Plant Gene 8 (2016) 9–15

Contents lists available at ScienceDirect

Plant Gene

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

The Onion Genomic Resource: A genomics and bioinformatics driven resource for onion breeding

Shantanu Shukla, M.A. Iquebal, Sarika Jaiswal, U.B. Angadi, Samar Fatma, Neeraj Kumar, Rahul Singh Jasrotia, Yasmin Fatima, Anil Rai, Dinesh Kumar *

Centre for Agricultural Bioinformatics, ICAR-Indian Agricultural Statistics Research Institute, New Delhi 110012, India

ARTICLE INFO

Article history:
Received 8 June 2016
Received in revised form 5 September 2016
Accepted 11 September 2016
Available online 17 September 2016

Keywords:
Allium
Expressed sequence tags
Gene prediction
Markers
miRNA
Onion

ABSTRACT

Onion (*Allium cepa* L.), often regarded as a crop having an antediluvian coexistence with humans, is by far one of the most challenging plant species to be worked on, especially with respect to delineating its genomic information. It is considered as a plant of immense culinary and medicinal importance. However, there are very limited genomic ventures that have so far been established that could shed light on some of the most captivating aspects of the onion genome. Onion Genomic Resource (OGR), with *three-tier architecture*, is the first of its kind, comprehensive web-resource/database, built in MySQL database and PHP that catalogues the genomic developments specific to onion. It houses information of assembly of 20,204 publicly available onion expressed sequence tags (ESTs), available 20,755 assembled transcripts and 249,987 unigenes from *Allium cepa* transcriptome shotgun assembly (TSA) along with their annotations and functional significance. A total of 1915 SSRs from Onion ESTs and 123,282 SSRs from Onion TSA data have been catalogued in OGR. Also, 135,424 SNPs and 11,891 Indels identified from Onion TSA data as well as 15 and 13 SNPs and Indels identified, respectively from Onion ESTs have been put in database. The resource also contains information of gene annotations, linked with KEGG pathways and 7 previously reported and 1 predicted onion miRNAs with their associated targets, which range from cytoplasmic globular proteins to membrane ion channels. Additionally, gene prediction was carried out for the unannotated sequences, of which few were observed to harbor coding regions for novel protein coding genes and transcripts that so far have not yet been identified. Over 200 different ready-to-use experimentally validated molecular markers were also mined from the existing literature to further enrich the lab-based studies targeting variety improvement. The OGR can be useful for onion molecular breeders as well as a valuable tool for confirmation of predicted ORF once the whole genome of onion is sequenced. The OGR is an open resource freely accessible at http://webtom.cabgrid.res.in/ogr/.

1. Introduction

Onion (*Allium cepa* L.) ($2n = 2\times = 16$), which is the largest genome (16 GB) of all the vegetable crops (Kuhl et al., 2004), belongs to the richly diversified genus *Allium* that encapsulates around 600 species, including garlic, shallot, and bunching onion, and has a distribution span over the continents of North America, North Africa, Europe and Asia (Fenwick et al., 1985). After potato and tomato, onion is the third most important horticultural crop, with over 140 countries cultivating it worldwide (FAOSTAT, 2013). Considering the current worldwide onion production which was estimated to be 78.31 million ton with an average productivity of 19.79 t/ha, India ranks first in terms of total area under onion cultivation, which is estimated to be 1.09 million hectares and is the second largest onion producer with 15.88 million tons, after China 22.46 million tons. However, in terms of productivity India (14.35 t/ha) stands at 90th position (FAOSTAT, 2013).

Annual production of onion is often hit by several abiotic and biotic factors. Major diseases that infest the crop include predominant fungal infections, like Downy Mildew (*Peronospora destructor (Berk.) Casp.*), White Rot *(Sclerotium cepivorum Berk.)*, Onion Smudge *(Colletotrichum circinans (Berk) Vogl)*, Onion Smut *(Urocystis cepulae Frost)*; bacterial pathogens like Soft Rot (*Erwinia carotovora* Holland); and viral diseases like Aster yellows which causes implacable damage to the onion crop (Fenwick et al., 1985). The abiotic stresses like high-temperature, uneven rainfall, soil nitrogen deficiency supplemented with improper irrigation also affect the production. Therefore, the given importance of onion as a crop, it becomes evident that a proper understanding of the onion genome can render the much needed information for the cultivar

* Corresponding author.
E-mail addresses: shaanbdu@gmail.com (S. Shukla), jiqubal@gmail.com (M.A. Iquebal), aijaiswal@gmail.com (S. Jaiswal), angadiub@gmail.com (U.B. Angadi), fatma.samar807@gmail.com (S. Fatma), endy171@gmail.com (N. Kumar), rahuljasrotia86@gmail.com (R.S. Jasrotia), yasmin389@gmail.com (Y. Fatima), anilrai645@gmail.com (A. Rai), dineshkumarbhu@gmail.com (D. Kumar).

http://dx.doi.org/10.1016/j.plgene.2016.09.003
2352-4073/© 2016 Elsevier B.V. All rights reserved.

improvement and towards understanding the genetic modulation of the abiotic and biotic stress responses.

Various biochemical and morphological markers are reported in literature (Cramer and Havey, 1999; Kim et al., 2014). Significant initial efforts were made by King et al. to bring about the first publically available RFLP based genetic map of onion (King et al., 1998). Additionally, first onion EST resource and a PCR marker based map were developed for aiding the initial *in silico* genomic data analysis (Kuhl et al., 2004; Martin et al., 2005). Subsequently, several other genetic maps of onion using AFLP (van Heusden et al., 2000a,b) and RAPD (Shigyo et al., 1997; Friesen and Klaas, 1998) markers have also been reported. More recently, researchers have started selectively targeting some of the genes affecting prominent trait loci, like for carbohydrate assimilation (Masuzaki et al., 2006; McCallum et al., 2006; Yaguchi et al., 2008), flavonoid biosynthesis (Masuzaki et al., 2006), lachrymatory factor synthesis (Masamura et al., 2012), bulb color (Kim et al., 2004; Kim et al., 2005a,b; Khar et al., 2008), pungency (McCallum et al., 2007, 2011; Thomas et al., 2011; McManus et al., 2012) and bulb formation and flowering (Lee et al., 2013). Markers are also reported for differentiation of three different cytoplasm required for hybrid production in onion (Kim et al., 2009).

All valuable information of genomic resource of onion are scattered and there is no freely accessible database of it. Initial efforts succeeded in bringing about a collective information resource on predominant *Allium* species in the form of AlliumMap (McCallum et al., 2012) but are not freely accessible.

Similar work has been reported in garlic that includes a well concerted analysis of garlic ESTs to predict their functional attributes in the form of GarlicESTdb (Kim et al., 2009). However, both these *Allium* genus databases are confined to limited information on SNP and SSR markers and genic region variations by RNA seq data is yet to be covered. Such database is not only required as research tool but also by molecular breeders if it is having ready to use genotyping information of markers.

