, the genotyping data for the SNP markers studied were used to draw the neighbor-joining tree on the basis of the bootstrap value in DARwin software. The principal coordinate analysis (PCA) displayed similar grouping of accessions with some minor deviation (Fig. 2).

The neighbor-joining dendrogram considering the bootstrap values classified all the 148 accessions into

two major clusters, A and B (Fig. 3). Cluster A can be divided into two parts, A1 and A2: sub-cluster A1 is composed of 35 accessions (23.6 %) and A2 of 21 accessions (14.9 %), whereas the major cluster B comprises three sub-clusters, B1, B2 and B3, composed of 52 (35.1 %), 14 (9.4 %) and 26 (17.5 %) accessions, respectively, out of the total of 148 accessions of *J. curcas*. Genetic relationships between the 148 exotic and indigenous genotypes were resolved further by principal coordinate analysis (PCA), which grouped all these accessions into two major distinct groups and five sub-groups. In the PCA plot (Fig. 2) NBRI 054 (Coimbatore, India) and NBJC 186 (Africa) are distinctly separated and form a separate cluster. The low genetic diversity in *J. curcas* among Indian and exotic accessions could probably be due to limited propagation of the crop through vegetative cutting. However, the study using the AFLP-primer sets revealed that maximum genetic similarity existed among the genotypes collected from Madhya Pradesh, Gujarat and Uttar Pradesh (Tatikonda et al. 2009).

Interestingly, our study showed that the accessions from Madhya Pradesh had maximum similarity with the accession from Andhra Pradesh. The genetic closeness of the accessions in a cluster from different regions indicates the lack of gene flow between adjacent populations in each region. The results of PCA analysis were comparable to the cluster analysis with little deviation. However, the cluster analysis showed a complex pattern of genetic relationships among 148 genotypes, structured in a number of distinct small breeding lineages. A high level of similarity was observed among these genotypes. The overall grouping pattern of the PCA corresponded well with the clustering pattern of the dendrogram. However, distance-based methods introduce simplification and distortion among the members of a large cluster.

