Genomics 112 (2020) 99–107

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Transcriptome analysis of Snow Mountain Garlic for unraveling the organosulfur metabolic pathway

Rukmankesh Mehra[a,b], Rahul Singh Jasrotia[c], Ankit Mahajan[b], Deepak Sharma[b], Mir Asif Iquebal[c], Sanjana Kaul[b], Manoj Kumar Dhar[a,b,*]

[a] Bioinformatics Centre, School of Biotechnology, University of Jammu, Jammu 180006, India
[b] Genome Research Laboratory, School of Biotechnology, University of Jammu, Jammu 180006, India
[c] Centre for Agricultural Bioinformatics (CABin), ICAR-Indian Agricultural Statistics Research Institute, New Delhi 110012, India

ARTICLE INFO

Keywords:
Snow Mountain Garlic
Allium porrum
Transcriptome
Sequencing
SNP
Organosulfur compounds

ABSTRACT

Snow Mountain Garlic grows in the high altitudes of the Himalayas under low temperature conditions. It contains various bioactive compounds whose metabolic pathways have not been worked out at genomic level. The present work is the first report on the transcriptome sequencing of this plant. > 43 million paired-end reads (301 × 2) were generated using Illumina Miseq sequencing technology. Assembling of the sequencing data resulted in 326,785 transcripts. Differentially expressed genes between the clove and leaf tissues were identified and characterized. Besides, greater emphasis was laid on the genes, which were highly expressed in clove since the latter is assumed to contain high content of the bioactive compounds. Further analysis led to the identification of the genes plausibly involved in the organosulfur metabolism. We also identified several simple sequence repeats and single nucleotide polymorphism. These constitute valuable genetic resource for research and further genetic improvement of the plant.

1. Introduction

Genus *Allium* L. of family Alliaceae represents an interesting assemblage of about 850 species, which are perennial and characterized by underground rhizomes or bulbs [1–3]. Most of these plants have high economic value and include vegetables such as onion, garlic, leek *etc.*, ornamentals or medicinal plants. Biochemically, *Allium* species have good amounts of carbohydrates, however, these also contain proteins, pectin, minerals and few polyamines. The most distinguishing character of Alliums is the presence of sulfur containing compounds in the form of non-protein amino acids [4]. The latter, odourless, stable and non-volatile compounds known as S-alk(en)yl cysteine sulfoxides serve as precursors of volatile flavour or neutraceutical compounds. Four different types of precursors have been identified in *Allium* species [5]. It has been argued that the differences in the flavour characteristics among species and cultivars may be dependent on the sulfur uptake and its processing during various steps of the flavour biosynthetic pathway [3]. The present study aims to identify some of the genes of the organosulfur pathway in one of the important species of *Allium* growing in the Himalayan region.

Snow Mountain Garlic or Kashmiri garlic is a purified form of garlic. It is known by various popular names such as Kashmiri Lahsun, Ek Pothi Lahsun and One Pod Garlic. The latter two names are due to its characteristic feature of forming single clove per bulb. It is palatable in raw form without harming stomach or intestine. This variety grows in cold climate of the Himalayas at 6000 m above the sea level [6]. It can survive at very low temperature (-10 °C) and oxygen conditions [6,7].

Snow Mountain Garlic has been previously described either as a subspecies of *Allium sativum* [7] or as *Allium schoenoprasum* [6,8]. However, the bulbs of Snow Mountain Garlic are not quite similar to either *A. sativum* or *A. schoenoprasum*. *A. sativum* generally possess multiple cloves serially arranged in one bulb whereas *A. schoenoprasum*, which is known as chive, possesses rhizomes and is therefore different. Therefore, we further examined the taxonomic identity of this crop. Snow Mountain Garlic produces single clove, which is a similar feature to that of *Allium ampeloprasum* [9,10]. However, it is not a strict criteria and is debated in the past – whether the single bulb producing garlic is *A. sativum* or *A. ampeloprasum* [11]. Figliuolo and Di Stefano [11] described the single bulb Chinese garlic as *A. sativum*. Nevertheless, *A. ampeloprasum* is very diverse and is widely spread either as wild or cultivated form [10,12].

Gohil and Koul [13] have described Snow Mountain Garlic from Kashmir as *Allium porrum* based on the detailed cytological characterization. *A. porrum* and *A. ampeloprasum* are many times treated as

[*] Corresponding author at: Bioinformatics Centre, School of Biotechnology, University of Jammu, Jammu 180006, India.
E-mail address: manojkdhar@rediffmail.com (M.K. Dhar).

https://doi.org/10.1016/j.ygeno.2019.07.014
Received 11 February 2019; Received in revised form 5 July 2019; Accepted 25 July 2019
Available online 26 July 2019

synonyms having very similar karyotype; however, there are certain distinctive features, which make them different [13]. They described this plant as totally sterile, and we also observed no flowering in it. Further, the morphology of Snow Mountain Garlic shows that the bulbs are scaly with membranous scales, ovoid bulbs, up to 2 cm long and 0.5 cm broad [6,8]. The leaves are fistular, cylindrical and glabrous with membranous sheaths, 10–15 cm in length and 1–2 mm broad. Surprisingly, other features of Snow Mountain Garlic also match with *A. ampeloprasum* and *A. porrum*. Similar to these plants, Snow Mountain Garlic is milder than *A. sativum* and is palatable in raw form [9]. The leaves of Snow Mountain Garlic are flat, very similar to those of leeks. These observations strongly suggest that Snow Mountain Garlic belongs to *A. porrum*. Since the material of Snow Mountain Garlic was obtained from the same region from where Gohil and Koul [13] had collected, therefore, in this paper we have treated Snow Mountain Garlic as *A. porrum*.

Snow Mountain Garlic has promising medicinal properties and has shown some degree of remedial effects against hypertension, anthero-sclerosis, diabetes and cancer [14–16]. It also possesses some degree of immunomodulatory activity [16–18]. *A. ampeloprasum and A. porrum* contain several sulfur containing bioactive components such as *S*-methyl cysteine sulfoxide, S-propyl cysteine sulfoxide, S-propenyl cysteine sulfoxide, N-(γ-glutamyl)-S-(E-1-propenyl) cysteine, dimethyl disulfide, propyl propenyl disulfide, methyl propenyl disulfide, dimethyl trisulfide, methyl propenyl trisulfide and methyl propyl trisulfide [9,19]. Not surprisingly, we found several genes metabolizing these compounds abundant in clove of Snow Mountain Garlic (described in results). It has been reported that the potential medicinal properties of garlic growing at higher altitude increases by seven times due to an elevated level of organosulfur content [6].

We observed that although, information on transcriptomes of other garlic types was available, there were no reports on transcriptomics of Snow Mountain Garlic. Hence, the present work on the transcriptome analysis was undertaken to fill this lacuna. Since clove is an important part of garlic and the organosulfur compounds are perhaps responsible for its promising bioactivity, the main aim was to study the expression pattern of the organosulfur metabolizing enzymes in clove *versus* other tissues. Therefore, we sequenced the clove and leaf tissues in two biological replicates, assembled them *de novo* and characterized differentially expressed genes. This comparative transcriptomics helped us to identify distinct gene families active in leaf and clove.

2. Materials and methods

2.1. Raising plant material and total RNA extraction

The cloves of Snow Mountain Garlic were planted during October 2016 in earthen pots maintained in the Greenhouse. Green leaves were harvested first and later when the cloves developed, they were also harvested. The harvested leaves and cloves were immediately frozen in the liquid nitrogen and later stored at -80 °C until use. The leaf and clove tissues were further studied in two biological replicates. Trizol method [20] and HiPurA RNA isolation kit were used to extract the total RNA from the plant tissues. The purified RNA samples were diluted with DEPC treated water to make a final volume of 25 to 40 µl. The quality and concentration of the total RNA were analyzed using Nanodrop spectrophotometer [21,22] and Bioanalyzer [23,24].

