PLOS | ONE

RESEARCH ARTICLE

De Novo Transcriptome Analysis of *Allium cepa L.* (Onion) Bulb to Identify Allergens and Epitopes

Hemalatha Rajkumar[1]*, Ramesh Kumar Ramagoni[1], Vijayendra Chary Anchoju[1], Raju Naik Vankudavath[2], Arshi Uz Zaman Syed[3]

1 Department of Microbiology & Immunology, National Institute of Nutrition, Indian Council of Medical Research, Hyderabad, Telangana State, 500 007, India, **2** Biomedical Informatics Center, National Institute of Nutrition, Indian Council of Medical Research, Hyderabad, Telangana State, 500007, India, **3** Nucleome Informatics Pvt. Ltd., Hyderabad, Telangana State, 500 049, India

* rhemalathanin@gmail.com

Citation: Rajkumar H, Ramagoni RK, Anchoju VC, Vankudavath RN, Syed AUZ (2015) *De Novo* Transcriptome Analysis of *Allium cepa L.* (Onion) Bulb to Identify Allergens and Epitopes. PLoS ONE 10(8): e0135387. doi:10.1371/journal.pone.0135387

Editor: Turgay Unver, Cankiri Karatekin University, TURKEY

Received: March 30, 2015

Accepted: July 21, 2015

Published: August 18, 2015

Data Availability Statement: The raw sequence datasets of Allium cepa (onion) bulb were submitted to the primary NGS data archive, known as Sequence Read Archive (SRA) at NCBI (www.ncbi.nlm.nih.gov/sra). The BioProject accession number for the onion bulb transcriptome is PRJNA248253 and SRA experimental accession number is SRX542992. The filtered and assembled data was also submitted to an archive of computationally assembled sequences called TSA (Transcriptome Shotgun Assembly Sequence Database); accession number: 248253 TSA.

Abstract

Allium cepa (onion) is a diploid plant with one of the largest nuclear genomes among all diploids. Onion is an example of an under-researched crop which has a complex heterozygous genome. There are no allergenic proteins and genomic data available for onions. This study was conducted to establish a transcriptome catalogue of onion bulb that will enable us to study onion related genes involved in medicinal use and allergies. Transcriptome dataset generated from onion bulb using the Illumina HiSeq 2000 technology showed a total of 99,074,309 high quality raw reads (~20 Gb). Based on sequence homology onion genes were categorized into 49 different functional groups. Most of the genes however, were classified under 'unknown' in all three gene ontology categories. Of the categorized genes, 61.2% showed metabolic functions followed by cellular components such as binding, cellular processes; catalytic activity and cell part. With BLASTx top hit analysis, a total of 2,511 homologous allergenic sequences were found, which had 37–100% similarity with 46 different types of allergens existing in the database. From the 46 contigs or allergens, 521 B-cell linear epitopes were identified using BepiPred linear epitope prediction tool. This is the first comprehensive insight into the transcriptome of onion bulb tissue using the NGS technology, which can be used to map IgE epitopes and prediction of structures and functions of various proteins.

Introduction

Allium cepa (onion) is one of the most important vegetable crops in the *Allium* family known for its nutritional and medicinal properties [1], [2]. Onion perhaps is among the earliest cultivated vegetables and dates back to 5000 years [3]. Onions have antioxidant, anti-cholesterol anticancer properties and anti-inflammatory activitydue to phenolics compounds and flavonoids; but can induce symptoms of food allergy such as asthma, rhino conjunctivitis and

Funding: The authors acknowledge the Department of Biotechnology, Ministry of science and technology for funding the study (Grant no. BT/PR-11489/FNS/20/391/2008). The authors thank the National Institute of Nutrition, ICMR for supporting this work. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. Nucleome Informatics Pvt. Ltd., provided support in the form of a salary for author Syed Arshi Uz Zaman, but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific role of this author is articulated in the author contribution section.

Competing Interests: Syed Arshi Uz Zaman is employed by Nucleome Informatics Pvt. Ltd. The authors did not receive any financial support from Nucleome Informatics Pvt. Ltd, though it is a commercial company. The authors have taken only bioinformatics and technical support from that organization due to the unavailability of technical facilities in their organization. There are no patents, products in development or marketed products to declare. This does not alter the authors' adherence to all the PLOS ONE policies on sharing data and materials, as detailed online in the guide for authors.

contact dermatitis in sensitised individuals [4–7]. Diallyl disulphide is a major allergen in garlic and onion that is known to cause contact dermatitis [1], [8].

In food allergic individuals, exposure to allergenic epitopes from food proteins causes production of specific IgE antibodies that bind to the surface of tissue bound mast cells or circulating basophils, which ultimately releases inflammatory mediators. The actions of these mediators cause the clinical signs and symptoms of food allergy [9], [10]. Quite a few diagnostic techniques are available for the diagnosis of IgE mediated food allergy such as double-blind placebo-controlled food challenge (DBPCFC), skin prick test (SPT), specific IgE antibody test, case history etc. [11], [12]. However, diagnosis of food allergy is complicated with low or no detectable levels of specific IgE in patients with symptoms of IgE mediated food allergy [13], [14] and detection of specific IgE is not necessarily associate with clinical symptoms [15–17]. Novel diagnostic methods such as focusing on protein and epitope specificity are under investigation. In this scenario, epitopes on onion bulb allergens were predicted as onion is widely consumed and is known to cause IgE mediated allergy [1], [18–20].

A. *cepa* is a diploid (2n = 16) plant and has approximately 16.4 giga (billion) bases per 1C; it's nuclear genome is one of the largest nuclear genomes among all diploids and is over six times greater than maize or humans [21], [22]. Though onion has been used extensively in the past for cytogenetic studies, molecular analysis has been lacking.

Although modern sequencing and proteomic technologies permit the ready detection of numerous protein sequence variants in any organisms, for onions, there are no allergenic proteins and genomic data available. Onion is an example of an under-researched crop which has a complex heterozygous genome. Whereas genome based research has previously been hindered by limited sequence resources and allergenic protein information.

In recent years several studies have successfully reported the generation of transcriptome data and its analysis as an effective tool to study gene expression in specific tissues and also provide a platform to address comparative genomics for gene discovery in non-model plants in which no reference genome sequences are available [23–25]. Next-generation sequencing (NGS) technologies, such as Illumina, Genome Sequencer FLX system (GS FLX), and ABI SOLiD, allow analysis of the transcriptome because of increased throughput and reduced sequencing cost [25–29]. Due to the availability of quick, low cost sequencing and high quality annotation using different assembly tools, it has become possible to analyze and understand the transcriptome of onion plant [30].

This study describes the generation, *de novo* assembly and annotation of transcriptome dataset derived from cDNAs obtained from onion bulb using the Illumina HiSeq 2000 sequencing technology. The final assembly was functionally annotated, allowing for the identification of putative genes responsible for allergenicity and other similar sequences. Identification of these genes leads to testable hypotheses concerning their conserved function and to rational strategies to improve allergen free (genetically modified) onions.

