Development and validation of microsatellite markers from *de novo* transcriptome assembly of eggplant (*Solanum melongena* L.) and its putative progenitor *S. incanum* L. cultivars

Shailesh K. Tiwari^{1@}, Pallavi Mishra^{1,2}, Sakshi Singh³, Vinay K Singh³, Sarvesh P Kashyap¹, Major Singh^{1,4}, Kavindra N Tiwari², Prakash S Naik¹ and Bijendra Singh¹

¹*Division of Crop Improvement, ICAR-Indian Institute of Vegetable Research, Varanasi, Uttar Pradesh, India*

²*Department of Botany, MMV, Banaras Hindu University, Varanasi, Uttar Pradesh, India*

³*Centre for Bioinformatics, School of Biotechnology, Banaras Hindu University, Varanasi, Uttar Pradesh*

⁴*Current Address: ICAR-Directorate of Onion and Garlic Research Rajgurunagar, Pune Maharashtra, India*

@ Corresponding Author: tiwarishailu@gmail.com; Mobile: +91-9473877501

E-mail addresses:

tiwarishailu@gmail.com (SK Tiwari); pallavimishra23@gmail.com (Pallavi Mishra); (Sakshi Singh); officerbiotech.1@gmail.com (Vinay K Singh); singhsarvesh10@gmail.com (SP Kashyap); kntiwari@bhu.ac.in (KN Tiwari); singhvns@gmail.com (Major Singh); naikps1952@gmail.com (PS Naik); bsinghiivr@gmail.com (B Singh)

Abstract

An elite cultivar of eggplant, Ramnagar Giant (*Solanum melongena* L.) and W-4 (*S. incanum* L.) with contrasting horticultural traits were used as parental lines to develop a mapping population of RILs. To accelerate breeding programs and to develop large scale SSR markers to be used in QTL mapping, RNASeq libraries from different tissues of both the parental plants were deep sequenced and assembled into representation of a high quality *de novo* transcriptome using Illumina-based Next Generation Sequencing technology. 99.99% of high quality bases were obtained from all the tissues and deposited in TSA database at the NCBI link. Total 3, 156 and 3, 196 SNVs were detected in *S. melongena* and *S. incanum*, respectively. In *S. melongena*, 11, 262 SSR while in *S. incanum* 11, 829 SSR containing regions were identified. Based on functional annotation, 21, 914 unique genes could be identified for *S. melongena*, 21,706 unique genes for *S. incanum* and overall, 60 different transcription factors were identified in both the lines. Further, a total of 536 SSR markers were designed and screened for polymorphism of which, 157 markers produced polymorphism between the parental lines. The polymorphic SSRs shall be used for genotyping of RILs to map QTLs for various horticultural traits in eggplant and identification of candidate genes in response to biotic and abiotic stress.

Key words: *de novo* Transcriptome Assembly, Ramnagar giant, SNVs, SSR, sequencing, *Solanum melongena*, *Solanum incanum*, unigenes.

Introduction

Eggplant (*Solanum melongena* L., Solanaceae; $2n = 2x = 24$) commonly called brinjal or aubergine, is the third most important vegetable crop after tomato and potato in the genus *Solanum* [1]. While most of the other species of this genus are representatives of new world, eggplants are phylogenetically distinct and are indigenous to the old world with most related species reported to be originated in India, Africa, and South America [2, 3]. FAOSTAT data reveals that eggplant

production has increased up to three folds in the last two decades, with China, India, Egypt, and Turkey being the top four eggplant producing countries across the world [4]. The berries are rich in naturally occurring water-soluble flavonoids [5], vitamins and minerals [6], polyphenol derivatives [7, 8], antioxidants like chlorogenic acid (CGAs) [9], caffeic acid [10, 11], and many other phytochemical compounds of human health benefits [12, 13]. Further, the low calorie, high potassium and high dietary fiber content makes it a vegetable of choice for the obese and diabetes patients [14].

High morphological diversity is observed among different cultivars and landraces of eggplant forms available in the market [15]. Common domesticated traits such as colour, shape, size, and chemical composition of fruits are highly specific to eggplant and important for breeding and marker-assisted selection (MAS) [16, 17]. The genetic variation in different morphological forms are a result of quantitative trait loci or QTLs which determine either complete or a portion of the total variation, and enables identification of large number of genes whose expression controls quantitative traits of agricultural importance such as plant yield, weight, fruit texture, quality, shape, color, chemical composition and plant resistance to diseases and pathogens [18]. Eggplant and their wild relatives have been an interesting subject of classical genetical analysis and research to the interested plant breeders [19]. Number of QTLs and mapping studies are reported for morphological traits in eggplant like fruit shape, size, colour, and prickliness [16, 17, 20], yield-related traits [21, 22], biochemical traits like location of CGA and polyphenol oxidase genes [9, 23], resistance to *Ralstonia solanaceae* [24], resistance to fungal wilts [25], resistance against Fusarium wilt [26], tolerance to bacterial wilt [27], and so on. Detailed biochemical and genetic information has been accumulated in eggplant for assessment of genetic diversity, QTL detection, and mapping based studies using range of molecular markers such as RAPD, AFLP, ISSR, SSR, SRAP, ESTs, and SNP [9, 16, 17, 26, 28-37]. These reports provide useful grounds for deciphering mechanisms underlying genetic control of the complex traits which are fundamental to the eggplant research program.

The progression of next-generation sequencing (NGS) technology has provided new ways for sequencing of plants especially for non-model species [38-40]. As compared to the traditional Sanger's method for DNA sequencing, this method is cost-effective, fast, and useful in investigating the genome elements and molecular profile in comparatively easy ways [41]. NGS platform supports the view to sequence nucleotides frequently and allow massive sequencing of short sequence reads [42, 43] and large-scale development of molecular markers that can be helpful in analysing genetic diversity within large populations, validation of foreign gene introgression, linkage mapping, QTL mapping, *etc.* [41, 44]. Due to rapid advancements of whole-genome sequencing (WGS) and transcriptome methodologies in the last decade, assaying and gene expression profiling of many crops from the family *Solanaceae* were fastened to generate extensive molecular information. However, while the WGS of the model crops like tomato, potato, and pepper started early in 1990s and late 2000 [45-53], the eggplant genome has been merely addressed. Only a draft genome sequence of an eggplant cultivar, *Nakate-shinkuro* is published till date [54], which reflects that there is lack of a comprehensive dataset for eggplant genetic resources that could provide complete genetic information and makes the eggplant genome lagged behind with its natives with only little explored at genome and proteome level. The transcriptome assembly information in eggplant is also limited with only fewer reports published till date [1, 55-61]. The obtained information from published reports show that transcriptome data has wide applications in comprehensive characterization of entire genome, targeted sequencing of desired mRNA, discovery of novel transcription factor and identification of their binding sites, and expression profiling of non-coding mRNA sequences [44]. However, the genomic architecture and whole genome sequence information in eggplant is still lagging despite of extensive efforts to collect genetic and molecular information for this species.