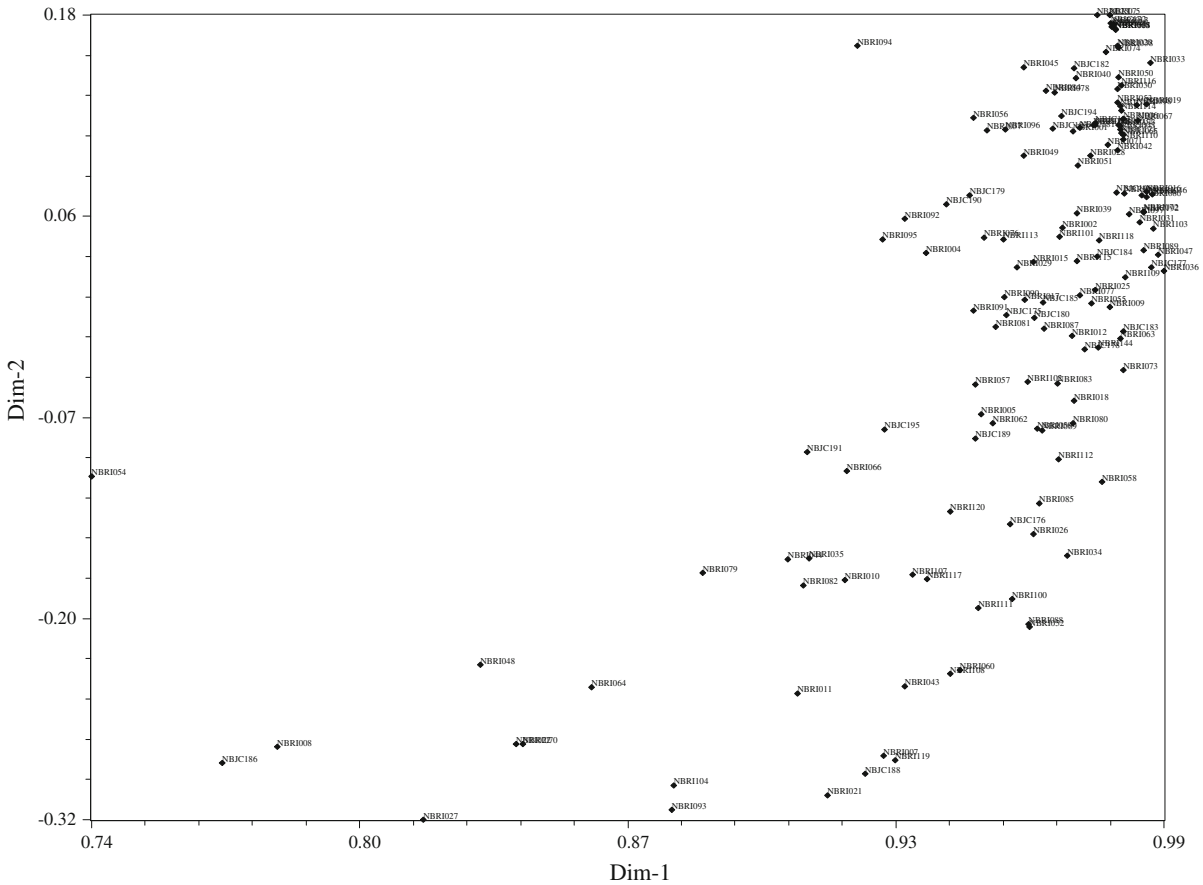


**Fig. 2** Two-dimensional plot of 148 accessions of *J. curcas* by principal component analysis using Jaccard's similarity coefficients

For this reason, accessions were grouped by a model-based clustering method (Pritchard et al. 2000). The model applied in our study allowed for better representation of the relationships between the main gene pools present in the collection and for clarification of the mixed ancestries of several accessions and even breeding groups.

## Population stratification of *J. curcas* genotypes

Bayesian-based population structure analysis implemented in Structure was used to determine the underlying groups ($K$) in the collection. The value of $K$ determines the number of sub-populations. The biggest initial increase in likelihood was for $K = 2$ and then progressively increased, until at optimal $K = 5$ and subsequent higher values of $K$, the group split progressively into more population lines. Simulation studies had shown that the LnP(D) value continued to increase and form a cluster until $K = 5$, indicating that $K = 5$ provided a good fit to the data (Fig. 4).

The population structure at $K = 5$ consistently showed five groupings. The nuclear SNP genotypes of the worldwide collection of samples ($n = 148$) were best described by ive clusters. However, the groupings were not consistent with the pattern of origin of the genotypes. This pattern of clustering is supported by various computational software analyses. Consistent with the rest of the analysis, the genotype NBJC 186 (Togo, Africa) formed a cluster with the Indian accessions NBRI 054 (Tamil Nadu, India) and NBRI 022 (Bihar). Furthermore, African accessions clustered with NBRI 039 (Haryana), NBRI 006 (Uttaranchal), NBRI 099 (Uttaranchal) and NBRI 036 (Punjab). The groupings of African *J. curcas* with native *J. curcas* are highly consistent with the analysis at $K = 4$ and $K = 5$. This observation also supports the findings made by Sato et al. (2011) that a significant phylogenetic relationship exists between

