De novo transcriptomic analyses to identify and compare allergens in foods

Shaymaviswanathan Karnaneedi[1-3], Vachiranee Limviphuvadh[4, 5], Sebastian Maurer-Stroh[4-6], and Andreas L. Lopata[1-3,7]

[1] Molecular Allergy Research Laboratory, College of Public Health, Medical and Veterinary Sciences, James Cook University, Townsville, Queensland, Australia;

[2] Australian Institute of Tropical Health and Medicine, James Cook University, Townsville, QLD, Australia;

[3] Centre for Food and Allergy Research, Murdoch Children's Research Institute, Melbourne, VIC, Australia;

[4] Biomolecular Function Discovery Division, Bioinformatics Institute, Agency for Science, Technology and Research, Singapore;

[5] IFCS Programme, Singapore Institute for Food and Biotechnology Innovation, Agency for Science, Technology and Research, Singapore;

[6] Department of Biological Sciences, National University of Singapore, Singapore.

[7] Tropical Futures Institute, James Cook University Singapore, Singapore;

SK: shaymaviswanathan.karnaneedi@my.jcu.edu.au

VL: vachiraneel@bii.a-star.edu.sg

SM: sebastianms@bii.a-star.edu.sg

AL: andreas.lopata@jcu.edu.au

***Corresponding author:**

Shaymaviswanathan Karnaneedi

Email: shaymaviswanathan.karnaneedi@my.jcu.edu.au

Running Head: Transcriptomic analyses to identify new food allergens

Abstract

Food allergens have been traditionally identified using biomolecular and immunological approaches. However, the techniques used in extracting proteins from the food source to be analysed may hinder the availability of all proteins when assessing immunological allergenicity. Additionally, depending on the number and pool of patient sera used to detect the IgE antibody-binding allergens, some allergens may not be detected if not all the patients in the pool are sensitised to all the allergens.. To overcome these limitations, we describe an additional approach before the *in vitro* approaches, by analysing the transcriptome *in silico* for all putative allergens within the analysed food source.

Key words: Food allergen, allergen identification, homologous allergen, transcriptomic, bioinformatics, transcriptome assembly, RNA-Seq, allergen database, BLAST search, Tropomyosin.

1. Introduction

Allergens in foods have been traditionally identified and characterised using bimolecular and immunological methods, which include protein extraction from a particular food source, protein profiling, and immunoblot technique applied on the protein extracts using serum immunoglobin E (IgE) antibodies from individuals with known food allergy to the analysed food source [1, 2]. This traditional approach that has been used for decades, however, has several limitations that are often overlooked. The biggest limiting factor in using the biomolecular and immunological approach is the limitation posed by the extraction methods [3]. Depending on the protocol and extraction buffer used, some proteins with low solubility, high isoelectric point, or low abundance may be missed during the extraction, and these proteins' potential allergenicity would not be included in the profile of allergenic proteins [4]. Additionally, the use of a small number of patients may hinder the possibility of detecting all possible allergens within the food source. Allergic patients are often not sensitised to all allergens of a particular food source, therefore, some allergens may be missed depending on the number and pool of patient sera used for identifying the allergens [3].

Whilst the traditional allergen identification pipeline involving biomolecular and immunological methods is crucial in identifying the potential allergenicity of new putative allergens, we suggest the inclusion of an additional approach before using the traditional methods, to ensure that all potentially allergenic proteins are included in the allergen identification methods.

Studies have found that homologous allergens (allergens that are the same protein from different species) have high amino acid (AA) sequence identity [5]. Therefore, analysing and comparing all proteins within the subject/analysed organism with all known and registered allergens from various food sources will identify all known as well as putative allergens in the

food source to be analysed [5, 6]. However, to obtain the proteome of any organism, extraction processes still play a major role in limiting the proteins that are extracted. Therefore, we suggest the use of the transcriptome (complete messenger RNA sequences) of an organism that will be translated to proteins downstream [7].

