Indian J. Genet., 79(4) 732-740 (2019)
DOI: 10.31742/IJGPB.79.4.12

Deciphering the unique SNPs among leading Indian tomato cultivars using double Digestion Restriction Associated DNA sequencing

Reddaiah Bodanapu, Sreehari V. Vasudevan, Navajeet Chakravartty, Krishna Lalam, Sivarama Prasad Lekkala, Boney Kuriakose[1], Chetan Bhanot, Navitha Kokkonda[1], George Thomas[1], Saurabh Gupta*, Vijaya Bhasker Reddy Lachagari*

AgriGenome Labs Pvt. Ltd., BTIC, MN iHub, Genome Valley, Hyderabad 500 078; [1]AgriGenome Labs Pvt. Ltd., Infopark Road, Kakkanad 682 042, Kerala

(Received: May 2019; Revised: July 2019; Accepted: September 2019)

Abstract

Worldwide grown and consumed tomato (*Solanum lycopersicum* L.) is used as model crop for new cultivar and fruit development. Genetic and genomic studies on Indian tomato cultivars will provide an insight that will enable development of breeding strategies and crop improvement. The present study aims to identify the high quality common and unique SNPs and INDELs, present in 9 different Indian tomato cultivars using double digest restriction site-associated DNA sequencing (ddRAD-seq). A total of 36.8 million raw reads were generated for selected cultivars and an average of 94% high quality reads of each were uniquely aligned to the reference tomato genome (SLv3.0). Out of 6,957 SNPs and 188 INDELs, we found 1,165 SNPs and 68 INDELs in genic regions.The genetic relationship among these cultivars suggested 4 well-differentiated groups of cultivars. Similarly, 7 and 33 SNPs were identified in chloroplast and mitochondrial genomes of tomato. SNPs markers were identified for common and specific genes associated with different pathways and their gene ontology (GO) annotated. These SNPs/INDELs could be useful as markers for variety identification for genetic purity analysis. Findings from this work will be useful to the research community, particularly plant breeders as a resource for SNP marker development.

Key words: Tomato, SNPs, INDELs, ddRAD-seq, Genomic resources

Introduction

Tomato (*Solanum lycopersicum* L.) a highly consumed vegetable across the world, originated from South and Central America and spread to the rest of the world with accompanying morphological diversification. India is the second largest producer of tomato after China. Notably, the fruit colours, sizes, shapes, tastes and flavours of various cultivars are being associated with local environments and gastronomies (Nag et al. 2017; Khan et al. 2017). Among the Indian states, Andhra Pradesh holds the top position in production of tomato (Monthly Report Tomato January 2018, http://agriculture.gov.in/). Varieties such as Pusa-120, Pusa Ruby, HS-101, HS-102, Hisar Arun, and Hisar Lalit are endorsed for cultivation across India. In addition, numerous other cultivars i.e., Kashi Aman, Kashi Abhiman, Arka Ananya, Arka Vardhan, Arka Saurabh, Arka Meghali, Arka Vishal, Arka Abhijit, Arka Ahuti (Sel-11) Arka Vikas (Sel-22), Arka Abha (BWR-1), Arka Alok (BWR-5), Kashi Amrit, Kashi Anupam, Kashi Sharad, Pusa Sheetal, Pusa Gaurav, Pusa Rohini, Pusa Hybrid-2, Pusa Hybrid-4 and Pusa Hybrid-8 are widely used for cultivation in different parts of India (Singh et al. 2016). Genome sequencing of 84 tomato accessions including wild species of *S. lycopersicon*, *S. arcanum*, *S. eriopersicon* and *S. neolycopersicon* have been conducted by the 100 tomato genome projects. Recently, the 150 tomato genome ReSequencing project (https://www.tomatogenome.wur.nl/) was established and started to explore the in-depth genetic variation available in tomato (Ezura et al. 2016; Aflitos 2014; Causse et al. 2013; Lin et al. 2014). These attempts were aimed to categorizing genomewide Single Nucleotide Polymorphisms (SNPs) within *S. lycopersicum* along with shared polymorphisms among closely related species (Shirasawa et al. 2013). Additionally, numerous interspecific genetic linkage maps have been constructed between well known cultivars of tomato and were used to identify the responsible genes for

*Corresponding author's e-mail: vb.reddy@aggenome.com; saurabh.g@aggenome.com

Published by the Indian Society of Genetics & Plant Breeding, A-Block, F2, First Floor, NASC Complex, IARI P.O., Pusa Campus, New Delhi 110 012; Online management by www.isgpb.org; indianjournals.com

interspecific and intraspecific phenotypic variations (DeVicente et al. 1993; Lin et al. 2014).

High through put sequencing and genotyping methods have been playing a crucial role in the progress of genomics and genetics. Rapid progress in next-generation sequencing (NGS) technologies have made it easy to generate large number of SNPs in both model, non-model crop and vegetable plant species (Cloutier et al. 2016; Davey et al. 2011). Congruently, double digest restriction site-associated DNA sequencing (ddRAD-Seq) technology has recently become popular due to its flexibility, low cost, and advantage over genotyping by sequencing (GBS) and restriction site-associated DNA sequencing (RAD-Seq) (Elshire et al. 2011; Baird et al. 2008). In ddRAD-Seq two restriction enzymes are employed for digestion of genomic DNA which reduces the time and cost to prepare the sequence libraries, allows paired-end sequencing (Peterson et al. 2012). In present study, ddRAD-Seq was performed for nine well known tomato cultivars *viz*., Arka Abha (BWR-1), Arka Ahuti (Sel-11), Arka Alok, Arka Ashish, Arka Saurabh, Arka Meghali, Arka Vikas (Sel-22), Arka Vikas-vir and Periyakulam 1 (PKM1) to annotate and investigate the unique variants across afore mentioned lines to identify individual markers. Subsequently, the effects of these variants on different gene functions were also investigated and explored for marker assisted selection and breeding of the tomato cultivars.

