Please cite this article as: Michal *et. al.*, (2016). Experimental Pipeline for SNP and SSR Discovery and Genotyping Analysis of Mango (*Mangifera indica* L.), *Bio-protocol* 6 (16): e1910. DOI: 10.21769/BioProtoc.1910.

Experimental Pipeline for SNP and SSR Discovery and Genotyping Analysis of Mango (*Mangifera indica* L.)

Michal Sharabi-Schwager[#], Mor Rubinstein[#], Mazal Ish shalom, Ravit Eshed, Ada Rozen, Amir Sherman, Yuval Cohen and Ron Ophir*

Department of Fruit Trees Sciences, Institute of Plant Sciences, Agricultural Research Organization, Volcani Center, Rishon Lezion, Israel

*For correspondence: ron@agri.gov.il

[#]Contributed equally to this work

[Abstract] Establishing a reservoir of polymorphic markers is an important key for marker-assisted breeding. Many crops are still lack of such genomic infrastructure. Single nucleotide polymorphisms (SNPs) and simple sequence repeats (SSRs) are useful as markers because they are widespread over the genome and many technologies were developed for high throughput genotyping. We present here a pipeline for developing a reservoir of SNP and SSR markers for *Mangifera indica* L. as an example for fruit tree crops having no genomic information available. Our pipeline includes *de novo* assembly of reference transcriptome with MIRA and CAP3 based on reads produced by 454-GS FLX technology; Polymorphic loci discovery by alignment of Illumina resequencing to the transcriptome reference; Identifying a subset of loci that are polymorphic in the entire germplasm collection for downstream diversity analysis by genotyping with Fluidigm technology.

[Background] Considerations of high-throughput sequencing: This pipeline does not include RNA/DNA extraction and other molecular biology lab protocols for next generation sequencing (NGS). It is common to outsourcing NGS. Therefore, it includes DNA preparation for genotyping only. Before describing the pipeline below, we would like to comment about the considerations regarding the sequencing.

Assumption: In this pipeline, we assume a non-model organism which has no genomic infrastructure at all. For marker discovery, one will need a reference and resequencing to discover the polymorphism. The ultimate reference is a genome. However, due to the fact that having a good draft or a complete reference genome is still expensive task our recommendation is to sequence a reference transcriptome from a pool of tissues. The pool of tissues should compensate the unequal gene representation as a result of tissue-specific expression.

Technology: For the purpose of a reference transcriptome sequencing, 454-GS Flx Titanium or any long reads NGS technology is preferred. For marker discovery by resequencing, a pool of genomic DNA (gDNA) from the population under study is a cost-effective solution. Polymorphic loci in such pool are a representative sample of the polymorphic loci in the population. Here the important factor is the reads' depth which should strive to an average coverage of 50x and no less than 20x. In a case of large genomes the choice of gDNA resequencing might be too expensive to get coverage of 50x.

Please cite this article as: Michal *et. al.*, (2016). Experimental Pipeline for SNP and SSR Discovery and Genotyping Analysis of Mango (*Mangifera indica* L.), *Bio-protocol* 6 (16): e1910. DOI: 10.21769/BioProtoc.1910.

Alternatively, mRNA extraction of a pool of tissues and population individuals would be a cheaper option.

The aim of this protocol is to provide a pipeline (Figure 1) for the bioinformatics and genomics support unit that assist the breeder of a crop which has no genomic information to establish a set of polymorphic SNP and SSR markers. This set can be used for marker-assisted breeding studies as well as for exploring the diversity in the crop's germplasm collection diversity.

