Sherman et al. BMC Plant Biology (2015) 15:277
DOI 10.1186/s12870-015-0663-6

BMC Plant Biology

RESEARCH ARTICLE Open Access

Mango (*Mangifera indica* L.) germplasm diversity based on single nucleotide polymorphisms derived from the transcriptome

Amir Sherman[1*], Mor Rubinstein[1], Ravit Eshed[1], Miri Benita[1], Mazal Ish-Shalom[1], Michal Sharabi-Schwager[1,2], Ada Rozen[1], David Saada[1], Yuval Cohen[1] and Ron Ophir[1*]

Abstract

Background: Germplasm collections are an important source for plant breeding, especially in fruit trees which have a long duration of juvenile period. Thus, efforts have been made to study the diversity of fruit tree collections. Even though mango is an economically important crop, most of the studies on diversity in mango collections have been conducted with a small number of genetic markers.

Results: We describe a *de novo* transcriptome assembly from mango cultivar 'Keitt'. Variation discovery was performed using Illumina resequencing of 'Keitt' and 'Tommy Atkins' cultivars identified 332,016 single-nucleotide polymorphisms (SNPs) and 1903 simple-sequence repeats (SSRs). Most of the SSRs (70.1 %) were of trinucleotide with the preponderance of motif (GGA/AAG)n and only 23.5 % were di-nucleotide SSRs with the mostly of (AT/AT)n motif. Further investigation of the diversity in the Israeli mango collection was performed based on a subset of 293 SNPs. Those markers have divided the Israeli mango collection into two major groups: one group included mostly mango accessions from Southeast Asia (Malaysia, Thailand, Indonesia) and India and the other with mainly of Floridian and Israeli mango cultivars. The latter group was more polymorphic ($F_S = -0.1$ on the average) and was more of an admixture than the former group. A slight population differentiation was detected ($F_{ST} = 0.03$), suggesting that if the mango accessions of the western world apparently was originated from Southeast Asia, as has been previously suggested, the duration of cultivation was not long enough to develop a distinct genetic background.

Conclusions: Whole-transcriptome reconstruction was used to significantly broaden the mango's genetic variation resources, i.e., SNPs and SSRs. The set of SNP markers described in this study is novel. A subset of SNPs was sampled to explore the Israeli mango collection and most of them were polymorphic in many mango accessions. Therefore, we believe that these SNPs will be valuable as they recapitulate and strengthen the history of mango diversity.

Keywords: Mango, Genetic diversity, Transcriptome, SNP, SSR

* Correspondence: asherman@volcani.agri.gov.il; ron@agri.gov.il
[1]Department of Fruit Trees Sciences, Institute of Plant Sciences, Agricultural Research Organization, Volcani Center, Rishon Lezion, Israel
Full list of author information is available at the end of the article

Background

The origin of *Mangifera indica* L. species which includes all commercial cultivars is still undetermined. The genus *Mangifera* has approximately 70 members which are located mostly on the Malay peninsula, in the Indonesian archipelago, in Thailand and in the Philippines [1, 2]. Some of these species have edible fruit which are locally cultivated. Mango cultivation began a few thousand years ago in India. It first spread from Southeast Asia, only several hundred years ago, with the Portuguese and Spaniards to Africa, Central and South America. In recent years mango has become common in most tropical and subtropical regions. India together with several other countries in Southeast Asia is the main growing and production center for mango. Hundreds of known cultivars has been isolated in the last few hundred years in several mango growing countries, mainly in India, and in the Pacific islands [2]. A secondary mango center flourished in Florida during the late nineteenth century and early twentieth century, and many new Floridian cultivars were promoted [3]. These cultivars are adapted to the taste of the Western consumer by breeding to a red blush coloration, mild taste and mild aroma idoetype. However, there is still some demand for cultivar improvement, and several breeding programs are active in Australia, South Africa, Brazil and Israel [4].

Germplasm collections are important for genotypic and phenotypic analyses, and as a genetic resource in breeding programs. Knowledge of the diversity and the genetic structure of these collections is a fundamental for association studies and controlled breeding [5]. Despite the mango economic importance, the available genetic and genomic resources for mango cannot support modern breeding or the study of the molecular mechanisms underlying mango's physiology. A limited genetic map with very low resolution has been created for mango [6]. A few studies have attempted to decipher relationship among mango cultivar collections worldwide [7–14]. Twenty five Floridian accessions from the USDA collection were grouped into two clusters based on 28 random amplified polymorphic DNA (RAPD) markers [15]. One cluster was comprised of a group of Floridian accessions that are closely related to the Indian cultivar 'Mulgoba' whereas the other cluster contained a group of more distant accessions. A sample from the Floridian groups was also included in a work on the relationship of 22 mango accessions from the Thai mango collection. The variability of the Thai accessions was high and they were not distinguishable from the Floridian accessions, apparently because most of them were seedlings [8]. The Pakistanian collection mostly included Indian-originated mango accessions. Based on RAPD analysis of 44 loci due to high diversity in mango, only the southern Indian accessions could be separated from northern and eastern ones [10]. Two other studies investigated the association between genetic diversity and geographical properties of accessions in India [16, 17]. Those studies weakly separated the northern and eastern accessions from the southern and western ones. A Spanish research group showed that 16 simple-sequence repeats (SSRs) can differentiate the Floridian cultivars from the Indian and the Filipino ones in the 28 accessions of a Spanish collection [18].

Recently, the genetic diversity of mangoes originated from Andhra Pradesh, the major mango breeding area in India, was studied based on 106 polymorphic SSRs. Accessions of the same ideotype (juicy, pickle, table) were more related to each other but did not show any significant differentiation [19]. Further support for the high diversity of mango came from a study of six Colombian cultivar groups showing that the diversity within the six groups is as high as the diversity between them, which indicating very minor population divergence [11]. A broader survey of mango collection, including many geographical locations, was performed in Australia with the caveat of a low number of markers (11 SSRs) [13, 20]. The mangoes were successfully classified into five geographical origins however an attempt to classify the accessions by mono- or polyembryonic phenotype was unsuccessful.

Molecular efforts to create wide genomic and genetic data for mango are in their initial stages. These efforts have included establishment of a leaf transcriptome [21] and fruit transcriptomes at different developmental and ripening stages [22–25]. Next generation sequencing (NGS) technologies are excellent tools for genome-wide marker discovery and exposing genetic variation [26]. *De novo* transcriptome sequencing is one solution for marker discovery, gene expression analysis and exposing genetic variability in organisms with no genomic infrastructure such as olives, Chinese chestnut, carrot and pomegranate [27–31]. Large scale sets of genetic markers can be used to establish genetic maps. These maps can then be utilized for plant breeding and be utilized for anchoring in *de novo* genome assemblies. Moreover, studying the genetic variation of the germplasm collection can give insights into the historical basis of the diversity and can additionally be used for genome wide association studies in order to identify markers linked with important horticultural traits for plant breeding [32].

In the present work we describe our effort to broaden the transcriptome resources for mango by sequencing RNA from various tissues and fruit stages. Using 454-GS FLX Titanium technology we reconstructed a large portion of the 'Keitt' mango transcriptome and used it as a reference for aligning resequencing results. Resequencing of the 'Keitt' transcriptome itself as well as another mango accession,'Tommy Atkins', by Illumina HiSeq 2000 was used to discover a large set of genetic variation. A subset of that variation was utilized in order to explore the Israeli mango collection which comprises cultivars from several world regions.