Materials and Methods

Plant material

Fresh onion bulb (Pusa Madhavi light red onion variety) was collected from our institutional vegetable garden and the bulb was cut into small pieces in RNA later solution to extract RNA.

RNA isolation and quality control

Total RNA was isolated using the High Pure RNA Isolation Kit (Roche Life sciences,Basel, Switzerland) following the manufacturer's instructions. The integrity was checked with bioanalyzer (Agilent 2100 Technologies, USA) to yield high quality RNA for further processing.

The mRNA was extracted and purified from the total RNA by using TruSeq stranded mRNA HT sample preparation kit (Illumina Inc., U.S.A), followed by purity check with Qubit fluorometer before proceeding to cDNA synthesis.

Paired-end library preparation and Illumina sequencing

The cDNA was synthesized and libraries were prepared with the above-mentioned TruSeq Stranded mRNA HT Sample prep kit according to manufacturer's protocol. After PCR amplification the libraries were purified on agarose gel. The quality of the library was assessed by Qubit 2.0 fluorometer (Life technologies, USA.) and inserts with 150–250 bp were selected. Clustering was performed using TruSeq PE cluster kit v3-cBot-HS on cBot (Illumina Inc, USA). The samples were run (2x100) on Illumina HiSeq 2000 instrument using TruSeq SBS kit v3-HS (200 cycles) (Illumina Inc, USA) following manufacturer's instructions. The sequence data generated in this study have been deposited at NCBI, Sequence Read Archive (SRA) database under the accession number PRJNA248253 (BioProject) with experiment accession number SRX547958 for PE reads.

Raw data pre-processing and *de novo* transcriptome assembly

The sequenced raw data was processed to obtain high quality clean reads, we used the open-source software package Trimmomatic v0.32 to identify and to trim nucleotides falling below the established quality threshold (minimum 20 phred score) as well as to trim adapter sequences [31]. A minimum length of 50 nt (nucleotide) after trimming was applied. Orphaned reads were assigned as single-end reads. Onion bulb whole transcriptome assembly was carried out in three different stages. At first stage of assembly, the high quality clean reads were assembled by using Velvet (Version 1.2.10) followed by Oases (Version 0.2.8) to built k-mer specific transcripts from k-mer 27 to k-mer 63 (S1 Table), both Velvet and Oases are meant for assembly of short reads [32], [33]. The second stage of assembly was performed to obtain merged assembly of transcripts obtained from first stage i.e., from k-mer 27 to k-mer 63 using Velvet and Oases long read option. The third/final stage of assembly was performed by using CD-HIT which produces a set of 'non-redundant' (nr) onion bulb transcriptome catalogue [34]. The GC content analysis of onion transcripts was performed by using in-house perl scripts.

Similarity search and functional annotation

To deduce the putative function, onion transcripts were used as queries to search protein databases using BLASTx (Basic Local Alignment Search Tool) programme against the NCBI non-redundant (nr) green plant database, homology searches were performed with an e-value cut-off of 1E-06. Gene Ontology (GO) classification was carried out by using BLASTx results in Blast2GO software for functional classification of GO terms [35]. GO terms were obtained from nr hits using Blast2GO software with default parameters during mapping and an e-value cut-off of 1E-06 was used for BLAST hits in the BLAST2GO annotation step.

Identification of allergens and homologous sequences

For the identification of food and other allergens and their families represented in onion transcriptome, the onion transcripts were searched against the allergen data collected using BLASTx with an E-value cut-off of 1E-05 and the default parameters for BLASTx were used: Expected value is 10, word size is 3, scoring matrix is BLOSUM62, gap penalty existence is 11, gap penalty extension is 1 and "low complexity region" was not selected.

Total number of 553 allergen groups which has protein sequences were collected from the allergen data source, AllergenOnline, version 11, 1491 sequences (553 groups, 265 species) which is maintained by Food allergy research and resource programme (FARRP) allergen protein sequences, University of Nebraska–Lincoln [36]. For allergen identification contiguous 80 amino acid sequence length transcripts were separated. E values and identity percentages were evaluated to estimate potential cross-reactivity that may endow with an immunological target for IgE antibodies.

Identification of epitopes on allergenic proteins

Contigs that had shown best E-Values in the BLASTx top hits were considered for epitope prediction. Accelrys Discovery Studio Visualizer (version 4.0) was used to translate contigs in nucleotide sequences to protein sequences [37]. These translated protein sequences were submitted to antibody epitope prediction tool by "BepiPred linear epitope prediction" method [38]. Internally BepiPred use Parker hydrophilicity scale, Levitt secondary structure scale, and Hidden Markov Models [39–41]. The default values were maintained for threshold (0.350) and window size (7). The predicted epitope length with start and end positions were mentioned including residue scores.

Results

Sequencing and *de novo* transcriptome assembly to build onion transcriptome catalogue

A total of 99,074,309 high quality raw sequence reads (~20Gb) were generated using Illumina HiSeq 2000 technology (Table 1). The raw data filtration was carried out by using Trimmomatic v0.32 which resulted in high quality 83,046,820 paired-end (PE), 1,384,023 single-end forward and 93 single-end reverse reads (Table 2). After quality assessment and data filtering, the high quality clean reads were used for *de novo* assembly of the onion bulb transcriptome in three different stages. In the first stage of assembly, the high quality clean reads were assembled using Velvet/Oases assembler at different k-mer lengths (S1 Fig, S1 Table). At second stage of assembly, we merged the transcripts which were obtained from first stage of assembly into a consensus transcriptome using Velvet and Oases long read option (S2 Table). The non-redundant (nr) representative sequences were generated by the CD-HIT algorithm at the third and final stage of assembly (S2 Table), which resulted in 293,475 transcripts of 280,882,036 bp in size. On analyzing the final assembly only transcripts with more than 100 bp length were considered for functional annotation (S2 Fig). GC content (Guanine and Cytosine ratio) of the onion bulb transcriptome catalogue was determined using in-house perl script and the average GC content was found to be 38% (Table 2).

Table 1. Summary of raw reads of *Allium cepa* bulb transcripts.

Raw reads	99,074,309 (~20Gb)
Clean PE reads	83,046,820
Clean SE forward reads	1,384,023
Clean SE reverse reads	93
G+C%	43%

Gb, Giga bases; PE, paired end; SE, single end; G+C, Guanine+Cytocine; %, percentage

doi:10.1371/journal.pone.0135387.t001

Table 2. Statistics of non-redundant set of *Allium cepa* bulb transcripts obtained from final stage of assembly.