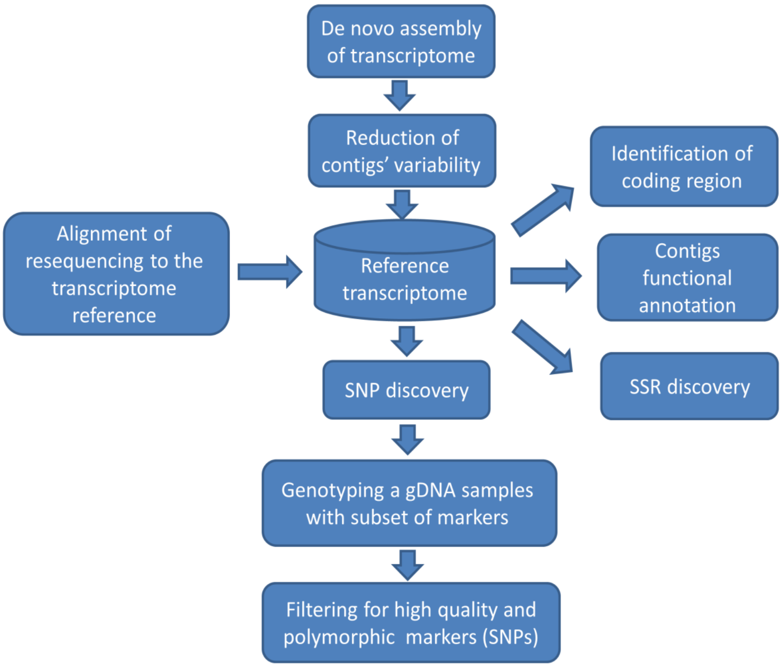
De novo assembly of transcriptome
Reduction of contigs' variability
Alignment of resequencing to the transcriptome reference
Reference transcriptome
Identification of coding region
Contigs functional annotation
SSR discovery
SNP discovery
Genotyping a gDNA samples with subset of markers
Filtering for high quality and polymorphic markers (SNPs)

Figure 1. Flowchart of a pipeline for marker discovery. The reference transcriptome here (represented as a database shape) is the link connecting function annotation with genetic variation.

Materials and Reagents

1. 50 ml Falcon tube
2. Young leaf tissue
3. Tris (Amresco, catalog number: 77-86-1)
4. EDTA (Sigma-Aldrich, catalog number: E5134)
5. NaCl (Sigma-Aldrich, catalog number: S3014)
6. 3% CTAB (Hexadecylrimethylammonium bromide) (Sigma-Aldrich, catalog number: H5882)
7. 2% polyvinylpyrolidone (PVP) (MW 40,000) (Amresco, catalog number: 9003-39-8)
8. 1% β-mercaptoethanol (Sigma-Aldrich, catalog number: M3148)
9. 5 M ammonium acetate (Sigma-Aldrich, catalog number: A1542)
10. Chloroform:isoamyl alcohol mix [24:1 (v:v)]
11. Isopropanol (stored at -20 °C)

12. Ethanol
13. RNase A (> 70 Kunit/mg protein, > 20 mg protein/ml) (Sigma-Aldrich, catalog number: R4642)
14. Extraction buffer (see Recipes)
15. TE buffer (see Recipes)

Equipment

1. 65 °C water bath
2. 37 °C water bath/block
3. IKA-A11 analytical grinding mill (IKA®-Werke GmbH & Co. KG)
4. Cooled centrifuge (Sorvall RC5plus) with Fixed Angle Rotor (Fiberlite™ F13-14 x 50cy) (Thermo Fisher Scientific, catalog number: 096-1450).
5. Agarose gel apparatus
6. Nanodrop spectrophotometer
7. Recommended hardware specifications (for bioinformatics pipeline)
   a. CPU
      Architecture: x86_64
      CPU op-mode(s): 64-bit, 8 cores, Thread(s) per core: 2
      Vendor ID: GenuineIntel
      CPU MHz: 1596.000
   b. Memory
      MemTotal: 48 GB
      SwapTotal: 4GB

Software

1. "Sff_extract" (https://bioinf.comav.upv.es/sff_extract/) – Converting and preprocessing, *e.g.*, adapter removal and base-call clipping 454-GS FLX raw files to text formats (fasta and quality).
   Note: Sff_extract is now part of the tool set seq_crumbs (https://bioinf.comav.upv.es/seq_crumbs/)
2. MIRA (https://sourceforge.net/projects/mira-assembler/) – A multi-pass DNA sequence data assembler/mapper for whole genome and/or transcriptome projects. MIRA is a multi-platforms assembler capable assembling reads from a combination of platforms or from each platform separately.
3. CAP3 (http://seq.cs.iastate.edu/cap3.html) – CAP3 is for small-scale assembly of sequences with or without quality values.
4. Trimmomatic (http://www.usadellab.org/cms/?page=trimmomatic) – Trimmomatic is a fast, multi-threaded command line tool that can be used to trim and crop Illumina (FASTQ) data as well as to remove adapters.

5. FASTX (http://hannonlab.cshl.edu/fastx_toolkit/) – Preprocessing, *e.g.*, adapter removal and base-call clipping, short reads (fastq files).