Sequencing of onion genome is challenging because of largest size and poor density of one gene per 168 KB (Jakse et al., 2008). Since whole genome sequence of onion is yet to be done, thus it will be a valuable tool for confirmation of predicted ORF in the days to come.

The present work aims at development of web genomic resources using all available EST and RNA Seq data along with SNP and SSR marker discovery. The work also aims at miRNA and its target prediction along with gene annotation.

2. Materials and method

2.1. Data collection and pre-processing

A total of 20,204 EST sequences corresponding to *Allium cepa*, were retrieved from the EST database of the NCBI. The redundancy in these sequences were removed and used as input for contig and singleton generation to the EGassembler (Masoudi-Nejad et al., 2006) which is an automated tool for handling large sets of EST data that performs sequence cleansing, repeat masking, vector trimming, organelle masking, clustering and fragment assembly. Repeat masking was performed against the RepBase repeats library and organelle masking was performed with highly conserved plastid library of *Arabidopsis*. Vector Masking was done against the non-redundant core vector library of NCBI. An overlap percent identity cutoff of 80 was implemented to get the best possible contigs and singletons with minimum false positives. Also, 20,755 assembled transcripts were obtained from Transcriptome Shotgun Assembly (TSA) sequence database of NCBI (accession JR842819-JR863573). Also the 249,987 onion transcriptome assembly data was retrieved from NCBI (NCBI-SRA accession SRR1293377; GenBank: GBJZ01000001-GBJZ01249987) (Rajkumar et al., 2015). Annotation, SSRs and SNPs markers' identification from these data has been done. The workflow was followed as illustrated in Fig. 1. Provision of these ESTs to be categorised tissue has been provided in database under the tab “Tissue” of OGR.

2.2. Functional annotation

All generated 3754 contigs, 6534 singletons, 13,378 transcripts and 123,282 unigenes were considered for functional annotation using the Blast2GO Pro (Conesa et al., 2005). These were subjected to a BLASTX (Altschul et al., 1990) with a high ExpectValue cutoff of 1.0E-3, against non-redundant protein database of NCBI. This was followed by GO mapping for the identification of inherent ontologies and annotation. The automated Blast2GO mapping of the contigs, singletons, transcripts and unigenes were done against the Gene Ontology Database and the ontology assignments were verified against the Evidence Code (Consortium, 2004). Annotation was further done for these mapped entries using default parameters. InterProScan (Zdobnov and Apweiler, 2001) was performed for deciphering associated enzymes using the GO-EnzymeCode Mapping module of Blast2GO.

2.3. Novel gene prediction

Since available genomic databases have limited onion sequences and information, thus novel gene prediction was followed for contigs without any hits. Contig set was subjected to *in silico* ORF prediction using MEGANTE (Numa and Itoh, 2014) and the consensus gene structure was assembled using JIGSAW (Allen et al., 2006). *Arabidopsis thaliana* genome was chosen as the reference set for ORF prediction.

2.4. miRNA identification and target prediction

The onion ESTs were also mined for the presence of any possible miRNA sequences as described by Zhang and co-workers (Zhang et al., 2005). The entries that were predicted as potential miRNAs were further evaluated for structure by RNAfold server (Hofacker, 2003). Since miRNAs play a key role in gene regulation and complement overall development of the plant by complementarily binding to their targets, we performed target prediction of the predicted miRNA on the annotated onion data set using the psRNATarget server (Dai and Zhao, 2011). More flexible parameters were used with an ExpectValue of 5.0 and a UPE of 20.0 keeping other parameters as default to identify all possible target set among the annotated onion data set.

2.5. Computational marker identification

SSR markers for the retrieved onion ESTs and transcriptome data were identified using Microsatellite Identification tool (MISA; http://pgrc.ipk-gatersleben.de/misa) and primers were generated using Primer3 (Rozen and Skaletsky, 1996). To identify the inherent single nucleotide variations in the onion ESTs, we subjected the EST sequences to SNPsFinder (Song et al., 2005) against the *A. thaliana* gene sets that were accessed from unigene data repository of NCBI. Once the SNPs and Indels were identified in the sequences related primer sets were generated for them by SNPsFinder. For the transcriptome data of *Allium cepa*, SNPs and Indels were mined using the BWA (Li and Durbin, 2009) and SAMtools (Li et al., 2009) pipeline.

2.6. Literature mining for identification of molecular markers with experimental verification in onion

Intensive literature mining was done to identify molecular markers such as AFLP, SSRs and SNPs in onion. The markers were extracted along with their GenBank associations, if existing, as mentioned in the literature. The data were arranged based on the chromosome to which these markers were tagged to as well as on the basis of their cluster group.

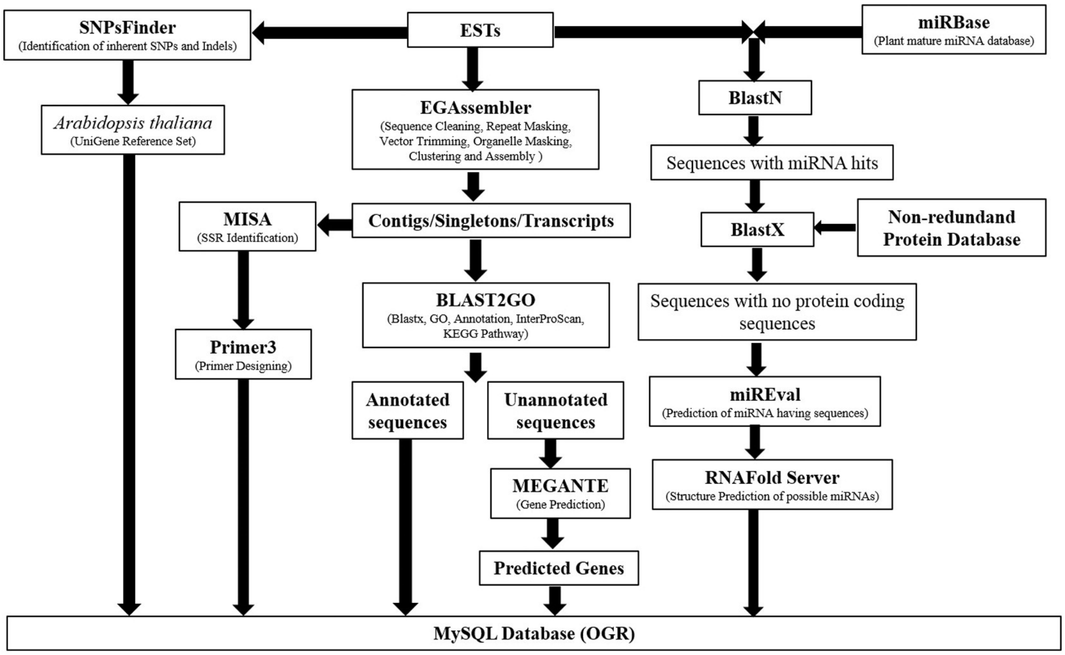

SNPsFinder
(Identification of inherent SNPs and Indels)
ESTs
miRBase
(Plant mature miRNA database)
Arabidopsis thaliana
(UniGene Reference Set)
EGAssembler
(Sequence Cleaning, Repeat Masking, Vector Trimming, Organelle Masking, Clustering and Assembly)
BlastN
Sequences with miRNA hits
MISA
(SSR Identification)
Contigs/Singletons/Transcripts
BlastX
Non-redundand Protein Database
BLAST2GO
(Blastx, GO, Annotation, InterProScan, KEGG Pathway)
Sequences with no protein coding sequences
Primer3
(Primer Designing)
Annotated sequences
Unannotated sequences
miREval
(Prediction of miRNA having sequences)
MEGANTE
(Gene Prediction)
RNAFold Server
(Structure Prediction of possible miRNAs)
Predicted Genes
MySQL Database (OGR)

Fig. 1. The overall methodology followed during the construction of the Onion Genomic Resource.