2.2. Library preparation and sequencing

Poly-A containing mRNAs were isolated from the four total RNA samples. For each sample, double-stranded DNA was synthesized using TruSeq stranded mRNA LT-Set A kit of Illumina. The cDNA libraries were prepared according to the manufacturer's instructions (Illumina, San Diego, CA). The concentration of cDNA libraries was analyzed using high sensitivity (HS) protocol of Qubit (High Sensitivity, Invitrogen). The quality and quantity of the four cDNA libraries were analyzed using Agilent 2100 Bioanalyzer (Agilent [23]). The sequencing data were generated using Illumina MiSeq paired-end sequencing technology (Illumina, San Diego, CA). The experiment was performed as guided by the manufacturer.

2.3. Pre-processing and de novo transcriptome assembly

The quality of the sequencing data was checked using FastQC tool (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) [25]. The preprocessing of the data was performed using Trimmomatic tool [26]. The low-quality reads with phred-score ≤ 20 and reads containing ambiguous bases N were removed. The adaptor sequences and the low-quality bases from 5 and 3 were trimmed to improve the quality of data. *De novo* assembly of the filtered and clean transcriptome data was performed using Trinity software, which is a short-read assembly program [27].

2.4. Identification of differentially expressed genes

The high-quality reads of each sample were mapped onto the reference *de novo* assembly using Bowtie program [28] to find the read density. The gap free alignment of the reads was obtained using a criterion to map on the assembly with a maximum of two mismatches. The calculation of the abundance and expression values was carried out using RSEM (RNA-Seq by Expectation–Maximization) tool [29]. These values were calculated for each transcript using fragments per kilobase of exon per million mapped reads (FPKM). The differential expression of the genes was identified using the Empirical analysis of Digital Gene Expression in R (EdgeR) package as implemented in Bioconductor [30]. In order to select the significant differentially expressed genes (DEGs), the threshold for the False Discovery Rate (FDR) values was kept relatively low to 0.01 and log fold change was set relatively high to 5.

2.5. Functional annotation and categorization of the differentially expressed genes

The homology-based search of the identified DEGs was carried out against the non-redundant database of NCBI using BLAST program [31]. The e-value threshold was set to 1e-05. Further, the functional categorization of the genes was performed by Gene Ontology terms [32], which classify the genes under three categories, namely, biological process, molecular function and cellular component using Blast2Go software [33]. The annotation of the DEGs for the enzyme classification number (EC) and metabolic pathways was carried out by searching against the KEGG database [34] using Blast2Go. The domains present in these genes were characterized by searching against the InterProScan [35] using Blast2Go.

2.6. Detection of SSR markers and SNPs

The mining of simple sequence repeat (SSRs) markers was performed using the MISA (MIcroSAtellite identification tool) Perl script [36]. For the detection of single nucleotide polymorphisms (SNPs), reads from each sample were mapped onto the *de novo* transcriptome assembly using BWA (Burrows Wheeler Aligner) tool [37]. SAMtools package [38] was then used for the identification of SNPs and Indels (insertion/deletion) from the aligned reads. The parameters used for SNP mining included the quality score of 20, depth of 5× and a maximum of two SNPs were permissible within 50 base-pair on either side.

3. Results and discussion

3.1. Preprocessing and de novo assembly

In total, 43,516,300 paired-end reads (301 bp × 2) were generated

Table 1
Statistics of the transcriptome data.

Samples	Input read pairs	Both surviving	Forward only surviving	Reverse only surviving	Dropped
A_leaf-Rep1	5,052,640	3,037,530 (60.12%)	1,919,264 (37.99%)	14,607 (0.29%)	81,239
A_leaf-Rep2	4,391,422	2,753,885 (62.71%)	1,559,514 (35.51%)	14,254 (0.32%)	63,769
B_clove-Rep1	6,422,850	3,895,287 (60.65%)	2,404,055 (37.43%)	21,649 (0.34%)	101,859
B_clove-Rep2	5,891,238	3,313,831 (56.25%)	2,479,193 (42.08%)	14,958 (0.25%)	83,256

*“A_leaf-Rep1” is the paired-end sequencing data of biological replicate 1 of the leaf sample, “A_leaf-Rep2” is the paired-end sequencing data of biological replicate 2 of the leaf sample, “B_clove-Rep1” is the paired-end sequencing data of biological replicate 1 of the clove sample, “B_clove-Rep2” is the paired-end sequencing data of biological replicate 2 of the clove sample.

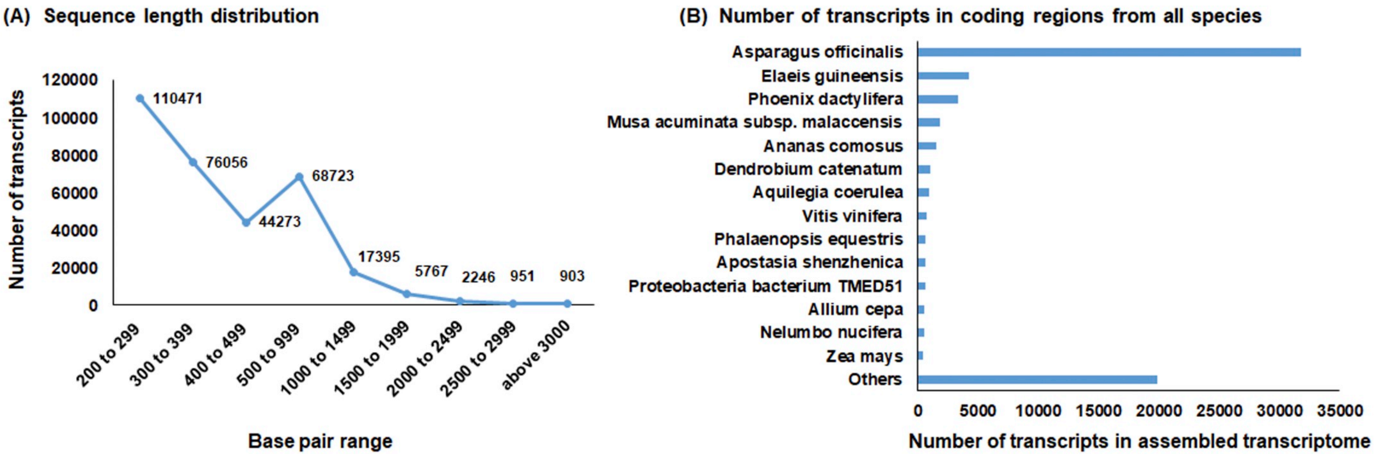

Fig. 1. (A) Sequence length distribution of the transcripts in the assembled transcriptome. (B) Species hit distribution in the likely coding regions of the assembled transcriptome.

from the two tissue samples of Snow Mountain Garlic *viz.*, clove and leaf (two biological replicates each). After pre-processing, 330,123 low-quality reads were removed and finally high-quality reads (43,186,177, ~43 million) of both the samples were pooled for the *de novo* transcriptome assembly (Table 1). Trinity generated 326,785 transcripts with N50 of 549 bp and GC content of 39.84%. In the transcriptome assembly, the shortest read of 200 bp was obtained and the average length was 493.67 bp. In addition, 110,471 transcripts had the read length ranging between 200 and 299, which constituted around 33.8% of the whole assembly, while 6 transcripts had the length > 10,000 bp (Fig. 1A). The sequencing data has been submitted to NCBI repository under the Bioproject – PRJNA489986 (SRA accession: SRP160459).

The 326,785 transcripts (from 43 million 301 bp paired end reads) generated here is a robust data to gain an effective insight into the transcriptome. Previously, several transcriptomic studies have been performed to understand other *Allium* types. In a study, 72.53 million 100 bp paired end reads were generated to analyze the sucrose metabolism in *Allium cepa* [39]. Liu et al. generated about 128.9 million 125 bp paired end reads of *Allium porrum* [40]. Kamenetsky and coworkers assembled > 32 million 250 bp paired end reads of fertile garlic [41]. Liu et al. used 69.7 million 125 bp paired end reads for garlic analysis [42]. In addition, Shemesh-Mayer et al. generated six libraries of 17–21 million 100 bp one-end clean reads [43]. The number of reads generated was comparatively high, however, the reads were one-end and read length was also small. In our study, we focused on generating longer paired end reads (300 bp) to possibly get a uniform coverage, detect splice isoforms and to possibly avoid repeat sequences and include missing insertions. Therefore, the data assembled by us is a good representative for the detailed analysis of Snow Mountain Garlic.