   Note: FASTX-Toolkit is a collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing.
6. Bowtie2 (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml) – Alignment of short reads to a reference genome/transcriptome.
7. Samtools (http://www.htslib.org/) – SAM (Sequence Alignment/Map) format is a generic format for storing large nucleotide sequence alignments. SAMTools provide various utilities for manipulating alignments in the SAM format, including sorting, merging, indexing and generating alignments in a per-position format.
8. Getorf (http://emboss.sourceforge.net/download/) – EMBOSS tool for identification of open reading frame ORF in mRNA sequence.
9. MIcroSAtellite (MISA) identification tool (http://pgrc.ipk-gatersleben.de/misa) – This tool allows the identification and localization of perfect microsatellites as well as compound microsatellites which are interrupted by a certain number of bases.
10. SciRoKo (http://kofler.or.at/bioinformatics/SciRoKo/index.html) – A tool for fast whole-genome microsatellite search. For example, the whole rice genome may be searched in 55 sec.
11. VarScan (http://varscan.sourceforge.net/) – VarScan is a platform-independent mutation caller for targeted, exome, and whole-genome resequencing data generated on Illumina, SOLiD, Life/PGM, Roche/454, and similar instruments.

Data analysis

A. Data

1. 454-GS FLX Titanium mRNA contigs were deposited in transcriptome shotgun assembly (TSA) repository of NCBI: accession No. GBJO00000000 (Sherman *et al.*, 2015).
2. Illumina short reads were deposited in short reads archive (SRA) of NCBI: experiment accession No. SRX651793 GBJO00000000 (Sherman *et al.*, 2015).

B. *De novo* transcriptome assembly

1. Raw sequence reads of the 454-GS FLX Titanium platform were pre-processed by “Sff_extract” (https://bioinf.comav.upv.es/sff_extract/) and arguments for removing the adaptors and clipping the poly-A were applied.

```
# Command line example:
sff_extract -i "pool.sff{species:mango} " -s mango_in.454.fasta -q mango_in.454.fasta.qual -x
mango_traceinfo_in.454
```

2. *De novo* assembly with MIRA 3.2 (Chevreux *et al., 2004*)

```
#Command line example
mira -project=mango -job=denovo,est,normal,454 -CO:mr=yes:asir=yes 454_SETTINGS -ED:ace=no
-LR:mxti=yes -CL:cpat=on -CO:fnicpst=yes:mrpg=8:mnq=25:mgqrt=30

# Mira (for version 4 or later) manifest file for de novo assembly of the 454 reads only
project = mango
job = denovo,est,normal,454
parameters = COMMON_SETTINGS -CO:mr=yes:asir=yes 454_SETTINGS -ED:ace=no -LR:mxti=yes
-CL:cpat=on -CO:fnicpst=yes:mrpg=8:mnq=25:mgqrt=3
readgroup = group_name
data = /path/to/SP1_in.454.fasta
data = /path/to/SP1_in.454.fasta.qual
technology = 454
```

3. Reduction of contig variability (merging transcript variants) by running Cap3 and creating super-contigs
   Note: Cap3 is downloaded separately from MIRA (see Software list section).

```
#Command line example
cap3 mango.fasta mango.qual
```

Note: mango.fasta and mango.qual are output files of MIRA, created in the mango_d_results directory.

4. Filtering out contigs with length less than 200 bp
   Refer to the fasta file from here on as reference.transcriptome.contigs.fasta

C. Functional annotation

1. Identifying the coding region to annotate marker position, *i.e.*, inside or outside coding sequence. Finding open reading frames (ORFs) by the "getorf" program of the EMBOSS package (Rice *et al*., 2000). The longest ORF with start and stop codons was chosen for each contig (-find 1) with a minimum cutoff of 50 amino acids (-minisize 150).
   Note: The argument (-minimize) is given in base pairs (50 bp x 3 = 150 bp).

```
#Command line example
getorf -sequence reference.transcriptome.contigs.fasta -minisize 150 -find 1 -outseq
reference.transcriptome.proteins.fasta
```

2. Reference transcriptome contigs annotation to connect variability with functionality using Blast 2GO (Gotz *et al.,* 2008).