Of the several species originated in Africa, *S. incanum*, a wild ally of eggplant is believed to be putative progenitor of *S. melongena* [3, 34-36], and together they both belongs to the complex

92 *melongeneae*. They are readily crossable with each other to produce highly fertile interspecific
 93 hybrids and therefore are objects of interests in eggplant breeding and development of experimental
 94 populations. Therefore, with the objective of large-scale development of SSR markers for QTL
 95 mapping of domestication traits in eggplant and to collect more genomic information, we performed
 96 *de novo* transcriptome sequencing of two accessions of eggplant, Ramnagar Giant (*S. melongena*
 97 L.), a local cultivar restricted to Ramnagar and Varanasi regions of eastern Uttar Pradesh, and W-4,
 98 a genotype /line, a semi-wild close relative and its putative progenitor of eggplant *S. incanum* L., at
 99 the ICAR-IIVR Varanasi. The two accessions have been used as parents (Ramnagar Giant X W-4)
 100 to develop a mapping population of 114-F8 recombinant inbred lines (RILs). The parental lines are
 101 morphologically diverse from each other and exhibit distinct variations in traits of commercial and
 102 horticultural importance. W-4 is a spiny, wild form of eggplant showing resistance to some
 103 common pests hampering eggplant cultivation like *Fusarium oxysporum* f. sp. *Melongenae* [62, 63],
 104 *Phomopsis vexans* [64], and *Leucinodes orbonalis* [65]; while Ramnagar Giant is a local cultivar
 105 from Varanasi, the plants with dark green broad foliage bearing huge size rounded fruits of light
 106 green color with 3-4 fruits per plant and fruit weight 0.80-2.5 kg, and reported to possess moderate
 107 resistance to the *Phomopsis blight* of eggplant [66]. While the transcriptome assembly information
 108 of *S. incanum* has been previously reported [55], this is the first comprehensive report of *de novo*
 109 transcriptome sequencing of Ramnagar Giant which is an indigenous eggplant cultivar. Here, we
 110 successfully developed high quality SSR markers to be used in QTL mapping of horticultural traits
 111 and also described a set of differentially expressed annotated transcripts that shall further be used in
 112 discovery of novel gene families and transcription factors (TFs) associated with biosynthesis of
 113 active plant ingredients in response to biotic or abiotic stress and in defense related mechanisms in
 114 eggplant.

Materials and methods

Plant tissue collection and mRNA library preparation

De novo transcriptome sequencing of *S. melongena* and *S. incanum* (Fig 1) was performed in the present study. Both the lines have been maintained at the experimental farm of ICAR-Indian Institute of Vegetable Research, Varanasi. For deep mRNA sequencing, total RNA was extracted from following plant tissues each from *S. melongena* and *S. incanum*; root, stem, leaves, peduncle, floral buds, flower, calyx, and fruits. The tissues collected were immediately washed in 95% ethanol, hashed into small pieces, suspended in solution of 15 ml RNeasy lysis reagent (Ambion Cat#7020) and left for 3-4 h at 4°C. The samples were then transferred to 50 ml screw-capped falcon vials and stored at -80°C. The isolated RNA was quantified using Nanodrop spectrophotometer while other qualitative estimations were performed by Agilent High-sensitivity Bioanalyzer Chip. RNA Integrity Number (RIN) was calculated and finally, Illumina HiSeq-2000 platform was used to sequence the cDNA synthesised from high-quality RNA of both the genotypes.

Deep RNA sequencing and raw data quality control

Paired-end mRNA library was constructed from the total RNA as per the manufacturer's protocol of Illumina based Next generation sequencer HiSeq-2000 to obtain high quality reads for their representation in the form of de novo assembly of the eggplant transcriptomes. Two cDNA libraries (read_1 and read_2) were constructed for the two genotypes on Illumina platform for transcriptome sequencing. More than 98% of high quality bases, both in forward and reverse (paired-end) reads were obtained from all the plant tissues. Based on this, the best quality reads (>200bp) were used for their assembly into contigs and further represented in the form of transcripts. All the sequencing related process such as quality testing of raw reads, filtering of High Quality (HQ) reads, trimming of adapter and low quality reads was done by SeqQC V2.0. The average Phred scaled quality score was found to be >20 which is an indication of a good quality of sequencing run.

Transcriptome sequencing and NGS data assembly

Assembling of sequences through NGS technique allows one to obtain the high throughput fragments (reads) from the molecular (RNA) sequences which are further assembled and overlapped together to construct cDNA and can further be sequenced from both the directions to obtain paired-end reads [67]. The cDNA library from the pooled total RNA in equal quantity from all the plant tissues were sequenced from both the lanes to generate 100 bp paired-end reads with a coverage of 68.97x. *De novo* assembly of the obtained reads was accomplished with Velvet 1.2.10 and Oases v.0.2.08 using de Bruijn graph algorithm, and Oases (version 0.2.01) to obtain best quality raw data [68, 69]. The filtered transcripts (>200bp) were successfully deposited in Transcriptome Shotgun Assembly (TSA) database at the National Centre for Biotechnology Information (NCBI) portal with primary accessions GAYR000000000 and GAYS000000000 for *S. melongena* and *S. incanum*, respectively.

SSR mining, SSR distribution, GC content, and SNV's detection

SSR prediction was performed by the tool MISA v1.0 (MicroSAteellite; <http://pgrc.ipk-gatersleben.de/misa>) [70], and analysis of GC content was carried out using SSR locator. Because SSR markers show greater degree of polymorphism and have co-dominant character, they are a good choice for diversity analysis and genetic mapping studies. The maximum fragment length considered for SSR prediction was 100 bp. For SNV's detection and variation study, bowtie2 [71] and VarScan (v2.3.4) packages were used.

Identification of TF classes

For identification of TF classes in *S. melongena* and *S. incanum*, sequence similarity search was done using PlantTFDB (plant transcription factor database), NCBI-CDD (conserved domain database) and tBLASTn algorithm taking *S. lycopersicum* TF sequences as the reference. For this,

related protein sequences of a TF class was identified from Plant TFDB and aligned against the TSA of *S. melongena* (taxid: 4111) and *S. incanum* (taxid: 329799) using tBLASTn. Presence of conserved domains were further confirmed with NCBI-CDD (<https://www.ncbi.nlm.nih.gov/Structure/cdd/>)

Similarity search, Functional annotation, and pathway analysis

Functional annotation of unigenes remains a difficult and challenging endeavour particularly for the non-model plant species such as eggplant due to limited stock of reference plant genomes at the public databases. For sequence similarity search and to predict annotation, the publically available databases BLASTx and TBLASTx search were performed to match the NCBI's non-redundant (nr) plant sequences from the species including *S. lycopersicum*, *S. tuberosum*, *Vitis vinifera*, *Glycine max*, *Medicago truncatula* and *Ricinus communis* [72]. Homology search and motif search of the transcripts was executed using BLASTx. GO database and KAAS (KEGG Automatic Annotation Server: <http://www.genome.jp/kegg/kaas/>) server was used for gene annotation and pathways analysis, respectively [73, 74].