In this methods chapter, we describe how to assemble a transcriptome *de novo* from an organism to be analysed and identify new putative allergens within the organism using AA sequence similarity search with all known and registered allergens to date. Our approach will identify putative allergens that can be ensured to enter the traditional biomolecular and immunological approach of allergen identification. For example, if a putative allergen identified within the transcriptomic analysis is missed in the extraction process, a different extraction method could be utilised or a recombinant form of the allergen could be generated to assess immunological allergenicity. In this chapter, we detail an example study which analyses the transcriptomes of five shrimp species and identifies new and putative allergens within shrimps [7]. Whilst ten allergenic proteins have been identified in shrimps [8], this study aimed to identify the complete allergen repertoire (allergome) of shrimps by especially focusing on putative allergens that are often missed in the biomolecular and immunological methods.

2. Materials

2.1. Bioinformatics tools

1. High-performance computing (HPC) resources and a command line interface (CLI) terminal to run command scripts for the bioinformatics tools used. This requires a computer with:

a. MAC OS with 64 logical processors and 32 GB RAM using the embedded terminal.
b. Windows 10 operating system with 64 logical processors and 32 GB RAM using PuTTY (https://www.putty.org/), a terminal emulator.

2. RNA-Seq data of organism(s) to be analysed downloaded from public repositories such as National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/).
3. FASTA files downloaded from NCBI (https://www.ncbi.nlm.nih.gov/) and UniProt-KB (http://www.uniprot.org/).
4. BLAST tool downloaded from NCBI website (https://blast.ncbi.nlm.nih.gov/Blast.cgi).
5. Rcorrecter software downloaded from https://github.com/mourisl/Rcorrector
6. Trinity RNA-Seq *de novo* transcriptome assembler program downloaded from https://github.com/trinityrnaseq/trinityrnaseq
7. Transrate software downloaded from https://github.com/Blahah/transrate
8. BUSCO software downloaded from https://busco.ezlab.org/
9. Geneious bioinformatics software downloaded from https://www.geneious.com/
10. Microsoft Excel spreadsheet software from Microsoft 365, downloaded from https://www.microsoft.com/en-us/microsoft-365/excel

3. Methods

3.1. RNA-Seq data acquisition

1. Transcriptomic analyses usually require a transcriptome to be assembled *de novo* from an RNA-Seq dataset, as studies generally only submit their raw RNA-Seq datasets and not their assembled transcriptomes in public repositories (*see* **Note 1**).

2. RNA-Seq raw data of the organism to be analysed can be downloaded from the NCBI platform (https://www.ncbi.nlm.nih.gov/) within the Short Read Archive (SRA) database.
3. Alternatively, if there is a specific set of RNA-Seq data required from a particular project or target publication that has been deposited in the NCBI repository, the search within NCBI can be conducted by using the BioProject or BioSample code followed by downloading a sequence data from the SRA experiments (*see* **Note 2**).
4. Some model organisms will have multiple entries in the SRA database and any of the data in the SRA database can be selected to be analysed provided they do not include a modified or treated organism.
5. Ensure that it is an RNA-Seq dataset with no modulation or treatment done on the species analysed prior to obtaining the RNA-Seq data. This detail can be identified from the "Study:" field. Also, ensure that the RNA-Seq data are generated from Illumina (most reads in NCBI are Illumina reads) for processing by Rcorrecter (Section 3.2.)
6. Download the RNA-Seq data in FASTA format. The downloaded RNA-Seq file from NCBI SRA Database will be a zipped file (.gz) and therefore will need to be unzipped.
7. The RNA-Seq dataset should be in a FASTA format, and downstream methods require the use of bioinformatics tools/programs and thus require the use of command prompts in command-line interfaces (CLI) through a terminal.

3.2. RNA-Seq correction

1. Quality metrics of the RNA-Seq data are the major factor in determining the quality of the *de novo* assembled transcriptome.

