Biomedical Data Mining

Extended Abstract

Insilco studies of mango genome cultivars and development of analysis tool

Rabia Faizan

Abstract

Mango is one of the acclaimed and fifth most significant subtropical/tropical organic product crops worldwide with the creation focused in India and South-East Asia. As of late, there has been an overall interest in mango genomics to deliver devices for Marker Assisted Selection and attribute affiliation. There are no electronic investigated genomic assets accessible for mango especially. Subsequently a total mango genomic asset was needed for development in examination and the executives of mango germplasm. In this venture, we have done near transcriptome investigation of four mango cultivars for example cv. *Langra*, cv. *Zill*, cv. *Shelly* and cv. *Kent* from Pakistan, China, Israel, and Mexico individually.

The crude information is acquired through De-novo arrangement get together which produced 30,953-85,036 unigenes from RNA-Seq datasets of mango cultivars. The undertaking is meant to give mainstream researchers and overall population a mango genomic asset and permit the client to look at their information against our examined mango genome data sets of four cultivars (cv. *Langra*, cv. *Zill*, cv. *Shelly* and cv. *Kent*). A mango web genomic asset MGdb, depends on 3-level engineering, created utilizing Python, level record information base, and JavaScript. It contains the data of anticipated qualities of the entire genome, the unigenes explained by homologous qualities in different species, and GO (Gene Ontology) terms which give a brief look at the characteristics wherein they are included. This web genomic asset can be of colossal use in the evaluation of the exploration, improvement of the prescriptions, getting hereditary qualities and gives valuable bioinformatics answer for examination of nucleotide succession information. We report here world's first webbased genomic asset especially of mango for hereditary improvement and the board of mangogenome.

Keywords: RNA-seq, marker assisted selection, annotation, gene ontology, phylogeny.

Introduction

As a member of the family Anacardiaceae, mango (Mangifera Indica Linn.) ranks second among 3 tropical fruit crops after banana due to its rich sensational taste, color, aroma and huge 4 economics significance (Litz RE, 2009; Srivastava S, 2016). According to Food and Agriculture Organization of the United Nations (FAO), India holds the 1st position in mango production followed by China, whereas Pakistan and Mexico rank 5th and 6th position respectively.

In the Ayurvedic and indigenous medical systems, it counts for an important herb for nearly over 4000 years (KA Shah, 2010). The substances in mango have high therapeutic potential. According to Ayurveda, mango tree comprises of different medicinal properties could be used to treat tumor, rabid dog, piles, toothache, diarrhea, cough, insomnia, hypertension, asthma, anemia, hemorrhage, dysentery, heat stroke, blisters, miscarriage, liver disorders, tetanus, wounds in the mouth and excessive urination. Phytomedicines should be sufficiently regulate based on this knowledge (Kalita, 2014).

Mango has a small genome size of 439 Mb (2n = 40) and a diploid fruit tree with 20 pairs of chromosomes (Arumuganathan, 1991). There are total 72 species of genus Mangifera from which most of them surviving in the rain forests of Malaysia and Indonesia (Nagendra K

Rabia Faizan
Sir Syed University of Engineering & Technology Karachi, Pakistan, E-mail: rabiatabassumkhi@gmail.com

Singh, 2016). Mango has been widely cultivated in India and Southeast Asia for thousands of years. It

has now grown throughout the world (tropical and sub-tropical) in 99 countries with a total fruit production of 34.3 million tons of fruit per annum (Galán Saúco, 2013). Asia is the largest producer of mangoes, 76% of world production comes from Asia with the second and third largest producers Americas (12%), and Africa (11.8%) (Galán Saúco, 2013).

Mango is a rich source of vitamins (vitamin A and vitamin C) and minerals and popular for its attractive color, texture and juicy flavor covers the total area of 2,297,000 ha in 2010 to 2011 (Hiwale1, 2015). The chemical analysis of mango pulp evinces that it has a relatively high content of calories (60 Kcal/100 g fresh weight) and is an important source of potassium, fiber, and vitamins (Marianna Lauricella, 2017). The nutritive value of mango is listed in (Table 1) reporting some data from the National Nutrient Database for Standard Reference (United States Department of Agriculture) (USDA, 2016).

This project aims to build a successful response to the difficulties currently faced by fruit science agencies in meeting the need of gradually increasing fruit demands. The project intended to contribute to the development of a framework designed to support the analysis of the vast data for the availability of mango genomic information. It will also provide a platform for genetic information of different varieties of Mangifera Indica from different countries to make the estimated data available for the scientific community and general public.