2.7. Database and Web interface development

We designed comprehensive relational schema using MySQL to catalog the generated information using onion ESTs and transcriptome data. The whole schema summarizes tables firmly connected using the primary key and foreign key constructs of the MySQL database. To facilitate the data retrieval and to enhance the ease of access, a supporting web interface was developed using PHP.

3. Results and discussion

3.1. Annotation of onion EST and transcriptome data

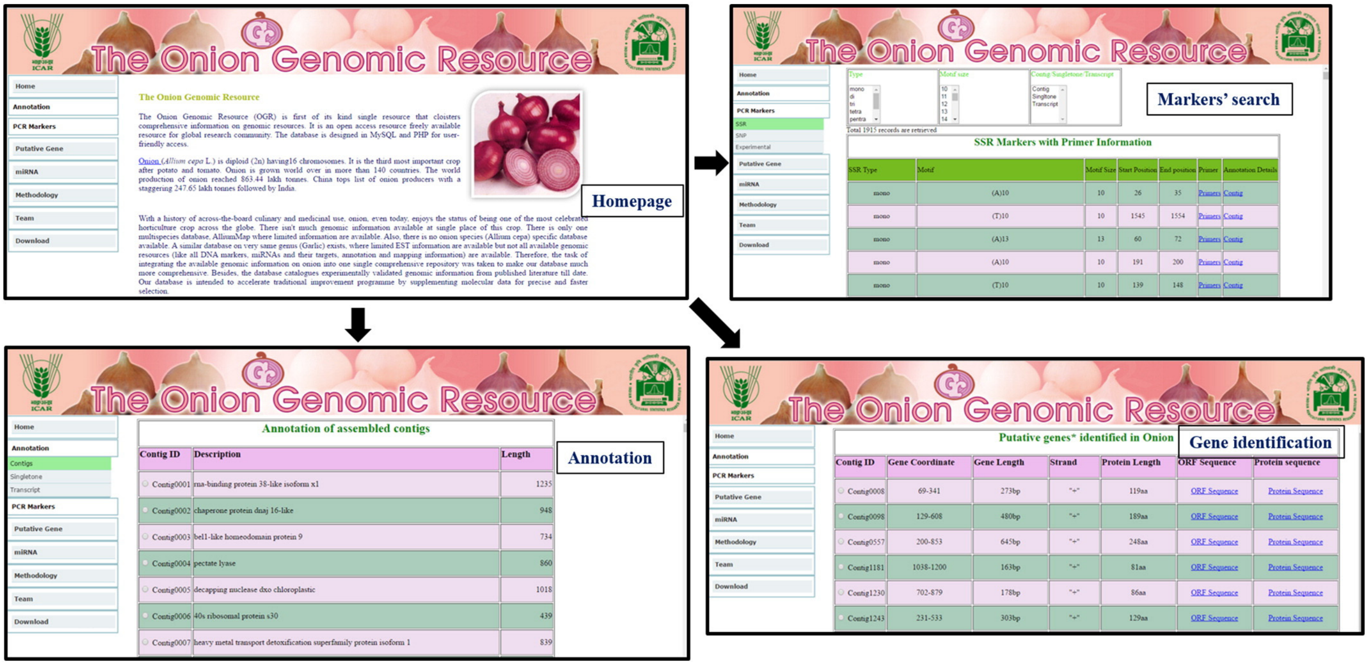
3.1.1. Data pre-processing and assembly of EST dataset

The entire set of 20204 ESTs was pre-processed and assembled into contigs and singletons. The sequence pre-processing included sequence cleansing, repeat masking, vector masking and organelle masking. Sequence cleansing validated a total of 20170 sequences as contaminant free and deleted 34 EST sequences as contaminated, with 13 of them being extremely low quality sequences. Further, a total of 99846 bases were removed out using the RepeatMasker. Subsequent steps of vector masking and organelle masking trashed 18 and 6 sequences respectively. The final assembled contig set consisted of 3754 consensus contigs and 6534 singletons. The consensus contigs and singletons accounted for about 19% and 41% respectively of the total EST set. Together, the entire non-redundant contig set contained 12121 sequences which was approximately 60% of the overall EST number.

3.1.2. Annotation

The achieved assembled sequences were subjected to Blastx against the non-redundant protein database of NCBI that returned hits for 3587 contig sequence and 167 contigs could not get annotated. The possible reason that 167 contigs remained unannotated was due to the high Expect Value of 1.0e − 3. Further, the Blast2GO mapping identified gene ontologies for 3243 contigs as some of the Blast hits with low sequence similarity were not mapped due to a stringent mapping cutoff. Subsequently, InterProScan was initiated to complement the mapped data with the structural similarity search in terms of protein family, protein domains, repeat regions, binding site, active site and motifs. In the entire set of 3754 contigs, InterProScan predicted a total of 53196 structural entities out of which about 49% is constituted by domains followed by 40% of the entities classified as family, rest all the other structural entities had a marginal share. The final annotation was then run for all the mapped contigs which qualified the annotation cutoff and the results were integrated with the results of InterProScan. Additionally, to improve the overall annotation and to identify the inherent pathways of the annotated enzyme, a KEGG mapping was also performed. Similar procedure was followed for singletons, transcripts and the unigenes dataset. A Total of 826 enzymes were identified in a set of 129 pathways with starch and sucrose metabolism pathway containing as many as 27 enzymes followed by amino sugar and nucleotide sugar metabolism pathway with 25 enzymes in EST data. Abundance of carbohydrate metabolizing enzymes are expected in onion hosts where sugar entities are transformed into other forms especially fructan stored in the bulb (Sinclair et al., 1995; Gennaro et al. 2002). We identified a total of 9 and 19 enzymes in sulfur metabolism pathway, cysteine and methionine metabolism pathway respectively that directly influence the pungency by forming important intermediates required for the formation of volatile compounds from EST assembly dataset (McCallum et al., 2007).