Results and discussion

Genic variation is a very useful resource for marker assisted selection (MAS) and association studies. Therefore RNA samples of two mango accessions, 'Tommy Atkins' and 'Keitt', were obtained from a pool of tissues (young leaves, young inflorescences, young fruit, flesh and peels of mature fruit) as a representative transcriptome (hereafter Pool transcriptome). By pooling we expected to compensate for tissue-specific gene expression. Variation discovery in the transcriptome was performed in two steps. First, *de novo* assembly of the whole transcriptome was performed by 454-GS FLX Titanium sequencing of 'Keitt'. Second, resequencing of both mango cultivars, 'Tommy Atkins' and 'Keitt', was aligned to 'Keitt' *de novo* assembly contigs to obtain high coverage and therefore high accuracy of allele identification [33].

Assembly and annotation of the reference transcriptome

The sequencing of 'Keitt' using 454-GS FLX Titanium was yielded 1,329,313 reads. After filtering out low quality and empty reads, *de novo* assembly was performed on 1,113,875 reads resulting in 60,997 contigs. These contigs were then reassembled into super-contigs using the CAP3 program [34]. Ten percent of the contigs (6396) were assembled into super-contigs most (90 %) of which comprised 2 to 3 contigs. Altogether, the assembled 'Keitt' transcriptome contained 47,956 singleton contigs and super-contigs (hereafter mango contigs). We compared the results of the assembly in this work with two additional published assemblies that were based on a different sequencing strategy [21, 23]. Those transcriptomes were sequenced from RNA samples of leaf (hereafter Leaf transcriptome) [21] and fruit peel (hereafter Peel transcriptome) [23] tissues using Illumina technology followed by *de novo* assembly of short reads. Ninety percent of the contigs were 412, 219, and 223 bp or longer and half were at least 757, 321, 438 bp long for Pool, Leaf and Peel transcriptome assemblies respectively (Fig. 1). Both statistics suggested that the contigs of the Pool transcriptome are twice as long as those of the Leaf and Peel transcriptome assemblies [21, 23]. Obviously, the novel Pool transcriptome in this study significantly contributes to the length of available transcripts.

Functional annotation was also performed. First, the functional annotation of the Pool transcriptome resulted in a successful list of 40,971 hits (85 %) as a result of similarity searches against 'Gene Bank', 'TAIR', and 'UniProt' protein databases (Table 1). Second, by comparison to Leaf and Peel transcriptomes, we could investigate what are the common functionalities between leaves and fruit peel and assess whether novel transcriptome information was revealed in the Pool transcriptome. A reciprocal blast was run between the Pool and Peel transcriptomes, and between Pool and Leaf transcriptomes. The best hits were taken as the homologous transcripts. The number of Pool transcripts that were homologous to the Peel transcripts only was 10,251 whereas 3860 Pool transcripts were homologous to Leaf transcripts only. The common subset of transcripts, i.e., the intersection of the Peel, Leaf, and Pool transcriptomes, included 8660 transcripts (Additional file 1: Table S1). Half of the transcripts in the Pool transcriptome (21,880; 49 %) had no homolog in either the Peel or Leaf transcriptomes. The excess of transcripts in the

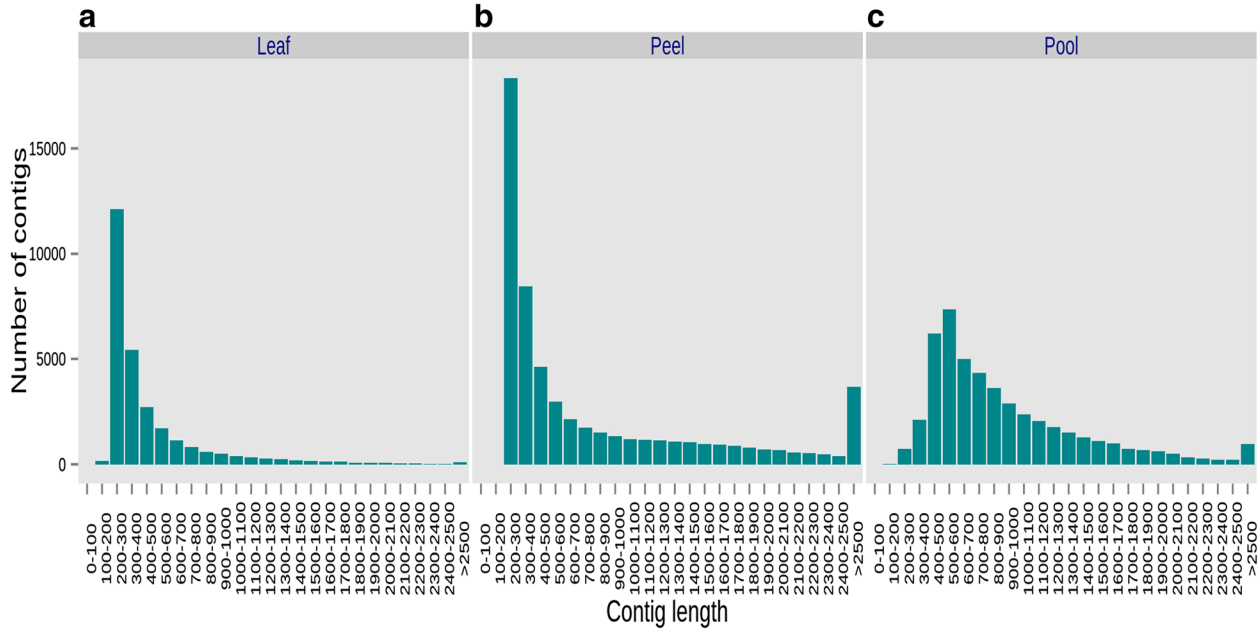

Fig. 1 Distribution of contig lengths and comparison with two published mango transcriptomes. The distribution of contig lengths from three assemblies was plotted: Leaf (**a**), Peel (**b**), and Pooled (**c**) of tissues. The distribution of consensus contig lengths is drawn as 100-bp long bins

Table 1 Number of mango contig homologous hits

	Non -redundant GeneBank proteins (nr)	TAIR	UniProt	Union of three database hits
Pool	40,795	34,918	30,684	40,971
Pool and peel intersection	17,366	16,079	13,173	17,423
Pool and leaf intersection	12,022	11,351	9390	12,038
Pool, peel and leaf intersection	8371	8074	6669	8380

Pool transcriptome relative to the Peel and Leaf transcriptomes could reveal new functionalities. Therefore a comparison of gene ontology (GO) functional categories between the common subset of transcripts and the rest of the transcripts might reveal whether or not new functionalities have been rendered. Figure 2 illustrates the distribution of GO-slim categories in the Pool transcriptome. In general, most of the GO-slim categories existed in both subsets of the Pool transcriptome. However, three transcripts related to cell communication category in the biological process ontology appeared exclusively in the Pool transcriptome as were five other transcripts related to the extracellular space.