Materials and methods

Genomic DNA isolation and double enzyme digestion

Eight Arka series of lines were obtained from IIHR, Bengaluru, India and PKM1 was obtained from Tamil Nadu Agricultural University, Coimbatore, Tamil Nadu, India. These nine tomato cultivars were cultivated in the fields of AgriGenome Labs Pvt. Ltd. at Hyderabad, India. Three-week-old plant's leaves were collected and genomic-DNA (gDNA) was isolated using the DNeasy Plant Mini Kit (Qiagen, Hilden, Germany). The gDNA was quantified were checked using agarose gel electrophoresis,BioSpectrometer fluorescence (Eppendorf, Germany) and Qubit fluorometer (Life Technologies, Carlsbad, CA, USA). The quantified gDNA was used for restriction enzyme digestion.

MluCI and *SphI* enzyme pairis capable of producing a large number of fragments in a broad range of plant species (Burns et al. 2017). Hence same enzyme combination was used to generate high number of fragments from the gDNA of selected cultivars. 1μg gDNA of each cultivar was double digested with *MluCI* and *SphI* restriction enzymes by incubating at 37°C for 16-20 hours.The ligation of P1 and P2 (Barcoded) adapters was carried out using T4 DNA ligase. Ligated product was fractionated by 2% SybrSafe agarose gel electrophoresis to select the product size in between 250-350 bp (Peterson et al. 2012). Cleaning was performed using AMpure XP beads (Beckman Coulter Genomics).

Data generation and variants annotation

Quality and quantity check of ligated samples were performed and to increase the concentration of sequencing libraries, PCR amplification (8-12 cycles) was performed by adding Index 1 and Index 2 (8nt long) for multiplexing using Phusion^{TM}Taq DNA polymerase kit. Products of PCR amplification was analysed in Agilent bioanalyzer to determine molarity and fragment size distribution (Peterson et al. 2012). For each sample, individual libraries were prepared and included in one lane. The sequencing was done based on V4 chemistry on the HiSeq-4000 platform.

The reads were filtered based on presence of specific RAD tags, followed by base trimming using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and adapter trimming using cutadapt (v1.8.1) (Martin 2011). The filtered data was aligned to tomato nuclear genome of *S. lycopersicum* (v3.0) obtained from Sol Genomics (https://solgenomics.net/) and with its mitochondria and chloroplast genomes downloaded from NCBI using Bowtie2 (v2.2.2.9) (Langmead et al. 2012). The variant calling was performed based on aligned file of the samples with the reference genome using SAM tools (v0.18) (Li et al. 2009). Variants were filtered based on read depth (RD), minimum alleles frequency (MAF). In addition, polymorphic homozygous markers were identified between the cultivars using an in-house PERL scripts. Further, functional annotation of the identified variants associated genes was performed using SnpEff(v3.1) (http://snpeff.sourceforge.net/).

Zygosity, diversity, phylogenetic and kinship analysis

The zygosity analysis, diversity analysis, phylogenetic and kinship analysis, were carried out for nine cultivars based on the genotype data at RD $\geq$ 10, MAF $\leq$ 0.05. The dendrogram was constructed based on the genotype data using similarity matrix generated by Neighbor Joining (NJ) module. The degree of kinship among individuals was estimated based on the genotype using VanRaden's method (VanRaden 2008). Kinship analysis was studied using Centered-IBS matrix using Trait Analysis by Association, Evolution and Linkage (TASSEL5) v5.2.28 (Bradbury et al. 2007). Further, functional annotation and gene ontology analysis of the genes associated with common and unique variants were identified using an in-house annotation pipeline and required information were fetched from NCBI and UniProt database.

Results and discussion

Datageneration analysis and alignments

Genomic DNA sampleof each cultivar was digested with *MluCI* and *SphI* restriction enzymes having different frequencies of recognition sites. Quality check of fragmented samples indicates that all fragments conformed to screening criteria. Furthermore, quality check, screening and filtering of raw data reveals different read statistics as depicted in Table 1 (Lachagari et al. 2019; Gupta et al. 2018). Highest and lowest number of reads with RAD Tag was found in Arka Abha (BWR-1), and Arka Vikas-Vir respectively. The percentage of uniquely aligned reads from Arka Abha (BWR-1), Arka Ahuti (Sel-11), Arka Alok (BWR-5), Arka Ashish, Arka Meghali, Arka Saurabh, Arka Vikas (Sel-22), PKM1 and Arka Vikas-vir with reference tomato genome was reported as 93.56, 94.38, 92.86, 94.19, 94.42, 93.22, 93.36, 93.30, and 94.72%, respectively. Average of 3.97% and 3.80%of unique reads were mapped to mitochondrial and chloroplast genomes respectively (Table 1). Density plot of uniquely aligned reads with reference tomato genome indicates their distribution over all 12 chromosomes of tomato.

Identification of polymorphic variants and cultivar specific markers

A total of 6,957 SNPs and 188 INDELs were observed collectively from nine cultivars when compared with the nuclear genome. The variants were found in both genic and intergenic regions of the nuclear genome and showed distribution over all chromosomes (Table 2). Out of 1,165 SNPs in genic regions, the highest number of SNPs we found to be associated with Arka Ashish (460) and Arka Alok (BWR-5) (416) while Arka Ahuti (185) shared the lowest contribution (Supplementary Table S1). All Supplementary Tables (Table S1 to S5) are available in the haward dataserve and can be accessed through https://doi. org/10.7910/DVNLLXNVGE. Similarly, out of 188 INDELs found in genic regions, shows that Arka Ashish (42) shared the highest contribution (Fig. 1 and Table 1). Distribution of SNPs unique to cultivar or cultivar specific genic regions were also identified *viz*., Arka Abha (BWR-1) (50), Arka Ahuti (Sel-11) (29), Arka Alok (BWR-5) (73), Arka Ashish (140), Arka Meghali (26), Arka Saurabh (21), Arka Vikas (Sel-22)(30), PKM1(47) and Arka Vikas-vir (44) (Supplementary Table S2). Similarly, cultivar specific INDELs i.e., Arka Ahuti (Sel-11) (1), Arka Ashish (1), PKM1(2) and Arka Vikas-vir(1) were also reported (Supplementary Table S3). Annotation of SNPs were categorised into missense, 3' or 5'-UTR, splice region, up- and down-stream gene variants and missense mutations which are responsible for phenotypic trait change.