Total number of transcripts	293,475
G+C%	38%
Total transcriptome length	280,882,036 bp
Average transcript size	957.1 bp
Transcript N50	1,594 bp
Max. transcript size	12,638 bp

G+C, Guanine+Cytocine; %, percentage; Max, maximum; bp, base pairs

doi:10.1371/journal.pone.0135387.t002

Functional annotation and characterization of onion transcripts

The assembled onion bulb transcript set was analysed for similarity/sequence conservation against the NCBI green plant nr-database using BLASTx search. To define a significant hit, an E-value cut-off threshold of ≤1E-06 was considered. From the total of 293,475 onion transcripts, 15,434 transcripts were annotated by BLASTx search and 4,780 transcripts were mapped against nr database by using BLAST2GO tool. In addition, of the 293,475 transcripts, a total of 115,251 transcripts had unique matches, whereas 178,224 transcripts had no blast hits (S3 Fig). Overall, a total of 115,251 (40%) transcripts exhibited significant similarity with at least one of the predicted protein from sequenced plants in the database. Likewise, we analysed the sequence conservation of onion transcripts with proteomes of all sequenced plant species. From the 115,251 transcripts, the highest matching of onion transcripts was observed with *Vitis vinifera* (14%) followed by *Oryza sativa* (8%), *Theobroma cacao* (5.4%) and other species (Fig 1). However, only 1,086 (1%) sequences were found as onion specific.

Functional classification of onion transcriptome by GO

In order to assign putative functions, transcripts from onion were compared against the GenBank's nr-protein sequences of green plants and UniProt database using BLASTx algorithm. The associated hits were searched for their respective GO. Based on sequence homology *Allium cepa* genes were categorized into 49 different functional groups and were categorized as biological process, cellular component and molecular function (S4 Fig). Most of the genes however, were classified under 'unknown groups' in all three GO categories. Of the categorized genes, 61.2% showed metabolic functions followed by other functions and cellular components such as binding (58.5%), cellular processes (57.6%), catalytic activity (47.7%) and cell and cell part (45.6%).

Identification of allergens & epitopes of onion transcripts with allergen database

Onion allergen species distribution was identified with BLASTx top hits. Most of the onion transcripts were categorized under "unknown species" against allergen database. A total of 2,511 homologous allergenic sequences were found in onion transcriptome, which had 37 to 100% similarity with existing allergen database. A huge number of onion transcripts have shown homology to *Cryptomeria japonica* (Japanese cedar, 9%) allergens, followed by *Blattella Germania* (German cockroach, 7%) and *Corylus avellana* (Common hazel, 6%) (Fig 2). Of the 2,511 homologous allergenic sequences identified, when nucleotide bases of 240 or more (equal to 80 amino acid length or more) were considered for blast top hits, 682 transcripts from onion bulb were found to be similar to published allergen sequences, which were

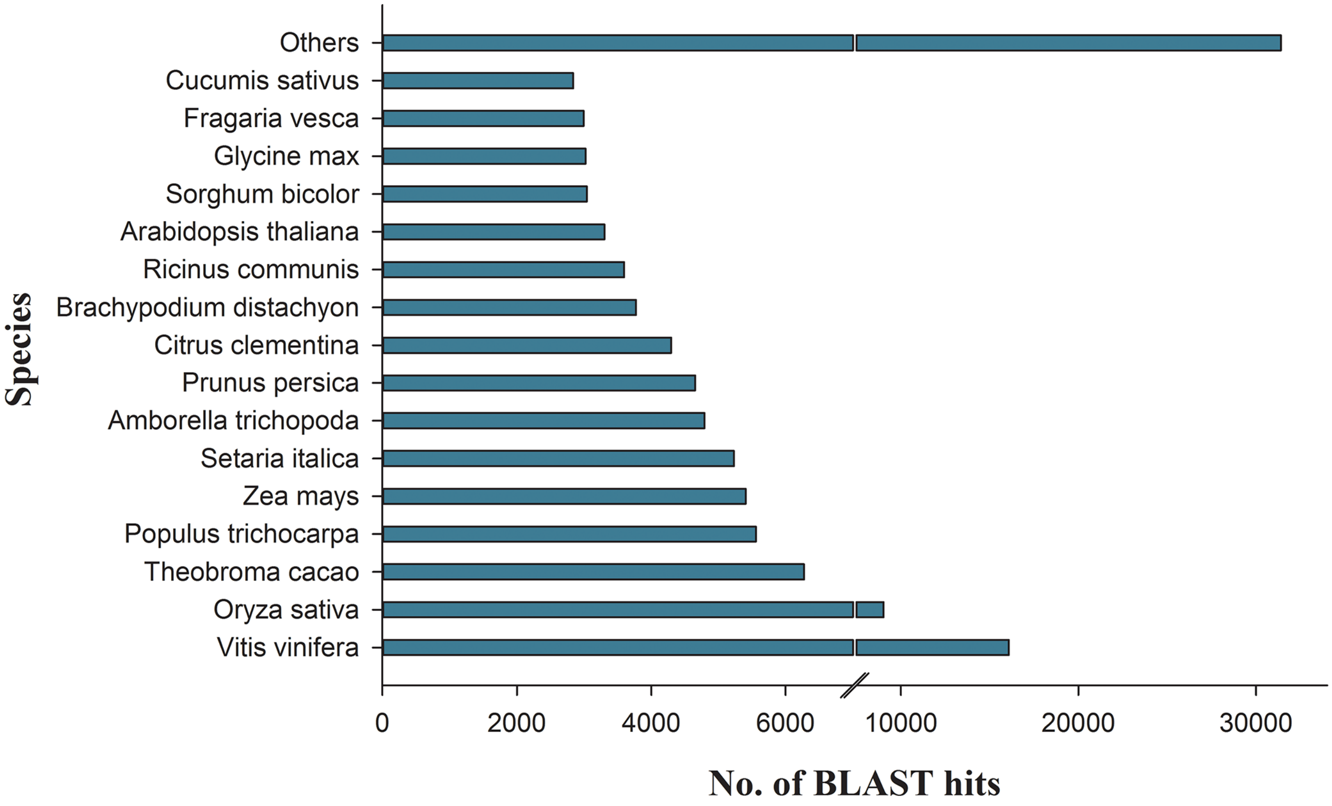

Fig 1. Annotations and BLAST top hits of onion transcripts with other species. This figure shows species distribution of onion transcripts with other species by BLASTx. Most of the onion sequences were homologous to *Vitis vinifer*a (14%) followed by *Oryza sativa* (8%) and *Theobroma caca*o (5.4%).

doi:10.1371/journal.pone.0135387.g001

altogether 46 different types. Thus, 46 allergens were identified from onion bulb whole transcriptome (Fig 3).

From the BLASTx results the onion transcripts contig 116485 showed E-value of '0', bit-score of 525.013 and homology of 75% with Enolase 1, 2-phospho-D-glycerate hydro-lyase. The onion contig 165540 showed E-value of 4.93E-169, bit-score of 483.411 and homology of 76% with ASP f 23 allergen (Table 3). From the 46 identified allergens (BLASTx results) epitopes were predicted using BepiPred linear epitope prediction tool. A total of 521 B cell linear epitopes were identified from 46 allergens of onion bulb when amino acid length of six or more was considered (S3 Table).