Materials and Methods

Retrieval of Mango unigenes

Data Mangifera Indica Linn. usually known as mango, is a species of the flowering plant in the sumac and poison ivy family Anacardiaceous. In this project four mango cultivars cv. Langra from Pakistan, cv. Zill from China, cv. Kent from Mexico and cv. Shelly from Israel transcriptomic sequences or the unigenes which are generated by processing the RNA-seq reads and transcriptome de novo assembly reported in 2014 (Azim MK, 2014) are retrieved, analyzed and presented in a database. We obtained four unigenes datasets corresponding to four mango cultivars from Jamilur-Rahman Center for Genome Research, Dr. Panjwani Center for Molecular Medicine and Drug Research, International Center for Chemical and Biological Sciences, University of Karachi.

BLAST Analysis of Unigenes

The Blast Analysis or sequence comparison of unigenes of four mango cultivars were done against two databases: non–redundant (NR) protein sequence and Swiss–Prot. First, all unigenes of Pakistan dataset (cv. Langra) was aligned against non–redundant (NR) protein sequence database using BLASTX (E value cutoff $\leq$ 1e−3 and 10 hits per sequence) in Blast2GO java application to retrieve common transcripts. Obtained results were then subjected to mapping and annotation for functional annotation and Gene Ontology (GO) assignments. This analysis was done only with cv. Langra from Pakistan for the Gene search module of the Website. Secondly, the four unigenes datasets are filtered for redundant sequences using CD-HIT (Cluster Database at High Identity with Tolerance) (Fu L, 2012) at 90% identity threshold. The resultant non-redundant unigenes were further analyzed for coding regions using TransDecoder. Obtained coding sequences (CDS) were then subjected to homology search and BLASTed against Swiss–Prot databases by BLASTP with an E value cutoff $\leq$ 1e−3 to find the common and unique transcripts. These BLAST results are then used to modify the FASTA header of each sequence (unigene) of four assembled datasets (cv. *Langra*, cv. *Kent*, cv. *Shelly* and cv. *Zill*) (Figure 2). Following steps have been taken to modify the FASTA headers:

1. Extract the sequence, name and length columns from the original file.

2. Extract the description of the species with the name from the resultant file.

3. Merge the two files with the common column name.

Results

The four mango cultivars from cv. *Langra*, cv. *Zill*, cv. *Shelly* and cv. *Kent* shows they have 30,953, 57,544, 58,797 and 85,036 unigenes respectively generated by transcriptomic de novo assembly resulted from Trinity (Grabherr MG, 2011; Waqasuddin Khan 2017). These unigenes were further characterized for functional annotations using BLAST (Jones P, 2014). BLAST homology search of cv. *Langra* showed 83% unigene sequences have 45-100% similarity with sequences in Nr database in which 59% sequences assigned with GO terms, 63% get mapped and remaining 17% are the novel sequences. The CDS analysis identified 14066 (~45%), 34893 (~59%), 58614 (~69%) and 35364 (~61%) protein coding sequences in cv. Langra, cv. Zill, cv. Shelly and cv. Kent unigenes datasets respectively. The homology search of these ORFs of four mango cultivars showed 57-100% similarity with sequences in Swiss-Prot database. The calculated probabilities for a variety of parameter choices by using the random model and scores are recorded (Altschul SF, 1990).

Discussion

Mango is one of the most important fruits of the tropical ecological region of the world, well known for its nutritive value, aroma, and taste (Iquebal et al., 2017). Transcriptomic sequences of different mango cultivars provided a wealth of data related to protein-coding sequences. This study and work resulted in an 'analyzed web-based mango genomic resource' from four mango cultivars grown in Pakistan, China, Israel and Mexico (Waqasuddin Khan 2017). Initially, we retrieved the mango genome assembled unigenes data of four different cultivars (cv. *Langra* (Pakistan), cv. *Zill* (China), cv. *Shelly* (Israel) and cv. *Kent* (Mexico)) from the international center for chemical and biological sciences located at the University of Karachi in Pakistan. These transcriptome sequences were obtained from RNA-Seq experiments using Illumina NGS technology (Waqasuddin Khan 2017). To analyze and annotate the data of mango transcriptome, total assembled unigenes from four different countries (cv. *Langra* (Pakistan), cv. *Zill* (China), cv. *Shelly* (Israel) and cv. *Kent* (Mexico)) were processed using three methods: comparative searching, mapping, and annotation.

Acknowledgment

We would like to express our sincere gratitude to Dr. Kamran Azim (project advisor) and Safina Abdul Razzak from Muhammad Ali Jinnah University for their invaluable advice, guidance and enormous patience throughout the development of the project. We also acknowledge the support of Dr. Zaheer-ul-haq Qasmi from Dr. Panjwani Center for Molecular Medicine and Drug Research, International Center for Chemical and Biological Sciences, University of Karachi in the analysis of the data.

Rabia Faizan
Sir Syed University of Engineering & Technology Karachi, Pakistan, E-mail: rabiatabassumkhi@gmail.com

This work is partly presented at 9th International Conference on Bioinformatics & System Biology Singapore City, Singapore March 20-21, 2019