Genetic relationship analysis among tomato cultivars

This analysis identified 6,957 SNPs sitesacross nine cultivars when compared with tomato reference genome. Genotyping data revealed 139,140 gametes at read depth 10 with MAF $\leq$ 0.05. Arka Ahuti (Sel 11) showed the lowest percentage heterozygosity (7.863), followed by PKM1(11.8), Arka Vikas-vir (12.10), Arka Meghali (22.66), Arka Vikas (Sel-22) (24.78), Arka Saurabh (26.46), Arka Ashish (26.82) Arka Alok (BWR-5) (31.594) and Arka Abha (BWR-1) (36.136). This analysis indicated less heterozygosity resulting in more homologues to reference genome. Phylogenetic analysis among selected cultivars and reference tomato genome showed their classification into 4 main clades on the basis of SNPs clustering: clade 1 consists of *S. lycopersicum* (v3.00), Arka Ahuti (Sel-11), Arka Abha (BWR-1) and Arka Ashish; clade 2 consists of Arka Saurabh, Arka Vikas (Sel-22), PKM1; clade 3 has Arka Vikas-vir, Arka Meghali and clade 4 has only Arka Alok (BWR-5) (Fig. 2). The larger clades, *viz*., clade1, clade 2, contained two major subclades showing their genetic relatedness. Similar types of analyses in onion inbred lines (Lee 2018) and grape cultivars (Laucou et al. 2018).

Identification of SNPs in chloroplast and mitochondrial genome

A total of 7 and 33 SNPs (RD $\geq$ 10) were identified while comparing with chloroplast and mitochondrial genome respectively (Supplementary Tables S4 and S5). Annotation of SNPs shows only one SNPs at position 2,239 in *Mttb* gene with variation A/T, found in cultivar Arka Meghali, PKM1, whereas A/G in Arka

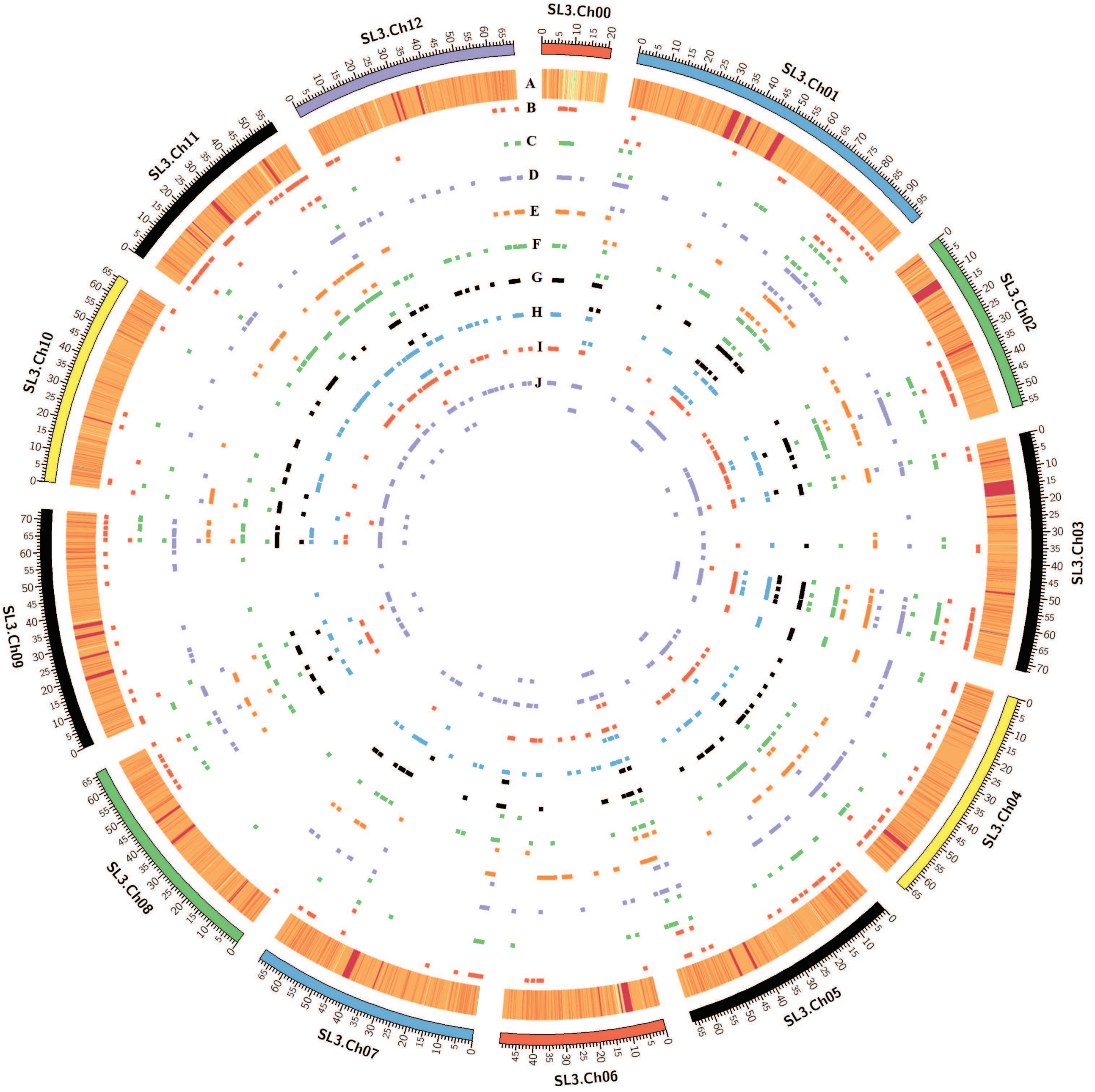


Fig. 1. Chromosomal distribution of generated reads and identified SNPs and INDELs in each cultivar of tomato in a Circos plot. (A) number of generated base pair density in a chromosomal location shown heatmap. Distribution of identified SNPs (up) and INDELs (down) via scatter plots for all nine cultivars i.e. (B) ArkaAbha (BWR-1), (C) Arka Ahuti (Sel-11), (D) Arka Alok (BWR-5), (E) Arka Ashish (F) Arka Meghali (G) Arka Saurabh, (H) Arka Vikas (Sel-22), (K) Arka Vikas-vir and (J) PKM1

Saurabh. *Mttb* (Acc. No. NC_035963.1) is a *SecY*-independent transporter protein that helps in proton motive force dependent protein transmembrane transporter activity. Remaining identified SNPs are part of the non-coding, chloroplast and mitochondrial genomes, may be contributing to specific phenotypic traits of each cultivar.