The overall classification of these sequences was done using GO, categorizing them according to their biological functions, cellular component and cellular function. Segregating the contigs based on the biological function returned 117 different functions wherein the maximum number of contig sequences were involved in cellular processes (2049 contigs) closely followed by metabolic processes (1975 contigs) suggesting a huge share of genes controlling the conversion of metabolites especially sugar and sulfur-containing ones. On the basis of cellular localization maximum contigs were assimilated inside the cell cytoplasm and intracellular organelles while a significant amount of sequences contributed to the maintenance of cell wall integrity. Finally, on the account of their molecular functions it was observed that the majority of the contiguous depicted catalytic activity and a significant



The Onion Genomic Resource

The Onion Genomic Resource (OGR) is first of its kind single resource that cloisters comprehensive information on genomic resources. It is an open access resource freely available resource for global research community. The database is designed in MySQL and PHP for user-friendly access.

Onion (*Allium cepa* L.) is diploid (2n) having16 chromosomes. It is the third most important crop after potato and tomato. Onion is grown world over in more than 140 countries. The world production of onion reached 863.44 lakh tonnes. China tops list of onion producers with a staggering 247.65 lakh tonnes followed by India.

With a history of across-the-board culinary and medicinal use, onion, even today, enjoys the status of being one of the most celebrated horticulture crop across the globe. There isn't much genomic information available at single place of this crop. There is only one multispecies database, AlliumMap where limited information are available. Also, there is no onion species (Allium cepa) specific database available. A similar database on very same genus (Garlic) exists, where limited EST information are available but not all available genomic resources (like all DNA markers, miRNAs and their targets, annotation and mapping information) are available. Therefore, the task of integrating the available genomic information on onion into one single comprehensive repository was taken to make our database much more comprehensive. Besides, the database catalogues experimentally validated genomic information from published literature till date. Our database is intended to accelerate traditional improvement programme by supplementing molecular data for precise and faster selection.

Type: mono, di, tri, tetra, pentra
Motif size: 10, 11, 12, 13, 14
Contig/Singletone/Transcript: Contig, Singltone, Transcript

Total 1915 records are retrieved

SSR Markers with Primer Information

SSR Type	Motif	Motif Size	Start Position	End position	Primer	Annotation Details
mono	(A)10	10	26	35	Primers	Contig
mono	(T)10	10	1545	1554	Primers	Contig
mono	(A)13	13	60	72	Primers	Contig
mono	(A)10	10	191	200	Primers	Contig
mono	(T)10	10	139	148	Primers	Contig

Annotation of assembled contigs

Contig ID	Description	Length
○ Contig0001	rna-binding protein 38-like isoform x1	1235
○ Contig0002	chaperone protein dnaj 16-like	948
○ Contig0003	bel1-like homeodomain protein 9	734
○ Contig0004	pectate lyase	860
○ Contig0005	decapping nuclease dxo chloroplastic	1018
○ Contig0006	40s ribosomal protein s30	439
○ Contig0007	heavy metal transport detoxification superfamily protein isoform 1	839

Putative genes* identified in Onion

Contig ID	Gene Coordinate	Gene Length	Strand	Protein Length	ORF Sequence	Protein sequence
○ Contig0008	69-341	273bp	"+"	119aa	ORF Sequence	Protein Sequence
○ Contig0098	129-608	480bp	"+"	189aa	ORF Sequence	Protein Sequence
○ Contig0557	200-853	645bp	"+"	248aa	ORF Sequence	Protein Sequence
○ Contig1181	1038-1200	163bp	"+"	81aa	ORF Sequence	Protein Sequence
○ Contig1230	702-879	178bp	"+"	86aa	ORF Sequence	Protein Sequence
○ Contig1243	231-533	303bp	"+"	129aa	ORF Sequence	Protein Sequence

Fig. 2. The database search for markers, genes and annotation.

amount of sequences also showed activity as DNA-binding transcription factors which might have a direct role in the regulation of gene expression. The search for annotated sequences is delineated in Fig. 2.

Out of 249,987 unigenes, 123,282 showed similarity with known genes. Annotations of 6534 singletons, 13,378 transcripts and 123,282 unigenes retrieved as discussed in earlier section were done and

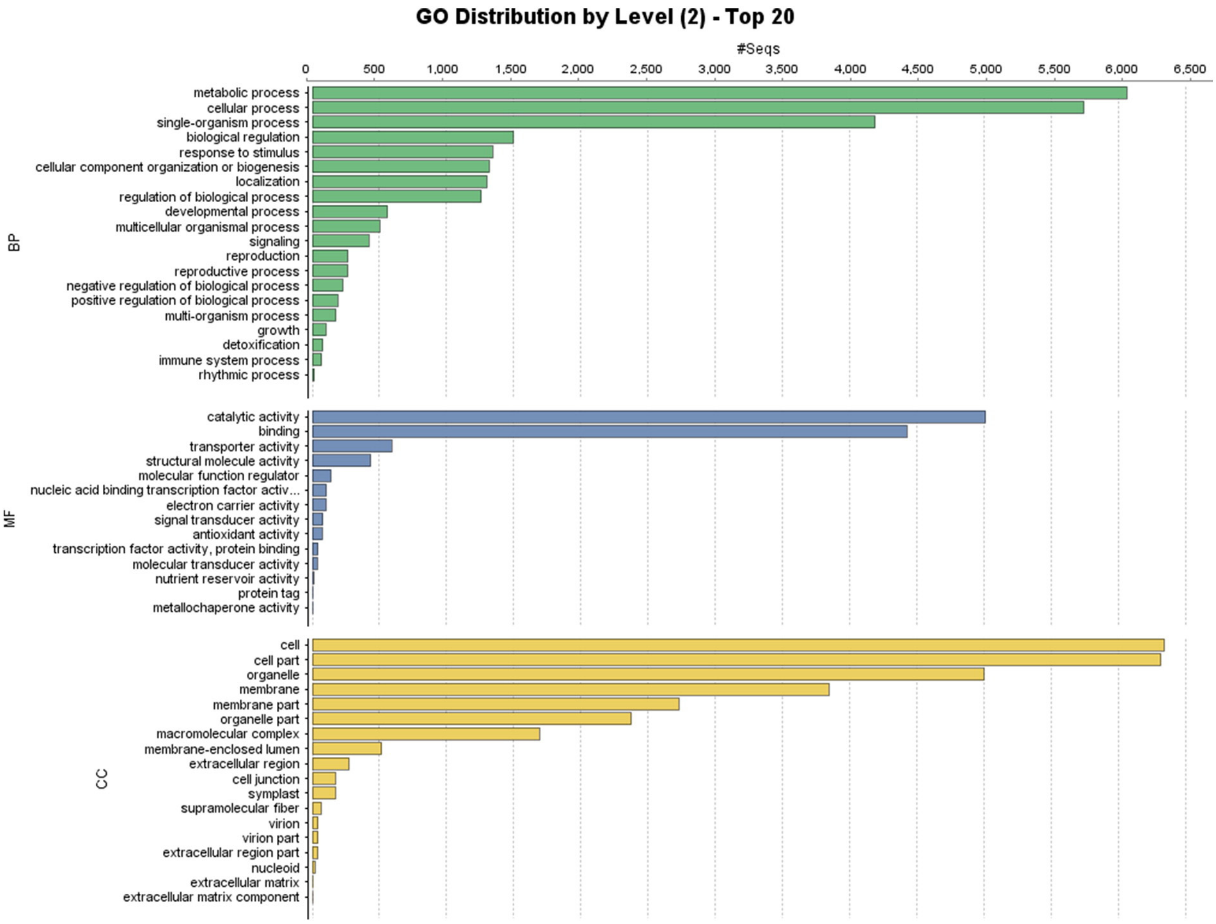


Fig. 3. GO-term annotation for biological process (BP), Molecular Function (MF) and Cellular component (CC) for transcript data.