Transcriptome variation

SSRs and single-nucleotide polymorphisms (SNPs) are highly useful in plant genetics and breeding for the construction of linkage maps and MAS [35, 36]. Therefore, we focused on the repertoire of SSRs and SNPs in the mango transcriptome. The number of SSRs found in the transcriptome was 1903. The SSRs were discovered in 4 % (1787) of all transcripts of the Pool transcriptome (Additional file 2: Table S2). The lengths of the SSR motifs ranged from 1 to 6. Most of the SSRs are trinucleotides (70.1 %) followed by dinucleotide (23.5 %) (Fig. 3a). The most frequent dinucleotide motif was (AT/AT)n with a frequency of 166 out of 590 followed by (GA/TC)n, (AG/CT)n, and (TA/TA)n (Fig. 3b). The least frequent motifs (only 10 %) were (CA/TG)n and (AC/GT)n. The three most frequent trinucleotide motifs are (GAA/TTC)n, (AAG/CTT)n, and (AGA/TCT)n with the proportions of 12, 10 and 10 % of all trinucleotide motifs, respectively (Fig. 3c). The novel SSRs, in this study, are expected to greatly enrich the mango community reservoir of SSRs that have already been reported [8, 9, 13, 14, 18, 20, 37, 38]. The SSRs in those studies were used as a genetic tool to investigate diversity in local germplasm collections. In general, those studies were based on a few SSRs and the frequency of the SSR motifs in the genome was not reported. Thus the SSR motifs could not be compared. However, the pattern of SSR motifs is known to be species-specific [39, 40]. Thus, a discovery of the same pattern of SSR motifs in the same species can strengthen the SSRs' reliability. Previously reported SSR motifs were congruent with the motifs that were reported here verifying their

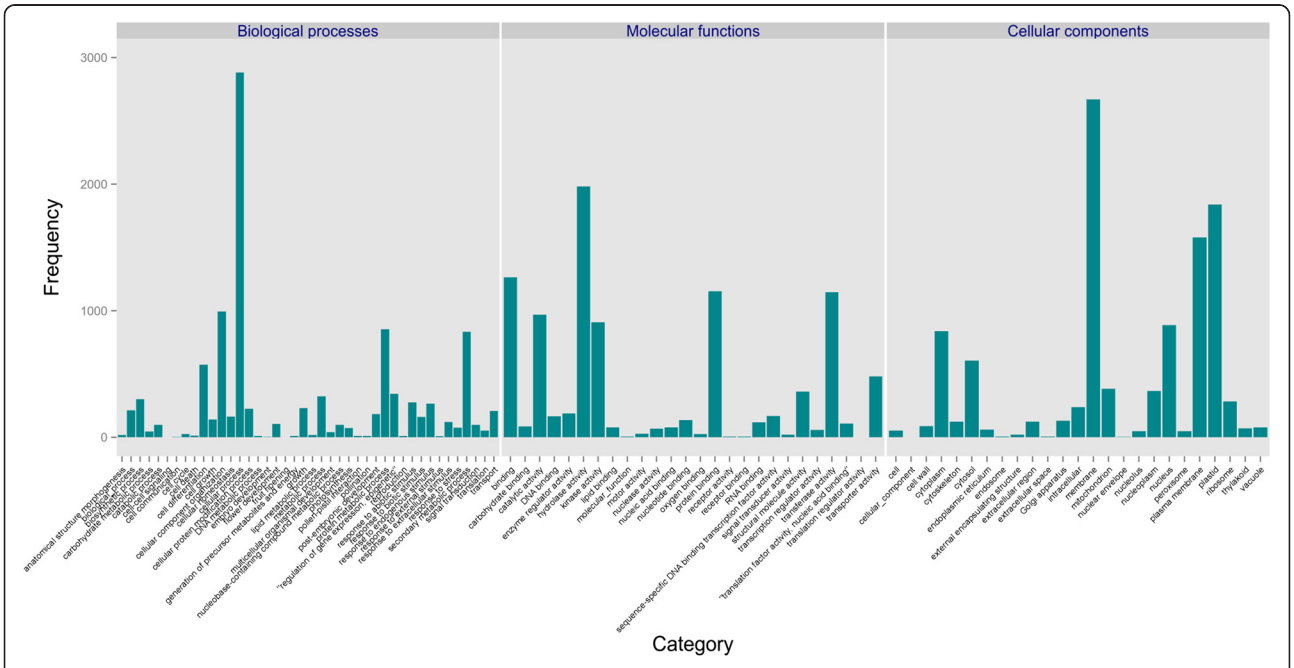

Fig. 2 Comparison of mango gene ontology categories in three transcriptome assemblies. Contigs were annotated by running blast search against 'nr' database and then performing mapping to Slim-GO categories by Blast2GO. The distribution of contigs of the three ontologies, biological processes, molecular functions, and cellular components was plotted for transcripts that were included exclusively in the transcriptome from the pool (Pool only) of tissues (root, leaf, flower and fruit developmental stage 3; turquoise bars)

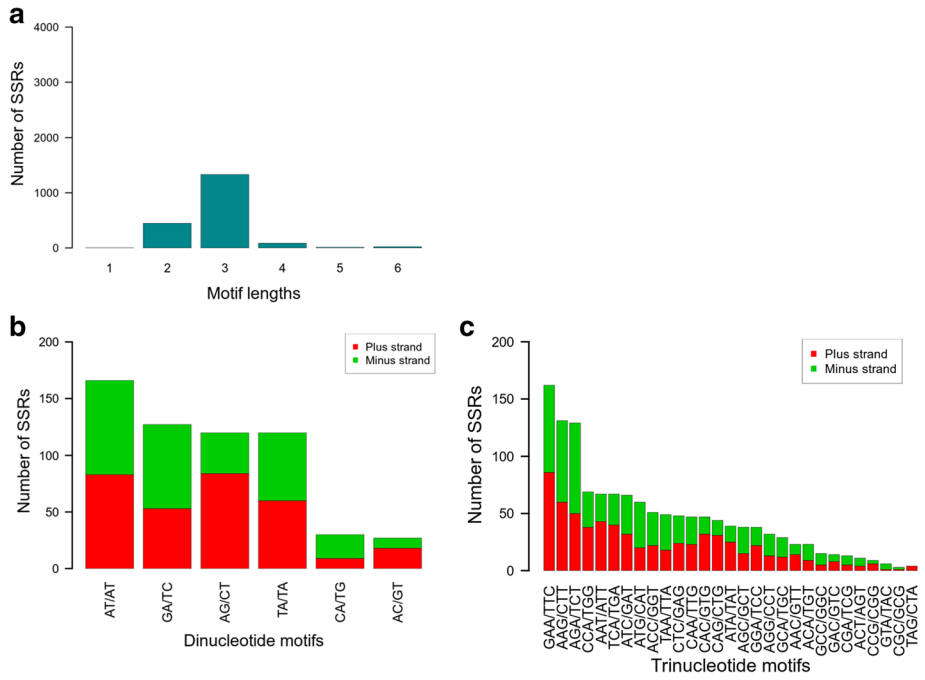

Fig. 3 SSR length and motif distribution. The number of mono- to hexanucleotide SSR motifs was counted (**a**). The nucleotide compositions of the most frequent motifs (di- and trinucleotide motifs) were determined for each type and are illustrated in a bar plot for dinucleotide (**b**) and trinucleotide motifs (**c**). Motifs that are reverse-complementary were plotted as stacked bars: "plus strand" (*red*) and "minus strand" (*green*)

reliability. For example, a study of Australian collection's diversity identified 100 SSRs within approximately 24K expressed sequence tags (ESTs) [20]. The trinucleotide motifs were more frequent than dinucleotide motifs in both the Australian collection in the present study. Moreover, the motif patterns that were reported as the preponderant ones were congruent with our observations. The trinucleotide motif, (AAG/CTT)n, was ranked as the most and second most frequent in the Australian study and in our study, respectively, and the dinucleotide motif, (AG/CT)n, was ranked as the most and third most frequent, respectively. The list of SSRs discovered might be useful for MAS and genetic surveys. However, in spite of the fact that NGS can be used for SSR discovery, high-throughput technologies (microarrays and NGS) are more available for SNPs [26, 35, 41]. Therefore, in terms of parallel genotyping the available technologies tilt the balance in favor of using SNPs as markers rather than SSRs.