3.2. Analysis of the assembled transcriptome

In the assembled transcriptome, likely coding regions were identified using TransDecoder v5.5.0 program (https://github.com/TransDecoder/TransDecoder/wiki) and then single best open reading frame (ORF) was selected per transcript for analysis, as done in the previous studies [44–48]. Out of the 326,785 assembled transcripts, 78,880 putative full-length coding genes were identified (Supplementary Coding file.txt). Homology search of these genes against NCBI NR protein database using Blastx showed similarity of 68,174 coding genes in the database (Supplementary Table S1). However, no hit was found for 10,706 genes. The species distribution of all these genes showed the highest number of hits with *Asparagus officinalis* (31,774 genes) followed by *Elaeis guineensis* (4246 gene), *Phoenix dactylifera* (3321 genes), *Musa acuminata* subsp. *malaccensis* (1888 genes) and *Ananas comosus* (1575 genes) as shown in Fig. 1B and Supplementary Table S2. Notably, 998 hits (genes) were found against 17 *Allium* species and subspecies, out of which *Allium cepa* and *Allium sativum* showed the highest hits, *i.e.* 565 and 265 genes respectively. Interestingly, each *Allium ampeloprasum* and *Allium fistulosum* showed 38 hits. However, some hits were also found against some bacteria, which may be the hits by chance.

Comparatively very high number of hits (maximum) were found with *Asparagus officinalis* (Fig. 1B). It is a perennial flowering plant that is also related to the *Allium* species (onion and garlic) and was once classified in the Liliaceae family [49]. However, later this family was divided into Asparagaceae (*Asparagus* plant) and Amaryllidaceae (onion-like). Due the high similarity between the *Allium* and *Asparagus* genera, and due to the availability of the genomic data of *Asparagus*, it was highly expected to get maximum matches with *Asparagus*. On the other hand, despite a good similarity with the transcripts, *Elaeis guineensis* and *Phoenix dactylifera* did not show any close relation with *Allium.*

3.3. DEGs, homology search and genes involved in organosulfur metabolism

In order to identify the genes (of clove) involved in the organosulfur metabolism, reads of all the samples were mapped onto *de novo* transcriptome assembly for calculating expression values in the form of

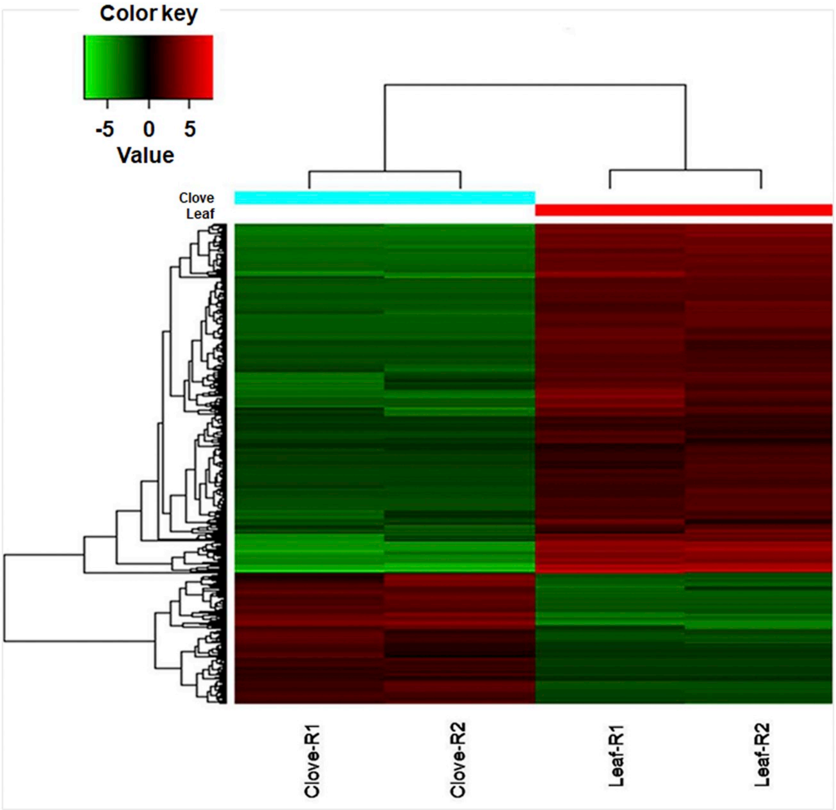

Fig. 2. Heat map of the differential expressed genes obtained in the clove *versus* leaf samples. The red color represents the upregulated and the green color shows the downregulated genes in the clove and leaf samples. Clove-R1 and Clove-R2 are the two biological replicates of clove, and leaf-R1 and leaf-R2 are the two replicates of leaf. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

FPKM, which was further used for the identification of the differential expressed genes (DEGs). A total of 1885 DEGs were obtained from the clove *versus* leaf samples (DEG sequences supplied as a separate file – Supplementary DEG file.txt). This included 510 upregulated and 1375 downregulated genes in the clove sample in comparison to the leaf and *vice versa* as shown in the heat map Fig. 2.

Further, the homology search of the DEGs was performed against the NCBI non-redundant database using Blastx program (Fig. 3 and Supplementary Tables S3 and S4). In total 1645 DEGs showed similarity with other known proteins in the database, while 43 transcripts did not show any similarity. The maximum similarity hits were found with *Asparagus officinalis i.e.*, 842 transcripts, followed by 129 and 72 transcripts with *Elaeis guineensis* and *Phoenix dactylifera,* respectively. Notably, the top five species against which the highest number of transcripts were found in the coding genes of the whole transcripts (Fig. 1B) and in the DEGs (Fig. 3A) are the same. Interestingly, 40, 14 and 2 transcripts showed similarity hits with *Allium cepa*, *Allium sativum* and *Allium ampeloprasum* respectively, which belong to the same genus as Snow Mountain Garlic (Fig. 3A and Supplementary Tables S3 and S4). These small numbers of hits may be due to the non-availability of their genomic data.

As Snow Mountain Garlic is a bulbous plant, we also analyzed the DEG hits found with the other bulbous species (Fig. 3B and Supplementary Table S4). For this, we collected a list of bulbous genera from the online resource – “Pacific Bulb Society” (www.pacificbulbsociety.org), and used this for the identification of bulbous hits in our DEGs. A total of 136 transcripts (BLAST hits) were found to match with 24 bulbous species that belong to 15 genera. Noticeably, we found a good number of hits against some bulbous species, such as *Allium cepa* (40 hits), *Allium sativum* (14 hits), *Manihot esculenta* (14 hits) and *Nelumbo nucifera* (11 hits). A small number of transcripts (< 10 hits) also matched against 20 other bulbous species, such as *Narcissus tazetta* var. *chinensis* (9 hits), *Hyacinthus orientalis* (6 hits) and *Lycoris radiate* (6 hits). Interestingly, two hits were also found with *Allium ampeloprasum*. As already discussed, this small number was because of the lack of its genomic data. Some of the organosulfur metabolizing genes that we discussed later were proposed based on the similarity with these bulbous species, such as allinase (based on similarity with *A. cepa* and *A. fistulosum*), peroxidase (based on *A. cepa*) and glutathione gamma-glutamylcysteinyltransferase (based on *A. sativum*).

From the *in silico* annotation of the DEGs, several genes were found to be probably involved in the organosulfur metabolism in Snow Mountain Garlic. We mainly analyzed the genes that showed > 6-fold higher expression in clove than in leaf. The highly expressed gene in clove was alliinase having about 12-fold higher expression than in leaf tissue. Alliinase is implicated in the conversion of alliin to allicin, which is an organosulfur metabolite [50–53]. Other highly expressed genes in clove that are the candidates for the organosulfur metabolism include plant cysteine oxidase 2-like, cysteine protease XCP1, serine carboxypeptidase-like 45, flavonol sulfotransferase-like, caffeoyl-CoA *O*-methyltransferase-like, glutathione S-transferase T1 and glutathione gamma-glutamylcysteinyltransferase 1-like isoform X1.