2. However, despite the advances in sequencing technology, raw RNA-Seq datasets generally contain various random sequencing errors throughout the dataset, resulting from the experimental methods utilised to obtain RNA and the sequencing methods applied.
3. To remove random sequencing errors, Song and Florea (2015) developed a software known as "Rcorrecter" that utilises a *k*-mer-based method designed to correct random sequencing errors specifically in organisms with complex transcriptomes; and specifically for Illumina RNA-Seq reads [9].
4. Rcorrecter can be downloaded from https://github.com/mourisl/Rcorrector and run within the terminal on the previously downloaded RNA-Seq dataset, following the software designer's instructions. Downloading and usage instructions can be found at https://github.com/mourisl/Rcorrector and may vary as updates are introduced to the software.

3.3. *De Novo* Transcriptome assembly

1. Sequencing error-corrected RNA-Seq dataset can now be assembled into a transcriptome *de novo* using the Trinity RNA-Seq *de novo* transcriptome assembly program. The program utilises a combination of three software modules: "Inchworm", "Chrysalis" and "Butterfly" [10, 11].
2. The Trinity RNA-Seq *de novo* transcriptome assembly program can be downloaded from https://github.com/trinityrnaseq/trinityrnaseq and run within the terminal following the software designer's instructions. Various updates occur and therefore, please follow the instructions specific to the downloaded version.

a. The Trinity program will begin running the embedded modules. Firstly, the "Inchworm" uses another embedded software called "Jellyfish" to extract and count *k*-mers and then proceeds to efficiently assemble the short RNA reads by choosing the most abundant *k*-mers to extend the sequences into linear contigs [10].

b. The next module, "Chrysalis", looks for overlaps among the contigs and integrates the linear contigs to generate a de-Bruijn graph for each cluster. Unlike "Inchworm", "Chrysalis" takes into account alternative variants of contigs and places them within the same cluster [10].

c. The third and final module, "Butterfly", resolves the de-Bruijn graph and provides a set of contigs/transcripts, which makes up the transcriptome (the sum total of all sequences of mRNA molecules expressed) [10].

3. In the example of Table 1, the assembly of five different shrimp species resulted in transcriptomes with numbers of contigs ranging from 28,101 to 42,510; and assembly sizes ranging between 21.7Mb to 38.1Mb (Table 1).

Table 1: Trinity transcriptome assembly metrics, TransRate score, and BUSCO scores (C: complete, F: fragmented, M: missing) of five shrimp species and their three biological replicates (1-3). Trinity transcriptome assembly metrics are the number of contigs and assembly size for each transcriptome.

Shrimp species		No. of contigs	Assembly size	TransRate score	BUSCO scores
Litopenaeus vannamei	1	32,302	28.6Mb	0.413	C:56%, F:21%, M:23%
	2	33,574	29.4Mb	0.401	C:56%, F:23%, M:21%
	3	28,101	22.7Mb	0.419	C:48%, F:25%, M:27%
Penaeus monodon	1	41,971	37.9Mb	0.387	C:66%, F:20%, M:14%
	2	40,927	36.5Mb	0.364	C:66%, F:19%, M:14%
	3	42,510	38.1Mb	0.390	C:64%, F:21%, M:14%
Melicertus latisulactus	1	29,743	21.9Mb	0.391	C:49%, F:24%, M:27%
	2	37,128	25.6Mb	0.410	C:46%, F:26%, M:27%
	3	28,125	21.7Mb	0.411	C:43%, F:25%, M:32%
Fenneropenaeus merguiensis	1	37,572	31.4Mb	0.385	C:64%, F:17%, M:19%
	2	41,336	34.8Mb	0.385	C:67%, F:16%, M:17%
	3	38,638	33.5Mb	0.389	C:65%, F:19%, M:16%
Metapenaeus endeavouri	1	35,407	25.9Mb	0.374	C:48%, F:25%, M:27%
	2	30,879	23.2Mb	0.399	C:48%, F:24%, M:27%
	3	38,204	25.5Mb	0.355	C:49%, F:26%, M:25%

3.4. *De Novo* Transcriptome assembly quality analysis

1. Transrate can be used to determine the quality of the *de novo* assembled transcriptome. Transrate is a bioinformatics tool designed to analyse transcriptomes which are assembled *de novo* and provide a quality metric between 0.0 – 1.0, known as the Transrate assembly score.