Primer designing for polymorphism and QTL mapping in eggplant

MISA v1.0 was used for identification of SSR containing regions within 94, 838 sequences (spanning in length~73 Mb) in *S. melongena* and 95, 096 sequences (length~73 Mb) in *S. incanum*. A set of 536 SSR markers were designed using SSR locator and custom synthesized for detection of polymorphism between the parental lines (*S. melongena* and *S. incanum*), and further validation on the mapping population of RILs.

DNA extraction and PCR protocol for newly synthesized SSR markers

Total genomic DNA was isolated from 20 days young leaves of *S. melongena* and *S. incanum* by CTAB method [75]. After isolation, 5 µl of DNA sample was mixed with 5 µl of 6X loading dye

and run on an 0.8% agarose gel to access the DNA quality. DNA quantity was measured on Nanodrop spectrophotometer with peak absorbance at 260 nm. To check the amplification and validation of SSR primers upon parental lines and RILs, PCR reaction was carried out in a Biometra thermal cycler taking 50 ng genomic DNA as template mixed with 10 X Taq buffer, 2.5 mmol/l MgCl₂, 0.2 mmol/l dNTPs, 0.5 mmol/l each of the SSR primer and 1 U Taq DNA Polymerase and final volume was made of 25 µl. The cyclic reactions involved 4 min at 94°C, 36 cycles of 1 min at 94°C, 1 min at 48°C, 2 min at 72°C, and finally 10 min at 72°C. The amplified products were separated on polyacrylamide gel electrophoresis (PAGE). For this, 8% non-denaturing gel was prepared in gel casters using 4 ml of 29:1 Acrylamide/Bis (% w/v), 8 ml distilled water, 1.68 gm Urea, 1.4 ml of 5 X TBE, 0.168 µl of 10% APS (ammonium persulphate) and 10 µl TEMED and left at room temperature for 45-60 minutes for polymerization to occur. The comb was slowly removed and gel was dipped in electrophoretic tank containing 10 X TBE buffer (pH 7.0-8.5) prepared at a working strength of 1 X. 10 µl PCR product (of both the parental lines and RILs) was loaded and separated for 3 h at 90 Volts. After separation, the gel was taken out carefully and stained with ethidium bromide in a separate tank and visualized under Alpha Imager™ 3400 for genotyping of RILs and for further analysis.

Results and discussion

De novo transcriptome sequencing of *S. melongena* and *S. incanum*

After filtering adapter sequences and low quality sequences, a total of 101060300 numbers of raw reads were perpetuated from the raw data for sequencing, of which 50901025 number of HQ reads were obtained for *S. melongena* and 50159275 number of HQ reads were obtained for *S. incanum* (Tables 1-3). The raw paired-end read sequence data of FASTQ file size of 3.8 Gb were obtained for *S. melongena* and *S. incanum*. On an average, the overall data quality was very good with more than 98% of HQ paired-end (PE) bases in both genotypes (Figs 2A and 2B). To assemble

transcripts from short read sequences, Velvet and Oasis graph based assemblers were used [76]. 100.58 million (98.8%) HQ Paired reads were obtained for *S. melongena* and 100.32 million (98.86%) HQ paired reads were obtained for *S. incanum* (**Fig 3**). Bases having ≥ 20 phred score was considered for QC parameters in both the species. See raw and filtered data quality control report in **S1 Table**. The sequencing informations of *S. melongena* and *S. incanum* were submitted to the Sequence Read Archive (SRA) of NCBI portal with primary accessions SRX365148 and SRX365146, respectively.

SSR identification in *S. melongena* and *S. incanum*

Using the SSR locator and MISA v1.0 tool, SSR containing regions were identified within 94,838 sequences (spanning in length~73 Mb) in *S. melongena* and 95,096 sequences (spanning in length~73 Mb) in *S. incanum*. Maximum number of bases interrupting two consecutive SSRs in a compound microsatellite was found to be 100 bp. Total numbers of 11,262 and 11,829 SSRs could be identified in *S. melongena* and *S. incanum*, respectively. These SSRs were distributed in 9,839 numbers of sequences in *S. melongena* and 10,235 sequences in *S. incanum*. In *S. melongena*, highest percentage of SSR distribution was accounted for mononucleotides (~58%), followed by trinucleotides (~22%), dinucleotides (~15%) and complex, tetra- and penta-nucleotides approximately 4%, 1% and 0.12%, respectively and GC content was found to be 43%. In *S. incanum*, mononucleotides accounted for ~54%, followed by trinucleotides (~21%), dinucleotides (~14%) and complex, tetra- and penta-nucleotides approximately 5%, 3% and 0.8%, respectively and 43% GC content (**Fig 4 and Tables 4 and 5**). The distribution of SSRs and their motifs are summerized in **S2 and S3 Tables**. Through SSR locator, a total of 536 new SSR markers were designed and custom synthesized from the transcriptome sequence of *S. melongena* and *S. incanum*. (Details and sequences of all primers are given in **S2 Table**).

232 **SNVs detection**

233 Bowtie2 tool was used for SNV's detection and VarScan package (v2.3.4) was used for variation
 234 analysis to identify allelic variations in *S. melongena* and *S. incanum*. We reported a total number of
 235 3, 156 SNVs in *S. melongena* and 3, 196 SNVs in *S. incanum*. Within these SNVs, 227
 236 homozygous SNVs (~7%) were detected in *S. melongena* and 233 homozygous SNVs (~7%) were
 237 detected in *S. incanum*. The number of heterozygous SNVs detected were 2, 929 (~93%) and 2, 964
 238 (~93%) in *S. melongena* and *S. incanum*, respectively (**Fig 5**). SNVs were further distinguished into
 239 separate classes based on within sequence variations (**Fig 6**). For other details, see **S4 Table**.

240 **Transcripts annotation of *S. melongena* and *S. incanum***

241 Nucleotide non-redundant data sets from plant species like *S. lycopersicum*, *S. tuberosum*, *Vitis*
 242 *vinifera*, *Glycine max*, *Medicago truncatula* and *Ricinus communis* were used to annotate the
 243 clustered transcripts to represent in the form of a library using BLASTx. A transcript coverage of cut
 244 off > 70% (E-value<= 0.001) was considered for homology prediction and functional annotation of
 245 eggplant transcripts. In *S. melongena*, total 94,838 predicted transcripts were examined to match
 246 against the Plant GDB for annotation of which 26, 914 transcripts (28.37%) could be annotated.
 247 Similarly, in *S. incanum*, 95, 096 predicted transcripts were examined to match against the Plant
 248 GDB of which 26, 660 transcripts (28.03%) could be annotated (**Fig 7**). Besides these, 21, 914
 249 unique genes for *S. melongena* and 21, 706 unique genes for *S. incanum* were also identified (**S6**
 250 **and S7 Tables**).