Functional annotation of missense SNPs associated genes

Distribution of identified synonymous, missense, intron, downstream gene, upstream gene variants, splice region, 5'- and 3'- prime UTR and stop gain SNPs of each cultivaris depicted in Fig. 3. Intron variants was observed across all cultivars as opposed

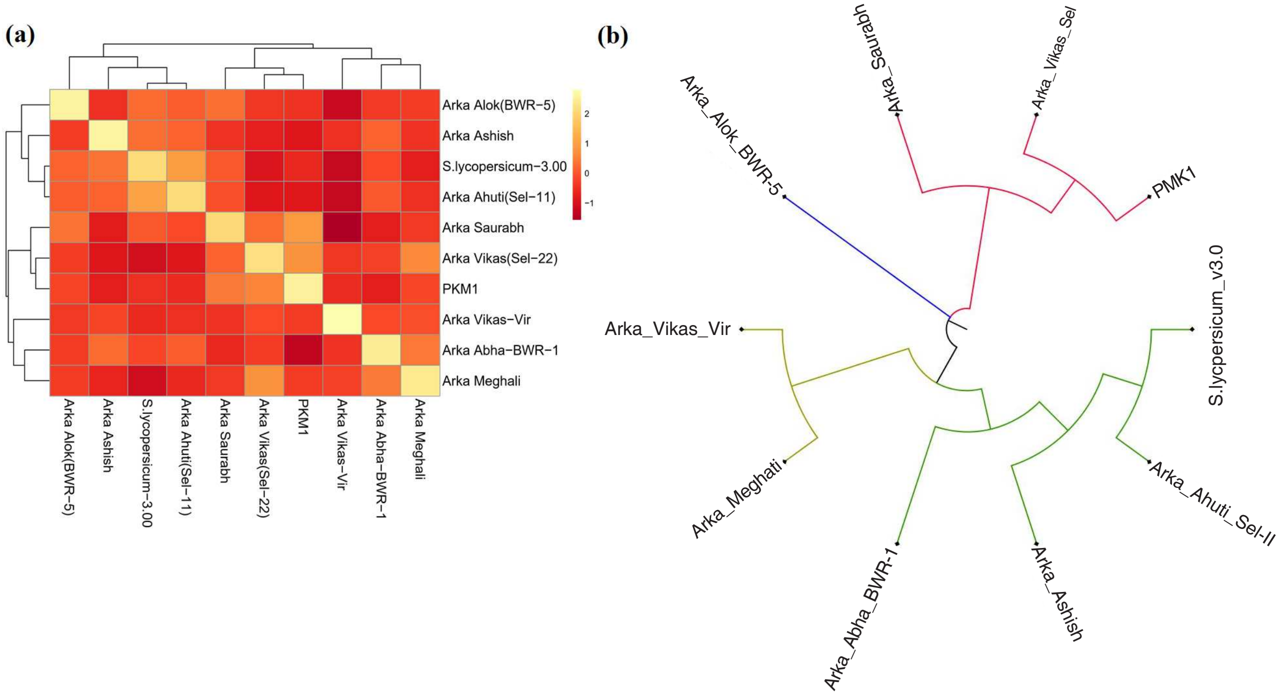


Fig. 2. **Diversity and genetic relationship analysis among all selected cultivars with reference *S. lycopersicum* (v3.00) genome. (a) Heatmap of kinship analysis identified SNPs across all cultivar and (b) Circular phylogenetic plot shows evolutionary divergence among the cultivars**