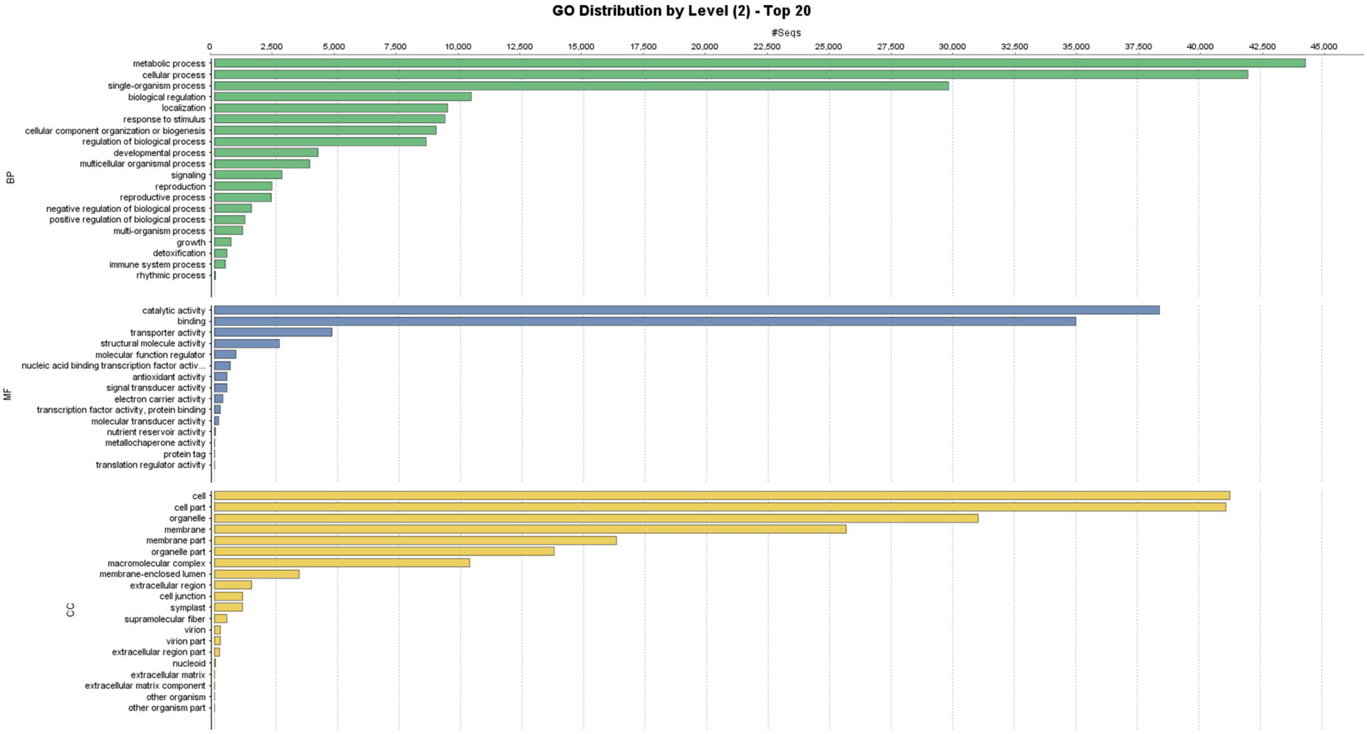


Fig. 4. GO-term annotation for biological process (BP), Molecular Function (MF) and Cellular component (CC) for unigenes data.

available at OGR. Fig. 3 shows the GO-term annotation distribution for biological process (BP), Molecular Function (MF) and Cellular component (CC) of the transcript data. Similarly, Fig. 4 delineates the gene ontology distribution for the unigenes data.

3.1.3. Gene prediction

The set of sequences that did not generate any Blast hit were further processed for the presence of any putative genes that contain the CDS and could possibly code for hypothetical proteins. Hence, we searched for gene signatures for the unannotated sequences against the unique gene set of *Arabidopsis thaliana*. We predicted twenty putative coding regions that did not code for any known proteins with high confidence. The coding regions were compared to the UniProt databases but they did not return any significant matches with an existing annotated protein sequence. This however, needs to be further validated experimentally if these genes and their protein products really exist in onion genome. The database search for predicted genes can be followed as in Fig. 2.

3.2. miRNA identification and target prediction

Experimental identification of miRNAs is very tedious with ample chances of contamination and therefore quite recently computational prediction techniques are used for the purpose of miRNA identification. A very limited number of microRNAs (miRNAs), i.e., only 7 have been predicted in onion using computational pipeline (Zhang et al., 2005). Recently, there have been attempts to identify plant miRNAs based on the EST sequences available in the public data repositories like in case of garlic (Panda et al., 2014), wheat (Jin et al., 2008), cotton (Qiu et al., 2007; Zhang et al., 2007), rapeseed (Xie et al., 2007) etc. Following the similar protocol one more miRNA is added to the existing list of 7 miRNAs using the EST data of onion (Fig. 5). The identified miRNA i.e. ace-MIR444a, belongs to the miRNA family MIR444 and is 21 bases long which is similar to the size of previously identified miRNAs in onion. It was identified in the EST sequence gi34464571 which showed an identity of 95% with osa-miR444b.1 from *Oryza sativa* with a low expect of 2e − 004. There were no gaps observed between the target and the query sequence. The identified miRNA had a minimum folding energy (MFE) of −77.85 Kcal/mol and a minimum folding energy index of 0.810 which substantiated the fact that the miRNA structure was very stable with a perfectly folded stem loop structure (Fig. 3). The structure satisfied all the parameters necessary for a sequence to be computationally characterized as a miRNA suggested by Zhang and coworkers (Zhang et al., 2005). The reason that only one miRNA was isolated from the entire set of EST sequences can be attributed to the fact that a very parsimonious parameter set was applied to negate the appearance of false positives, thereby restricting the prediction to only ones that satisfied all the constraints.

Subsequently, the identified miRNA was subjected to a target prediction against the available set to determine its possible regulatory role. We identified around 19 target sequences that traverse both cytoplasmic enzymes and membrane bound proteins therefore, signifying the importance of the identified novel miRNA and its regulatory mechanism especially related to abiotic stress response. The parameters were relaxed in case of target identification as we wanted to observe all the potential targets that might get regulated by the identified miRNA. The search for miRNA and its target in the onion web resource is shown in Fig. 6.