251 **DEGs, Gene ontology, Unigene identification and pathway analysis**

252 It was found that 379 genes were up-regulated, 368 genes were down-regulated, while 747 genes
 253 were found to be differentially expressed between the transcriptome assembly of *S. melongena* and
 254 *S. incanum*, respectively (**Fig 8**). A total of 21, 914 and 21, 706 unique genes were mapped in *S.*

melongena and *S. incanum*, respectively. Analysis for identification of transcripts involved in cellular and biochemical pathways was carried out was based on KAAS (KEGG Automatic Annotation Server: <http://www.genome.jp/kegg/kaas/>) server. GO database was exploited for gene annotation in which, predicted transcripts with gene ontology annotation was 5, 087 out of 9, 085 annotated transcripts in *S. melongena* and 5, 064 out of 8, 968 annotated transcripts in *S. incanum*. The GO annotation summary of biological processes, cellular components, and molecular functions of *S. melongena* and *S. incanum* are described separately (**Fig 9 and S9 Table**). It was found that 1, 211 (out of total 9,085 annotated transcripts) in *S. melongena* and 937 (out of 8,938 annotated transcripts) in *S. incanum* could be assigned with KEGG pathway. The top ten KEGG pathways were identified in both the species based on KEEG ortholog, which play a key role in biochemical and cellular metabolism (**S8 and S9 Tables**).

Identification of TF classes in *S. melongena* and *S. incanum*

Overall, 60 different TF classes were identified in responses to biotic and abiotic stresses in transcriptome sequences of *S. melongena* and *S. incanum* which play a crucial role in different expression and developmental pathways. Each TF class represented number of genes identical to those available in PlantTFDB. Based on similarity search, some of the important TF associated genes such as *AP2*, *ARF*, *ARR-B*, *bHLH*, *bZIP*, *CAMTA*, *C2C2*, *Dof*, *ERF*, *GATA*, *WRKY*, *NAC* and *YABBY* were successfully identified in both the species (**Fig 10 and S10 Table**).

Parental screening for presence of polymorphism and validation of polymorphic SSRs on RILs

536 newly synthesized SSR primers were tested on parental lines to check the amplification of which 502 primers (93.65%) produced amplicons and 157 markers (31.27%) detected polymorphism between the parental lines through PAGE. The polymorphic SSRs were further validated on RILs for genotyping through PAGE (**Figs 11A and 11B**). The details of primers pairs

used for detection of polymorphism between parental lines and obtained polymorphic primers are given in **S2 Table**, and have been used with suffix ‘*Sm*’ and ‘*Si*’ for primer sequence IDs of *S. melongena* and *S. incanum*, respectively for further studies. From the obtained results, it can be inferred that NGS-derived markers are more useful and informative in detecting genetic polymorphism between related and highly inbred species. Further, the NGS data can be exploited in much broader sense to collect more genomics and proteomics information for non-model species like eggplants.

Conclusion

Next generation sequencing is a powerful tool for rapid collection of vast amount of RNASeq data from organisms and detailed analysis of the complete set of transcripts for understanding their developmental stages and physiological conditions [76-78]. The analysis workflow enables the discovery of unique and candidate genes having significant roles in the molecular and biological processes particularly for the non-model species like eggplant. Moreover, a large number of SSR and SNP markers can also be generated from the assembly [79, 80]. Since the genome of several species of eggplant still remains lagged behind the other vegetable crops like tomato and potato, RNASeq technique would bring this crop a stage up towards deciphering the molecular mechanism underlying defence, developmental and physiological responses through the structural, functional, and phylogenomics approach.

The present report is the first transcriptome sequencing assembly information of the cultivar Ramnagar Giant which is least characterized species from *melongenae* complex of eggplant. The cultivar can be useful for breeding purpose for development of high yielding varieties. The Ramnagar Giant x W-4 developed RILs population would be further utilized for QTL mapping of horticultural and domestication traits in eggplant. High morphological variation was observed among the RILs. Analysis of polymorphism and variation in this population would be useful for

tracing the significant QTLs distributed along the eggplant genome. Moreover, a huge number of SSR markers developed through *RNASeq* in the present study will be useful to provide a good platform for mapping and diversity studies on intra- and inter-specific population of eggplant genotypes. Furthermore, the markers will also be useful for elucidating genetic relationship among cultivated eggplant and their wild relatives. Lastly, the results assembled through the *RNASeq* of the two species are enriched with high accuracy and reliable for future for identification of unique and novel candidate genes and development of more SSR and SNP markers with implications to modern marker-assisted breeding in eggplant. Overall, the comprehensive datasets obtained through the Ramnagar Giant and W-4 transcriptomes will be used for QTL mapping and serve as a reference for further analysis of disease resistance candidate genes and broaden the understanding of this agriculturally important vegetable crop.

Acknowledgements

Indian Council of Agricultural Research (ICAR), New Delhi is acknowledged for funding this research. Authors are thankful to the Director, ICAR-IIVR, Varanasi, Head, Division of Crop Improvement, ICAR-IIVR, Varanasi and The Coordinator, School of Biotechnology, BHU, Varanasi for providing necessary facilities to carry out the present analysis.

Author Contributions

Conceptualization of experiment: SKT, MS, PSN and BS Formal analysis: SKT, PM, SS, VKS. Experimental validation: PM and SPK. Contributed reagents/materials/analysis tools: SKT, VKS, MS, PSN and BS. Supervision: SKT. Original draft writing: PM, SPK and SKT. Review and editing of original draft: VKS, and KNT.

324 **References**

- 325 1. Yang X, Cheng Y-F, Deng C, Ma Y, Wang Z-W, Chen X-H, et al. Comparative
326 transcriptome analysis of eggplant (*Solanum melongena* L.) and turkey berry (*Solanum torvum*
327 Sw.): phylogenomics and disease resistance analysis. BMC Genomics. 2014;15(1):412.
- 328 2. Sękara A, Cebula S, Kunicki E. Cultivated eggplants—origin, breeding objectives and
329 genetic resources, a review. Folia Horticulturae. 2007;19(1):97-114.
- 330 3. Knapp S, Vorontsova MS, Prohens J. Wild relatives of the eggplant (*Solanum melongena*
331 L.: Solanaceae): new understanding of species names in a complex group. PloS ONE.
332 2013;8(2):e57039.
- 333 4. Food, Agriculture Organization of the United N. FAOSTAT statistics database. [Rome?] :
334 FAO, 1998.
- 335 5. Aza-Gonzalez C, Herrera-Isidró L, Núñez-Palenius H, De La Vega OM, Ochoa-Alejo N.
336 Anthocyanin accumulation and expression analysis of biosynthesis-related genes during chili
337 pepper fruit development. Biologia Plantarum. 2013;57(1):49-55.
- 338 6. Gebhardt S, Lemar L, Haytowitz D, Pehrsson P, Nickle M, Showell B, et al. USDA national
339 nutrient database for standard reference, release 21. United States Department of
340 AgricultureAgricultural Research Service. 2008.
- 341 7. Stommel JR, Whitaker BD. Phenolic acid content and composition of eggplant fruit in a
342 germplasm core subset. Journal of the American Society for Horticultural Science.
343 2003;128(5):704-710.
- 344 8. Docimo T, Francese G, Ruggiero A, Batelli G, De Palma M, Bassolino L, et al.
345 Phenylpropanoids accumulation in eggplant fruit: characterization of biosynthetic genes and
346 regulation by a MYB transcription factor. Frontiers in Plant Science. 2016;6:1233.
- 347 9. Gramazio P, Prohens J, Plazas M, Andújar I, Herraiz FJ, Castillo E, et al. Location of
348 chlorogenic acid biosynthesis pathway and polyphenol oxidase genes in a new interspecific

349 anchored linkage map of eggplant. BMC Plant Biology. 2014;14:350. doi: 10.1186/s12870-014-
350 0350-z.