Alliin is a derivative of cysteine amino acid [19,52,54–56]. Therefore, it can be interpreted that the enzymes of the cysteine metabolism may conceivably be involved in the alliin biosynthesis. These enzymes include plant cysteine oxidase and cysteine protease that cause the oxidation and proteolysis processes, respectively.

Two popular proposed pathways of the alliin biosynthesis are the serine route and the glutathione route [52,55]. The amino acid serine contributes to the alliin biosynthesis [55] by reacting with allyl group (unknown source) to form S-allyl cysteine, which further catalyzes to form alliin. The enzyme serine carboxypeptidase is a protease that causes the hydrolysis of a peptide bond at the carboxyl-terminal of serine. Further, glutathione is also implicated in the alliin biosynthesis through glutathione route [52,55,57]. The enzymes glutathione S-transferase and glutathione gamma-glutamylcysteinyltransferase show possible involvement in the glutathione metabolism. Glutathione gamma-glutamylcysteinyltransferase is an aminoacyltransferase enzyme that belongs to the transferases family. It is involved in the catalysis of gluthathione and [Glu(-Cys)]n-Gly to form Gly and [Glu(-Cys)]

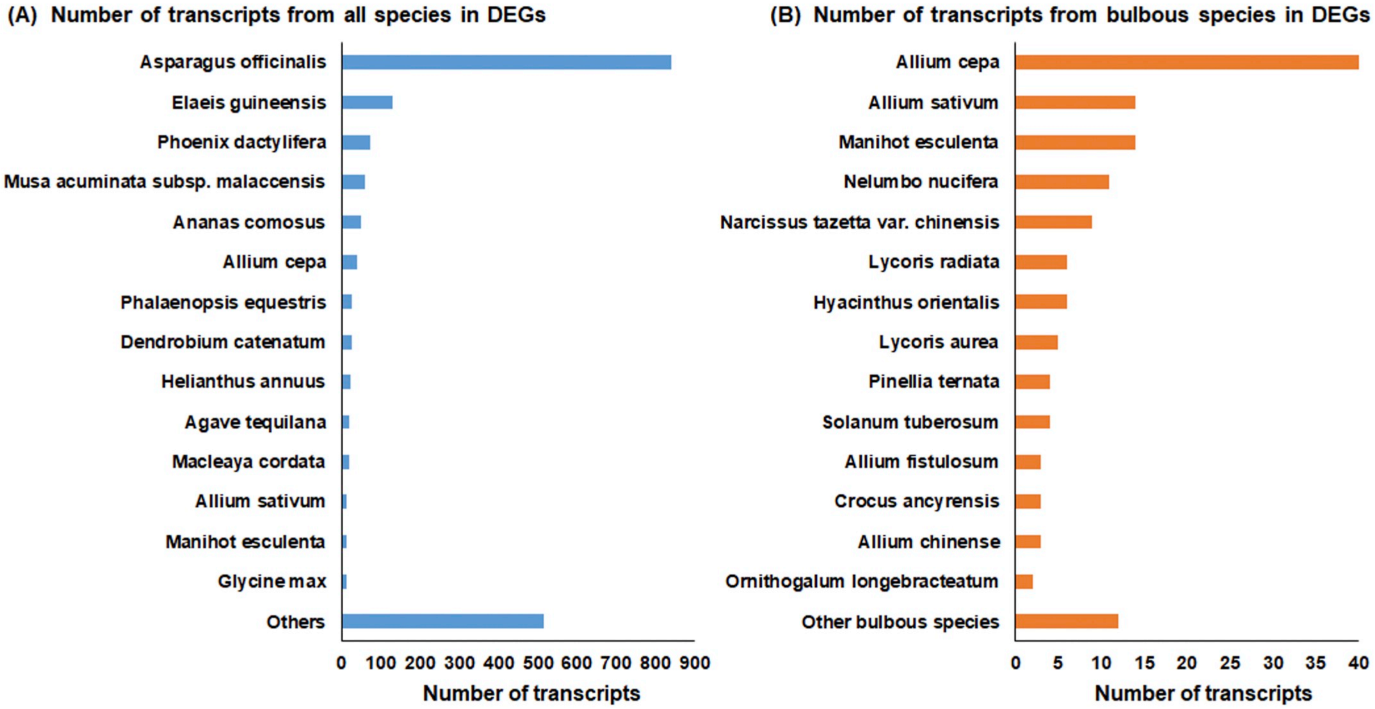

Fig. 3. Species distribution in the DEG hits found in Blastx homology search. (A) Number of transcripts found from all the species. (B) Number of transcripts specific to the bulbous plants.

n + 1-Gly. Glutathione S-transferase causes the conjugation of reduced gluthathione to the xenobiotic substrates.

The enzyme caffeoyl-CoA *O*-methyltransferase showed > 6 fold increased expression in clove than in leaf. It causes the conversion of S-adenosyl-L-methionine and caffeoyl-CoA to form S-adenosyl-L-homocysteine and feruloyl-CoA. The product of this reaction S-adenosyl-L-homocysteine may play a role in the alliin biosynthesis, as it has been reported that many plant metabolites are produced by involving this step [58]. The enzyme flavonol sulfotransferase showed > 9 fold higher expression in clove than in leaf. It involves the transfer of sulfur from 3′-phosphoadenylyl sulfate to flavonol to form adenosine 3′,5′-bisphosphate and flavonol 3-sulfate, and therefore, it may also play a role in the organosulfur metabolism in Snow Mountain Garlic.

The above analysis shows that several candidate enzymes of the organosulfur metabolism are present in the DEGs. These genes are highly expressed in clove, which is expected, as the organosulfur compounds are supposed to be abundant in clove than in leaf.

3.4. DEGs highly expressed in leaf

In addition to the genes highly expressed in clove, we also analyzed the highly expressed genes in leaf (Supplementary Table S3). The top highly expressed gene showed > 14 fold high expression than in clove and was involved in chlorophyll binding function. As expected, most of the highly expressed genes were involved in chloroplastic, photosystem, carbonic anhydrase and other photosynthetic processes. However, we also observed few genes with possible functions in the organosulfur metabolism that include serine-glyoxylate aminotransferase, chloroplastic thioredoxin and cysteine protease. It may be possible that the precursors of some sulfur compounds are synthesized in leaf and then transported to other parts including clove. Since we were not interested in the analysis of the other genes, we restricted our further analysis to the organosulfur metabolic genes.

3.5. Gene ontology and protein domains

Using Blast2Go tool, mapping of the transcripts, their functional annotation, domain/family search and the KEGG pathway search were performed. A total of 176 transcripts could be assigned the gene ontology (GO) terms. The transcripts were categorized into three sub-categories: biological process (17 GO terms), molecular function (8 GO terms) and cellular component (9 GO terms). In biological process GO terms, the maximum number of transcripts were found to involve in the cellular processes (132 transcripts), while binding (82 transcripts) and cell part (137 transcripts) functions were reported for the maximum number of transcripts in the molecular function and cellular component, respectively (Fig. 4).

This shows that in biological process, these transcripts are majorly involved in the cellular processes, followed by single-organism and metabolic processes. > 80 genes are involved in metabolism, which shows that these DEGs also contain pathway genes, some of which may possibly be implicated in the organosulfur metabolism. In molecular function, the presence of > 80 genes in the binding processes suggests the possibility of their participation in catalysis. However, these DEGs also contains proteins that either form a part of a cell or its extracellular region (in cellular component).