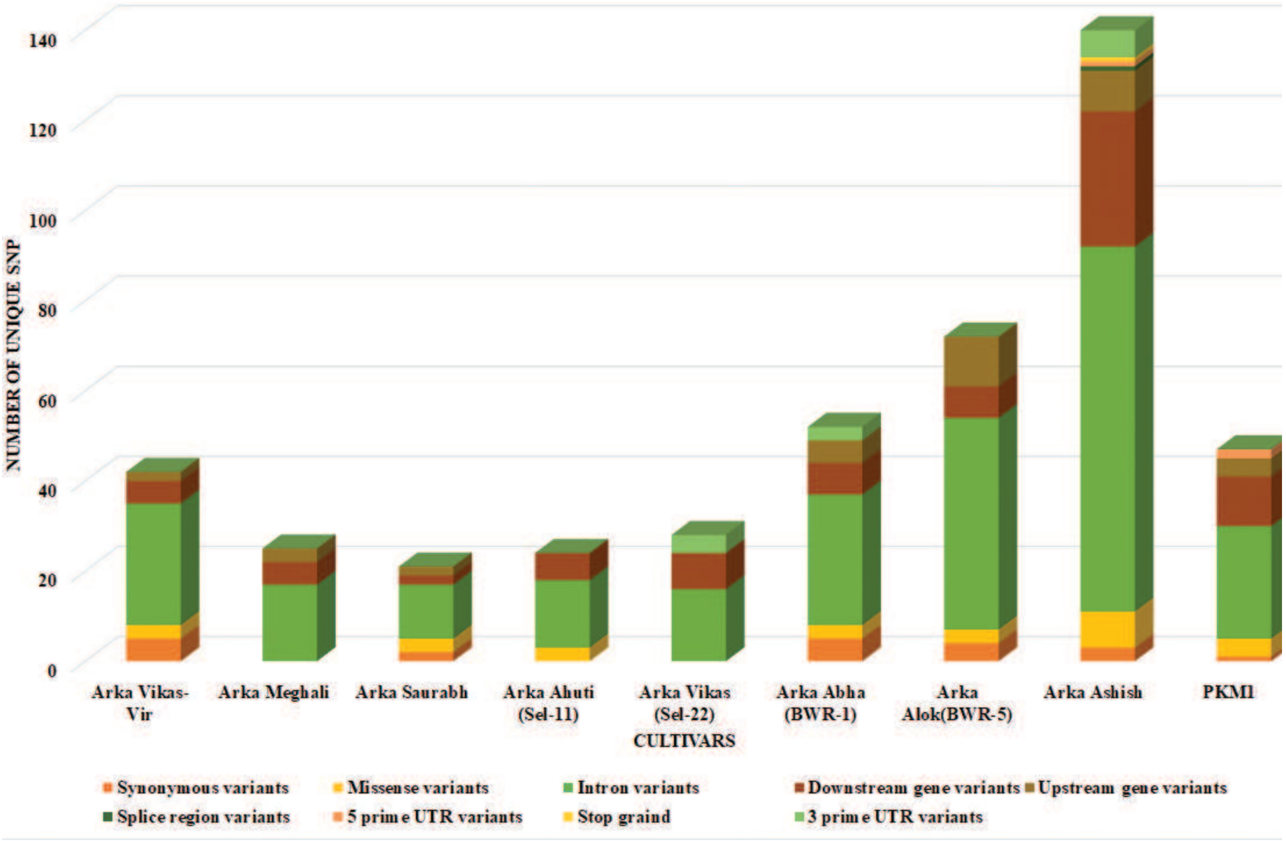


Fig. 3. Cultivar-wise distribution of identified unique SNPs with their effects

Table 1. Sequenced reads measurements across selected 9 cultivars along with statistics genome-wide alignment and annotated SNPs and INDELs at read depth (RD) $\geq$10

Sample names	# of reads with RAD-Tag in (R1+R2)	# of high-quality reads	# of aligned reads	% of aligned reads	% of uniquely aligned reads with nuclear genome	# of aligned reads and % of mapping with Mitochondrial genome	# of aligned reads and % of mapping with Chloroplast genome	# of SNP	# of genic SNPs	# of cultivar specific unique genic SNPs	# of INDELs	# of genic INDELs	# of cultivar specific unique genic INDELs
ArkaAbha (BWR-1)	6,404,140	6,404,090	5,439,634	84.94	93.56	225,520(3.52%)	273,094(4.26%)	2,764	396	50	84	38	0
ArkaAhuti(Sel-11)	3,666,454	3,666,384	3,286,179	89.63	94.38	155,528(4.24%)	157,718(4.30%)	732	185	29	37	30	1
Arka Alok (BWR-5)	4,126,266	4,125,944	3,510,353	85.08	92.86	145,866(3.53%)	161,490(3.91%)	2,198	416	73	48	33	0
Arka Ashish	5,282,126	5,281,902	4,708,815	89.15	94.19	238,658(4.51%)	251,490(4.76%)	2,260	460	140	66	42	1
ArkaMeghali	4,386,412	4,386,366	3,907,374	89.08	94.42	183,282(4.17%)	226,516(5.16%)	2,199	357	26	76	34	0
Arka Saurabh	4,529,000	4,528,704	3,809,998	84.13	93.22	156,462(3.45%)	182,570(4.03%)	1,875	349	21	56	35	0
Arka Vikas(Sel-22)	3,619,712	3,619,558	3,129,469	86.46	93.36	125,388(3.46%)	153,646(4.24%)	2,117	340	30	49	37	0
PKM1	4,109,130	4,108,926	3,510,666	85.44	93.30	137024(3.334%)	56,338(1.37%)	1,862	302	47	67	40	2
Arka Vikas -Vir	725,242	725,218	678,659	93.58	94.72	39,948(5.50%)	158,59(2.18%)	725	256	44	13	11	1

Note: # indicates Number and % indicates Percentage

to other types of variants. Important missense SNPs associated with the genes were also identified (Supplementary Tables S1 and S2). Seven missense SNPs of Arka Ahuti (Sel-11) associated with 7 genes in which three SNPs are unique to this cultivar. *Auxin transport protein*, *TCP transcription factor 4* and *Gibberellin receptor GID1A* have unique mutations K2704T, L515S and W289C, respectively; possibly these mutations enhance the essential hormones that regulates growth and development (Murase 2008; Gil 2001; Gupta et al. 2018a). A total of 17 and 18 SNPs found in Arka Vikas (Sel-22) and Arka Vikas-vir respectively, sharing some common SNPs, while they also have one and three unique SNPs, respectively. Three unique SNPs of Arka Vikas-virare associated with 30S ribosomal protein S14 (T22A and R8H), and DNA repair endonuclease UVH1 (G866S), while in case of Arka Vikas (Sel-22), the genesare not functionally annotated. Twenty-one missense SNPs were observed in Arka Saurabh cultivar in which 3 SNPs are unique to this variety. These mutations found in *Serine/threonine-protein kinase Nek1* (H199N), *Nucleotidyl transferase* family (H322R), *Programmed cell death 7* (S313N), *Cation/*H^+*antiporter* (N620D), *RNA polymerase-associated protein RTF1* (K185N), *Dentin sialophospho protein* (Q428P), *DnaJ* (A1224S), *VHS domain-containing protein* (G661D), *Exostosin* (F307S), *BTB/POZ domain-containing protein* (D245E), *Transducin/WD40 repeat-like superfamily protein* (R672C), Intracellular protein transporter *USO1-like protein* (R240W), *ARM repeat superfamily protein* (N770S), *Calcium-dependent lipid-binding-like protein* (V2933F), *Poly(A)-RNA polymerase cid14* (R1188H), *Tetratricopeptide repeat (TPR)-like superfamily protein* (V71D), *Pentatricopeptide repeat-containing protein* (L238V) and *Pleiotropic drug resistance ABC transporter* (A515P), may be responsible for various stress tolerance and resistance phenotypic traits of this cultivar (Gupta et al. 2018b; Gupta et al. 2017b). Out of 19 missense SNPs, three unique SNPs are found in ArkaAbha (BWR-1) and mutating the disease resistance protein RPP5 (D22E), *Alpha/beta-Hydrolases superfamily protein* (V306I) and *Proteasome-associated ECM29-like protein* (A784D) enhancing the disease tolerance, aligning with this cultivar characteristics (Belkhadir et al. 2004). Eight unique missense SNPs out of 22