Discussion

Onion is one of the most important vegetable crops in the *Allium* family, known for its nutritional and medicinal properties. Vegetables of the genus *Allium* are widely consumed, but remain poorly understood genetically. *Allium* species are notable for their very large genomes, and typically ranges from10–20 Gbp, which have made it difficult for genomic studies and therefore have precluded genome sequencing to date [42].

Very few genomic resources including EST sequences and molecular markers are available for onion bulb compared to other vegetable species [43]. Transcriptome catalogue of onion bulb was generated for the first time, using which allergens and epitopes were identified in the current study. Transcriptome analysis is particularly useful for revealing relationships between plant gene expression and phenotypes. To build the transcriptome, the sequencing lengths of

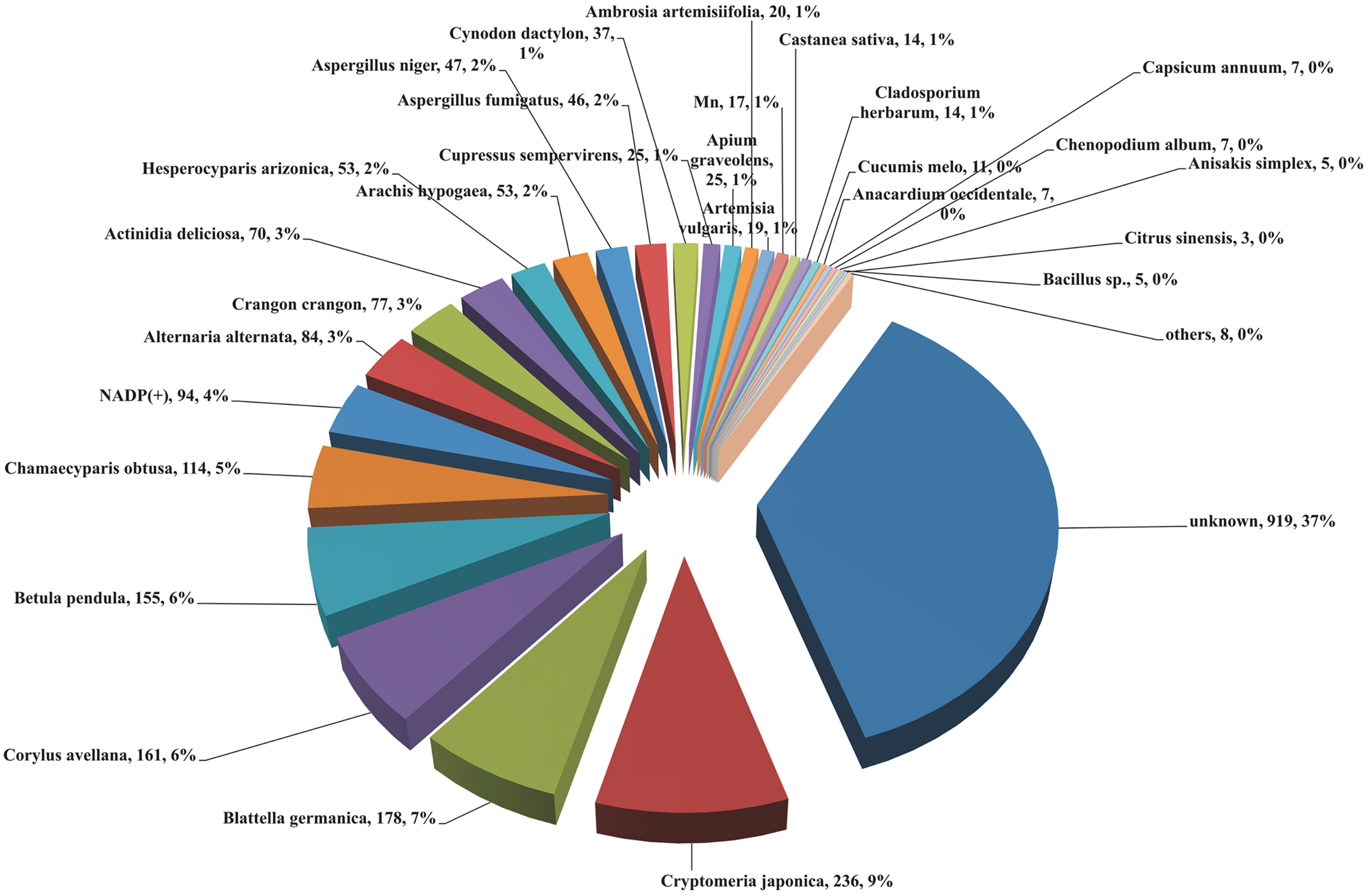

Fig 2. Onion allergen species distribution identified with BLASTx. The data represents number of transcripts and percentage of species distribution of onion bulb. For each species that was matching with onion, total number of hits is given along with relative percentage of homology. The comma (,) is separating the number of transcritpts from the % of species distribution. Highest number of onion transcripts have shown homology to *Cryptomeria japonica* allergens (9%), followed by *Blattella Germania* (7%) and *Corylus avellana* (6%).

doi:10.1371/journal.pone.0135387.g002

2×100 bp paired-end sequencing was opted using Illumina HiSeq 2000 V3 chemistry, which is suitable for analysis of 200–5,000-bp lengths. This system is the most efficient and accurate approach to demarcate the boundaries of transcription units of genes and complements other methods for transcriptome studies [44].

In this study, Velvet and Oases assemblers were used in the first two stages for assembly of short read sequences. This technique yielded long N50 lengths and therefore we could obtain better quality assembly of short reads. An earlier study by Rohini G et al. also showed the *de novo* assembly by Velvet followed by Oases programmes and validated these programs for better assembly [45]. In the final stage of assembly we had used CD-HIT algorithm to reduce the redundancy.

Nearly 100 million sequence reads were generated for onion bulb in the current study and 293,475 non-redundant set of transcripts were found (Tables 1 and 2). Only few transcripts were annotated, and complete information on gene functionality could be obtained for few genes. Interestingly, half of the amount of onion transcripts did not show significant homology with existing sequences, suggesting novel genes that may perform species specific functions in onion bulb; nevertheless, from the functional annotation a large number of onion transcripts