In the recent years, with the evolution of next generation sequencing, many studies have developed SNP markers for marker-assisted breeding [32, 42–45]. NGS has leveraged the genome-wide SNP discovery in non-model organisms such as spruce [46], apple [47, 48], and pomegranate [31]. However, no study of SNP development for mango has been reported yet. In the present work, two mango accessions' transcriptomes ('Keitt', 'Tommy Atkins') were resequenced and aligned to a *de-novo* assembled transcriptome as a reference. The analysis resulted in the discovery of 332,016 SNPs (Additional file 3: Table S3) using VarScan [49]. The polymorphism type of those loci for the two accessions' transcriptomes can be either polymorphic, i.e., heterozygous (He) or non-polymorphic, i.e., homozygous (Ho). The possible combinations of the genotype calls for the two transcriptomes fall into four categories: both transcriptomes are homozygous (HoHo), 'Keitt' is heterozygous and 'Tommy Atkins' is homozygous (HeHo), 'Keitt' is homozygous and 'Tommy Atkins' is heterozygous (HoHe), and both transcriptomes are heterozygous (HeHe). Note that if both transcriptomes are homozygous, they are homozygous for different alleles. The distribution of SNPs into these categories was 24,136, 33,554, 164,454, and 109,872, respectively. Thus 'Tommy Atkins' is more polymorphic than 'Keitt'. As expected more SNPs were discovered in the flanking regions of the open reading frames (ORFs; hereafter outORF), than within them (hereafter inORF). The ratio of outORF to inORF SNPs was 2.18 on the average. This ratio was uniformly maintained in all SNP categories except in the HoHo category where the ratio of outORF to inORF SNPs is two and it was found to be significantly smaller (χ^2 test; df =3; p-value <0.001) than 2.18 as a result of a slight increase of inORF SNPs. Herein we described the first set of SNP markers for mango. The closest fruit tree relative of mango with a published genome, *Citrus sinensis*, is as polymorphic as mango [50]. The genome project of the sweet orange reported 1.06 million

SNPs in the entire genome with one-third are in genic regions [51]. Like orange, 70 % of the transcripts included at least one SNP while only 63 % included at least one SNP in the exonic regions. Other studies of fruit trees reported much less polymorphism in expressed regions: 6500, 71,482 and 23,742 in pomegranate [31], apple [52], and eucalyptus [27], respectively. These findings confirm previous results that mango is a highly heterozygous (or polymorphic) species [7, 12, 53].

Germplasm kinship

An overwhelming number of SNPs derived from the genic region of the genome may be useful in the future for genome-wide association studies (GWAS). However as a preliminary step to such studies, a survey of the structure and diversity of the mango collection is required [54]. A subset of 239 high quality SNPs was used for genotyping 74 accessions of the Israeli mango collection, one SNP per contig. The SNPs subset was not biased toward "inORF" or "outORF" types of SNPs (χ^2-test; df = 1; p-value = 0.74) and was therefore representative. As reported in previous studies of collections, mangoes are highly polymorphic [13, 19]. The median polymorphism information content (PIC) was 0.4 whereas less than 1 % of the applicable SNPs were of minor allele frequency (MAF) value <0.05. Thus most of the SNPs were polymorphic in the Israeli mango collection although they were discovered in only two accessions, 'Keitt' and 'Tommy Atkins'. That is reasonable presuming that mango is highly polymorphic.

The Israeli germplasm collection comprises cultivars that were originated from India, Southeast Asia, South America and the Pacific islands, Florida, Australia, and from elite local hybrids. A dendrogram based on the proportion of shared alleles distance classified the accessions in the mango collection into two genetic subgroups. The dendrogram (Fig. 4a) split the mango collection into two major clusters: 1) a small one that comprises most of the Indian accessions clustered together with accessions from Southeast Asia (SP1), 2) and a larger one which comprised of the Floridian, South African, Australian, local (Israeli) and South American accessions (SP2). This division separates Indian and Southeast Asian accessions from the rest. In other words, the mango accessions that are cultivated in the western part of the world can be genetically separated from those that are cultivated in its eastern part of the world. Due to the fact that the origin of mango has been suggested to be from the eastern part of the world [55], SP1 might be more related to the landrace mangoes.

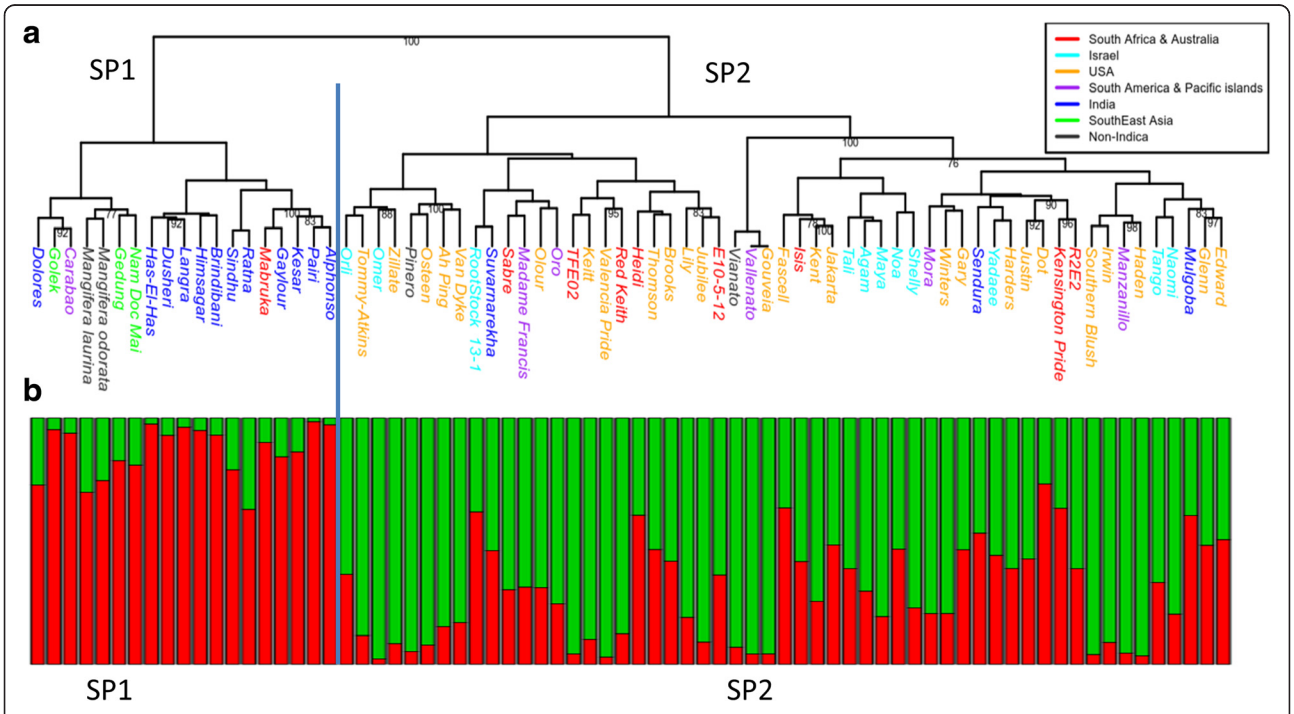

Fig. 4 Dendrogram and genetic structure of 74 accessions in the Israeli mango germplasm collection. Genotyping of 74 mango accessions from the Israeli mango collection was performed with 239 SNPs. The genotyping results were used to classify the accessions into sub-populations and reveal their genetic structure. **a** Classification was performed by drawing a dendrogram based on 1- proportion of shared alleles (PSA) as a genetic distance. Only confident branches with bootstrap values ≥90 were assigned. The two major groups are notated as subpopulation 1 (SP1) and as subpopulation 2 (SP2). **b** Genetic structure was revealed by STRUCTURE program with K = 2 as found by simulation and ΔK likelihood method. The division of STRUCTURE's Q-value bar plot into two (*vertical blue line*) corresponds to the two major significant clusters in the dendrogram. Note that the Y-axis (not plotted) scale is between 0 and 1 and represents proportion