   Blast2GO GUI options:

   a. Start → load sequences (*e.g.,* fasta)
   b. Blast → Run Blast Description Annotator
   c. Mapping → Run mapping
   d. Annot → Run annotation
   e. InterPro → Run interproscan

```
#Command line example:
blastall –p blastx –i contigs.fasta.file –d blast.formated.nr.database –e 1e-5 –m 7 –o local_blast.xml
# Alternatively, the new blast example
blastx -db   blast.formated.nr.database -outfmt 5 -evalue 1e-5   -out local_blast.xml -query
contigs.fasta.file
#Blast2GO run
java -cp *:ext/*: es.blast2go.prog.B2GAnnotPipe -in BlastResults.xml -out results/myproject -prop
b2gPipe.properties -v -annot -dat -img -ips ipsr -annex -goslim -wiki html_template.html
```

D. SNP and SSR discovery

Adapter removal and low-quality base pairs clipping are performed by Trimmomatic (Bolger *et al.*, 2014) and FASTX (http://hannonlab.cshl.edu/fastx_toolkit/).

Note: Optional but highly recommended if the alignment is performed on RNA-Seq.

```
#Command line example
A. Quality and adapters trimming with Trimmomatic

java -jar /path/to/trimmomatic-0.33.jar PE /path/to/sample_name_L001_R1.fastq path/to/sample_name
_L001_R2.fastq /path/to/sample_name_pair_L001_R1.fastq
/path/to/sample_name_sngl_L001_R1.fastq /path/to/sample_name_pair_L001_R2.fastq
/path/to/sample_name_sngl_L001_R2.fastq
ILLUMINACLIP:/path/to/adapters/adapters.fasta:2:30:10:1:true LEADING:3 TRAILING:3
MAXINFO:30:0.8 MINLEN:30

B. Clipping first 3 nt (if required) with FASTX

fastx_trimmer -Q33 -v -f 4 -i /path/to/sample_name_pair_L001_R1.fastq   -o
/path/to/sample_name_pair_trim_L001_R1.fastq
```

```
fastx_trimmer -Q33 -v -f 4 -i /path/to/sample_name_pair_L001_R2.fastq -o
/path/to/sample_name_pair_trim_L001_R2.fastq
```

1. Use R1 and R2 of trimmed pair files, *i.e.*, sample_name_pair_L001_R1.fastq, sample_name_pair_L001_R2.fastq for downstream analysis.
2. Alignment of resequencing of Illumina HiSeq-2000 reads to the transcriptome reference with bowtie2 (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml).

```
#Command line example
bowtie2-build -f reference.transcriptome.contigs.fasta   ref
#Note: reference.transcriptome.contigs.fasta is the reference file name. ref is the reference name as it should be referred later in downstream analysis.
bowtie2 -q --phred33 -t -x ref -1 /path/to/sample_name_pair_L001_R1.fastq -2
/path/to/sample_name_pair_L001_R2.fastq   -S sample_name_ref.sam
```

3. Running samtools (http://www.htslib.org/) and VarScan (Koboldt *et al., 2009)* for SNP discovery
   Note: The criteria for selecting an SNP subset are dependent on the project. However, a few criteria are advisable to ensure confident SNP loci in any project:
   a. No other SNPs in the flanking regions (100 bp each side) to enable of primer design for further analyses (Absolute differences of the SNP value between the previous and next values in 'Position' column should > 100).
   b. Only one SNP per reference-transcriptome contig (unique value at 'Chrom' column).
   c. Bi-allelic confidence ('SamplesHet' value > 0).

```
#Command line example
samtools faidx reference.transcriptome.contigs.fasta
samtools view -bS S sample_name_ref.sam   -o S sample_name_ref.bam
samtools sort READS_ref.bam sample_name_ref_sorted
samtools index sample_name_ref_sorted.bam
samtools mpileup –f   reference.transcriptome.contigs.fasta sample_name_ref_sorted.bam –o ref_
sample_name_ref.mpileup
java -jar VarScan.v2.3.9.jar mpileup2snp ref_READS.mpileup > X.varscan &
```

4. SSR discovery within the contigs the reference transcriptome.
   MIcroSAtellite (MISA) identification tool (http://pgrc.ipk-gatersleben.de/misa) and SciRoKo (Kofler *et al.*, 2007) are run with default parameters.

```
# MISA command line
perl misa.pl reference.transcriptome.contigs.fasta
```