351 10. Cao G, Sofic E, Prior RL. Antioxidant capacity of tea and common vegetables. Journal of
352 Agricultural and Food Chemistry. 1996;44(11):3426-3431.

353 11. Co LL. Common medicinal plants of the Cordillera region (northern Luzon, Philippines).
354 1989.

355 12. Mwonjoria JK, Ngeranwa JJ, Kariuki HN, Githinji CG, Sagini MN, Wambugu SN. Ethno
356 medicinal, phytochemical and pharmacological aspects of *Solanum incanum* (lin.). International
357 Journal of Pharmacology and Toxicology. 2014;2(2):17-20.

358 13. Auta R, James S, Auta T, Sofa E. Nutritive value and phytochemical composition of
359 processed *Solanum incanum* (Bitter garden egg). Science World Journal. 2011;6(3):5-6.

360 14. Prabhu M, Natarajan S, Veeraragavathatham D, Pugalendhi L. The biochemical basis of
361 shoot and fruit borer resistance in interspecific progenies of brinjal (*Solanum melongena*). EurAsian
362 Journal of BioSciences. 2009;3.

363 15. Portis E, Cericola F, Barchi L, Toppino L, Acciarri N, Pulcini L, et al. Association mapping
364 for fruit, plant and leaf morphology traits in eggplant. PLoS One. 2015;10(8):e0135200.

365 16. Nunome T, Ishiguro K, Yoshida T, Hirai M. Mapping of fruit shape and color development
366 traits in eggplant (*Solanum melongena* L.) based on RAPD and AFLP markers. Breeding Science.
367 2001;51(1):19-26.

368 17. Nunome T, Negoro S, Kono I, Kanamori H, Miyatake K, Yamaguchi H, et al. Development
369 of SSR markers derived from SSR-enriched genomic library of eggplant (*Solanum melongena* L.).
370 Theoretical and Applied Genetics. 2009;119(6):1143-1153.

371 18. Paran I, Zamir D. Quantitative traits in plants: beyond the QTL. Trends in Genetics.
372 2003;19(6):303-306.

- 373 19. Daunay M, Lester R, Gebhardt C, Hennart J, Jahn M, Frary A, et al. Genetic resources of
374 eggplant (*Solanum melongena* L.) and allied species: a new challenge for molecular geneticists and
375 eggplant breeders. Nijmegen, The Netherlands: Nijmegen University Press; 2001.
- 376 20. Doganlar S, Frary A, Daunay M-C, Lester RN, Tanksley SD. Conservation of gene function
377 in the Solanaceae as revealed by comparative mapping of domestication traits in eggplant. Genetics.
378 2002;161(4):1713-1726.
- 379 21. Portis E, Barchi L, Toppino L, Lanteri S, Acciarri N, Felicioni N, et al. QTL mapping in
380 eggplant reveals clusters of yield-related loci and orthology with the tomato genome. PLoS ONE.
381 2014;9(2):e89499. doi: 10.1371/journal.pone.0089499.
- 382 22. Frary A, Frary A, Daunay M-C, Huvenaars K, Mank R, Doğanlar S. QTL hotspots in
383 eggplant (*Solanum melongena*) detected with a high resolution map and CIM analysis. Euphytica.
384 2014;197(2):211-228.
- 385 23. Toppino L, Barchi L, Lo Scalzo R, Palazzolo E, Francese G, Fibiani M, et al. Mapping
386 quantitative trait loci affecting biochemical and morphological fruit properties in eggplant (*Solanum*
387 *melongena* L.). Frontiers in Plant Science. 2016;7:256. doi: 10.3389/fpls.2016.00256.
- 388 24. Salgon S, Jourda C, Sauvage C, Daunay M-C, Reynaud B, Wicker E, et al. Eggplant
389 resistance to the *Ralstonia solanacearum* species complex involves both broad-spectrum and strain-
390 specific quantitative trait loci. Frontiers in Plant Science. 2017;8(828). doi:
391 10.3389/fpls.2017.00828.
- 392 25. Barchi L, Toppino L, Valentino D, Bassolino L, Portis E, Lanteri S, et al. QTL analysis
393 reveals new eggplant loci involved in resistance to fungal wilts. Euphytica. 2018;214(2):20. doi:
394 10.1007/s10681-017-2102-2.
- 395 26. Miyatake K, Saito T, Negoro S, Yamaguchi H, Nunome T, Ohyama A, et al. Detailed
396 mapping of a resistance locus against Fusarium wilt in cultivated eggplant (*Solanum melongena*).
397 Theoretical and Applied Genetics. 2016;129(2):357-367.

- 398 27. Patil V. Molecular characterization of natural diversity and identification of a trait specific
399 marker for tolerance to bacterial wilt in brinjal (*Solanum melongena* L.) (Doctoral dissertation)
400 2016.
- 401 28. Fukuoka H, Miyatake K, Nunome T, Negoro S, Shirasawa K, Isobe S, et al. Development of
402 gene-based markers and construction of an integrated linkage map in eggplant by using *Solanum*
403 orthologous (SOL) gene sets. *Theoretical and Applied Genetics*. 2012;125(1):47-56.
- 404 29. Baysal Ö, Siragusa M, Gümrükcü E, Zengin S, Carimi F, Sajeva M, et al. Molecular
405 characterization of *Fusarium oxysporum* f. *melongenae* by ISSR and RAPD markers on eggplant.
406 *Biochemical Genetics*. 2010;48(5-6):524-537.
- 407 30. Mutlu N, Boyacı FH, Göçmen M, Abak K. Development of SRAP, SRAP-RGA, RAPD and
408 SCAR markers linked with a *Fusarium* wilt resistance gene in eggplant. *Theoretical and Applied*
409 *Genetics*. 2008;117(8):1303.
- 410 31. Nunome T, Suwabe K, Iketani H, Hirai M, Wricke G. Identification and characterization of
411 microsatellites in eggplant. *Plant Breeding*. 2003;122(3):256-262.
- 412 32. Demir K, Bakir M, Sarikamis G, Acunalp S. Genetic diversity of eggplant (*Solanum*
413 *melongena*) germplasm from Turkey assessed by SSR and RAPD markers. *Genetics and Molecular*
414 *Research*. 2010;9(3):1568-1576.
- 415 33. Furini A, Wunder J. Analysis of eggplant (*Solanum melongena*)-related germplasm:
416 morphological and AFLP data contribute to phylogenetic interpretations and germplasm utilization.
417 *Theoretical and Applied Genetics*. 2004;108(2):197-208. doi: 10.1007/s00122-003-1439-1.
- 418 34. Karihaloo J, Brauner S, Gottlieb L. Random amplified polymorphic DNA variation in the
419 eggplant, *Solanum melongena* L.(Solanaceae). *Theoretical and Applied Genetics*. 1995;90(6):767-
420 770.