Three accessions from India fell within the Floridian-Israeli (SP2) cluster. 'Mulgoba' was reported as the parent of the Floridian cultivar 'Haden' and as a putative parent of other Floridian variants [3]. Moreover, 'Haden' has been suggested to be the parent of many other Floridian accessions [3]. Thus, 'Mulgoba' is the ancestor of most Floridian accessions. Recently a new study was published and reported about 387 mango accessions from all over India. In that study, the cultivar 'Suvarnarekha' was reported from South India as was 'Mulgoba' and they both were clustered together in a dendrogram by their geographical origin [14]. To the best of our knowledge no record exists of the origin and the genetic similarity of the third Indian accession, 'Sendura'. Moreover, the number of subpopulations estimated by Evanno's method [56] was K = 2. Most of the accessions from India which were clustered together were genetically homogeneous (Fig. 4b; mostly red bars), while the three accessions from India that were included in the cluster with the Floridian and Israeli accessions are highly admixed (Fig. 4b; red/green bars). Ravishankar et al. [14] showed that the Indian collection can be divided into two subpopulations corresponding to the geographical classification of south/west and north/east. Moreover, the south/west can be further divided into two sub-populations. It is not clear whether the additional genetic division is correlated with south and west geographical regions. However assuming this correlation would explain the fact that in the SP1 cluster, the Indian accessions were from north, east, and two from west, while the Indian accessions from the south were included separately in the SP2 cluster (Fig. 4a).

In contrast to the mango accessions' origin, there was no clear division observed between poly- and monoembryonic accessions in the SP1 and SP2 clusters. SP1 comprised 13 and five mono- and polyembryonic accessions respectively (one was undefined). SP2 comprised 40 and 11 mono- and polyembryonic accessions respectively (four were undefined). No significant difference (Fisher's exact test; p-value = 0.47) was observed between the proportions of poly- and monoembryonic accessions in the two clusters.

Mango diversity

The two major clusters in the dendrogram were compared for their genetic diversity. The expected heterozygosity of SP1 (median = 0.28) is significantly smaller (Wilcoxon test, p-value <0.001) than the expected heterozygosity of SP2 (median = 0.43). The accessions in SP1 are in Hardy-Weinberg equilibrium (HWE) with a median F_S of -0.05 (Wilcoxon test, p-value = 0.13). In contrast, slight outbreeding was estimated for the accessions in SP2 with a median of F_S = -0.1 (Wilcoxon-test; p-value <0.001). Both F_S values were close to zero and slightly negative, suggesting that mango accessions in these clusters are not prone to inbreeding. The SP1 cluster that was enriched in accessions from Southeast Asia and India, i.e., suggested mango's origin [55], and its accessions had probably been under cultivation longer duration than the accessions in SP2. Therefore the result that they were in HWE is acceptable. In contrast the SP2 cluster deviated from the HWE. One explanation is that as a group, the accessions in the cluster as a group appeared to be under shorter duration of cultivation. Alternatively, one might suggest that SP1 is comprised of accessions that are more related to landraces (note that non-indica mangoes are included). The SP2 cluster comprises of accessions that were subjected to breeding efforts. This may be one of the reasons that SP1 is under HWE while SP2 deviates from it. A supportive evidence that SP2 is a young subpopulation lies in the estimation of a F_{ST} value that is only slightly greater than zero (median = 0.03; Wilcoxon-test; p-value <0.001) which suggests that SP2 is only in the beginning of its differentiation. Small F_{ST} values, such as the one shown in this study, were previously suggested by three other studies [11, 18, 57] of the Indian and Colombian mango collections using SSR and RAPD markers respectively. The SP2 cluster is also more diverse than the SP1 cluster. The genetic structure analysis (Fig. 4b) illustrated that accessions in the SP1 cluster have come from a narrow genetic background whereas the Indian-originated accessions in the SP2 cluster are more admixed. An optional explanation for this might relate to the possibility that the founder of the SP2 subpopulation ('Mulgoba') was probably a hybrid of the two subpopulations described in Ravishankar and colleagues' study [14] and therefore heterozygous.

Finally, two non-indica species of the genus *Mangifera* were included in this study (*Mangifera laurina* and *Mangifera odorata*); they clustered together with SP1 subpopulation that contained mainly accessions that are cultivated in Southeast Asia and India. This supports the claim that the Southeast Asia and India is the origin of *Mangifera indica* [55] and that the accessions in SP1 are closer to landraces than the ones in SP2.

Conclusions

We have established a sequence for the mango transcriptome from a pool of tissues. This transcriptome was not reconstructed to study expression but rather served as a reduction in complexity for variation discovery. It was used as a reference to align resequencing of two commercially important mango accessions, 'Keitt' and 'Tommy Atkins', constituting a resource for genetic variation discovery. The annotation and the SSR motifs were congruent with the existing knowledge in the literature. The discovered SNPs established a large pool genetic variation in mango. A subset of this pool was shown to be applicable for studying diversity in the Israeli mango collection and for dividing it into two subpopulations,

i.e., two genetic groups. The SP1 cluster comprised a Southeast Asian and Indian accessions and was suggested to arise from a narrow genetic background. Yet it was found to be in HWE, probably due to a long duration of cultivation. In contrast the SP2 cluster comprised mainly accessions cultivated in the western world except for three Indian accessions, one of which had been reported to be the ancestor of many Floridian mangoes. The structure analysis based on the SNP markers suggested that the three Indian mango accessions are an admixture. Consequently, most of the descendent cultivars are admixtures as well. In contrast to SP1 accessions, those in SP2 were not in HWE. We suggest that the different results are probably due to difference in duration of cultivation, although this was not strongly supported by the results. We believe that the novel set of SNPs is valuable for mango because that they have been polymorphic in the Israeli mango collection and they enabled us to recapitulate the mango's diversity.

Methods

Plant material

Mango accessions from the Israeli mango germplasm collection were used in this study. The collection is comprised of accessions from different regions of the world as well as promising lines identified through the Israeli breeding program. A list of the accessions that were included in this study is provided in (Additional file 4: Table S4). All accessions were 15–20 years old, grafted on the 13/1 rootstock. Trees were grown in sandy soil at the Volcani Experimental Orchard in Volcani Center, Israel. All samples were collected, immediately frozen in liquid nitrogen and stored at −80 °C until use.

RNA isolation

RNA was purified from several tissues of 'Tommy Atkins' and 'Keitt' trees (young leaves, young inflorescences, fruitlets, flesh and peel of mature fruit). Total RNA was isolated using a hexadecyltrimethyl ammonium bromide (CTAB)-based method [58]. Tissue (2–3 g) was ground in liquid nitrogen and extracted in 20 ml extraction buffer (0.1 M Tris, 25 mM EDTA, 2 % (w/v) CTAB, 0.2 % polyvinylpyrrolidone [PVP], 2 M NaCl, 0.2 % ß-mercaptoethanol, pH 8.0) pre-warmed to 65 °C. After two phenol:chloroform extractions, RNA was precipitated with 2.5 M LiCl, and resuspended with 1 ml SSTE (0.5 % SDS, 1 M NaCl, 10 mM Tris, pH 8, 1 mM EDTA, pH 8). RNA was re-extracted twice with phenol:chloroform and precipitated in 70 % ethanol. Purified RNA was treated with RQ1 RNase-free DNase I (Promega) according to the manufacturer's instructions, followed by another extraction and precipitation. The RNA was assessed for integrity and quantified on a NanoDrop spectrometer and by separation on a 1.2 % agarose gel.