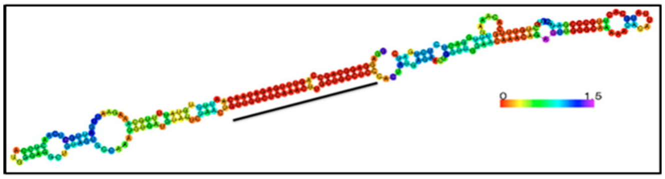
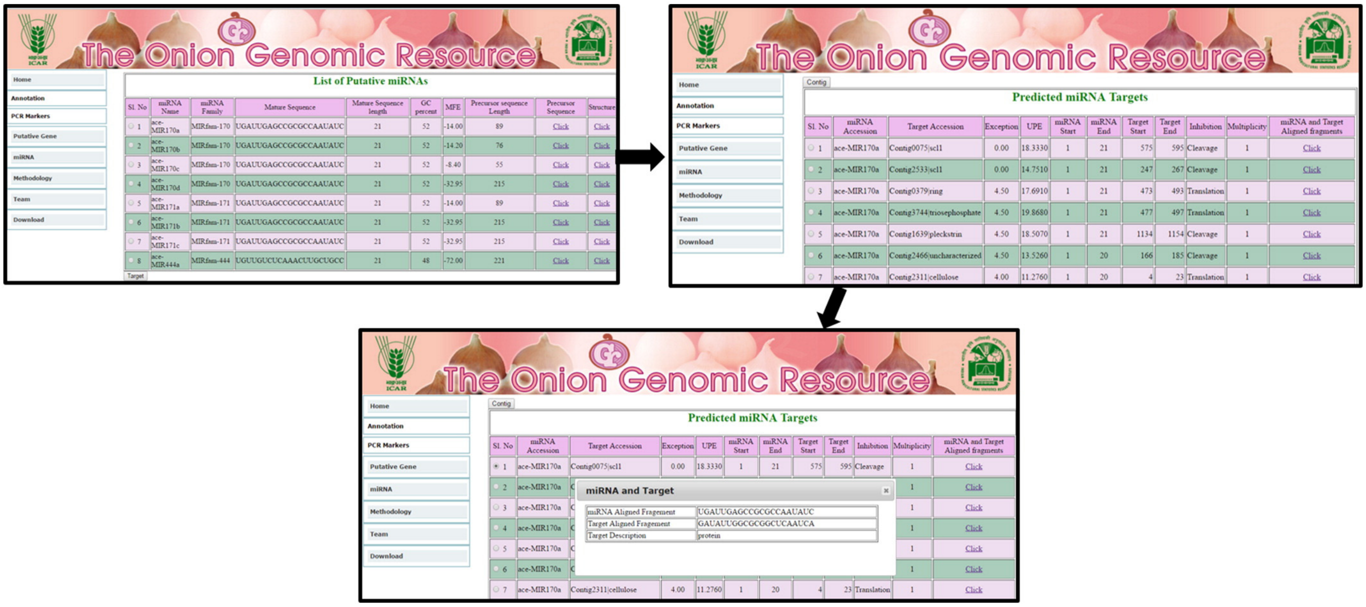


Fig. 5. The predicted miRNA ace-MIR444a in *Allium cepa* (Rregion coding for the miRNA is shown by the black line. The image was generated in accordance with the positional entropy of the individual residues, which varies between 0 and 1.5).



List of Putative miRNAs

Sl. No	miRNA Name	miRNA Family	Mature Sequence	Mature Sequence length	GC percent	MFE	Precursor sequence Length	Precursor Sequence	Structure
○ 1	ace-MIR170a	MIRfam-170	UGAUUGAGCCGCGCCAAUAUC	21	52	-14.00	89	Click	Click
○ 2	ace-MIR170b	MIRfam-170	UGAUUGAGCCGCGCCAAUAUC	21	52	-14.20	76	Click	Click
○ 3	ace-MIR170c	MIRfam-170	UGAUUGAGCCGCGCCAAUAUC	21	52	-8.40	55	Click	Click
○ 4	ace-MIR170d	MIRfam-170	UGAUUGAGCCGCGCCAAUAUC	21	52	-32.95	215	Click	Click
○ 5	ace-MIR171a	MIRfam-171	UGAUUGAGCCGCGCCAAUAUC	21	52	-14.00	89	Click	Click
○ 6	ace-MIR171b	MIRfam-171	UGAUUGAGCCGCGCCAAUAUC	21	52	-32.95	215	Click	Click
○ 7	ace-MIR171c	MIRfam-171	UGAUUGAGCCGCGCCAAUAUC	21	52	-32.95	215	Click	Click
○ 8	ace-MIR444a	MIRfam-444	UGUUGUCUCAAACUUGCUGCC	21	48	-72.00	221	Click	Click

Predicted miRNA Targets

Sl. No	miRNA Accession	Target Accession	Exception	UPE	miRNA Start	miRNA End	Target Start	Target End	Inhibition	Multiplicity	miRNA and Target Aligned fragments
○ 1	ace-MIR170a	Contig0075|scl1	0.00	18.3330	1	21	575	595	Cleavage	1	Click
○ 2	ace-MIR170a	Contig2533|scl1	0.00	14.7510	1	21	247	267	Cleavage	1	Click
○ 3	ace-MIR170a	Contig0379|ring	4.50	17.6910	1	21	473	493	Translation	1	Click
○ 4	ace-MIR170a	Contig3744|triosephosphate	4.50	19.8680	1	21	477	497	Translation	1	Click
○ 5	ace-MIR170a	Contig1639|pleckstrin	4.50	18.5070	1	21	1134	1154	Cleavage	1	Click
○ 6	ace-MIR170a	Contig2466|uncharacterized	4.50	13.5260	1	20	166	185	Cleavage	1	Click
○ 7	ace-MIR170a	Contig2311|cellulose	4.00	11.2760	1	20	4	23	Translation	1	Click

Predicted miRNA Targets

Sl. No	miRNA Accession	Target Accession	Exception	UPE	miRNA Start	miRNA End	Target Start	Target End	Inhibition	Multiplicity	miRNA and Target Aligned fragments
● 1	ace-MIR170a	Contig0075|scl1	0.00	18.3330	1	21	575	595	Cleavage	1	Click
○ 2	ace-MIR170a	C								1	Click
○ 3	ace-MIR170a	C								1	Click
○ 4	ace-MIR170a	C								1	Click
○ 5	ace-MIR170a	C								1	Click
○ 6	ace-MIR170a	C								1	Click
○ 7	ace-MIR170a	Contig2311|cellulose	4.00	11.2760	1	20	4	23	Translation	1	Click

miRNA and Target ×

miRNA Aligned Fragement	UGAUUGAGCCGCGCCAAUAUC
Target Aligned Fragement	GAUAUUGGCGCGGCUCAAUCA
Target Description	protein

Fig. 6. Web interface and additional query features for miRNA and its target prediction.