- 421 35. Tiwari SK, Karihaloo J, Hameed N, Gaikwad AB. Molecular characterization of brinjal
422 (*Solanum melongena* L) cultivars using RAPD and ISSR markers. Journal of Plant Biochemistry
423 and Biotechnology. 2009;18(2):189-195.
- 424 36. Karihaloo J, Kaur M, Singh S. Seed protein diversity in *Solanum melongena* L. and its wild
425 and weedy relatives. Genetic Resources and Crop Evolution. 2002;49(6):533-539.
- 426 37. Gramazio P, Prohens J, Plazas M, Mangino G, Herraiz FJ, Vilanova S. Development and
427 genetic characterization of advanced backcross materials and an introgression line population of
428 *Solanum incanum* in a *S. melongena* background. Frontiers in Plant Science. 2017;8:1477. doi:
429 10.3389/fpls.2017.01477.
- 430 38. Long Y, Wang Y, Wu S, Wang J, Tian X, Pei X. De novo assembly of transcriptome
431 sequencing in *Caragana korshinskii* Kom. and characterization of EST-SSR markers. PloS ONE.
432 2015;10(1):e0115805.
- 433 39. Annadurai RS, Jayakumar V, Mugasimangalam RC, Katta MA, Anand S, Gopinathan S, et
434 al. Next generation sequencing and de novo transcriptome analysis of *Costus pictus* D. Don, a non-
435 model plant with potent anti-diabetic properties. BMC Genomics. 2012;13(1):663.
- 436 40. Sahebi M, Hanafi MM, Azizi P, Hakim A, Ashkani S, Abiri R. Suppression subtractive
437 hybridization versus next-generation sequencing in plant genetic engineering: challenges and
438 perspectives. Molecular Biotechnology. 2015;57(10):880-903.
- 439 41. Varshney RK, Nayak SN, May GD, Jackson SA. Next-generation sequencing technologies
440 and their implications for crop genetics and breeding. Trends in Biotechnology. 2009;27(9):522-30.
- 441 42. Collins LJ, Biggs PJ, Voelckel C, Joly S. An approach to transcriptome analysis of non-
442 model organisms using short-read sequences. Genome Informatics. 2008;21:3-14.
- 443 43. Surget-Groba Y, Montoya-Burgos JI. Optimization of de novo transcriptome assembly from
444 next-generation sequencing data. Genome Research. 2010;20(10):1432-1440.

- 445 44. Morozova O, Marra MA. Applications of next-generation sequencing technologies in
446 functional genomics. *Genomics*. 2008;92(5):255-264.
- 447 45. Consortium PGS. Genome sequence and analysis of the tuber crop potato. *Nature*.
448 2011;475(7355):189.
- 449 46. Consortium TG. The tomato genome sequence provides insights into fleshy fruit evolution.
450 *Nature*. 2012;485(7400):635.
- 451 47. Michaelson MJ, Price HJ, Ellison JR, Johnston JS. Comparison of plant DNA contents
452 determined by feulgen microspectrophotometry and laser flow cytometry. *American Journal of*
453 *Botany*. 1991;78(2):183-188.
- 454 48. Mueller LA, Tanksley SD, Giovannoni JJ, Van Eck J, Stack S, Choi D, et al. The tomato
455 sequencing project, the first cornerstone of the International Solanaceae Project (SOL).
456 *International Journal of Genomics*. 2005;6(3):153-158.
- 457 49. Mueller LA, Lankhorst RK, Tanksley SD, Giovannoni JJ, White R, Vrebalov J, et al. A
458 snapshot of the emerging tomato genome sequence. *The Plant Genome*. 2009;2(1):78-92.
- 459 50. Datema E, Mueller LA, Buels R, Giovannoni JJ, Visser RG, Stiekema WJ, et al.
460 Comparative BAC end sequence analysis of tomato and potato reveals overrepresentation of
461 specific gene families in potato. *BMC Plant Biology*. 2008;8(1):34.
- 462 51. Zhu W, Ouyang S, Iovene M, O'Brien K, Vuong H, Jiang J, et al. Analysis of 90 Mb of the
463 potato genome reveals conservation of gene structures and order with tomato but divergence in
464 repetitive sequence composition. *BMC Genomics*. 2008;9(1):286.
- 465 52. Budiman MA, Mao L, Wood TC, Wing RA. A deep-coverage tomato BAC library and
466 prospects toward development of an STC framework for genome sequencing. *Genome Research*.
467 2000;10(1):129-136.
- 468 53. Peterson DG, Pearson WR, Stack SM. Characterization of the tomato (*Lycopersicon*
469 *esculentum*) genome using in vitro and in situ DNA reassociation. *Genome*. 1998;41(3):346-356.

- 470 54. Hirakawa H, Shirasawa K, Miyatake K, Nunome T, Negoro S, Ohyama A, et al. Draft
471 genome sequence of eggplant (*Solanum melongena* L.): the representative solanum species
472 indigenous to the old world. DNA research : an international journal for rapid publication of reports
473 on genes and genomes. 2014;21(6):649-660. doi: 10.1093/dnares/dsu027.
- 474 55. Gramazio P, Blanca J, Ziarsolo P, Herraiz F, Plazas M, Prohens J, et al. Transcriptome
475 analysis and molecular marker discovery in *Solanum incanum* and *S. aethiopicum*, two close
476 relatives of the common eggplant (*Solanum melongena*) with interest for breeding. BMC Genomics.
477 2016;17(1):300.
- 478 56. Ramesh KR, Hemalatha R, Vijayendra CA, Arshi UZS, Dushyant SB, Dinesh KB.
479 Transcriptome analysis of *Solanum melongena* L.(eggplant) fruit to identify putative allergens and
480 their epitopes. Gene. 2016;576(1):64-71.
- 481 57. Zhou X, Bao S, Liu J, Zhuang Y. De novo sequencing and analysis of the transcriptome of
482 the wild eggplant species *Solanum aculeatissimum* in response to *Verticillium dahliae*. Plant
483 Molecular Biology Reporter. 2016;34(6):1193-1203.
- 484 58. Gramazio P, Prohens J, Borràs D, Plazas M, Herraiz FJ, Vilanova S. Comparison of
485 transcriptome-derived simple sequence repeat (SSR) and single nucleotide polymorphism (SNP)
486 markers for genetic fingerprinting, diversity evaluation, and establishment of relationships in
487 eggplants. Euphytica. 2017;213(12):264.
- 488 59. Chen X, Zhang M, Tan J, Huang S, Wang C, Zhang H, et al. Comparative transcriptome
489 analysis provides insights into molecular mechanisms for parthenocarpic fruit development in
490 eggplant (*Solanum melongena* L.). PloS ONE. 2017;12(6):e0179491.
- 491 60. Li J, He Y-J, Zhou L, Liu Y, Jiang M, Ren L, et al. Transcriptome profiling of genes related
492 to light-induced anthocyanin biosynthesis in eggplant (*Solanum melongena* L.) before purple color
493 becomes evident. BMC Genomics. 2018;19(1):201.