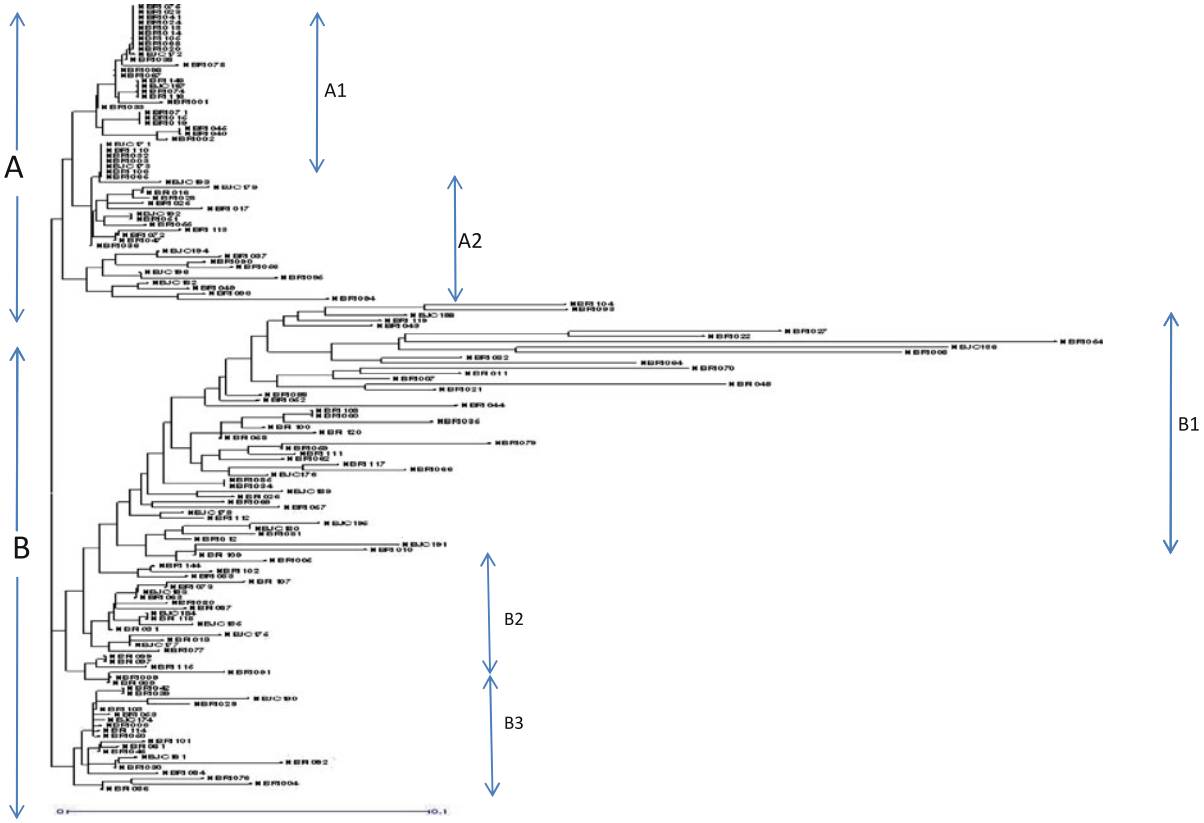


**Fig. 3** Genetic relationships between 148 accessions of *J. curcas* based on neighbor-joining tree constructed by DARwin at 1,000 bootstrap value

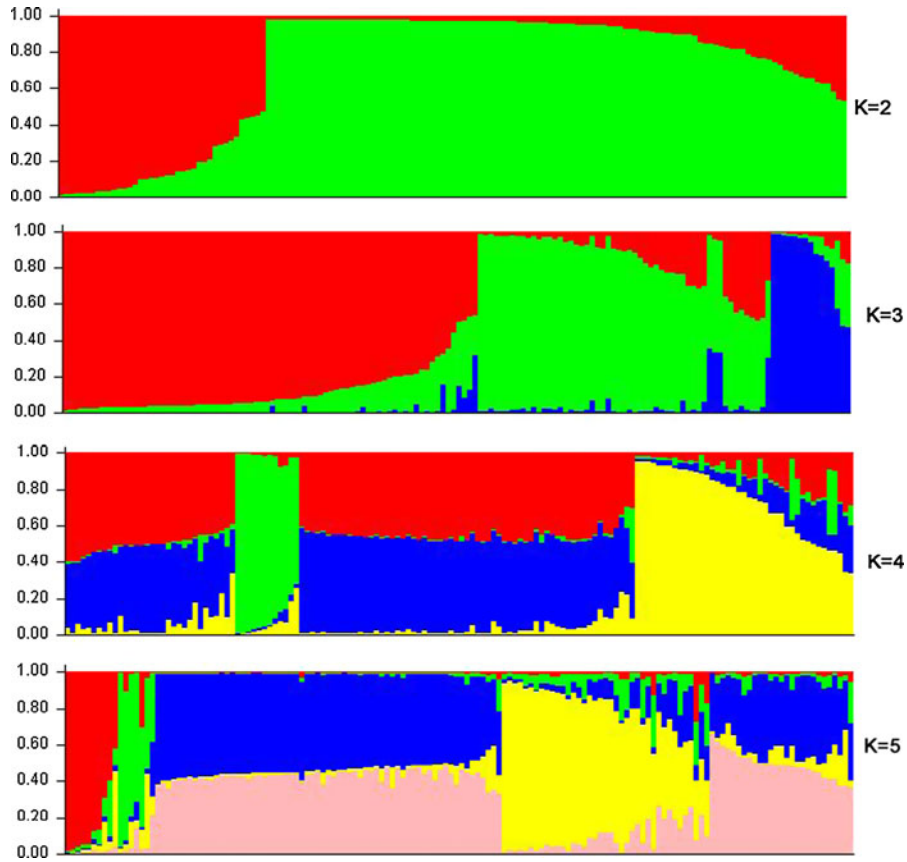


**Fig. 4** Population Structure of *J. curcas*. Values of $K$ (number of clusters) ranged from 2 to 5. A progressive increase in the grouping was observed from $K = 2$ to $K = 5$

the Asian and African lines when analyzed through SSR markers.