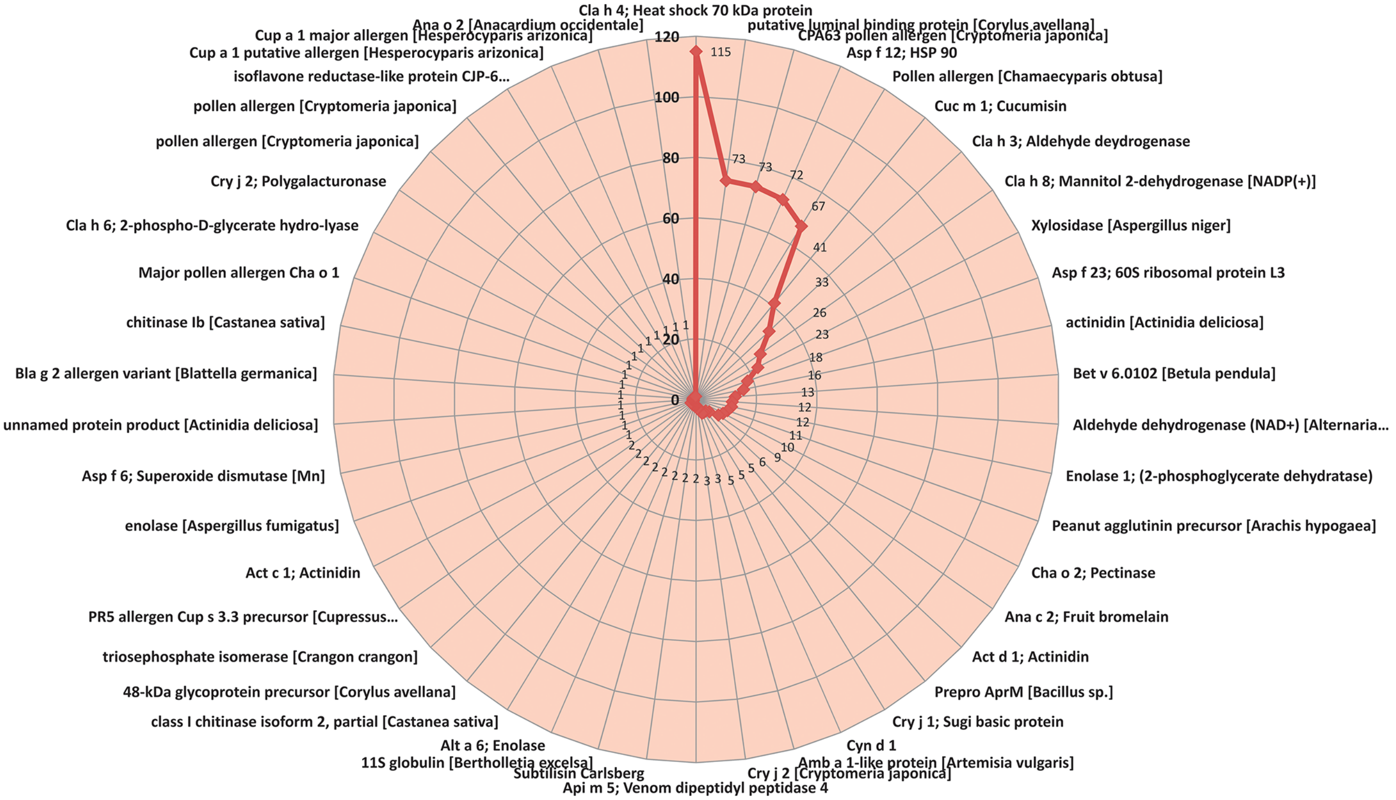

Fig 3. Onion allergens identified through transcriptome analysis. This radar graph depicts onion allergens identified through transcriptome analysis. Highest number of transcripts were homologous to Cla h4 (115) followed by putative luminal binding protein of *Corylus avellana* (73), CPA 63 pollen allergen *Cryptomeria japonica* (73), and Asp f12 of fungi (72). (Selected transcript sequence size: ≥80 amino acids).

doi:10.1371/journal.pone.0135387.g003

showed significant similarity with predicted proteins of plants. Examining a non-model vegetable plant such as onion can provide novel insights into the underlying mechanisms and diversity, however, to our surprise, only 1,086 (1%) sequences were identified as onion (*Allium cepa*) specific, thus signifying that limited information is available about the genomes or transcriptomes of onion and its related species. In the present study a large number of transcripts from *Allium cepa* were classified as unknown, suggesting the existence of massive potential for new gene identification.

Gene Ontology is an international standardized gene functional classification system that offers a strictly defined concept to comprehensively describe the properties of genes and gene products in any organism [35]. By using GO tool, we obtained a high number of onion bulb genes that were classified under either cell and cell part (cellular component) or metabolic processes (biological process) or binding activity indicating the dominance of genes that are responsible for cellular processes and metabolism, gene regulation and transcription factor binding processes (S4 Fig).

Base composition is a fundamental property of genomes and has a strong influence on various aspects related to gene function, regulation, structure (intron size and number), thermostability and species ecology [46–48]. However, the biological significance of GC content diversity in plants remains unclear because of a lack of sufficiently robust genomic data. The average GC content of the onion bulb transcripts in the present study was ~38%, which is very similar to what has been reported previously (Table 2) [49].

Table 3. The E-values and homology (%) of *Allium cepa* transcripts with matched allergen sequences.

Sequence Name	Allergen	Biochemical Function	Species	Common name	Type	Homology (%)	Seq. Len	Align. Len	E-Value	Bit-Score
Contig116485	N/A	Enolase 1, 2-phospho-D-glycerate hydro-lyase	N/A	N/A	N/A	75	1889	438	0	525.013
Contig165540	Asp f 23	60S ribosomal protein L3	Aspergillus fumigatus	A. Fumigatus	Fungi	76	1596	417	4.93E-169	483.411
Contig124037	Asp f 12	65 kDa IgE-binding protein; Heat shock protein hsp1	Aspergillus fumigatus	Fungi	Fungi	67	3117	680	1.46E-159	487.263
Contig147925	Alt a 11, Alt a 6, Alt a 5	Enolase; 2-phosphoglycerate dehydratase	Alternaria alternata	Fungi	Fungi	69	1980	431	9.71E-158	460.685
Contig147926	N/A	AF284645_1 enolase	Aspergillus fumigatus	Fungi	Fungi	69	1850	413	3.42E-149	437.187
Contig134658	Cuc m 1	Cucumisin, serine protease	Cucumis melo	Muskmelon	Plant	56	2397	735	9.67E-135	416.001
Contig121132	N/A	Xylosidase	Aspergillus niger	A. Niger	Fungi	53	2776	762	2.29E-133	418.313
Contig122480	N/A	Actinidin	Actinidia deliciosa	Fuzzy Kiwifruit	Plant	70	2423	360	2.09E-125	379.793
Contig192886	Cup a 1	Putative allergen Cup a 1	Hesperocyparis arizonica or Cupressus arizonica	Arizona cypress	Plant	63	1515	355	1.13E-111	334.339
Contig149472	N/A	Chitinase Ib	Castanea sativa	Sweet chestnut	Plant	74	1181	306	2.04E-111	327.791
Contig181997	N/A	Amb a 1-like protein	Artemisia vulgaris	Mugwort	Plant	63	1798	370	3.06E-110	334.724
Contig147924	Cla h 6	Enolase; 2-phospho-D-glycerate hydro-lyase; 2-phosphoglycerate dehydratase	Cladosporium herbarum	Fungi	Fungi	69	1739	303	1.22E-108	331.643
Contig185872	Cry j 1	Sugi basic protein	Cryptomeria japonica	Japanese cedar	Plant	61	1653	361	9.33E-108	326.25
Contig213065	N/A	Unnamed protein product	Actinidia deliciosa	Fuzzy Kiwifruit	Plant	67	1172	319	1.39E-107	320.087
Contig239367	N/A	11S globulin	Bertholletia excelsa	Brazil nut	Plant	61	1679	449	1.21E-104	321.627
Contig105897	Cha o 1	Major pollen allergen Cha o 1	Chamaecyparis obtusa	Japanese cypress	Plant	65	1915	355	3.22E-102	314.309
Contig87206	Act d 1	Actinidain	Actinidia deliciosa	Kiwi	Plant	66	1513	329	6.81E-102	309.686
Contig165488	Ana c 2	Fruit bromelain	Ananas comosus	Pineapple	Plant	65	1227	333	7.57E-100	300.056
Contig82715	Bet v 6.0102	Allergenic isoflavone reductase-like protein Bet v 6.0102	Betula pendula	Silver birch	Plant	68	1364	306	4.57E-99	298.13
Contig134058	CJP-6	Isoflavone reductase-like protein CJP-6	Cryptomeria japonica	Japanese cedar	Plant	64	1279	312	7.41E-97	291.197
Contig88117	N/A	Triosephosphate isomerase	Crangon crangon	Common shrimp	Crustacean	70	2141	248	3.05E-89	278.1
Contig97550	Cha o 2	Polygalacturonase; Major pollen allergen Cha o 2; Pectinase	Chamaecyparis obtusa	Japanese cypress	Plant	53	1956	466	1.48E-85	275.789
Contig181817	Cry j 2	Allergen Cry j 2	Cryptomeria japonica	Japanese cedar	Plant	60	1769	358	2.34E-84	270.781
Contig70943	Cla h 4	Heat shock 70 kDa protein	Cladosporium herbarum	Fungi	Fungi	48	3004	645	5.05E-84	281.952