Isolation of genomic DNA

Genomic DNA was isolated from young mango leaves (2 g) ground in liquid nitrogen and extracted with 15 ml of extraction buffer (100 M Tris, pH 8.0, 1.5 M NaCL, 3 % CTAB, PVP, 1 % ß-mercaptoethanol) and 15 ml of chloroform: isoamyl alcohol. Following a second extraction with chloroform: isoamyl alcohol, DNA was ethanol-precipitated, treated with 20 units of ribonuclease A (Sigma), precipitated and resuspended in water. The DNA was quantified in a NanoDrop spectrometer and by separation on a 0.8 % agarose gel.

High throughput sequencing

'Keitt' total RNA samples from the different tissues were mixed evenly and ran on one plate of the 454-Titanium platform. Construction of two cDNA libraries and 454 pyrosequencing were carried out at the W.M. Keck Center for Comparative and Functional Genomics, Roy J. Carver Biotechnology Center, University of Illinois at Urbana-Champaign. Briefly, mRNA was isolated from 20 μg of total RNA with the Oligotex kit (Qiagen, Valencia, CA). The mRNA-enriched fraction was converted to a primary cDNA library with adaptors compatible with the 454 system as previously published [59]. After library construction, the library was quantified using a Qubit fluorimeter (Invitrogen, CA) and average fragment sizes were determined by analyzing 1 μl of the samples on a Bioanalyzer (Agilent, CA) using a DNA 7500 chip. The libraries were diluted to 1×10^6 molecules/μl and pooled. Emulsion-based clonal amplification and sequencing on a full plate of the '454 Genome Sequencer FLX+' system were performed according to the manufacturer's instructions (454 Life Sciences, Branford, CT). Signal processing and base calling were performed using the bundled 454 Data Analysis Software v2.6. The read outcome was used to create a mango transcriptome as a reference for alignment of resequenced 'Keitt' and 'Tommy Atkins' total RNA mixture isolated from several tissues in equal amounts. Those RNA samples were prepared with Illumina's 'TruSeq RNAseq Sample Prep kit', quantified by qPCR, and sequenced for 100 cycles on a HiSeq 2000 using a 'TruSeq' SBS sequencing kit version 3. To get a lane-independent yields, 'Keitt' and 'Tommy Atkins' RNA samples were initially tagged and then mixed evenly and were run on two separate lanes. The sequence reads from those lanes were used for discovery of genetic variation.

De novo transcriptome assembly and functional annotation

Raw sequence reads of the 454-FLX GS Titanium platform were pre-processed by "SFF_extract" (http://bioinf.comav.upv.es/sff_extract/) and arguments for removing the adaptors and clipping the poly-A were applied. Reads were assembled by a stable version of MIRA, v3.2 [60].

For the MIRA run, we used the "Do-What-I-Mean" (DWIM) set of parameters as follows: "denovo, est, normal, 454", 'assume_snp_instead_repeat', 'clip_polyat' and 'force_nonIUPACconsensus_perseqtype' options on, and 'min_reads_per_group' = 8, 'min_neighbour_qual' = 25 and 'min_groupqual_for_rmb_tagging' = 30. One of MIRA features involves splitting mRNA unigenes into splice variants especially for polymorphic variants. Therefore a second assembly run on MIRA's contigs was performed by CAP3 [34] to produce super-contigs. Both super-contigs and the singletons, which are the MIRA's contigs were designated reference contigs. Contigs were deposited in the transcriptome shotgun assembly (TSA) sequence database [TSA: SAMN02905156, SAMN02947194].

All contigs were searched for open reading frames (ORFs) by the "getorf" program of the EMBOSS package [61]. The longest ORF with start and stop codons was chosen for each contig with a minimum cutoff of 67 amino acids.

A sequence-similarity search of contigs was run against the non-redundant (nr) protein database using blastx with a filter of e-value $<10^{-5}$. Best hits were further mapped to GO-slim by Blast2GO [62] and only hits with Blast2GO annotation score >55 were scored (Additional file 2: Table S2). Mapping of the mango peel transcripts [23] to the transcripts of the pooled tissues in this study was performed by blast search for all transcripts of the peel against pool and vice versa, and selecting the reciprocal best hits. Similar but separate mapping was performed with the transcripts of mango leaves [21].

SNP and SSR discovery

Read results from 'Keitt' and 'Tommy Atkins' mRNA resequencing using Illumina HiSeq 2000 were mapped to the 'Keitt' reference-transcriptome contigs using bowtie2 (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml). SNPs were discovered using Varscan [49].

SSR scanning was performed on the 47,956 reference contigs. MIcroSAtellite (MISA) identification tools (http://pgrc.ipk-gatersleben.de/misa/) and SciRoKo [63] were run with default parameters.

Genotyping assays

A subset of 472 SNPs was chosen for further analysis by maximizing sequence coverage of 1 SNP per contig. SNP assays for all 472 SNPs were developed by Fluidigm based on the genetic variation that was found between 'Keitt' and 'Tommy Atkins'. The assays were run according to the manufacture's instructions on an EP1 platform using '96 × 96' chips following standard Fluidigm protocols (http://www.fluidigm.com) with a minor modification of four no-template control (NTC) samples instead of one. The SNP assays were used to screen the 74 accessions' DNA samples by running on 'FR96.96' arrays of the EP1 Fluidigm platform according to the manufacturer's instructions (http://www.fluidigm.com).

Data analysis of mango diversity

To exclude bad samples and failed marker assays, samples that had more than 10 % "No Call" and assays with more than 30 % "No Call" were removed. The remaining subset was submitted for the downstream analysis. The PIC was calculated as [64].

$$\mathrm{PIC} = 1 - \sum \mathrm{p}_\mathrm{i}^2$$

where i is the i-th allele.

Germplasm accession classification and diversity

To assess the relationship between different mango accessions, we estimated the genetic distance as D = 1-proportion of shared alleles (PSA). PSA was calculated as

$$PSA = \frac{\sum_{i=1}^{L} PS_i}{2 * L}$$

where PS is the proportion of shared alleles for each locus and L is the total number of loci [65].

Hierarchical clustering was performed on a pairwise D distance matrix and the "ward" agglomerative method [66] was applied. The confidence limits of the tree topology were calculated by applying bootstrap method (1000 resampling of loci). To count the number of bipartitions fitting the tree we used the "ape" R-package [67] and presented the bootstrap values as percentages.

The subpopulation structure underlying the germplasm collection was estimated by running a simulation of STRUCTURE software v2.3.3 [68] with 5000 burn-in periods and 50,000 repetitions. The number of populations, K, was inferred by running the simulation of K = 1 to K = 10 (20 runs for each K) and using the likelihood method of ΔK [56].

The fixation indices F_S and F_{ST} [69] were calculated as

$$Fs = \frac{H_{exp} - H_{obs}}{H_{exp}}$$

where F_S is the fixation index of each subpopulation, H_{obs} is the observed heterozygous types and H_{exp} is the estimated heteozygosity under HWE,

$$F_{ST} = \frac{H_S - H_T}{H_T}$$

where F_{ST} is the genetic differentiation of a subpopulation due to genetic drift, H_S is the weighted average of all subpopulations' expected heterozygozity, and H_T is the expected heterozygosity in the entire population (germplasm collection).

Availability of supporting data

The dataset supporting the results of this article is available in the NCBI TSA (Transcriptome Shotgun Assembly Sequence Database, http://www.ncbi.nlm.nih.gov/genbank/tsa) repository under the accession numbers of BioProject: PRJNA254771, BioSample: SAMN02947194, and BioSample: SAMN02905156. These data can be found under a search in the Nucleotide repository at the NCBI site.