#misa 'ini' file

```
definition(unit_size,min_repeats):       1-10 2-6 3-5 4-5 5-5 6-5
interruptions(max_difference_between_2_SSRs):             100
```

```
# SciRoKo 3.4 run
1. Press button: LoadFastaFile → choose: reference.transcriptome.contigs.fasta
2. Press button: SaveResults → set an output filename for example: ssr_results.txt
3. Press button: StartSearch
4. Press button: LoadSSRs → choose: ssr_results.txt
5. Set 'Search Results' radio button
6. Copy and paste the table in the window to an excel file.
7. Split contig ID and contig description
```

5. Find the intersection between two tables by importing them into MS-Access or SQLite and run a SQL inner join command on contig-name, motif, and start-position.

```
#MS-Access SQL command example
SELECT SciRoKo_SSR.Sequence_Name AS ID, SciRoKo_SSR.Motif,
MISA_SSR.NumberOfRepeats, SciRoKo_SSR.SSR_Start AS SciRoKo_start,
SciRoKo_SSR.SSR_End AS SciRoKo_end, MISA_SSR.start AS misa_start, MISA_SSR.end AS
misa_end, MISA_SSR.size
FROM SciRoKo_SSR INNER JOIN MISA_SSR
ON
(SciRoKo_SSR.SSR_Start = MISA_SSR.start) AND (SciRoKo_SSR.Motif = MISA_SSR.Motif) AND
(SciRoKo_SSR.Sequence_Name = MISA_SSR.ID);
```

6. Genotyping with Fluidigm
   a. Large-scale genomic DNA extraction for sample genotyping isolated from young leaves
      i. Young developing Mango (*Mangifera indica* L.) leaves were collected from the orchard, frozen in liquid nitrogen and stored at -80 °C until used.
      ii. β-mercaptoethanol was added to extraction buffer which was pre-heated to 65 °C in a pre-warmed water bath.
      iii. 2 g of young leaf tissue was ground to a fine powder using IKA-A11 analytical grinding mill or with a mortar and pestle.
      iv. Ground tissue was transferred to a 50 ml Falcon tube and extracted with 15 ml pre-warmed extraction buffer. Extraction was performed by incubation for 30 min at 65 °C, with occasional mixing of the tube.

v. 15 ml of chloroform:isoamyl alcohol mix (24:1, v:v) was added to tubes. Samples were mixed and centrifuged at 17,000 *x g* for 10 min at 4 °C.

vi. The aqueous phase was transferred to a new 50 ml tube, and re-extracted with 15 ml of chloroform:isoamyl alcohol mix (24:1, v:v). Centrifugation was performed as above.

vii. The aqueous phase was transferred to a new tube. 1 volume of ice-cold isopropanol was added, tubes were mixed and DNA was precipitated by centrifugation at 17,000 *x g* for 20 min at 4 °C.

viii. Supernatant was carefully disposed of. Pellet was washed with 70% ice-cold ethanol, and centrifuged at 17,000 *x g* for 10 min at 4 °C.

ix. Supernatant was carefully disposed of. Pellet was left to dry at room temperature until it turns translucent), and suspended in 3 ml of TE buffer.

x. DNA solution was treated with 3 µl RNase A, and incubated for 30 min in 37 °C.

xi. DNA is precipitated by adding 1/10 volume of 5 M ammonium acetate, and 2/5 volumes of cold 100% ethanol. Tubes are mixed and centrifuged at 17,000 *x g* for 10 min at 4 °C.

xii. The supernatant was carefully disposed of. Pellet was washed with 70% ice-cold ethanol, and centrifuged at 17,000 *x g* for 10 min at 4 °C.

xiii. Pellet is air dried and final DNA is suspended in 200 µl of double distilled water or TE buffer.

xiv. DNA concentration and quality is analyzed on a 0.7% TAE agarose gel and with a Nanodrop spectrophotometer.

b. Genotyping on Fluidigm – EP1 Fluidigm standard protocols for FR96.96 chip with four no-template controls (NTCs) instead of one.

Briefly, the protocol is divided into two major sub-protocols – pre-amplification and the assay itself:

i. First, specific target amplification (STA) protocol is performed to have an approximately equal proportion from each target by running the following steps:

1) Preparing the 10x SNPtype STA. Primer pool for 96 assays.
2) Performing STA on a PCR machine.
3) Dilution of samples (the outcome will be used in stage 4 of the second part).

ii. Second, the assay of genotyping by specific target primers is performed in a Fluidigm 96.96 dynamic genotyping array on the EP1 platform as follow:

1) Priming the 96.96 Dynamic Array^{TM} IFC.
2) Preparing SNPtype assays mixes.
3) Preparing 10x Assays.
4) Preparing Sample Pre-Mix and Samples.
5) Loading the Chip.
6) Using the FC1^{TM} Cycler.
7) Using the EP1^{TM} Reader Data Collection Software.