(*Continued*)

Table 3. (*Continued*)

Sequence Name	Allergen	Biochemical Function	Species	Common name	Type	Homology (%)	Seq. Len	Align. Len	E-Value	Bit-Score
Contig116658	Cla h 10	Aldehyde dehydrogenase; Cla h 3	Cladosporium herbarum	Fungi	Fungi	50	2328	483	5.37E-60	207.608
Contig141069	N/A	Pollen allergen	Cryptomeria japonica	Japanese cedar	Plant	47	1672	313	2.23E-59	155.221
Contig128098	N/A	Class I chitinase isoform 2	Castanea sativa	Sweet chestnut	Plant	56	1528	247	7.32E-58	191.815
Contig116297	Act c 1	Actinidain	Actinidia deliciosa	Kiwi	Plant	50	1678	323	6.34E-55	187.193
Contig16398	Cup s 3.3 precursor	PR5 allergen Cup s 3.3 precursor	Cupressus sempervirens	Pencil pine or Mediterranean cypress	Plant	57	1445	241	2.31E-54	179.489
Contig122189	Asp f 6	Superoxide dismutase [Mn]	Aspergillus fumigatus	Fungi	Fungi	50	1281	241	3.89E-53	174.481
Contig86874	N/A	Aldehyde dehydrogenase (NAD+)	Alternaria alternata	Fungi	Fungi	45	2257	462	9.06E-53	186.422
Contig70938	N/A	Putative luminal binding protein	Corylus avellana	Common hazel	Plant	44	4214	721	4.39E-52	191.045
Contig152875	CPA63	Pollen allergen CPA63	Cryptomeria japonica	Japanese cedar	Plant	48	1590	437	6.92E-52	180.644
Contig175155	Api m 5	Venom dipeptidyl peptidase 4	Apis mellifera (Honeybee)	Honeybee	Insect	43	2585	622	8.42E-42	158.303
Contig176760	Cup a 1	Major allergen Cup a 1	Hesperocyparis arizonica or Cupressus arizonica	Arizona cypress	Plant	49	1577	253	2.10E-39	142.124
Contig198455	N/A	48-kDa glycoprotein precursor	Corylus avellana	Common hazel	Plant	44	1792	400	3.56E-36	135.191
Contig122991	Ana o 2	Legumin-like protein	Anacardium occidentale	Cashew tree	Plant	42	1378	401	2.51E-31	119.398
Contig178413	Cyn d 1	Major pollen allergen Cyn d 1	Cynodon dactylon	Bermuda grass	Plant	47	1150	242	7.30E-26	99.3673
Contig154078	N/A	Pollen allergen	Chamaecyparis obtusa	Japanese cypress	Plant	40	1895	435	9.82E-26	104.375
Contig181406	N/A	Subtilisin Carlsberg	Bacillus sp.	Bacillus	Bacteria	44	3643	246	1.49E-22	95.5153
Contig163717	N/A	Pollen allergen	Cryptomeria japonica	Japanese cedar	Plant	44	1663	260	4.98E-20	87.4261
Contig140011	N/A	Peanut agglutinin precursor	Arachis hypogaea	Peanut	Plant	48	2216	273	8.64E-19	81.6481
Contig155816	Bla g 2 allergen	Bla g 2 allergen variant	Blattella germanica	German cockroach	Insect	46	1842	240	3.19E-16	74.3294
Contig189186	Cla h 8	Mannitol 2-dehydrogenase [NADP(+)]	Cladosporium herbarum	Fungi	Fungi	43	1355	278	3.56E-15	69.3218
Contig157948	Cry j 2	Major pollen allergen; Polygalacturonase; Pectinase	Cryptomeria japonica	Japanese cedar	Plant	41	1014	249	1.47E-14	68.1662
Contig144878	Apr M	Prepro Apr M	Bacillus sp.	Bacillus	Bacteria	43	4317	253	1.08E-09	55.8398

Contig, Contiguous sequence; Seq. Len, Sequence length; Align. Len, Alignment length; %, percentage; E-value, Expect value; NA, not applicable

doi:10.1371/journal.pone.0135387.t003

Whole trascriptome sequencing may change the ways in which gene expression is studied, which is likely to have much future impact. The growing number of available allergen sequences together with the advancements of bioinformatics tools and methods will enable us to shed light on evolutionary and structural relationships between allergens from different sources [50].

Using bioinformatics approach we compared the onion bulb transcripts with existing allergen sequences in databases such as AllergenOnlline, FARRP, Allergome and Structural database for allergenic proteins (SDAP) to predict potential cross reactivity of onion bulb transcriptome sequences with available database [51], [52]. In our transcriptome data we have found more than 2,500 homologous sequences to 553 existing allergen groups containing 1,491 sequences of 265 different species. According to guidelines of Food and Agriculture Organisation/World health organization (FAO/WHO) and Codex (2003) the sequences which show more than 35% identity against amino acid sequences of expressed proteins (allergens) are considered as cross reactive allergens (>80 amino acid transcript sequences) [53]. By adopting these guidelines we have identified 682 homologous sequences and 46 different allergens in onion bulb. Transcript sequences generated from this study can be used to map epitopes of monoclonal antibodies and polyclonal sera from patients. With the support of total transcriptome of onion, the complete list of genes can be predicted based on which unknown protein structures may be modelled and novel diagnostic approach in food allergy and immunotherapeutic can be developed. Compared to the previously described random peptide libraries, the transcriptome RNA gene sequencing as performed in the current study offers good approach to identify epitopes. More than 500 B cell linear epitopes were identified from onion bulb when amino acid length of six or more was considered.