2. Transrate, designed by Smith-Unna (2016) provides this quality metric by comparing the RNA sequences of the contigs within the assembled transcriptome with the original RNA-Seq reads [12].
3. The Transrate software can be downloaded from https://github.com/Blahah/transrate or https://hibberdlab.com/transrate/installation.html, which also contains the updated installation and usage instructions. Running the Transrate software would result in various quality metrics. The most useful quality metric is the Transrate assembly score [12].
4. When examining the Transrate assembly score (0.0 – 1.0), the higher the score, the better the quality of the *de novo* assembled transcriptome. A meta-analysis of 155 published transcriptomes within the NCBI-TSA (Transcriptome Shotgun Assembly) database identified that a Transrate assembly score of 0.22 and above would indicate that the assembled transcriptome is superior in assembly quality than 50% of the 155 published transcriptomes included in the meta-analysis [12].

3.5. Transcriptome completeness analysis

1. Transcriptome completeness can be assessed by analysing the assembled transcriptome using a software known as BUSCO, which stands for Benchmarking Universal Single-Copy Orthologs [13].
2. BUSCO can be downloaded from https://gitlab.com/ezlab/busco/-/releases or https://busco.ezlab.org/, which also contains updated user guides based on the latest version.

3. BUSCO tool measures the presence of evolutionarily-informed expected genes within a specific lineage or phylum of the organism. These “expected genes” are single-copy orthologs that are near-universal and highly conserved genes within the phylum [13].
4. For each run of BUSCO software, the necessary datasets based on the lineage or phylum of the organism should be identified in the command line script or downloaded manually from https://busco-data.ezlab.org/v4/data/lineages/. For example, in this case, the Arthropoda dataset was used to measure the completeness of the five shrimp species.
5. Analysis by BUSCO results in scores in percentages of complete (C), fragmented (F) and missing (M) genes within the transcriptome, with results being ideally C > F or M genes.

3.6. Allergen Reference Database construction

1. In order to assess the assembled transcriptome for the presence of new and putative allergens, a database of reference allergens needs to be first established.
2. Ideally, to avoid missing out on any putative allergens, all known allergen amino acid (AA) sequences are to be compiled from various publicly available allergen databases into a manually constructed database in the format of a FASTA file.
3. However, it is also important to decide which publicly available allergen databases to retrieve allergen AA sequences from. The choice of allergen databases should be based on whether the allergens that are registered have enough evidence, supporting publications or peer review.
4. In the example used in this chapter, allergen AA sequences were retrieved from two public databases available online, namely the World Health Organization &

International Union of Immunological Societies (WHO/IUIS) Allergen Nomenclature database (www.allergen.org) [8], and AllergenOnline: Home of the FARRP (Food Allergy Research and Resource Program) Allergen Protein database (v.17) (www.allergenonline.org) [6, 14].

5. The latest version of AllergenOnline database is v.21, and the latest updates include all allergen AA sequences from WHO/IUIS Allergen Nomenclature database. Therefore, the use of only the AllergenOnline database would be sufficient.
6. To retrieve allergen AA sequences from AllergenOnline, the full list of allergens in PDF format can be downloaded from www.allergenonline.org, and the Accession number of all allergen AA sequences can be imported and downloaded from NCBI (https://www.ncbi.nlm.nih.gov/) and UniProt (https://www.uniprot.org/) database. The sequences need to be downloaded in FASTA format.
7. If the constructed allergen reference database is a combination of multiple databases, duplicates of allergen AA sequences need to be removed. Removing sequence duplicates can be performed by using Seqkit (https://bioinf.shenwei.me/seqkit/) using the command prompt 'rmdup' function. Download and usage instructions can be found at https://bioinf.shenwei.me/seqkit/.
8. The constructed allergen database can be viewed in text editor software such as Notepad, Notepad++, and other similar software.