8) Extracting the data for downstream bioinformatics analysis.

The full protocol description can be found at (http://www.mscience.com.au/upload/pages/fluidigmtech/fluidigm-snp-genotyping-user-guide-151112.pdf).

7. Filtering qualified SNPs for diversity analysis

Fluidigm genotype calls are divided into four categories by visual inspection:

a. Filtering out SNPs with a Category ≥ 2 (Table 1).
b. Filtering out SNPs with more than 10% no calls.
c. Filtering out samples with more than 33% no calls.
d. Filtering out markers with PIC < 0.1.
e. Filtering out markers with more than 90% of the samples have the same call, *i.e.*, segregating exactly the same.
f. Filtering out markers with less than 2 samples in each genotype.
g. Leaving only one marker from each pair of linked markers (R^2 > 0.7).
h. Leaving only one sample from group of sample having identical genotype. (Identity ≥ 0.95)

**Notes:*

a. These steps can be performed with any programming language, e.g., R, python, perl, C, etc. or SQL.

b. PIC is calculated as PIC = 1- ∑ p_i^2; i = a, A

*c. R^2 is calculated as r^2 = D^2/(p1*p2*q1*q2); D = (p11*p22)-(p12p21 p11,p22,p12,p21 are the proportions of all possible combinations of two bi-allelic loci.*

Table 1. Quality scores of each locus genotyping calls given by visual inspection.

Category	Category grade	Description
1	Excellent	Perfect- centered and separated 3 genotype clusters
2	Good	Clusters are separated but close to each other and at least one cluster is highly scattered
3	Moderate	Two genotype clusters, highly scattered, no separation between genotype
4	Declined	No calls (NC), high NTC

Recipes

1. Extraction buffer
   100 Tris, pH 8
   20 M EDTA
   1.5 M NaCl
   3% hexadecylrimethylammonium bromide (CTAB)
   2% polyvinylpyrolidone (PVP)
   1% β-mercaptoethanol
   All solution except β-mercaptoethanol are dissolved by stirring over several hours, and autoclaved. β-mercaptoethanol is added just prior to tissue extraction.
2. TE buffer
   10 mM Tris, pH 8.0
   1 mM EDTA

Acknowledgments

The protocol has been developed in a study which was supported by the Chief Scientist of Ministry of Agriculture and Rural Development [Grant No.: 203-0859-12].

References

1. Bolger, A. M., Lohse, M. and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15): 2114-2120.
2. Bowtie - An ultrafast memory-efficient short read aligner. *JOHNS HOPKINS University.*
3. Chevreux, B., Pfisterer, T., Drescher, B., Driesel, A. J., Muller, W. E., Wetter, T. and Suhai, S. (2004). Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res* 14(6): 1147-1159.
4. Fluidigm SNP Genotyping Guide. *Fluidigm.*
5. FASTX-Toolkit. *Hannonlab.*
6. Gotz, S., Garcia-Gomez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., Robles, M., Talon, M., Dopazo, J. and Conesa, A. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res* 36(10): 3420-3435.
7. Koboldt, D. C., Chen, K., Wylie, T., Larson, D. E., McLellan, M. D., Mardis, E. R., Weinstock, G. M., Wilson, R. K. and Ding, L. (2009). VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* 25(17): 2283-2285.
8. Kofler, R., Schlotterer, C. and Lelley, T. (2007). SciRoKo: a new tool for whole genome microsatellite search and investigation. *Bioinformatics* 23(13): 1683-1685.
9. MIcroSAtellite identification tool.

10. Reading/writing/editing/indexing/viewing SAM/BAM/CRAM format. *Samtools*.
11. Rice, P., Longden, I. and Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* 16(6): 276-277.
12. *Sff_extract. Bioinformatics at COMAV*.
13. Sherman, A., Rubinstein, M., Eshed, R., Benita, M., Ish-Shalom, M., Sharabi-Schwager, M., Rozen, A., Saada, D., Cohen, Y. and Ophir, R. (2015). Mango (*Mangifera indica* L.) germplasm diversity based on single nucleotide polymorphisms derived from the transcriptome. *BMC Plant Biol* 15: 277.