3.3. Computational marker identification from onion EST and transcriptome data

3.3.1. Identification of SSR markers

Simple sequence repeats (SSRs) are among the easily computationally identifiable markers that have direct implications in promoting marker/molecular-assisted selection (MAS) in plant breeding programs (Varshney et al., 2005; Collard and Mackill, 2008). A total of 96,474 SSR markers were mined from all the data. Of these, 122 and 349 were from assembled contigs and singletons, respectively from assembled EST sequences of onion. The most predominant among these were mononucleotide repeats. A total of 1444 SSRs and 94,559 SSRs were mined from transcripts and unigenes data, respectively. The trinucleotide motif AGC/CTG accounted for about 8% of the total SSRs identified and was the highest frequency motif amongst trinucleotide SSRs. We successfully generated primers for identified SSRs which can be readily used by molecular breeders. This information was duly integrated with the SSR information and is available on our database. The database search for markers is shown in Fig. 2.

3.3.2. Mining SNPs

We initially compared the EST set against *Arabidopsis thaliana* UniGene sequences as they were extensively refined and verified amongst all other plant species. However, we encountered only 15 small number of single nucleotide polymorphisms (SNPs) and total insertions and deletions (indel) count of 13. The reason attributed to such small number of SNPs and indels can be accounted on the fact that *A. thaliana* is a dicot and far distantly related to the monocot *A. cepa*. Therefore, we took barley (*Hordeum vulgare*) which is a close neighbor to *A. cepa* whose genome has recently been sequenced (The International Barley Genome Sequencing Consortium, 2012). This transition in the reference genome modestly improved the number of SNPs to 63 and number of Indels to 23. The highest number of SNPs in a single EST was observed in the anchor sequence gi34467659 with a total of 22 ESTs (34% of the total SNPs identified). Additionally, primers were generated to mark these EST-SNPs that could help in their experimental verifications. From the RNA Seq dataset, SNPs from *Allium cepa* were mined using *de novo* transcriptome assembly as a reference, where paired end reads were mapped with BWA tool and SAMtools was used for calling SNP and Indels. Total 135,424 SNPs and 11,891 indels were found using filtering criteria such as read depth greater four and quality score greater than 20.

3.4. Literature mining for the identification of molecular markers with experimental verification

Given the huge importance of DNA based molecular markers towards improving the efficiency of plant breeding, their extensive association with QTLs and in MAS, it becomes evident to incorporate information related to already identified markers in onion. Unlike the computationally identified markers which require an experimental verification, the previously reported markers in the literature are well identified and authenticated, thereby can be considered as ready-to use markers by molecular plant breeders. Therefore, we have compiled important literature related to reported markers in onion. The database so far contains a total of 209 experimentally identified marker sequences which will be constantly updated with new discoveries. As AlliumMap has one of the most recent published accounts of the existing experimentally validated molecular markers from the genus *Allium*, we have isolated a significant number of molecular markers from the provided list of marker sequence. The database predominantly has chromosome-wise marker information for SSR, SNP, and AFLP markers (Fig. 2). All the marker data have been duly linked to their supporting literature that can be used for inferring the role of these markers in the corresponding study.

3.5. Application of OGR

The OGR is the first single comprehensive onion specific resource that houses a large set of 3574, 6534 and 13,378 functionally annotated contigs, singletons and transcripts, respectively supplemented by other genomic information such as predicted genes miRNAs, SNPs, SSRs and markers. Focus was given in designing the OGR, especially in balancing the incorporation of functional genomic data and cataloging marker information from both experimental and computational sources that could help expand its usability. The OGR can be applied to research related to the marker assisted selection (MAS), comparative genomics, transgenic studies, pathway curation, metabolite engineering, abiotic and biotic stress response management, and gene regulation analysis.

4. Conclusion

We designed onion genomic resources (OGR) that catalogues comprehensive information of assembly publicly available onion ESTs and the available transcriptome data from *Allium cepa* along with their annotations and functional significance. We annotated 3754 contigs, 6534 singletons from onion EST sequences and 136,660 with transcripts/unigenes that were available in public domain. Additionally, information related to onion miRNAs and their targets are provided. A huge information of 96,474 SSR markers,135,439 SNPs and 11,904 indels have been catalogued along with over 200 experimental derived markers. It is a maiden effort to bridge the gap in the knowledge of this essential horticulture crop. This open resource freely accessible web resource will be useful to onion molecular breeders and assist in supplementing information once the whole genome of onion is sequenced.

Acknowledgement

Authors are thankful to Indian Council of Agricultural Research, Ministry of Agriculture and Farmers' Welfare, Govt. of India for providing all facilities. Authors are also thankful to National Agricultural Innovation Project funded by World Bank (30(68)/2009/Bio/Informatics/NAIP/O&M) for creating computational facility, ASHOKA at ICAR-IASRI, New Delhi, India where all analyses were performed.

References

Allen, J.E., Majoros, W.H., et al., 2006. JIGSAW, GeneZilla, and GlimmerHMM: puzzling out the features of human genes in the ENCODE regions. Genome Biol. 7 (Suppl 1), S9.

Altschul, S.F., Gish, W., et al., 1990. Basic local alignment search tool. J. Mol. Biol. 215 (3), 403–410.

Collard, B.C., Mackill, D.J., 2008. Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Philos. Trans. R. Soc., B 363 (1491), 557–572.

Conesa, A., Götz, S., et al., 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21 (18), 3674–3676.

Consortium, G.O., 2004. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 32 (suppl 1), D258–D261.

Cramer, C.S., Havey, M.J., 1999. Morphological, Biochemical, and Molecular Markers in Onion. HortSci. 34 (4), 589–593.

Dai, X., Zhao, P.X., 2011. psRNATarget: a plant small RNA target analysis server. Nucleic Acids Res. 39 (suppl 2), W155–W159.

FAOSTAT, 2013. Agriculture Organization of the United Nations.

Fenwick, G.R., Hanley, A.B., et al., 1985. The genus Allium—part 1. Crit. Rev. Food Sci. Nutr. 22 (3), 199–271.

Friesen, N., Klaas, M., 1998. Origin of some minor vegetatively propagated Allium crops studied with RAPD and GISH. Genet. Resour. Crop. Evol. 45 (6), 511–523.

Gennaro, L., Leonardi, C., et al., 2002. Flavonoid and carbohydrate contents in tropea red onions: effects of homelike peeling and storage. J. Agric. Food Chem. 50, 1904–1910.

Hofacker, I.L., 2003. Vienna RNA secondary structure server. Nucleic Acids Res. 31 (13), 3429–3431.

Jakse, J., Meyer, J.D.F., et al., 2008. Pilot sequencing of onion genomic DNA reveals fragments of transposable elements, low gene densities, and significant gene enrichment after methyl filtration. Mol. Gen. Genomics 280, 287.

Jin, W., Li, N., et al., 2008. Identification and verification of microRNA in wheat (*Triticum aestivum*). J. Plant Res. 121 (3), 351–355.

Khar, A., Jakse, J., et al., 2008. Segregations for onion bulb colors reveal that red is controlled by at least three loci. J. Am. Soc. Hortic. Sci. 133 (1), 42–47.

Kim, S., Binzel, M., et al., 2004. Pink (P), a new locus responsible for a pink trait in onions (*Allium cepa*) resulting from natural mutations of anthocyanidin synthase. Mol. Gen. Genomics. 272 (1), 18–27.