Conclusion

From the 99,074,309 high quality raw reads that were generated from onion bulb, 49 different functional groups of genes were identified. With BLASTx top hit analysis, a total of 2,511 homologous allergenic sequences were found, which had 37 to 100% similarity with 46 different types of allergens existing in the database. From the 46allergens, more than 500 B-cell linear epitopes were identified. This is the first comprehensive insight into the transcriptome of onion bulb tissue using the NGS technology, which can be used to map IgE epitopes and prediction of structures and functions of various proteins.

Supporting Information

S1 Fig. *De novo* assembly of onion bulb transcriptome. This figure represents the *De novo* assembly of onion bulb transcriptome generated by Velvet/Oases (1st and 2nd stage) and by CD-HIT (3rd stage). The bars indicate number of contigs (100 bp or longer). The lines indicate N50 length in bp (light green colour line with triangles) and average contig length (dark red colour line with rectangles). The left Y- axis indicates number of contigs and the right Y- axis indicates length in bp.
(TIF)

S2 Fig. *De novo* assembly length distribution. Frequency histogram showing number of transcripts as function of onion assembly read length distribution. The highest sequence length was 12,635 bp and only transcripts with ≥100 bp length sequences were considered for functional annotation.
(TIF)

S3 Fig. Onion BLASTx statistics. This figure shows the BLAST hits results including annotations and mapping.
(TIF)

S4 Fig. Histogram of onion Gene Ontology classification. GO terms obtained and classified into three major groups as Cellular Component, Molecular Function and Biological Process. Most of the onion GO terms observed as cell and cell part, metabolic processes, catalytic and binding processes.
(TIF)

S1 Table. Stage one onion (*Allium cepa L.*) bulb transcriptome assembly statistics. In stage 1 assembly Velvet/Oases tools were used to assemble high quality reads to generate kmer specific transcripts.
(XLS)

S2 Table. Assembly statistics of onion (*Allium cepa* L.) bulb transcriptome using Velvet/Oasis and CD-Hit tools. This table represents the statistics of assembled transcripts at second and third stages of assemblies.
(DOC)

S3 Table. The allergenic peptide or epitope prediction. The 46 identified allergens (BLASTx results) from onion bulb showed a total of 521 B cell linear epitopes when amino acid length of 6 or more was considered.
(XLS)

Acknowledgments

We thank the National Institute of Nutrition, ICMR for supporting this work.

Author Contributions

Conceived and designed the experiments: HR. Performed the experiments: RKR. Analyzed the data: RKR AUZS VCA RNV HR. Contributed reagents/materials/analysis tools: RKR VCA. Wrote the paper: HR RKR. Read and approved the final manuscript: HR RKR VCA RNV AUZS.

References

1. Bleumink E, Doeglas HM, Klokke AH, Nater JP (1972) Allergic contact dermatitis to garlic. Br J Dermatol 87: 6–9. doi: 10.1016/j.adengl.2012.07.006 PMID: 5043200
2. Gonzalez CA, Pera G, Agudo A, Bueno-de-Mesquita HB, Ceroti M, Boeing H, et al. (2006) Fruit and vegetable intake and the risk of stomach and oesophagus adeno carcinoma in the European Prospective Investigation into Cancer and Nutrition (EPIC-EURGAST). Int J Cancer 118: 2559–2566. doi: 10.1002/ijc.21678 PMID: 16380980
3. All about onions. National Onion Association. Retrieved 2015-01-20. Available at http://www.onions-usa.org/all-about-onions/nutritional-facts-vegetables.
4. Eric Block (2010) "Garlic and Other *Alliums*: The Lore and the Science". Cambridge: Royal Society of Chemistry.
5. Arochena L, Gámez C, Del Pozo V, Fernández-Nieto M (2012) "Cutaneous allergy at the supermarket". J Investig Allergol Clin Immunol 22: 441–442. PMID: 23101191
6. Moneret-Vautrin DA, Morisset M, Lemerdy P, Croizier A, Kanny G (2002) Food allergy and IgE sensitization caused by spices: CICBAA data (based on 589 cases of food allergy). Allerg Immunol (Paris) 34:135–140.
7. Valdivieso R, Subiza J, Varela-Losada S, Subiza JL, Narganes MJ, Martinez- Cocera C, et al. (1994) Bronchial asthma, rhinoconjunctivitis, and contact dermatitis caused by onion. J Allergy Clin Immunol 94: 928–930. doi: http://dx.doi.org/10.1016/0091-6749(94)90161-9 PMID: 7963160
8. Moyle M, Frowen K, Nixon R (2004) Use of gloves in protection from diallyl disulphide allergy. Australas J Dermatol 45: 223–225. PMID: 15527433
9. Sampson HA (2004) Update on food allergy. J Allergy Clin Immunol 113: 805–819. PMID: 15131561

10. Vijayendra CA, Hemalatha R, Narendra BK, Ramesh KR, Sathavahana C, Dinesh KB, et al,. (2012) CD23, Total IgE and Th1/Th2 Cytokines in Asthma Patients. Int J Pharm Pharm Sci 4: 175–177.

11. Sicherer SH, Sampson HA (2010) Food allergy. J Allergy Clin Immunol 125: S116–S125. doi: 10.1016/j.jaci.2009.08.028 PMID: 20042231

12. Bruijnzeel-Koomen C, Ortolani C, Aas K, Bindslev-Jensen C, Bjorksten B, Moneret Vautrin D, et al. (1995) Adverse reactions to food. European Academy of Allergology and Clinical Immunology Subcommittee. Allergy 50: 623–635. PMID: 7503398

13. Shek LP, Bardina L, Castro R, Sampson HA, Beyer K (2005) Humoral and cellular responses to cow milk proteins in patients with milk-induced IgE-mediated and non-IgE-mediated disorders. Allergy 60: 912–919. PMID: 15932382

14. Host A, Halken S (1990) A prospective study of cow milk allergy in Danish infants during the first 3 years of life. Clinical course in relation to clinical and immunological type of hypersensitivity reaction. Allergy 45: 587–596. PMID: 2288394

15. Asero R, Ballmer-Weber BK, Beyer K, Conti A, Dubakiene R, Fernandez- Rivas M, et al. (2007) IgE-mediated food allergy diagnosis: Current status and new perspectives. Mol Nutr Food Res 51: 135–147. PMID: 17195271

16. Hamilton RG, Macglashan DW Jr., Saini SS (2010) IgE antibody-specific activity in human allergic disease. Immunol Res 47: 273–284. doi: 10.1007/s12026-009-8160-3 PMID: 20066506