In order to identify the domains present in the DEGs, the search of the transcripts was performed against InterProScan. 926 transcripts were found to represent 468 domain, while 1437 transcripts were present in 681 protein families (Fig. 5 and Supplementary Table S5). The highest number of transcripts (64) comprised alpha crystallin/Hsp20 domain (IPR002068) followed by 58 transcripts that contained histone domain (IPR007125). In each of the remaining domains, < 25 transcript were present. Interestingly, 16 transcripts comprised alliinase C-terminal domain (IPR006948) and 12 transcripts comprised alliinase EGF-like domain (IPR006947). As discussed earlier, alliinase converts alliin to allicin, both of which are organosulfur compounds. Other domains with probable function in the organosulfur metabolism included glutathione S-transferase N-terminal domain (IPR004045), which was present in 11 transcripts and glutathione S-transferase C-terminal-like domains (IPR010987: in 8 transcripts, IPR004046: in 5 transcripts and IPR034347: in 2 transcripts), which were present in 15 transcripts. These domains are probably involved in the glutathione metabolism (discussed in detail above). The thioredoxin domain (IPR013766),

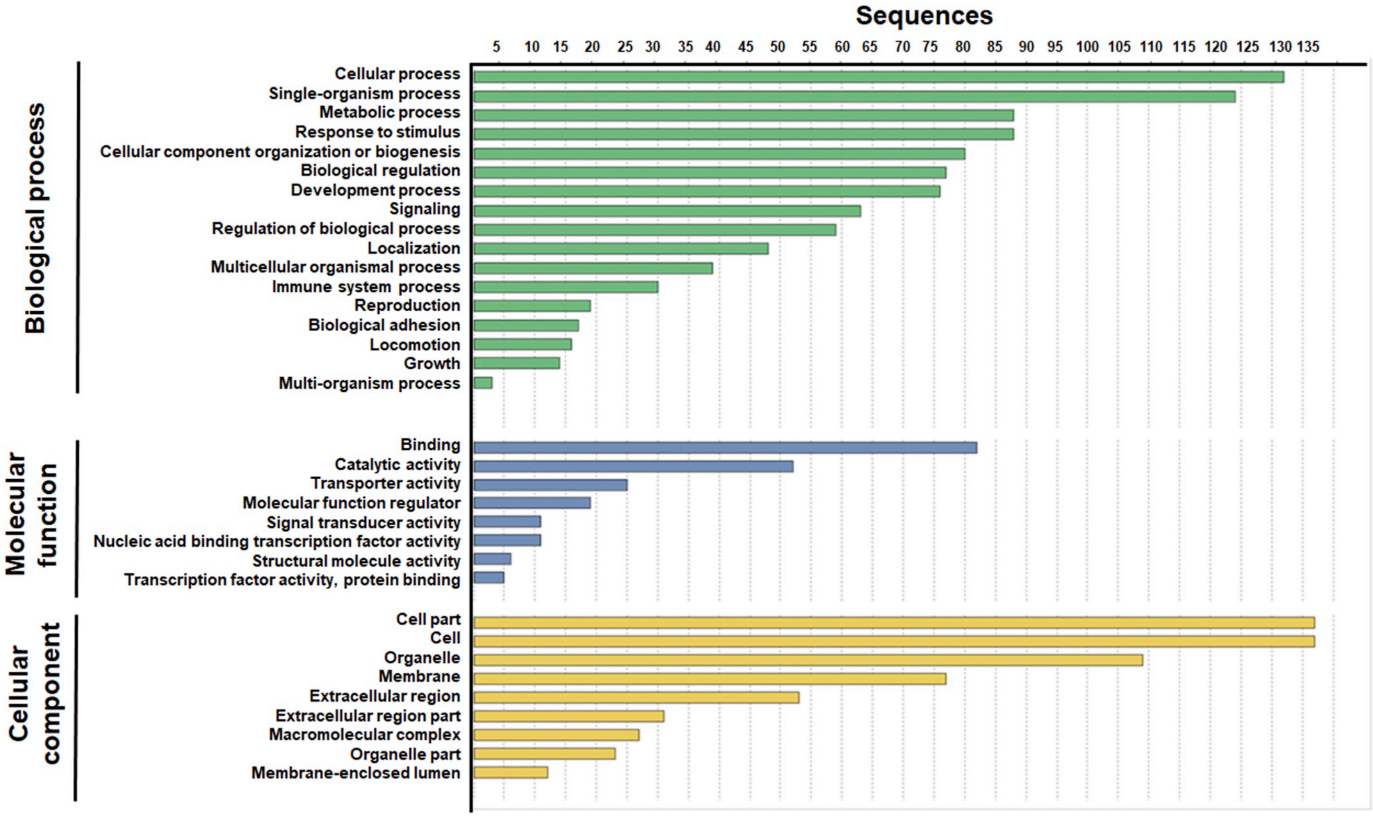

Fig. 4. Gene ontology of differential expressed genes, green colored bars represent biological process, blue bars represent molecular function and yellow represent cellular component. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

which causes the reduction of other proteins by the cysteine thiol-disulfide exchange [59], was present in 11 transcripts, and another domain known as methyltransferase (IPR013216) occurred in 4 transcripts. Twelve different InterProScan domains with peptidase functionality were found namely IPR000668 (8 transcripts), IPR033121 (6 transcripts), IPR000642 (4 transcripts), IPR022764 (4 transcripts), IPR010259 (4 transcripts), IPR001915 (1 transcript), IPR032632 (1 transcript), IPR000209 (1 transcript), IPR007863 (1 transcript), IPR032416 (1 transcript), IPR000994 (1 transcript) and IPR011546 (1 transcript). In addition, 10 transcripts also comprised Myc-type and basic helix-loop-helix (bHLH) domain (IPR011598), which performs regulatory function.

Similarly, the transcripts belonging to the protein families with possible functionality of the organosulfur metabolism were also found. These families included S-adenosyl-L-methionine-dependent methyltransferase (IPR029063: 13 transcripts), glutathione S-transferase C-

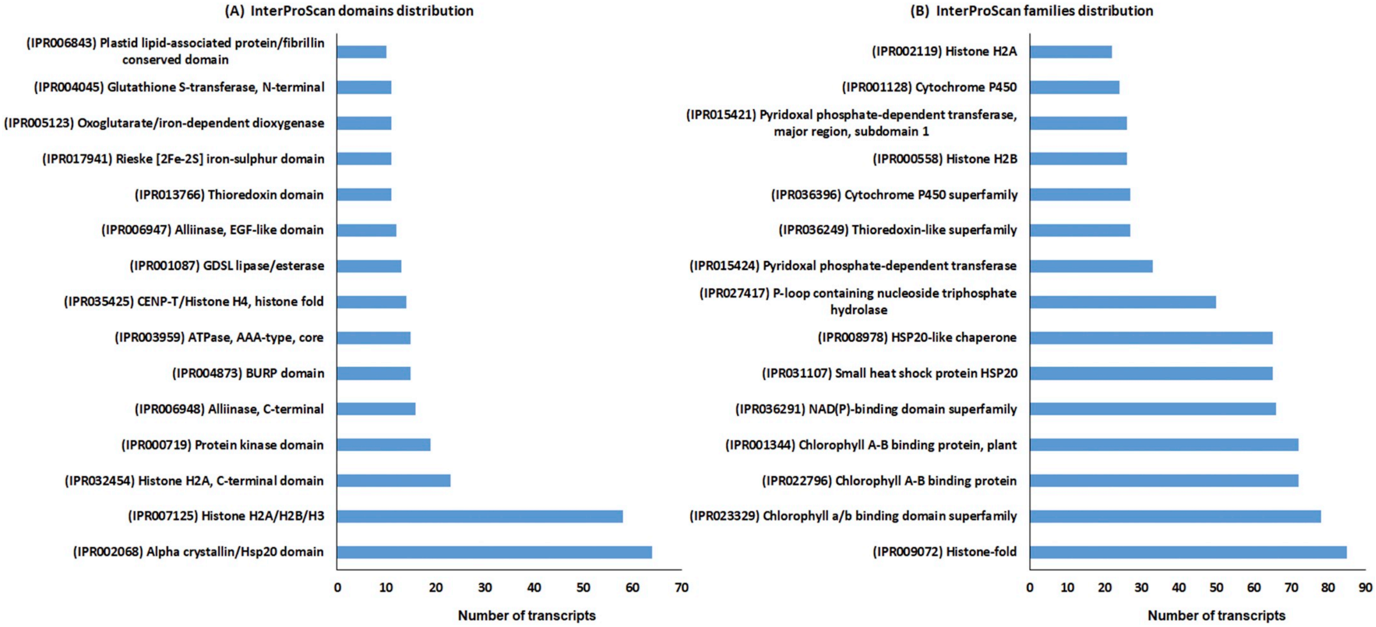

Fig. 5. InterProScan domains (A) and families (B) distribution. Only the top 15 hits matching with the maximum number of transcripts are shown.

terminal domain superfamily (IPR036282: 12 transcripts), alliinase N-terminal domain superfamily (IPR037029: 12 transcripts), serine carboxypeptidase (IPR001563:7 transcripts), thioredoxin (IPR005746: 9 transcripts) and thiolase-like (IPR016039: 5 transcripts).