Table 2. Chromosome wise distribution of identified SNPs and INDELs

Chromosome	# of SNPs Loci	# of genic SNPs Loci	# of INDELs Loci	# of genic INDELs Loci
SL3.0ch00	621	37	3	0
SL3.0ch01	309	96	16	10
SL3.0ch02	600	136	9	5
SL3.0ch03	263	102	11	9
SL3.0ch04	843	185	17	8
SL3.0ch05	507	76	12	7
SL3.0ch06	646	81	8	3
SL3.0ch07	132	58	14	4
SL3.0ch08	132	46	4	0
SL3.0ch09	289	97	11	3
SL3.0ch10	596	44	16	5
SL3.0ch11	1364	149	45	12
SL3.0ch12	655	58	22	2
TOTAL	6,957	1,165	188	68

Note: # indicates Number and % indicates Percentage

SNPs reported in Arka Ashish are associated with *Cell division cycle protein27* (K575T), *N-glycosylase/ DNA lyase* (R16H), *Carbohydrate-bindingprotein* (K1538R), *Plant cadmium resistance 10* (S9L), *Zinc finger C_2H_2 type* (T233A) and *Histidine-tRNA ligase* (T423P). Combined effect of these SNPs possibly participating in physiological traits and tolerant to powdery mildew characteristic provide this variety with survival advantages. Arka Alok (BWR-5) is bacterial wilt resistant variety with medium size fruit having 17 SNPs in which three unique SNPs cause mutations in *Pentatricopeptide repeat-containing protein* (E498G), *Mitochondrial trehalose-6-phosphate synthase-6* (V466A) and *GI protein* (A1011P). Plant specific nuclear *GI protein*is involved in variety functions *viz*., flowering time regulation, light signalling, herbicide tolerance, cold tolerance, drought tolerance, hypocotyl elongation, control of circadian rhythm, sucrose signalling,starch accumulation, chlorophyll accumulation, transpiration, and *miRNA* processing (Mishra et al. 2015; Gupta et al. 2019). Arka Meghali has 14 missense SNPs associated with *Serine/ threonine-protein kinase Nek1*, *Nucleotidyl transferase family protein*, *Programmed cell death 7 protein*, *Dentinsialophosphoprotein, VHS domain-containing protein, BTB/POZ domain containing protein, Intracellular protein transporter USO1-like protein, ARM repeat superfamily protein, Bidirectional sugar transporter SWEET, Peroxidase, Dynein-1-alpha* heavy chain, *flagellar inner arm I1 complex* and *ARM repeat superfamily proteins* with mutations of H199N, H322R, S313N, Q428P, G661D, D245E, K232I, N770S, N70K & P45A, K91M, A6V & F7C and T510A respectively, out of these, none of the missense SNPs is unique for this cultivar. Besides Arka varieties, we also identified 17 SNPs in PKM1 cultivar of which, four SNPs are unique and cause the mutations in proteins *Brefeldin A-inhibited guanine nucleotide-exchange protein 2* (K876R), *Mediator of RNA polymerase II transcription* (T2N), *Microtubule-associated protein* (A147P) and *Gibberellin receptor GID1A* (G301C) which may possibly control the regulation of transcription and nutrient transportation form the soil (Murase 2008; Blazek 2005; Gupta et al. 2017a; Gupta et al. 2017b; Mishra et al. 2019).

Functional annotation of INDELs containing genes

The INDELs that are unique to one cultivaralong with associated genes having functional annotations were identified (Supplementary Table S3). Arka Ashish (42) was having highest INDELs count, followed by Arka Abha (BWR-1) (38), Arka Vikas (Sel-22) (37), Arka Saurabh (35), Arka Meghali (34), Arka Ahuti (Sel-11) (29). Arka Ashish showedone unique insertion of 'TA' in place of 'T' in the gene encoding kinase family with *ARM repeat domain-containing protein*. Similarly, we also observed unique deletion of 'CTTTTTTTTTT' in gene encoding *importin subunit beta-1 protein* and 'CAAAAAAAAA' in gene encoding *Ubiquitin-like-specific protease ESD4 protein* of Arka Ahuti (Sel-11), and Arka Vikas-vir. PKM1 showed 40 INDELs in which most of which are like the Arka cultivar; only two INDELs were found unique to this cultivar. Identified unique and common deletions can be useful in determining genetic identity and mapping studies.

Gene ontology examination of genic region SNPs

The gene ontology (GO) analysis of genic region SNPs was identified and genes with variants belonging to various biological processes, molecular functions and cellular components.Biological processes GO terms showed that Transcription [GO:0006351] and Regulation of transcription [GO:0006355] are supported by highest number genes among this group (Supplementary Fig. S1a), while considering the molecular function: ATP binding [GO:0005524] GO term supported by 120 genes (Supplementary Fig. S1b). Similarly, GO terms associated with cellular components has been identified and found that 193 and 130 genes supporting integral components of membrane [GO:0016021] and nucleus [GO:0005634] respectively (Supplementary Fig. S1c).

Overall study provides cultivar specific as well as common SNPs and INDELs for nine leading tomato cultivars of India using ddRAD-seq analysis. Identified SNPs with their genetic characteristics and functional annotations that enabled design markers for cultivar specific SNPswill be useful for variety identification through genetic purity analysis. Besides, identified information can be useful for plant breeders to develop/improve tomato cultivars.

Author's contributions

Conceptualization of research (BR, VBRL, GT); Designingof the experiments (BR, LK, SPL, VBRL, SG); Executionof field/lab experiments and data generation (BR, LK, SPL, BK, NV); Execution *in silico* works (SV, SG, NC, CB); Analysis of data and interpretation (SV, SG, NC); Preparation ofmanuscript (SG, VBRL, SPLN, GT).

Declaration

The authors declare no conflict of interest. This research did not involve any experiment on humans or animals. Data generated in this work was related to rice plant and submitted in SRA databaseBioProject ID: PRJNA484084. Hence, all the authors declare that there is no non-compliance with ethical standards.

Acknowledgments

Authors acknowledgeICAR-Indian Institute of Horticulture Research for providing the Arka series tomato seeds and Tamil Nadu Agricultural University for providing PMK1 tomato seeds.Study was supported by SciGenom Research Foundation (SGRF). The preprint version of manuscript available in BioRxivhttps://www.biorxiv.org/content/10.1101/541227v1.