- 494 61. Xiao XO, qiu Lin W, Li K, Feng XF, Jin H, feng Zou H. Transcriptome analyses reveal
495 anthocyanin biosynthesis in eggplants. PeerJ Preprints, 2018; 2167-9843.
- 496 62. Boyaci F, Unlu A, Abak K, editors. Screening for resistance to Fusarium wilt of some
497 cultivated eggplants and wild Solanum accessions. XXVIII International Horticultural Congress on
498 Science and Horticulture for People (IHC2010): International Symposium on New Developments in
499 Plant Genetics and Breeding, 2010; 935:23-27.
- 500 63. Yamakawa K, Mochizuki H. Nature and inheritance of Fusarium-wilt resistance in eggplant
501 cultivars and related wild Solanum species. Yasai Shikenjo hokoku= Bulletin of the Vegetable and
502 Ornamental Crops Research Station Series A. 1979.
- 503 64. Collonnier C, Fock I, Kashyap V, Rotino G, Daunay M, Lian Y, et al. Applications of
504 biotechnology in eggplant. Plant Cell, Tissue and Organ Culture. 2001;65(2):91-107.
- 505 65. Behera T, Gyanendra S. Studies on resistance to shoot and fruit borer (*Leucinodes*
506 *orbonalis*) and interspecific hybridization in eggplant. Indian Journal of Horticulture.
507 2002;59(1):62-6.
- 508 66. Pandey K, Pandey P, Kalloo G, Chaurasia S. Phomopsis blight in brinjal and sources of
509 resistance. Indian Phytopathology. 2002;55(4):507-509.
- 510 67. Francis Y, Chin L, Henry C, Leung M, Yiu S. Sequence assembly using next generation
511 sequencing data--challenges and solutions. Science China Life Sciences. 2014;57(11):1140.
- 512 68. Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: robust de novo RNA-seq assembly
513 across the dynamic range of expression levels. Bioinformatics. 2012;28(8):1086-1092.
- 514 69. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn
515 graphs. Genome Research. 2008;18(5):821-829.
- 516 70. Thiel T, Michalek W, Varshney RK, Graner A. Exploiting EST databases for the
517 development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.).
518 Theoretical and Applied Genetics. 2003;106.

519 71. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nature Methods*.
520 2012;9(4):357-359.

521 72. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool.
522 *Journal of Molecular Biology*. 1990;215(3):403-410.

523 73. Consortium GO. The Gene Ontology (GO) database and informatics resource. *Nucleic*
524 *Acids Research*. 2004;32(suppl 1):D258-D261.

525 74. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. KAAS: an automatic genome
526 annotation and pathway reconstruction server. *Nucleic Acids Research*. 2007;35(suppl_2):W182-
527 W185.

528 75. Doyle JJ, Doyle JL. A rapid procedure for DNA purification from small quantities of fresh
529 leaf tissue. *Phytochemical Bulletin*. 1987;19:11-15.

530 76. Garg R, Patel RK, Tyagi AK, Jain M. De novo assembly of chickpea transcriptome using
531 short reads for gene discovery and marker identification. *DNA Research*. 2011;18(1):53-63.

532 77. Clarke K, Yang Y, Marsh R, Xie L. Comparative analysis of de novo transcriptome
533 assembly. *Science China Life Sciences*. 2013;56(2):156-162.

534 78. de Jong M, Rauwerda H, Bruning O, Verkooijen J, Spaink HP, Breit TM. RNA isolation
535 method for single embryo transcriptome analysis in zebrafish. *BMC Research Notes*. 2010;3(1):73.

536 79. Iorizzo M, Senalik DA, Grzebelus D, Bowman M, Cavagnaro PF, Matvienko M, et al. De
537 novo assembly and characterization of the carrot transcriptome reveals novel genes, new markers,
538 and genetic diversity. *BMC Genomics*. 2011;12(1):389.

539 80. Blanca J, Cañizares J, Roig C, Ziarsolo P, Nuez F, Picó B. Transcriptome characterization
540 and high throughput SSRs and SNPs discovery in *Cucurbita pepo* (Cucurbitaceae). *BMC*
541 *Genomics*. 2011;12(1):104.

Figure Legends

Fig 1. Ramnagar Giant (*S. melongena*) and W-4 (*S. incanum*) lines used for generating transcriptome assembly.

Fig 2A. Paired end reads statistics of *S. melongena*.

Fig 2B. Paired end reads statistics of *S. incanum*.

Fig 3. Graphical representation of total reads mapped summary.

Fig 4. Graph showing identified SSR with distribution.

Fig 5. Graph showing SNV distribution.

Fig 6. Graphical representation of variation classification.

Fig 7. Transcript annotation summary of *S. melongena* and *S. incanum*.

Fig 8. Differentially expressed genes details of transcripts.

Fig 9. GO summary of *S. melongena* and *S. incanum*.

Fig 10. Identified transcription factor detail statistics.

Fig 11A. PAGE images showing different transcriptome derived SSR markers used for screening polymorphism between parental lines.

Fig 11B. PAGE images showing validation of polymorphic SSR markers upon RILs.

Tables

Table 1. Assembly statistics of *S. melongena* and *S. incanum* transcriptomes.

Table 2. Raw Data Quality Statistics of *S. melongena*.

Table 3. Raw Data Quality Statistics of *S. incanum*.

Table 4. Statistics of SSR identified in *S. melongena* and *S. incanum*.

Table 5. Distribution of different SSR type classes in *S. melongena* and *S. incanum*.