The identification of the protein domains and families with plausible functionality in the organosulfur metabolism increases the probability of the involvement of the DEGs in its metabolism.

3.6. Pathway characterization

The pathway analysis using the KEGG database showed the involvement of 82 DEGs in 32 pathways (Fig. 6 and Supplementary Table S6). The maximum number of transcripts (*i.e.* 15) showed participation in the porphyrin and chlorophyll metabolic pathways, with very broad functional category of synthase (EC 4.2.1.24) and ceruloplasmin (EC 1.16.3.1). However, we also found the genes appear to be specifically involved in the organosulfur metabolism. Four transcripts, belonging to the three enzyme categories as dehydrogenase ($NADP^+$) (EC 1.1.1.49), thioredoxin peroxidase (EC 1.11.1.15) and transferase (EC 2.5.1.18), were found to involve in the glutathione metabolic pathway. The maximum number of enzymes (3 enzymes) were present in this pathway, in addition to the antibiotics biosynthetic pathway (3 enzymes). One enzyme (cytosine-5-)-methyltransferase (EC 2.1.1.37) was implicated in the cysteine and methionine metabolism. In addition, four transcripts with phosphatase function (EC 3.6.1.15) were present in the thiamine metabolism.

The above analysis suggests the possible involvement of some of the transcripts in the organosulfur metabolic pathways, the functions of which are similar to our BLAST analysis.

3.7. Molecular markers identification

In order to build a genomic resource for further genetic improvement of Snow Mountain Garlic, simple sequence repeat markers were identified in the *de novo* assembly as well as in DEGs (Supplementary Table S7). A total of 17,374 putative SSR were obtained from the *de novo* transcriptome assembly. 774 SSRs were involved in compound formation and 1127 transcripts contained > 1 SSR. Further, 211 SSRs were obtained from 1885 differentially expressed genes (Table 2). A total of 130 unique motifs were found and out of these, mononucleotide repeats A and T were highly abundant *i.e.* 5775 and 5233, respectively. Whereas in case of di repeats, TA and AT repeats were found the maximum number of times, *i.e.* 426 and 424, respectively, while AAG and TTC tri repeats were found 239 and 191 times, respectively.

These reported genic microsatellite markers can be used as the functional domain markers as well as in the ecological, quantitative trait loci (QTL) mapping, evolutionary, comparative genomics and genetic diversity studies. Due to the several advantages of the genic SSR repeats, such as transferability, functional diversity, gene function and linkage mapping, these have been preferred in various genomics programs [60]. The identification of the genic SSR markers from the transcriptomic data has been achieved in many other crops such as sugarcane [61], *Vigna mungo* [62], *Allium sativum* [63] and holy basil (tulsi) [64].

3.8. Variant calling

Single nucleotide polymorphism (SNP) and insertion/deletion (Indel) were identified from the two biological replicates of the clove and the leaf samples of Snow Mountain Garlic. A total 107,222 variants were found, which include 100,261 SNPs and 6961 Indels. In total, 68,246 and 32,553 variants were transition (Ts) and transversion (Tv), respectively with Ts/Tv ratio of 2:1. The maximum count of the variants were found in the transcripts *viz* TRINITY_DN51928_c0_g1_i1 (protein name: auxin transport protein BIG) *i.e.* 41, followed by 32 and 25 in TRINITY_DN51888_c0_g1_i1 (Protein name: uncharacterized protein) and TRINITY_DN51859_c2_g2_i2 (Protein name: polyprotein), respectively (Supplementary Table S8). Out of 107,222 variants, 101,193 and 94,374 variants were identified in the clove and leaf samples, respectively. In the leaf samples, 88,492 SNPs and 5882 Indels were observed, while 94,684 SNPs and 6509 Indels were identified in clove (Supplementary Table S8). These predicted genic variants can be used in various association studies as well as in the crop improvement programs. Such genic region variants have been reported in other crops for trait improvement programs such as wheat [65] and sorghum [66] with cold tolerance and wheat dormancy [67].

4. Conclusions

To our best knowledge, this study provides the first report on the transcriptomics and organosulfur metabolic genes of Snow Mountain Garlic. We generated > 43 million paired-end reads (301 × 2) using the Illumina MiSeq platform, which were assembled into 326,785 transcripts. We characterized the likely coding genes present in this assembly, and further, identified differentially expressed genes in clove *versus* leaf samples. These differentially expressed genes were functionally characterized using Blast, Gene Ontology terms, KEGG pathways, Pfam domains and families. As Snow Mountain Garlic is a bulbous species, we also analyzed these genes with respect to other bulbous species and found good number of hits with *Allium cepa* and

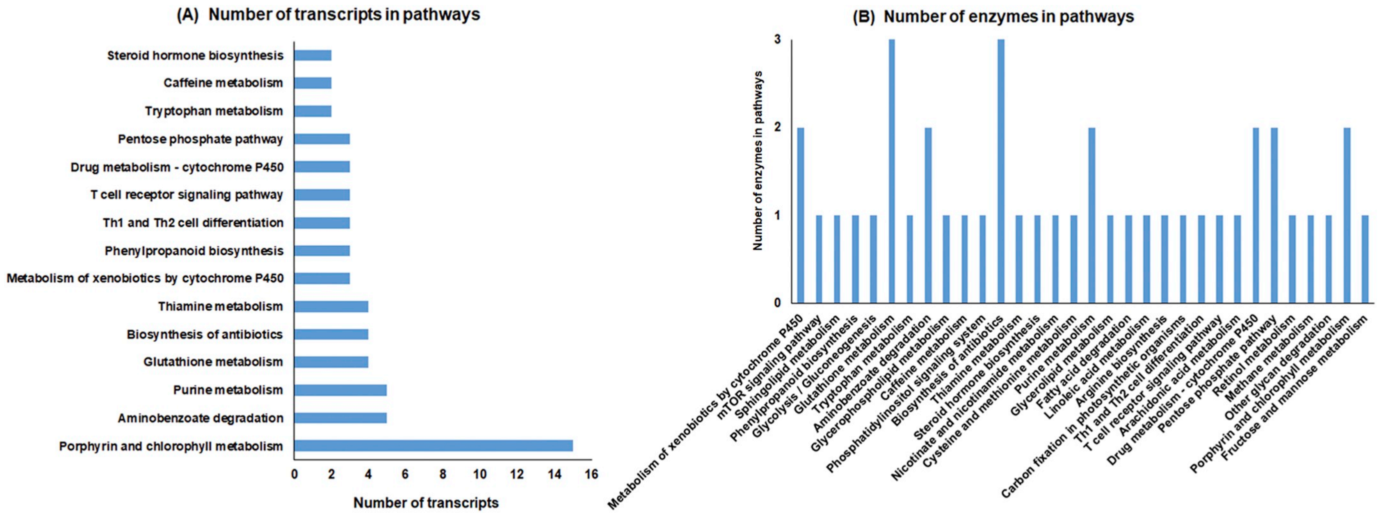

Fig. 6. (A) Number of transcripts belonging to the top 15 pathways. (B) Number of enzymes belonging to all the identified pathways.

Table 2
Molecular markers (SSRs) identified from the *de novo* assembly and DEGs.