References

Aflitos S., Schijlen E., Jong H., Ridder D., Smit S., Finkers R. and Bakker F. 2014. Exploring genetic variation in the tomato (*Solanum* section *Lycopersicon*) clade by whole genome sequencing. Plant J., **80**: 138-148. doi:10.1111/tpj.12616.

Baird N. A., Etter P. D., Atwood T. S., Currey M. C., Shiver A. L., Lewis Z. A. and Johnson E. A. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PloS One, **3**: e3376. doi: 10.1371/journal.pone.0003376.

Belkhadir Y., Subramaniam R. and Dangl J. L. 2004. Plant disease resistance protein signaling: NBS–LRR proteins and their partners. Curr. Opin. Plant Biol., **7**: 391-399. doi: 10.1016/j.pbi.2004.05.009.

Blazek E., Mittler G. and Meisterernst M. 2005. The mediator of RNA polymerase II. Chromosoma., **113**: 399-408.doi: 10.1007/s00412-005-0329-5.

Bradbury P. J., Zhang Z., Kroon D. E., Casstevens T. M., Ramdoss Y. and Buckler E. S. 2007. TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics., **23**: 2633-2635. doi: 10.1093/bioinformatics/btm308.

Burns M., Starrett J., Derkarabetian S., Richart C. H., Cabrero A. and Hedin M. 2017. Comparative performance of double-digest RAD sequencing across divergent arachnid lineages. Mol. Ecol. Resour., **17**: 418-430. doi: 10.1111/1755-0998.12575.

Cloutier S., Banks T. W. and Kumar S. 2016. SNP Discovery Through Next-Generation Sequencing and Its Applications. New Jersey 08758 USA: Apple Academic Press.

Davey J. W., Hohenlohe P. A., Etter P. D., Boone J. Q., Catchen J. M. and Blaxter M. L. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat. Rev. Genet., **12**: 499-510. doi: 10.1038/nrg3012.

Elshire R. J., Glaubitz J. C., Sun Q., Poland J. A., Kawamoto K., Buckler E. S. and Mitchell S. E. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One., **4**: e19379. doi: 10.1371/journal.pone.0019379.

Gil P., Dewey E., Friml J., Zhao Y., Snowden K. C., Putterill J. and Chory J. 2001. BIG: a calossin-like protein required for polar auxin transport in *Arabidopsis*. Genes Dev., **15**: 1985-1997. doi: 10.1101/gad.905201.

Gupta S., Gupta V., Singh V. and Varadwaj P. K. 2018a. Extrapolation of significant genes and transcriptional regulatory networks involved in *Zea mays* in response in UV-B stress. Genes Genomics, **40**: 973-990. doi: 10.1007/s13258-018-0705-1.

Gupta S., Singh Y., Kumar H., Raj U., Rao A. R. and Varadwaj P. K. 2018b. Identification of novel abiotic stress proteins in *Triticum aestivum* through functional annotation of hypothetical proteins-Interdiscip Sci., **10**: 205-220. doi: 10.1007/s12539-016-0178-3.

Gupta S., Mishra V. K., Kumari S., Chand R. and Varadwaj P. K. 2019. Deciphering genome-wide WRKY gene family of *Triticum aestivum* L. and their functional role in response to Abiotic stress. Genes Genomics, **41**: 79-94. doi: 10.1007/s13258-018-0742-9.

Gupta S., Yadav B. S., Raj U., Freilich S. and Varadwaj P.

K. 2017a. Transcriptomic analysis of soil grown T. aestivum cv. root to reveal the changes in expression of genes in response to multiple nutrients deficiency. Front Plant Sci., **8**: 1025. doi: 10.3389/fpls.2017.01025.

Gupta S., Kumari M., Kumar M. and Varadwaj P. K. 2017b. Genome-wide analysis of miRNAs and Tasi-RNAs in *Zea mays* in response to phosphate deficiency. Funct Integr Genomics., **17**: 335-351. doi: 10.1007/s10142-016-0538-4.

Khan M. A., Butt S. J., Nadeem F., Yousaf B. and Javed H. U. 2017. Morphological and physico-biochemical characterization of various tomato cultivars in a simplified soilless media. Ann. Agri. Sci., **62**: 139-143. doi: 10.1016/j.aoas.2017.10.001.

Lachagari V. R., Bodanapu R., Chakravartty N., Lekkala S. P., Lalam K., Kuriakose B. and Reddy A. R. 2019. Uncovering genome wide novel allelic variants for eating and cooking quality in a popular Indian rice cultivar, Samba Mahsuri. Curr. Plant Biol., **18**: 100111. doi: 10.1016/j.cpb.2019.100111.

Langmead B. and Salzberg S. L. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods., **9**: 357.doi: 10.1038/nmeth.1923.

Laucou V., Launay A., Bacilieri R., Lacombe T., Adam-Blondon A. F., Bérard A. and Le Paslier M. C. 2018. Extended diversity analysis of cultivated grapevine Vitis vinifera with 10K genome-wide SNPs. PloS one., **13**: e0192540. doi: 10.1371/journal.pone.0192540.

Lee J. H., Natarajan S., Biswas M. K., Shirasawa K., Isobe S., Kim H. T. and Nou I. S. 2018. SNP discovery of Korean short day onion inbred lines using double digest restriction site-associated DNA sequencing. PloS one, **13**(8): e0201229. doi:https://doi.org/10.1371/journal.pone.0201229.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N. and Durbin R. 2009. The sequence alignment/map format and SAMtools. Bioinformatics, **25**: 2078-2079. doi: 10.1093/bioinformatics/btp352.

Lin T., Zhu G., Zhang J., Xu X., Yu Q., Zheng Z. and Huang Z. 2014. Genomic analyses provide insights into the history of tomato breeding. Nat. Genet., **46**: 1220-6. doi: 10.1038/ng.3117.

Singh M., Prasanna H. C., Tiwari S., Gujar R. S. and Karkute S. G. 2016. Biology of Solanum lycopersicum (Tomato). New Delhi: Ministry of Environment, Forest and Climate Change Goverment of India.