17. Mari A (2002) IgE to cross-reactive carbohydrate determinants: analysis of the distribution and appraisal of the in vivo and in vitro reactivity. Int Arch Allergy Immunol 129: 286–295. PMID: 12483033

18. Enrique E, Malek T, De Mateo JA, Castello J, Lombardero M, Barber D, etal (2007) Involvement of lipid transfer protein in onion allergy. Ann Allergy Asthma Immunol 98: 202 PMID: 17304894

19. Arena A, Cislaghi C, Falagiani P (2000) Anaphylactic reaction to the ingestion of raw onion. A case report. Allergol Immunopathol (Madr) 28: 287–9

20. Asero R, Mistrello G, Roncarolo D, Amato S (2001) A case of onion allergy. J Allergy Clin Immunol 108: 309–10

21. Arumuganathan K, Earle ED (1991) Nuclear DNA content of some important plant species. Plant Mol Biol Rep 9: 208–218.

22. McCallum J, Samantha B, Masayoshi S, Yanbo D, Sjaak van H, Pither-joyce M, et.al. (2012) Allium-Map-A comparative genomics resource for cultivated *Allium* vegetables. BMC Genomics 13: 168. doi: 10.1186/1471-2164-13-168 PMID: 22559261

23. Parkinson J, Blaxter M (2009) Expressed sequence tags: An overview. Methods Mol Biol 533: 1–12. doi: 10.1007/978-1-60327-136-3_1 PMID: 19277571

24. Toledo-Silva G, Cardoso-Silva CB, Jank L, Souza AP (2013) *De novo* transcriptome assembly for the tropical grass *Panicum maximum Jacq*. PLoS ONE 8: e70781. doi: 10.1371/journal.pone.0070781 PMID: 23923022

25. Mine T, Kuaybe YK, Gabriel D, Baohang Z, Pilar H, Turgay U (2014) Sequencing of plant genomes–a review. Turk J Agric For 38: 1–16. doi: 10.3906/tar-1409-93

26. Nagalakshmi U, Wang Z, Waern K (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320: 1344–1349. doi: 10.1126/science.1158441 PMID: 18451266

27. Mardis ER (2008) Next-generation DNA sequencing methods, Annu. Rev. Genomics Hum Genet 9: 387–402.

28. Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotechnol 26: 1135–1145. doi: 10.1038/nbt1486 PMID: 18846087

29. Rowan BA, Weigel D, Koenig D (2011) Developmental genetics and new sequencing technologies: the rise of nonmodel organisms. Dev Cell 21: 65–76. doi: 10.1016/j.devcel.2011.05.021 PMID: 21763609

30. Collins LJ, Biggs PJ, Voelckel C, Joly S (2008) An approach to transcriptome analysis of non-model organisms using short-read sequences. Genome Inform 21: 3–14. PMID: 19425143

31. Anthony MB, Marc L, Bjoern U (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114–2120. doi: 10.1093/bioinformatics/btu170 PMID: 24695404

32. Daniel RZ, Velvet Ewan B. (2008) Algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Res 18: 821–829. doi: 10.1101/gr.074492.107 PMID: 18349386

33. Schulz MH, Zerbino DR, Vingron M, Birney E (2012) Oases: Robust *de novo* RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28: 1086–1092. doi: 10.1093/bioinformatics/bts094 PMID: 22368243

34. Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28: 3150–3152. doi: 10.1093/bioinformatics/bts565 PMID: 23060610

35. The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. Nat Genet 25: 25–29. PMID: 10802651

36. AllergenOnline database (version 11), allows online search of peer-reviewed allergen list available at http://www.allergenonline.org.

37. Accelrys Software Inc., (2013) Discovery Studio Modeling Environment, Release 4.0, San Diego: Accelrys Software Inc.

38. BepiPred Linear Epitope Prediction tool of Immune epitope database and analysis resource, available at http://tools.immuneepitope.org/bcell/

39. Parker JM, Guo D, Hodges RS (1986) New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites. Biochemistry 25: 5425–5432. PMID: 2430611

40. Levitt M (1978) Conformational preferences of amino acids in globular proteins. Biochemistry 17: 4277–4285. PMID: 708713

41. Larsen JE, Lund O, Nielsen M (2006) Improved method for predicting linear B-cell epitopes. Immunome Res 2: 2. PMID: 16635264

42. Ricroch A, Yockteng R, Brown S, Nadot S (2005) Evolution of genome size across some cultivated *Allium* species. Genome 48: 511–520. PMID: 16121247

43. Kuhl JC, Cheung F, Yuan Q, Martin W, Yayeh Z, McCallum J, 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*. The Plant Cell 16: 114–125. doi: 10.1105/tpc.017202 PMID: 14671025

44. Fullwood MJ, Wei CL, Liu ET, Ruan Y (2009) Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. Genome research 19: 521–532. doi: 10.1101/gr.074906.107 PMID: 19339662

45. Rohini G, Ravi PK, Akhilesh TK, Mukesh J (2011) *De novo* assembly of chickpea transcriptome using short reads for gene discovery and marker identification. DNA Res 18: 53–63. doi: 10.1093/dnares/dsq028 PMID: 21217129

46. Petr Š, Petr B, Lucie H, Ilia JL, Ladislav M, Ettore P, et.al. (2014) Ecological and evolutionary significance of genomic GC content diversity in monocots. Proc Natl Acad Sci 111: E4096–4102. doi: 10.1073/pnas.1321152111 PMID: 25225383

47. Li X-Q, Du D (2014) Variation, evolution, and correlation analysis of C+G content and genome or chromosome size in different kingdoms and phyla. PLoS ONE 9: e88339. doi: 10.1371/journal.pone.0088339 PMID: 24551092

48. Vinagradow AE (2003) DNA helix; the importance of being GC-rich. Nucleic Acids Res 31: 1838–1844. doi: 10.1093/nar/gkg296 PMID: 12654999

49. Chinnappareddy LRD, Khandagale K, Chennareddy A, Ramappa VG (2013) Molecular markers in the improvement of *Allium* crops. Czech J Genet Plant Breed 49: 131–139.

50. Radauer C, Breiteneder H (2007) Evolutionary biology of plant food allergens. J Allergy Clin Immunol 120: 518–525. doi: 10.1016/j.jaci.2007.07.024 PMID: 17689599

51. Mari A, Rasi C, Palazzo P, Scala E (2009) Allergen databases: current status and perspectives. Curr Allergy Asthma Rep 9: 376–383. PMID: 19671381

52. Ivanciuc O, Schein CH, Braun W (2003) SDAP: Database and Computational Tools for Allergenic Proteins. Nucleic Acids Res 31: 359–362. doi: 10.1093/nar/gkg010 PMID: 12520022

53. Report of a Joint FAO/WHO Expert Consultation on Allergenicity of Foods Derived from Biotechnology; Evaluation of Allergenicity of Genetically Modified Foods 2001.