565 **Supporting Information**

566 **S1 Table. Primary data statistics report.**

567 **S2 Table. *S. melongena* and *S. incanum* SSR details and primer sequences.**

568 **S3 Table. *S. melongena* and *S. incanum* SSR analysis data.**

569 **S4 Table. *S. melongena* and *S. incanum* SNV details.**

570 **S5 Table. *S. melongena* and *S. incanum* quantification data.**

571 **S6 Table. *S. melongena* transcripts annotation report.**

572 **S7 Table. *S. incanum* transcripts annotation report.**

573 **S8 Table. KEEG pathway analysis report.**

574 **S9 Table. Other analysed data.**

575 **S10 Table. *S. melongena* and *S. incanum* TFs summery.**

Fig 1

Fig 1a

Fig 2A

Fig 2B

Reads Mapped Summery of *S. melongena* and *S. incanum*

Fig 3

SSR Distribution in *S. melongena* and *S. incanum*

SNV distribution in *S. melongena* and *S. incanum*

**Classification of SNVs in
S. melongena and *S. incanum***

Fig. 6

Transcripts annotation summery of *S. melongena* and *S. incanum*

Fig. 8

GO Summary of *S. melongena* and *S. incanum*

Fig. 9

Fig. 10

Fig 11A PAGE gel images showing polymorphic pattern generated by transcriptome SSR primers; A-Primer *Si_234*, C-Primer *Si_269*, D-Primer *Si_270*, E-Primer *Si_289*, F-Primer *Si_297*, G-Primer *Si_680*, H-Primer *Si_333*, J-Primer *Si_355*. Lane 1: Ramnagar giant, Lane 2: W-4.

Fig 11B PAGE image showing amplification and polymorphism amongst 114 RILs population from the transcriptome derived SSR marker *Si_680*. Lane L: 1kb ladder; Lanes 1-114: RILs.

Table 1. Assembly statistics of *S. melongena* and *S. incanum* transcriptomes

Parameters	<i>S. melongena</i>	<i>S. incanum</i>
Total No. of Transcripts (>200bp)	145512	144948
Transcriptome Length(bp)	149498096	147728785
Min Transcript Length (bp)	201	201
Max Transcript Length (bp)	11255	12393
Average Transcript Length (bp)	1027.4	1019.2
N50 Contig Size (bp)	1737	1715

bioRxiv preprint doi: <https://doi.org/10.1101/560805>; this version posted February 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 2. Raw Data Quality Statistics of *S. melongena*

File Names	BRINJAL_1_R1	BRINJAL_1_R2
Minimum Length	100	100
Maximum Length	100	100
Average Length	100.00	100.00
Total Number of Reads	50901025	50901025
Total Number of Reads With Non-ATGC Bases	200000	37124
Percentage of Reads With Non-ATGC Bases	0.39%	0.07%
Total Number of Bases	5090102500	5090102500
Total Number of HQ Bases	5022256227	4992103422
Percentage of HQ Bases	98.67%	98.07%
Total Number of Non-ATGC Bases	234742	155062
Percentage of Non-ATGC Bases	<= 0.001%	<= 0.001%

HQ Paired End Filtered Data Quality Statistics

File Names	BRINJAL_1_R1_HQ_Reads	BRINJAL_1_R2_HQ_Reads
Minimum Length	100	100
Maximum Length	100	100
Average Length	100.00	100.00
Total Number of HQ PE Reads	50290793	50290793
Percentage HQ PE Reads	98.80%	98.80%
Total Number of Reads With Non-ATGC Bases	195584	33793
Percentage of Reads With Non-ATGC Bases	0.39%	0.07%
Total Number of Bases in HQ Reads	5029079300	5029079300
Total Number of HQ Bases in HQ Reads	4971914148	4948094496
Percentage of HQ Bases in HQ Reads	98.86%	98.39%
Total Number of Non-ATGC		

Bases in HQ Reads	228592	137464
Percentage of Non-ATGC Bases in HQ Reads	<= 0.001%	<= 0.001%

Read Distribution Summary

Total HQ Paired Reads (%)	100.58 Million (98.8 %)
Total HQ Unpaired Reads (%)	0.59 Million (0.57 %)
Total Low Quality Reads (%)	0.64 Million (0.62 %)

Table 3. Raw Data Quality Statistics of *S. sinuatum*

File Names	BRINJAL_2_R1	BRINJAL_2_R2
Minimum Length	100	100
Maximum Length	100	100
Average Length	100.00	100.00
Total Number of Reads	50735884	50735884
Total Number of Reads With Non-ATGC Bases	165813	38016
Percentage of Reads With Non-ATGC Bases	0.33%	0.07%
Total Number of Bases	5073588400	5073588400
Total Number of HQ Bases	5009410203	4981220984
Percentage of HQ Bases	98.74%	98.18%
Total Number of Non-ATGC Bases	178048	158746
Percentage of Non-ATGC Bases	<= 0.001%	<= 0.001%

HQ Paired End Filtered Data Quality Statistics

File Names	BRINJAL_2_R1_HQ_Reads	BRINJAL_2_R2_HQ_Reads
Minimum Length	100	100
Maximum Length	100	100
Average Length	100.00	100.00
Total Number of HQ PE Reads	50159275	50159275
Percentage HQ PE Reads	98.86%	98.86%
Total Number of Reads With Non-ATGC Bases	162235	34852
Percentage of Reads With Non-ATGC Bases	0.32%	0.07%
Total Number of Bases in HQ Reads	5015927500	5015927500
Total Number of HQ Bases in HQ Reads	4961753705	4939544661
Percentage of HQ Bases in HQ Reads	98.92%	98.48%
Total Number of Non-ATGC Bases in HQ Reads	173950	141882
Percentage of Non-ATGC Bases in HQ Reads	<= 0.001%	<= 0.001%

Read Distribution Summary	
Total HQ Paired Reads (%)	100.32 Million (98.86 %)
Total HQ Unpaired Reads (%)	0.55 Million (0.55 %)
Total Low Quality Reads (%)	0.6 Million (0.59 %)

Table 4. Statistics of SSR identified in *S. melongena* and *S. incanum*

Unit size of microsatellite	Minimum number of repeats
Mono-nucleotide repeats	10
Di-nucleotide repeats	6
tri-, tetra-, penta- & hexa-nucleotide repeats	5,4,4,4
Maximal number of bases interrupting 2 SSRs in a compound microsatellite	100

bioRxiv preprint doi: <https://doi.org/10.1101/160908>; this version posted February 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Statistics of microsatellite search for <i>S. melongena</i> and <i>S. incanum</i>		
Parameters / SSR motif length	<i>S. melongena</i>	<i>S. incanum</i>
Total number of sequences examined	94,838	95096
Total size of examined sequences (bp)	73634990 (~73MB)	73796620 (~73MB)
Total number of identified SSRs	11262	11829
Number of SSR containing sequences	9839	10235

Table 5. Distribution of different SSR type classes in *S. melongena* and *S. incanum*

SSR class	Number of SSRs Detected		% of SSRs Detected	
	<i>S. melongena</i>	<i>S. incanum</i>	<i>S. melongena</i>	<i>S. incanum</i>
Mononucleotides	6231	6123	57.80	54.62
Dinucleotides	1598	1607	14.82	14.33
Trinucleotides	2374	2402	22.02	21.43
Tetranucleotides	103	354	0.96	3.16
Pentanucleotides	14	90	0.13	0.80
Hexanucleotides	-	175	-	1.56
Complex	460	528	4.27	4.10