	De novo transcriptome assembly	Differential expressed genes
Total number of sequences examined	326,785	1885
Total number of identified SSRs	17,374	211
Number of SSR containing sequences	16,063	199
Number of sequences containing > 1 SSR	1127	12
Number of SSRs present in compound formation	774	5
Mono	11,717	135
Di	2585	19
Tri	2936	55
Tetra	100	1
Penta	20	1
Hexa	16	–

Allium sativum, which belong to the same genera as our plant. Using these DEGs, we proposed the genes of the organosulfur metabolism with specific reference to clove. These candidate genes include cysteine oxidase, cysteine protease, serine carboxypeptidase, flavonol sulfotransferase, caffeoyl-CoA *O*-methyltransferase, glutathione S-transferase and glutathione gamma-glutamylcysteinyltransferase. In addition to clove, we also analyzed the highly expressed genes in leaf, which mainly showed involvement in photosynthetic processes. We expect that this resource would contribute substantially in understanding the organosulfur metabolism in this plant. Using the sequencing data, we further identified single nucleotide polymorphisms and simple sequence repeat markers, which present probably the first substantial resource of this plant for genetic study and further crop improvement. In brief, this study provides a significant genetic resource of Snow Mountain Garlic, which has opened up new avenues for further molecular interventions.

The sequencing data used in this study are available in the NCBI repository under Bioproject: PRJNA489986 (SRA accession: SRP160459) and will be made open access after publication. A FASTA formatted file of 78,880 coding genes was supplied as “Supplementary Coding file.txt”.Another FASTA formatted file of the DEGs is provided as “Supplementary DEG file.txt”. Eight excel sheets are provided as supplementary tables. Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2019.07.014.

Funding source

The work is supported by the National Postdoctoral Fellowship (NPDF) award number: PDF/2016/000054.

Acknowledgement

The authors acknowledge the financial support provided under the “National Postdoctoral Fellowship” scheme of the Science & Engineering Research Board (SERB), Department of Science and technology, Government of India and support by the Department of Biotechnology, Government of India to the Bioinformatics Centre of the University. Financial support to MKD under SAP of University Grants Commission, New Delhi is also acknowledged. The authors acknowledge the valuable suggestions of the Reviewers especially with regard to the taxonomic classification of the plant.

Author contributions

RM and MKD planned and designed the work. RM performed the *in vitro* experiments and bioinformatics annotations. RJ and MAI performed the Blast annotation. RM and AM contributed in the cDNA sequencing. MKD analyzed and supervised the work. SK provided reagents for the study and helped in sequencing. RM drafted the manuscript. MKD and SK performed manuscript proofreading. RM, MKD, RJ and DS contributed in the manuscript revision.

References

[1] M. Abdelrahman, S. Hirata, S. Ito, et al., Compartmentation and localization of bioactive metabolites in different organs of *Allium roylei*, Biosci. Biotechnol. Biochem. 78 (2014) 1112–1122.

[2] R.M. Fritsch, F.R. Blattner, M. Gurushidze, New classification of Allium L. subg. Melanocrommyum (Webb & Berthel.) Rouy (Alliaceae) based on molecular and morphological characters, Phyton (Horn.) 49 (2010) 145–220.

[3] R. Kamenetsky, H.D. Rabinowitch, The genus *Allium*: a developmental and horticultural analysis, Horticult. Rev. 32 (2006) 329–378.

[4] W.M. Randle, J.E. Lancaster, 14 sulphur compounds in alliums in relation to flavour quality, Allium Crop Sci. (2002) 329.

[5] G.G. Freeman, R.J. Whenham, A survey of volatile components of some Allium species in terms of S-alk (en) yl-L-cysteine sulphoxides present as flavour precursors, J. Sci. Food Agric. 26 (1975) 1869–1886.

[6] M. Koul, S. Meena, A. Kumar, et al., Secondary metabolites from endophytic fungus *Penicillium pinophilum* induce ROS-mediated apoptosis through mitochondrial pathway in pancreatic cancer cells, Planta Med. 82 (2016) 344–355.

[7] R. Mahajan, *In vitro* and cryopreservation techniques for conservation of Snow Mountain garlic, in: S. Jain (Ed.), Protocols for *in vitro*Cultures and Secondary Metabolite Analysis of Aromatic and Medicinal Plants, Second Edition. Pp 335–346. Methods in Molecular Biology, vol. 1391, Humana Press, New York, NY, 2016.

[8] N.C. Shaw, Status of cultivated & wild *Allium* species in India: areview, Scitech J. 1 (2014) 28–36.

[9] P. Dey, K.L. Khaled, An extensive review on *Allium ampeloprasum*: a magical herb, Int. J. Sci. Res. 4 (7) (2015) 371–377.

[10] C. Guenaoui, S. Mang, G. Figliuolo, M. Neffati, Diversity in *Allium ampeloprasum*: from small and wild to large and cultivated, Genet. Resour. Crop. Evol. 60 (1) (2013) 97–114.

[11] G. Figliuolo, D. Di Stefano, Is single bulb producing garlic *Allium sativum* or *Allium ampeloprasum*? Sci. Hortic. 114 (4) (2007) 243–249.

[12] P. Hirschegger, J. Jakse, P. Trontelj, B. Bohanec, Origins of *Allium ampeloprasum* horticultural groups and a molecular phylogeny of the section *Allium* (*Allium*: Alliaceae), Mol. Phylogenet. Evol. 54 (2) (2010) 488–497.

[13] R.N. Gohil, A.K. Koul, Cytology of the tetraploid desynaptic *Allium porrum* [leeks], Nucleus 24 (1981) 79–83.

[14] H.P. Koch, L.D. Lawson, Garlic: The Science and Therapeutic Application of *Allium sativum* L. and Related Species, Williams & Wilkins, Baltimore, Maryland, 1996 (ISBN: 683181475).

[15] H.D. Reuter, *Allium sativum* and *Allium ursinum*: part 2 pharmacology and medicinal application, Phytomedicine 2 (1995) 73–91.

[16] D.R. Riggs, J.I. DeHaven, D.L. Lamm, *Allium sativum* (garlic) treatment for murine transitional cell carcinoma, Cancer 79 (1997) 1987–1994.

[17] M.S. Butt, M.T. Sultan, M.S. Butt, J. Iqbal, Garlic: nature's protection against physiological threats, Crit. Rev. Food Sci. Nutr. 49 (2009) 538–551.

[18] M. Iciek, I. Kwiecien, L. Wlodek, Biological properties of garlic and garlic-derived organosulfur compounds, Environ. Mol. Mutagen. 50 (2009) 247–265.

[19] R.M. Fritsch, M. Keusgen, Occurrence and taxonomic significance of cysteine sulphoxides in the genus *Allium* L.(Alliaceae), Phytochemistry 67 (11) (2006) 1127–1135.

[20] D.C. Rio, M. Ares, G.J. Hannon, T.W. Nilsen, Purification of RNA using TRIzol (TRI reagent), Cold Spring Harb Protoc 5 (2010) pdb–prot5439, https://doi.org/10.1101/pdb.prot5439.

[21] P. Desjardins, D. Conklin, NanoDrop microvolume quantitation of nucleic acids, J. Vis. Exp. 45 (2010) 2565, https://doi.org/10.3791/2565.

[22] ThermoScientific, NanoDrop: assessment of nucleic acid purity, Protoc. Prod. Manuals (2011) 1–2, https://doi.org/10.7860/JCDR/2015/11821.5896.

[23] Agilent Technologies, Agilent 2100 Bioanalyzer, (2007).

[24] A. Masotti, T. Preckel, Analysis of small RNAs with the Agilent 2100 Bioanalyzer, Nat. Methods 3 (2006).

[25] S. Andrews, FastQC: A Quality Control Tool for High Throughput Sequence Data, Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

[26] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics 30 (2014) 2114–2120.

[27] B.J. Haas, A. Papanicolaou, M. Yassour, et al., *De novo* transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and

analysis, Nat. Protoc. 8 (2013) 1494–1512.