## Conclusions

The study demonstrated that high-throughput sequencing on a 454 GS FLX sequencer of the complexity-reduced nuclear genomic DNA of *J. curcas* detects a large number of valid SNPs with great efficiency and accuracy at a much lower investment of time than the conventional methods. The genome of *J. curcas* showed relatively more SNPs than InDels, with a bias towards transition base substitution over transversion. The SNPs identified after complexity reduction of the genome function as an efficient method for identifying and estimating genetic diversity among a global population of *J. curcas*. A narrow level of genetic diversity is reported to exist among the various indigenous genotypes as compared to the exotic accessions. The diverse lines identified through these SNP markers are under further investigation for carrying out marker-assisted selection programs for genetic improvement of *J. curcas*.

**Acknowledgments** This work was supported under a supra-institutional programme of the Council of Scientific and Industrial Research (CSIR), Government of India. The Director, NBRI, is acknowledged for providing basic infrastructure facilities for carrying out the work.

## References

Antolin G, Tinaut FV, Briceno Y, Castano V, Perez C, Ramirez AI (2002) Optimisation of biodiesel production by sunflower oil transesterification. Bioresour Technol 83:111–114

Barker G, Batley J, O'Sullivan H, Edwards KJ, Edwards D (2003) Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. Bioinformatics 19:421–422

Batley J, Barker G, O'Sullivan H, Edwards KJ, Edwards D (2003) Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. Plant Physiol 132:84–91

Bohn M, Utz HF, Melchinger AE (1999) Genetic similarities among winter wheat cultivars determined on the basis of

RFLPs, AFLPs, and SSRs and their use for predicting progeny variance. Crop Sci 39:228–237

Bondioli P, Gasparoli A, Della Bella L, Tagliabue S, Toso G (2003) Biodiesel stability under commercial storage conditions over one year. Eur J Lipid Sci Technol 105:735–741

Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32:314–331

Breseghello F, Sorrells ME (2006) Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars. Genetics 172:1165–1177

Carvalho CR, Clarindo WR, Praça MM, Araújo FS, Carels N (2008) Genome size base composition and karyotype of Jatropha curcas L., an important biofuel plant. Plant Sci 174:613–617

Chabane K, Ablett GA, Cordeiro GM, Valkoun J, Henry RJ (2005) EST versus genomic derived microsatellite markers for genotyping wild and cultivated barley. Genet Resour Crop Evol 52:903–909

Cho YG, Ishii T, Temnykh S, Chen X, Lipovich L, McCouch SR, Park WD, Ayres N, Cartinhour S (2000) Diversity of microsatellites derived from genomic libraries and GenBank sequences in rice (*Oryza sativa* L.). Theor Appl Genet 100:713–722

Coryell VH, Jessen H, Schupp JM, Webb D, Keim P (1999) Allele-specific hybridization markers for soybean. Theor Appl Genet 98:690–696

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet 12:499–510

Dehgan B (1984) Phylogenetic significance of interspecific hybridization in Jatropha (Euphorbiaceae). Syst Bot 9:467–478

Drenkard E, Richter BG, Rozen S, Stutius LM, Angell NA, Mindrinos M, Cho RJ, Oefner PJ, Davis RW, Ausubel FM (2000) A simple procedure for the analysis of single nucleotide polymorphisms facilitates map-based cloning in Arabidopsis. Plant Physiol 124:1483–1492

El Bassam N (1998) Energy plant species: their use and impact on environment and development. James and James (Science Publishers)

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567–1587

Francis G, Edinger R, Becker K (2005) A concept for simultaneous wasteland reclamation, fuel production, and socio economic development in degraded areas in India: need, potential and perspectives of Jatropha plantations. Nat Resour Forum 29:12–24

Germano J, Klein AS (1999) Species-specific nuclear and chloroplast single nucleotide polymorphisms to distinguish *Picea glauca*, *P. mariana* and *P. rubens*. Theor Appl Genet 99:37–49

Heller J (1996) Physic Nut. *Jatropha curcas L.* Promoting the Conservation and use of underutilized and neglected crops. International Plant Genetic Resources Institute, Rome

Huang W, Marth G (2008) EagleView: a genome assembly viewer for next-generation sequencing technologies. Genome Res 18:1538–1543

Kaushik N, Kumar K, Kumar S, Roy S (2007) Genetic variability and divergence studies in seed traits and oil content of Jatropha (*Jatropha curcas* L.) accessions. Biomass Bioenergy 31:497–502