Additional files

Additional file 1: Table S1. Mango transcriptome annotation. (XLSX 16205 kb)

Additional file 2: Table S2. Mango simple sequence repeats. (XLSX 102 kb)

Additional file 3: Table S3. Mango SNP list. (XLSX 14669 kb)

Additional file 4: Table S4. List of accessions in the Israeli germplasm collection and metadata. (XLSX 11 kb)

Abbreviations

MAS: marker-assisted selection; NGS: next generation sequencing; ORF: open reading frame; PIC: polymorphism information content; PSA: proportion of shared alleles.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

RO: wrote the paper RO, AS and YC: Conceived and designed the experiments YC: contributed plant materials, manuscript discussion and review MR and MS: Analyzed the data – sequence annotations, clustering RO: Structure analysis and population genetics statistics RE and AR: operated the Fluidigm platform to produce genotype calls MB and MI: Perform molecular experiments – RNA and DNA extractions. DS: Field experiments that were generating the mango's populations. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by the Chief Scientist of Ministry of Agriculture and Rural Development [Grant No.: 203-0859-12].

Author details

[1]Department of Fruit Trees Sciences, Institute of Plant Sciences, Agricultural Research Organization, Volcani Center, Rishon Lezion, Israel. [2]The Robert H. Smith Institute of Plant Sciences and Genetics in Agriculture, Faculty of Agriculture, Food and Environment, The Hebrew University of Jerusalem, Rehovot, Israel.

Received: 27 April 2015 Accepted: 4 November 2015
Published online: 14 November 2015

References

1. Bompard J. Taxonomy and systematics. In: Litz RE, editor. The mango: Botany, production and uses. Wallingford: CAB International; 2009. p. 19–41.
2. Mukherjee S, Litz R. Introduction: botany and importance. In: Litz RE, editor. The mango: Botany, production and uses. 2nd ed. Wallingford: CAB international; 2009. p. 1–18.
3. Olano CT, Schnell RJ, Quintanilla WE, Campbell RJ. Pedigree analysis of Florida mango cultivars. Proc Fla State Hort Soc. 2005;118:192–7.
4. Bally IE, Lu P, Johnson P. Mango breeding. In: Jain SM, Priyadarshan PM, editors. Breeding plantation tree crops: tropical species. New York: Springer; 2009. p. 51–82.
5. Varshney RK, Graner A, Sorrells ME. Genomics-assisted breeding for crop improvement. Trends Plant Sci. 2005;10:621–30 [Trends in Plant Science 10th Anniversary Issue Feeding the World: Plant Biotechnology Milestones].
6. Kashkush K, Fang J, Tomer E, Hillel J, Lavi U. Cultivar identification and genetic map of mango (Mangifera indica). Euphytica. 2001;122:129–36.
7. Adato A, Sharon D, Lavi U, Hillel J, Gazit S. Application of DNA fingerprints for identification and genetic analyses of mango (Mangifera indica) genotypes. J Am Soc Hortic Sci. 1995;120:259–64.
8. Eiadthong W, Yonemori K, Sugiura A, Utsunomiya N, Subhadrabandhu S. Identification of mango cultivars of Thailand and evaluation of their genetic variation using the amplified fragments by simple sequence repeat-(SSR-) anchored primers. Sci Hortic. 1999;82:57–66.
9. Schnell RJ, Olano CT, Quintanilla WE, Meerow AW. Isolation and characterization of 15 microsatellite loci from mango (Mangifera indica L.) and cross-species amplification in closely related taxa. Mol Ecol Notes. 2005;5:625–7.
10. Ahmad Rajwana I, Tabbasam N, Malik AU, Malik SA, Mehboob-ur-Rahman, Zafar Y. Assessment of genetic diversity among mango (Mangifera indica L.) genotypes using RAPD markers. Sci Hortic. 2008;117:297–301.
11. Díaz-Matallana M, Schuler-García I, Ruiz-García M, Hodson de Jaramillo E. Analysis of diversity among six populations of Colombian mango (Mangifera indica L. cvar. Hilacha) using RAPDs markers. Electron J Biotechnol. 2009;12:1–2.
12. Hirano R, Htun Oo T, Watanabe KN. Myanmar mango landraces reveal genetic uniqueness over common cultivars from Florida, India, and Southeast Asia. Genome. 2010;53:321–30.
13. Dillon NL, Bally ISE, Wright CL, Hucks L, Innes DJ, Dietzgen RG. Genetic diversity of the Australian National Mango Genebank. Sci Hortic. 2013;150:213–26.
14. Ravishankar KV, Bommisetty P, Bajpai A, Srivastava N, Mani BH, Vasugi C, et al. Genetic diversity and population structure analysis of mango (Mangifera indica) cultivars assessed by microsatellite markers. Trees. 2015;1–9.
15. Schnell RJ, Ronning CM, Jr RJK. Identification of cultivars and validation of genetic relationships in Mangifera indica L. using RAPD markers. Theor Appl Genet. 1995;90:269–74.
16. Ravishankar KV, Anand L, Dinesh MR. Assessment of genetic relatedness among mango cultivars of India using RAPD markers. J Hortic Sci Biotechnol. 2000;75:198–201.
17. Karihaloo J, Dwivedi Y, Archak S, Gaikwad AB. Analysis of genetic diversity of Indian mango cultivars using RAPD markers. J Hortic Sci Biotechnol. 2003;78:285–9.
18. Viruel M, Escribano P, Barbieri M, Ferri M, Hormaza J. Fingerprinting, embryo type and geographic differentiation in mango (Mangifera indica L., Anacardiaceae) with microsatellites. Mol Breed. 2005;15:383–93.
19. Surapaneni M, Vemireddy LR, Begum H, Reddy BP, Neetasri C, Nagaraju J, et al. Population structure and genetic analysis of different utility types of mango (Mangifera indica L.) germplasm of Andhra Pradesh state of India using microsatellite markers. Plant Syst Evol. 2013;299:1215–29.
20. Dillon NL, Innes DJ, Bally IS, Wright CL, Devitt LC, Dietzgen RG. Expressed sequence tag-simple sequence repeat (EST-SSR) marker resources for diversity analysis of mango (Mangifera indica L.). Diversity. 2014;6:72–87.
21. Azim MK, Khan IA, Zhang Y. Characterization of mango (Mangifera indica L.) transcriptome and chloroplast genome. Plant Mol Biol. 2014;85:193–208.
22. Pandit SS, Kulkarni RS, Giri AP, Köllner TG, Degenhardt J, Gershenzon J, et al. Expression profiling of various genes during the fruit development and ripening of mango. Plant Physiol Biochem. 2010;48:426–33.
23. Luria N, Sela N, Yaari M, Feygenberg O, Kobiler I, Lers A, et al. De-novo assembly of mango fruit peel transcriptome reveals mechanisms of mango response to hot water treatment. BMC Genomics. 2014;15:957.
24. Wu H, Jia H, Ma X, Wang S, Yao Q, Xu W, et al. Transcriptome and proteomic analysis of mango (Mangifera indica Linn) fruits. J Proteomics. 2014;105:19–30.
25. Dautt-Castro M, Ochoa-Leyva A, Contreras-Vergara CA, Pacheco-Sanchez MA, Casas-Flores S, Sanchez-Flores A, et al. Mango (Mangifera indica L.) cv. Kent fruit mesocarp de novo transcriptome assembly identifies gene families important for ripening. Front Plant Sci. 2015;6:62 [*Plant Genetics and Genomics*].
26. Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet. 2011;12:499–510.
27. Novaes E, Drost DR, Farmerie WG, Pappas Jr GJ, Grattapaglia D, Sederoff RR, et al. High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome. BMC Genomics. 2008;9:312.
28. Barakat A, DiLoreto DS, Zhang Y, Smith C, Baier K, Powell WA, et al. Comparison of the transcriptomes of American chestnut (Castanea dentata) and Chinese chestnut (Castanea mollissima) in response to the chestnut blight infection. BMC Plant Biol. 2009;9:51.
29. Alagna F, Agostino ND', Torchia L, Servili M, Rao R, Pietrella M, et al. Comparative 454 pyrosequencing of transcripts from two olive genotypes during fruit development. BMC Genomics. 2009;10:399.