Causse M., Desplat N., Pascual L., Le Paslier M. C., Sauvage C., Bauchet G. and Bouchet J. P. 2013. Whole genome resequencing in tomato reveals variation associated with introgression and breeding events. BMC Genomics, **14**: 791. doi: 10.1186/1471-2164-14-791.

Mishra V. K., Gupta S., Chand R., Yadav P. S., Singh S. K., Joshi A. K. and Varadwaj P. K. 2019. Comparative transcriptomic profiling of High-and Low-grain Zinc and Iron containing Indian wheat genotypes. Curr. Plant Biol., **18**: 100105. doi: 10.1016/j.cpb.2019.100105.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMB net. J., **17**: 10-12.doi.org/10.14806/ej.17.1.200.

Mishra P. and Panigrahi K. C. 2015. GIGANTEA–an emerging story. Front Plant Sci., **26**: 6:8. doi: 10.3389/fpls.2015.00008.

Murase K., Hirano Y., Sun T. P. and Hakoshima T. 2008. Gibberellin-induced DELLA recognition by the gibberellin receptor GID1. Nature, **456**: 459-463. doi: 10.1038/nature07519.

Nag S. O. 2017. The World's Leading Producers of Tomatoes. Accese date april 25. https://www.worldatlas.com/articles/which-are-the-world-s-leading-tomato-producing-countries.html.

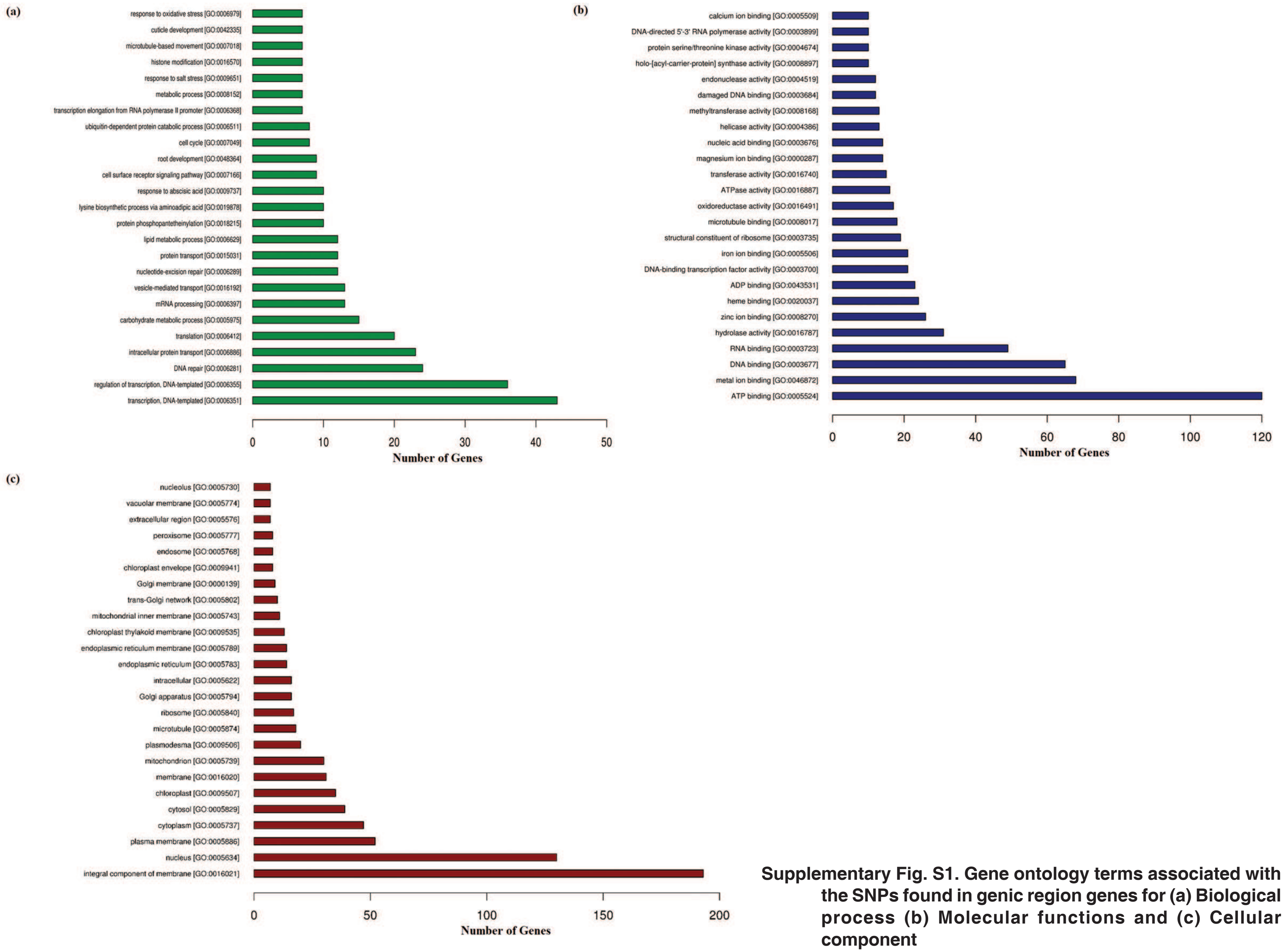
Peterson B. K., Weber J. N., Kay E. H., Fisher H. S. and Hoekstra H. E. 2012. Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. Plos One, **7**: e37135. doi: 10.1371/journal.pone.0037135.

Ezura H., Ariizumi T., Garcia-Mas J. and Rose J. (Eds.). 2016. Functional Genomics and Biotechnology in Solanaceae and Cucurbitaceae Crops (Vol. **70**). Springer.

Shirasawa K. and Hirakawa H. 2013. DNA marker applications to molecular genetics and genomics in tomato. Breed Sci., **63**: 21-30. doi: 10.1270/jsbbs.63.21.

DeVicente M. C. and Tanksley S. D. 1993. QTL analysis of transgressive segregation in an interspecific tomato cross. Genetics, **134**: 585-596.

VanRaden P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci., **91**: 4414-23. doi: 10.3168/jds.2007-0980.



Supplementary Fig. S1. Gene ontology terms associated with the SNPs found in genic region genes for (a) Biological process (b) Molecular functions and (c) Cellular component