[28] B. Langmead, Aligning short sequencing reads with Bowtie, Curr. Protoc. Bioinforma 32 (2010), https://doi.org/10.1002/0471250953.bi1107s32 11.7.1-11.7.14.

[29] B. Li, C.N. Dewey, RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome, BMC Bioinform. 12 (2011) 323.

[30] M.D. Robinson, D.J. McCarthy, G.K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data, Bioinformatics 26 (2010) 139–140.

[31] S.F. Altschul, W. Gish, W. Miller, et al., Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.

[32] M. Ashburner, C.A. Ball, J.A. Blake, et al., Gene ontology: tool for the unification of biology, Nat. Genet. 25 (2000) 25–29.

[33] S. Gotz, J.M. Garcia-Gomez, J. Terol, et al., High-throughput functional annotation and data mining with the Blast2GO suite, Nucleic Acids Res. 36 (2008) 3420–3435.

[34] H. Ogata, S. Goto, K. Sato, et al., KEGG: Kyoto encyclopedia of genes and genomes, Nucleic Acids Res. 27 (1999) 29–34.

[35] E. Quevillon, V. Silventoinen, S. Pillai, et al., InterProScan: protein domains identifier, Nucleic Acids Res. 33 (2005) W116–W120.

[36] S. Beier, T. Thiel, T. Munch, et al., MISA-web: a web server for microsatellite prediction, Bioinformatics 33 (2017) 2583–2585.

[37] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform, Bioinformatics 25 (2009) 1754–1760.

[38] H. Li, B. Handsaker, A. Wysoker, et al., The sequence alignment/map format and SAMtools, Bioinformatics 25 (2009) 2078–2079.

[39] C. Zhang, H. Zhang, Z. Zhan, Transcriptome analysis of sucrose metabolism during bulb swelling and development in onion (*Allium cepa* L.), Front. Plant Sci. 7 (2016) 1425.

[40] C. Liu, Q. Tang, C. Cheng, et al., Identification of putative CONSTANS-like genes from the *de novo* assembled transcriptome of leek, Biol. Plant. 62 (2018) 269–276.

[41] R. Kamenetsky, A. Faigenboim, E.S. Mayer, et al., Integrated transcriptome catalogue and organ-specific profiling of gene expression in fertile garlic (*Allium sativum* L.), BMC Genomics 16 (2015) 12.

[42] T. Liu, L. Zeng, S. Zhu, et al., Large-scale development of expressed sequence tag-derived simple sequence repeat markers by deep transcriptome sequencing in garlic (*Allium sativum* L.), Mol. Breed. 35 (2015) 204.

[43] E. Shemesh-Mayer, T. Ben-Michael, N. Rotem, et al., Garlic (*Allium sativum* L.) fertility: transcriptome and proteome analyses provide insight into flower and pollen development, Front. Plant Sci. 6 (2015) 271.

[44] A.V. Beletsky, M.A. Filyushin, E.V. Gruzdev, et al., De novo transcriptome assembly of the mycoheterotrophic plant *Monotropa hypopitys*, Genom. Data 11 (2017) 60–61.

[45] M. Carruthers, A.A. Yurchenko, J.J. Augley, et al., De novo transcriptome assembly, annotation and comparison of four ecological and evolutionary model salmonid fish species, BMC Genomics 19 (2018) 32.

[46] N. Ryder, K.M. Dorn, M. Huitsing, et al., Transcriptome assembly and annotation of johnsongrass (*Sorghum halepense*) rhizomes identify candidate rhizome-specific genes, Plant Direct. 2 (2018) e00065.

[47] M. Salem, B. Paneru, R. Al-Tobasei, et al., Transcriptome assembly, gene annotation and tissue gene expression atlas of the rainbow trout, PLoS One 10 (2015) e0121778.

[48] W.N.A.W. Zakaria, K.K. Loke, H.H. Goh, N.M. Noor, RNA-seq analysis for plant carnivory gene discovery in *Nepenthes* × *ventrata*, Genomics Data 7 (2016) 18.

[49] G.J.H. Grubben, O.A. Denton (Eds.), Plant Resources of Tropical Africa 2. Vegetables, PROTA Foundation, Wageningen; Backhuys, Leiden, CTA, Wageningen, 2004.

[50] J. Borlinghaus, F. Albrecht, M. Gruhlke, et al., Allicin: chemistry and biological properties, Molecules 19 (2014) 12591–12618.

[51] D. Ilic, V. Nikolic, L. Nikolic, et al., Allicin and related compounds: biosynthesis, synthesis and pharmacological activity, Facta Univ. 9 (2011) 9–20.

[52] M.G. Jones, J. Hughes, A. Tregova, et al., Biosynthesis of the flavour precursors of onion and garlic, J. Exp. Bot. 55 (2004) 1903–1918.

[53] M.E. Rybak, E.M. Calvey, J.M. Harnly, Quantitative determination of allicin in garlic: supercritical fluid extraction and standard addition of alliin, J. Agric. Food Chem. 52 (2004) 682–687.

[54] B. Iberl, G. Winkler, B. Muller, K. Knobloch, Quantitative determination of allicin and alliin from garlic by HPLC, Planta Med. 56 (1990) 320–326.

[55] M.G. Jones, H.A. Collin, A. Tregova, et al., The biochemical and physiological genesis of alliin in garlic, Med. Aromat Plant Sci. Biotechnol. 1 (2007) 21–24.

[56] P. Rose, M. Whiteman, P.K. Moore, Y.Z. Zhu, Bioactive S-alk (en) yl cysteine sulfoxide metabolites in the genus Allium: the chemistry of potential therapeutic agents, Nat. Prod. Rep. 22 (2005) 351–368.

[57] N. Yoshimoto, A. Yabe, Y. Sugino, et al., Garlic γ-glutamyl transpeptidases that catalyze deglutamylation of biosynthetic intermediate of alliin, Front. Plant Sci. 5 (2015) 758.

[58] W. Boerjan, J. Ralph, M. Baucher, Lignin biosynthesis, Annu. Rev. Plant Biol. 54 (2003) 519–546.

[59] L. Meng, J.H. Wong, L.J. Feldman, P.G. Lemaux, B.B. Buchanan, A membrane-associated thioredoxin required for plant growth moves from cell to cell, suggestive of a role in intercellular communication, Proc. Natl. Acad. Sci. U. S. A. 107 (2010) 3900–3905.

[60] R.K. Varshney, A. Graner, M.E. Sorrells, Genic microsatellite markers in plants: features and applications, Trends Biotechnol. 23 (2005) 48–55.

[61] S.K. Parida, A. Pandit, K. Gaikwad, et al., Functionally relevant microsatellites in sugarcane unigenes, BMC Plant Biol. 10 (2010) 251.

[62] R.S. Jasrotia, M.A. Iquebal, P.K. Yadav, et al., Development of transcriptome based web genomic resources of yellow mosaic disease in *Vigna mungo*, Physiol. Mol. Biol. Plants 23 (2017) 767–777.

[63] M. Ipek, N. Sahin, A. Ipek, et al., Development and validation of new SSR markers from expressed regions in the garlic genome, Sci. Agric. 72 (2015) 41–46.

[64] S. Gupta, R. Shukla, S. Roy, et al., *In silico* SSR and FDM analysis through EST sequences in *Ocimum basilicum*, Plant Omics 3 (2010) 121–128.

[65] D. Laudencia-Chingcuanco, S. Ganeshan, F. You, et al., Genome-wide gene expression analysis supports a developmental model of low temperature tolerance gene regulation in wheat (*Triticum aestivum* L.), BMC Genomics 12 (2011) 299.

[66] R. Chopra, G. Burow, C. Hayes, et al., Transcriptome profiling and validation of gene based single nucleotide polymorphisms (SNPs) in sorghum genotypes with contrasting responses to cold stress, BMC Genomics 16 (2015) 1040.

[67] J.M. Barrero, C. Cavanagh, K.L. Verbyla, et al., Transcriptomic analysis of wheat near-isogenic lines identifies PM19-A1 and A2 as candidates for a major dormancy QTL, Genome Biol. 16 (2015) 93.