30. Iorizzo M, Senalik DA, Grzebelus D, Bowman M, Cavagnaro PF, Matvienko M, et al. De novo assembly and characterization of the carrot transcriptome reveals novel genes, new markers, and genetic diversity. BMC Genomics. 2011;12:389.
31. Ophir R, Sherman A, Rubinstein M, Eshed R, Sharabi Schwager M, Harel-Beja R, et al. Single-nucleotide polymorphism markers from de-novo assembly of the pomegranate transcriptome reveal germplasm genetic diversity. PLoS ONE. 2014;9:e88998.
32. Chen H, Xie W, He H, Yu H, Chen W, Li J, et al. A high-density SNP genotyping array for rice biology and molecular breeding. Mol Plant. 2014;7:541–53.
33. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet. 2011;12:443–51.
34. Huang X, Madan A. CAP3: A DNA sequence assembly program. Genome Res. 1999;9:868–77.
35. McCouch SR, Zhao K, Wright M, Tung C-W, Ebana K, Thomson M, et al. Development of genome-wide SNP assays for rice. Breed Sci. 2010;60:524–35.
36. Zhang W-W, Pan J-S, He H-L, Zhang C, Li Z, Zhao J-L, et al. Construction of a high density integrated genetic map for cucumber (Cucumis sativus L.). Theor Appl Genet. 2012;124:249–59.
37. Ravishankar KV, Mani BH-R, Anand L, Dinesh MR. Development of new microsatellite markers from Mango (Mangifera indica) and cross-species amplification. Am J Bot. 2011;98:e96–9.
38. Tsai C-C, Chen Y-KH, Chen C-H, Weng IS, Tsai C-M, Lee S-R, et al. Cultivar identification and genetic relationship of mango (Mangifera indica) in Taiwan using 37 SSR markers. Sci Hortic. 2013;164:196–201.
39. von Stackelberg M, Rensing SA, Reski R. Identification of genic moss SSR markers and a comparative analysis of twenty-four algal and plant gene indices reveal species-specific rather than group-specific characteristics of microsatellites. BMC Plant Biol. 2006;6:9.
40. Sharma PC, Grover A, Kahl G. Mining microsatellites in eukaryotic genomes. Trends Biotechnol. 2007;25:490–8.
41. Cutler DJ, Zwick ME, Carrasquillo MM, Yohn CT, Tobin KP, Kashuk C, et al. High-throughput variation detection and genotyping using microarrays. Genome Res. 2001;11:1913–25.
42. Edwards JD, Janda J, Sweeney MT, Gaikwad AB, Liu B, Leung H, et al. Development and evaluation of a high-throughput, low-cost genotyping platform based on oligonucleotide microarrays in rice. Plant Methods. 2008;4:13.
43. Verde I, Bassil N, Scalabrin S, Gilmore B, Lawley CT, Gasic K, et al. Development and evaluation of a 9K SNP array for peach by internationally coordinated SNP detection and validation in breeding germplasm. PLoS ONE. 2012;7:e35668.
44. Xu Y, Lu Y, Xie C, Gao S, Wan J, Prasanna BM. Whole-genome strategies for marker-assisted plant breeding. Mol Breed. 2012;29:833–54.
45. Riedelsheimer C, Melchinger AE. Optimizing the allocation of resources for genomic selection in one breeding cycle. Theor Appl Genet. 2013;126:2835–48.
46. Pavy N, Gagnon F, Rigault P, Blais S, Deschênes A, Boyle B, et al. Development of high-density SNP genotyping arrays for white spruce (Picea glauca) and transferability to subtropical and nordic congeners. Mol Ecol Resour. 2013;13:324–36.
47. Chagné D, Crowhurst RN, Troggio M, Davey MW, Gilmore B, Lawley C, et al. Genome-wide SNP detection, validation, and development of an 8K SNP array for apple. PLoS ONE. 2012;7:e31745.
48. Troggio M, Gleave A, Salvi S, Chagné D, Cestaro A, Kumar S, et al. Apple, from genome to breeding. Tree Genet Genomes. 2012;8:509–29.
49. Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics. 2009;25:2283–5.
50. Malik SK, Rohini MR, Kumar S, Choudhary R, Pal D, Chaudhury R. Assessment of genetic diversity in sweet orange [Citrus sinensis (L.) Osbeck] cultivars of India using morphological and RAPD markers. Agric Res. 2012;1:317–24.
51. Xu Q, Chen L-L, Ruan X, Chen D, Zhu A, Chen C, et al. The draft genome of sweet orange (Citrus sinensis). Nat Genet. 2013;45:59–66.
52. Chagné D, Gasic K, Crowhurst RN, Han Y, Bassett HC, Bowatte DR, et al. Development of a set of SNP markers present in expressed genes of the apple. Genomics. 2008;92:353–8.
53. Chiang Y-C, Tsai C-M, Chen Y-KH, Lee S-R, Chen C-H, Lin Y-S, et al. Development and characterization of 20 new polymorphic microsatellite markers from Mangifera indica (Anacardiaceae). Am J Bot. 2012;99:e117–9.
54. Shriner D. Investigating population stratification and admixture using eigenanalysis of dense genotypes. Heredity. 2011;107:413–20.
55. Mukherjee SK. Origin of mango (Mangifera indica). Econ Bot. 1972;26:260–4.
56. Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software structure: a simulation study. Mol Ecol. 2005;14:2611–20.
57. Singh S, Bhat KV. Molecular characterization and analysis of geographical differentiation of Indian mango (Mangifera indica L.) germplasm. In: I International Symposium on Biotechnology of Fruit Species: BIOTECHFRUIT2008 839. 2008. p. 599–606.
58. Chang S, Puryear J, Cairney J. A simple and efficient method for isolating RNA from pine trees. Plant Mol Biol Report. 1993;11:113–6.
59. Lambert JD, Chan XY, Spiecker B, Sweet HC. Characterizing the embryonic transcriptome of the snail Ilyanassa. Integr Comp Biol. 2010;50:768–77.
60. Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Muller WE, Wetter T, et al. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Res. 2004;14:1147–59.
61. Rice P, Longden I, Bleasby A. EMBOSS: the European molecular biology open software suite. Trends Genet. 2000;16:276–7.
62. Gotz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, et al. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 2008;36:3420–35.
63. Kofler R, Schlotterer C, Lelley T. SciRoKo: a new tool for whole genome microsatellite search and investigation. Bioinformatics. 2007;23:1683–5.
64. Weir BS. Genetic data analysis. Methods for discrete population genetic data. Sunderland: Sinauer Associates, Inc. Publishers; 1990.
65. Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, Cavalli-Sforza LL. High resolution of human evolutionary trees with polymorphic microsatellites. Nature. 1994;368:455–7.
66. Odong TL, van Heerwaarden J, Jansen J, van Hintum TJ, van Eeuwijk FA. Determination of genetic structure of germplasm collections: are traditional hierarchical clustering methods appropriate for molecular marker data? Theor Appl Genet. 2011;123:195–205.
67. Paradis E, Claude J, Strimmer K. APE: analyses of phylogenetics and evolution in R language. Bioinformatics. 2004;20:289–90.
68. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155:945–59.
69. Wright S. Genetical structure of populations. Nature. 1950;166:247–49.