Kim, S., Jones, R., et al., 2005a. The L locus, one of complementary genes required for anthocyanin production in onions (*Allium cepa*), encodes anthocyanidin synthase. Theor. Appl. Genet. 111 (1), 120–127.

Kim, S., Yoo, K.S., et al., 2005b. Development of a PCR-based marker utilizing a deletion mutation in the dihydroflavonol 4-reductase (DFR) gene responsible for the lack of anthocyanin production in yellow onions (*Allium cepa*). Theor. Appl. Genet. 110 (3), 588–595.

Kim, D.-W., Jung, T.-S., et al., 2009. GarlicESTdb: an online database and mining tool for garlic EST sequences. BMC Plant Biol. 9 (1), 61.

Kim, Y.J., Seo, S.G., et al., 2014. Recovery effect of onion peel extract against H2 O2 - induced inhibition of gap-junctional intercellular communication is mediated through quercetin. J. Food Sci. 79 (5), 1011–1017.

King, J., Bradeen, J., et al., 1998. A low-density genetic map of onion reveals a role for tandem duplication in the evolution of an extremely large diploid genome. Theor. Appl. Genet. 96 (1), 52–62.

Kuhl, J.C., Cheung, F., et al., 2004. A unique set of 11,008 onion expressed sequence tags reveals expressed sequence and genomic differences between the monocot orders Asparagales and Poales. Plant Cell Online 16 (1), 114–125.

Lee, R., Baldwin, S., et al., 2013. Flowering locus T genes control onion bulb formation and flowering. Nat. Commun. 4.

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25 (14), 1754–1760.

Li, H., Handsaker, B., et al., 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25 (16), 2078–2079.

Martin, W.J., McCallum, J., et al., 2005. Genetic mapping of expressed sequences in onion and in silico comparisons with rice show scant colinearity. Mol. Gen. Genomics. 274 (3), 197–204.

Masamura, N., McCallum, J., et al., 2012. Chromosomal organization and sequence diversity of genes encoding lachrymatory factor synthase in *Allium cepa* L. Genes Genomes Genet. 2 (6), 643–651.

Masoudi-Nejad, A., Tonomura, K., et al., 2006. EGassembler: online bioinformatics service for large-scale processing, clustering and assembling ESTs and genomic DNA fragments. Nucleic Acids Res. 34 (suppl 2), W459–W462.

Masuzaki, S., Shigyo, M., et al., 2006. Direct comparison between genomic constitution and flavonoid contents in Allium multiple alien addition lines reveals chromosomal locations of genes related to biosynthesis from dihydrokaempferol to quercetin glucosides in scaly leaf of shallot (*Allium cepa* L.). Theor. Appl. Genet. 112 (4), 607–617.

McCallum, J., Clarke, A., et al., 2006. Genetic mapping of a major gene affecting onion bulb fructan content. Theor. Appl. Genet. 112 (5), 958–967.

McCallum, J., Pither-Joyce, M., et al., 2007. Genetic mapping of sulfur assimilation genes reveals a QTL for onion bulb pungency. Theor. Appl. Genet. 114 (5), 815–822.

McCallum, J., Thomas, L., et al., 2011. Genotypic variation in the sulfur assimilation and metabolism of onion (*Allium cepa* L.) I. Plant composition and transcript accumulation. Phytochemistry 72 (9), 882–887.

McCallum, J., Baldwin, S., et al., 2012. AlliumMap-A comparative genomics resource for cultivated Allium vegetables. BMC Genomics 13 (1), 168.

McManus, M.T., Joshi, S., et al., 2012. Genotypic variation in sulfur assimilation and metabolism of onion (*Allium cepa* L.) III. Characterization of sulfite reductase. Phytochemistry 83, 34–42.

Numa, H., Itoh, T., 2014. MEGANTE: a web-based system for integrated plant genome annotation. Plant Cell Physiol. 55 (1), e2.

Panda, D., Dehury, B., et al., 2014. Computational identification and characterization of conserved miRNAs and their target genes in garlic (*Allium sativum* L.) expressed sequence tags. Gene 537 (2), 333–342.

Qiu, C.X., Xie, F.L., et al., 2007. Computational identification of microRNAs and their targets in *Gossypium hirsutum* expressed sequence tags. Gene 395 (1), 49–61.

Rajkumar, H., Ramagoni, R.K., et al., 2015. De novo transcriptome analysis of *Allium cepa* L.(onion) bulb to identify allergens and epitopes. PLoS One 10 (8), e0135387.

Rozen, S., Skaletsky, H., 1996. Primer3: a software component for picking PCR primers. Source code available online at http://www-genome.wi.mit.edu/genome_software/other/primer3.html (verified 7 June 2004).

Shigyo, M., Miyazaki, T., et al., 1997. Assignment of randomly amplified polymorphic DNA markers to all chromosomes of shallot (*Allium cepa* L. Aggregatum group). Genes Genet. Syst. 72 (4), 249–252.

Sinclair, P.J., Blakeney, A.B., et al., 1995. Relationships between bulb dry matter content, soluble solids concentration and non-structural carbohydrate composition in the onion (Allium cepa). J. Sci. Food Agric. 69 (2), 203–209.

Song, J., Xu, Y., et al., 2005. SNPsFinder—a web-based application for genome-wide discovery of single nucleotide polymorphisms in microbial genomes. Bioinformatics 21 (9), 2083–2084.

The International Barley Genome Sequencing Consortium, 2012. A physical, genetic and functional sequence assembly of the barley genome. Nat. 491, 711–716.

Thomas, L., Leung, S., et al., 2011. Genotypic variation in sulphur assimilation and metabolism of onion (*Allium cepa L.*). II: Characterisation of ATP sulphurylase activity. Phytochemistry 72 (9), 888–896.

van Heusden, A., Shigyo, M., et al., 2000a. AFLP linkage group assignment to the chromosomes of *Allium cepa* L. via monosomic addition lines. Theor. Appl. Genet. 100 (3–4), 480–486.

van Heusden, A., Van Ooijen, J., et al., 2000b. A genetic map of an interspecific cross in Allium based on amplified fragment length polymorphism (AFLPTM) markers. Theor. Appl. Genet. 100 (1), 118–126.

Varshney, R.K., Graner, A., et al., 2005. Genic microsatellite markers in plants: features and applications. Trends Biotechnol. 23 (1), 48–55.

Xie, F.L., Huang, S.Q., et al., 2007. Computational identification of novel microRNAs and targets in *Brassica napus*. FEBS Lett. 581 (7), 1464–1474.

Yaguchi, S., McCallum, J., et al., 2008. Biochemical and genetic analysis of carbohydrate accumulation in *Allium cepa* L. Plant Cell Physiol. 49 (5), 730–739.

Zdobnov, E.M., Apweiler, R., 2001. InterProScan—an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17 (9), 847–848.

Zhang, B.H., Xiao Ping, P., et al., 2005. Identification and characterization of new plant microRNAs using EST analysis. Cell Res. 15 (5), 336–360.

Zhang, B., Wang, Q., et al., 2007. Identification of cotton microRNAs and their targets. Gene 397 (1), 26–37.