Kumar RS, Parthiban KT, Govinda Rao M (2009) Molecular characterization of Jatropha genetic resources through inter-simple sequence repeat (ISSR) markers. Mol Biol Rep 36:1951–1956

Liewlaksaneeyanawin C, Ritland CE, El-Kassaby YA, Ritland K (2004) Single-copy, species-transferable microsatellite markers developed from loblolly pine ESTs. Theor Appl Genet 109:361–369

Liu K, Muse SV (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics 21:2128–2129

Muminovic J, Melchinger AE, Luubberstedt T (2004) Genetic diversity in cornsalad (*Valerianella locusta*) and related species as determined by AFLP markers. Plant Breed 123:460–466

Openshaw K (2000) A review of *Jatropha curcas* L.: an oil plant of unfulfilled promise. Biomass Bioenergy 19:1–15

Pashley CH, Ellis JR, McCauley DE, Burke JM (2006) EST databases as a source for molecular markers: lessons from Helianthus. J Hered 97:381–388

Powell W, Morgante M, Andre C, Hanafey M, Vogel J, Tingey S, Rafalski A (1996) The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis. Mol Breed 2:225–238

Prevost A, Wilkinson MJ (1999) A new system of comparing PCR primers applied to ISSR fingerprinting of potato cultivars. Theor Appl Genet 98:107–112

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959

Rafalski A (2002) Applications of single nucleotide polymorphism in crop genetics. Curr Opin Plant Biol 5:94–100

Ram SG, Parthiban KT, Kumar RS, Thiruvengadam V, Paramathma M (2008) Genetic diversity among *Jatropha* species as revealed by RAPD markers. Genet Resour Crop Evol 55:803–809

Rohde K, Fuerst R (2001) Haplotyping and estimation of haplotype frequencies for closely linked biallelic multilocus genetic phenotypes including nuclear family information. Hum Mut 17:289–295

Rohlf FJ (1993) NTSYS-pc: numerical taxonomy and multivariate analysis system, version 2.11. Applied Biostatistics, Setauket, New York

Ronaghi M, Uhlen M, Nyren P (1998) A sequencing method based on real-time pyrophosphate. Science (Washington) 281:363–365

Sato S, Hirakawa H, Isobe S, Fukai E, Watanabe A, Kato M, Kawashima K, Minami C, Muraki A, Nakazaki N, Takahashi C, Nakayama S, Kishida Y, Kohara M, Yamada M, Tsuruoka H, Sasamoto S, Tabata S, Aizu T, Toyoda A, Shin-i T, Minakuchi Y, Kohara Y, Fujiyama A, Tsuchimoto S, Kajiyama S, Makigano E, Ohmido N, Shibagaki N, Cartagena JA, Wada N, Kohinata T, Atefeh A, Yuasa S, Matsunga S, Fukui K (2011) Sequence analysis of the genome of an oil-bearing tree, *Jatropha curcas* L. DNA Res 18:65–76

Singh P, Singh S, Mishra SP, Bhatia SK (2010) Molecular characterization of genetic diversity in *Jatropha curcas* L. Gene, Genomes Genomics 4:1–8

Tassell CPV, Smith TPL, Matukumalli LK, Taylor JF, Schnabel RD, Lawley CT, Haudenschild CD, Moore SS, Warren WC, Sonstegard TS (2008) SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat Meth 5:247–252

Tatikonda L, Wani SP, Kannan S, Beerelli N, Sreedevi TK, Hoisington DA, Devi P, Varshney RK (2009) AFLP-based molecular characterization of an elite germplasm collection of *Jatropha curcas* L., a biofuel plant. Plant Sci 176:505–513

Tenaillon MI, Sawkins MC, Anderson LK, Stack SM, Doebley J, Gaut BS (2002) Patterns of diversity and recombination along chromosome 1 of Maize (*Zea mays* ssp. *mays* L.). Genetics 162:1401–1413

Tsaftaris AS, Polidoros AN (1999) DNA methylation and plant breeding. Plant Breed Rev 18:87–176

Varshney RK, Salem KFM, Baum M, Roder MS, Graner A, Börner A (2008) SSR and SNP diversity in a barley germplasm collection. Plant Genet Resour 6:167–174

Weir BS (1996) Genetic data analysis II. Sinauer Associate Inc., Sunderlands

Wen MF, Wang HY, Xia ZQ, Zou ML, Lu C, Wang WQ (2010) Development of EST-SSR and genomic-SSR markers to assess genetic diversity in *Jatropha Curcas* L. BMC Res Notes 3:42