3.7. BLAST search for potential allergens

1. BLAST or Basic Local Alignment Search Tool is a bioinformatics tool freely available within the NCBI portal (https://blast.ncbi.nlm.nih.gov/Blast.cgi).

2. This widely used tool identifies similar sequence regions between protein or AA sequences ('pblast'), between nucleotide (DNA or RNA) sequences ('nblast'), and even between nucleotide and protein sequences ('blastx' or 'tblastn', depending on the query and reference sequences).
3. The BLAST tool requires a query set of sequences, which in this case is the assembled transcriptome, and a reference database, which is the constructed allergen reference database.
4. Since the query sequence is a set of nucleotide sequences (transcriptome composed of mRNA sequences) whilst the allergen reference database is a set of AA sequences, we utilise the 'blastx' function.
5. Download the BLAST tool from the NCBI website https://blast.ncbi.nlm.nih.gov/Blast.cgi and run it on the terminal. (*see* **Note 3**)
6. Alternatively, various GUI (Graphical User Interface) software exists, either free or paid, that can utilise the BLAST tool function, which may or may not be already installed within. For example, in the example in this methods chapter, the paid Geneious (https://www.geneious.com/) bioinformatics software was utilised; and already has the BLAST tool function installed.
7. When setting the BLAST search criteria, the query file and database (reference) file need to be identified, including the BLAST program to be utilised, which is 'blastx'. A more detailed choice of search criteria used in the Geneious software for this example is shown in Table 2 below.
8. Importantly, the maximum E-value should be set at 1e-7 or 1 x 10^{-7}. E-value, also known as the "Expect value", is a value that is inversely related to the degree of similarity between the transcriptome (nucleotide query sequence) and its

corresponding match within the allergen's AA sequence. Studies by Hileman et al. (2002) and Nugraha et al. (2017) identified that allergen detection using similar bioinformatics tools with an E-value smaller than 1e-7 are likely to be significant matches [3, 15].

9. If BLAST search was conducted within the terminal using command-line interfaces (CLI), search results appearing in a 'Hit Table' form can be downloaded and loaded in Microsoft Excel for downstream processing. Alternatively, if BLAST search was conducted in a GUI platform such as Geneious, search results in the form of 'Hit Table' can be viewed and processed within the software.
10. However, search results should still be downloaded and loaded in Microsoft Excel (https://www.microsoft.com/en-us/microsoft-365/excel) for ease of downstream processing as described further in this chapter.

Table 2: BLAST search criteria used when utilising the BLAST search tool within the Geneious™ software.

BLAST search criteria	
Query	Batch search of nucleotide sequences
Database	Allergens (AA)
Program	blastx
Results	Hit table
Retrieve	Matching regions with annotations
Maximum Hits	1
Low complexity filter	"*checked*"
Max E-value	1e-7
Word Size	3
Matrix	BLOSUM62
Number of CPUs	30
Gap cost (Open Extend)	11 1

3.8. Processing/Filtering BLAST search results

1. BLAST search results in the 'Hit Table' form will consist of a series of match results containing the contig (code) from the assembled transcriptome and reference allergen's AA sequence, with both sequences aligned (Fig. 1). The sequences aligned will be AA sequences as the 'blastx' function will translate the transcriptome's contig's nucleotide sequence into AA sequence.
2. The match results will also contain various metrics, of which the metric that is important for this chapter's downstream processing is the "% Pairwise Identity".
3. The list of BLAST search results needs to be sorted based on decreasing % Pairwise Identity (%PI). All sequence matches below 50% PI need to be removed, thus keeping all sequence matches more than 50% PI.
4. Next, the percentage of Subject Coverage needs to be calculated into a new column on Microsoft Excel using the metrics provided by the BLAST search tool, namely the "Sequence Length" and the "Subject Length". The term "Subject Coverage" is the percentage of the allergen sequence that is covered by the matching contig from the transcriptome [7], and requires:
    a) Sequence length: length of the matched consensus sequence
    b) Subject length: length of the allergen sequence from the constructed database
5. The sequence length can be found within the "Sequence Length" column in the BLAST search Hit Table results. The subject length will need to be exported from the allergen database. Subject coverage is calculated using the formula:

    % Subject coverage = Sequence length / Subject length x 100%

6. The list of BLAST search results needs to be sorted now based on decreasing % Subject Coverage. All sequence matches below 90% Subject Coverage need to be removed, thus keeping all sequence matches more than 90% Subject Coverage.
7. Finally, change the sorting in Microsoft Excel back to decreasing percentage PI, and add a second sorting factor based on the Accession number. This will result in duplicate allergen AA sequence matches to cluster together.
8. Remove duplicates by removing similar allergen matches, keeping only one Accession number of an allergen AA sequence which has the highest percentage PI and highest percentage subject coverage.
9. The end result will consist of allergens that are identified, based on AA sequence similarity, within the transcriptome of the analysed organism, with more than 50% PI and more than 90% coverage.

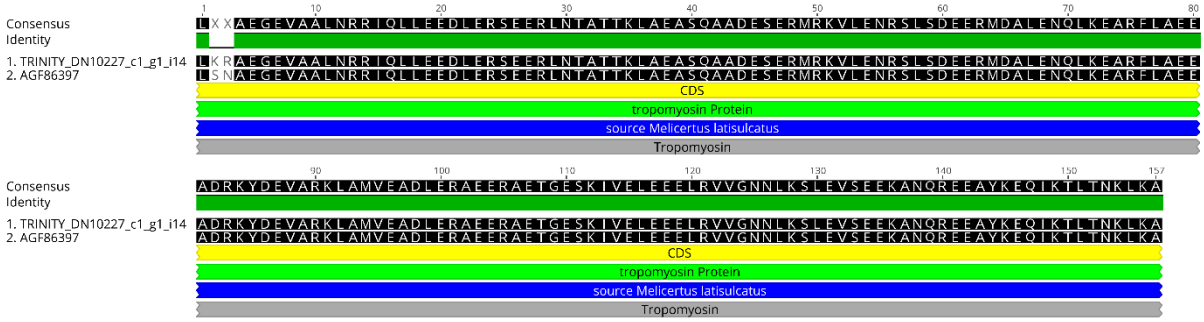

Fig. 1. Sequence alignment result produced by the BLAST search conducted in Geneious ™ software. The figure shows the aligned amino acid sequence of the tropomyosin allergen from *M. latisulcatus* (Genbank accession ID: AGF86397) which matched with one of the transcript/contig (TRINITY_DN10227_c1_g1_i14) within the assembled transcriptome of *P. monodon*.

3.9. Grouping allergens into known allergens, highly likely allergens, and likely allergens.

1. Among the list of allergens that are identified, known allergens of the analysed organism, if some have been previously identified, will and should be present. This is an indication of the validity of this approach.
2. In the circumstance where there are no previously identified allergens, the presence of known allergens belonging to species similar to the analysed organism or within the same genus or phylum should also be noted.
3. All other allergens identified are putative or potential new allergens, and this list of allergens can be classified based on their percentage PI into highly likely allergens (>70% PI) and likely allergens (>50% PI). Depending on the depth of analyses, focus can be given to the highly likely putative allergens or both.
4. The reason for 50% PI as the cut-off limit is due to information provided by Aalberse (2000) and substantiated by Goodman et al. (2016) that AA pairwise identity of more than 50% between homologous protein and a known allergen is the most predictive metric to determine whether the protein is likely to be an allergen, and identify the likelihood of cross-reactivity between the proteins [5, 6]. Aalberse (2000) also notes that the likelihood is higher when the AA pairwise identity is more than 70% [5].

3.10. Categorizing new allergens based on the origin organism

1. The identified putative new allergens can be classified into different groups, based on the organism from where the reference allergen originates, into any preferred phylogenetic classification. In the example used in this chapter, the new potential allergens were grouped into allergens previously identified in “mites”, “insects”, “fish”, “fungi”, “plants” and “other” [7].

2. All allergens within these classifications can be distributed within a pie chart to show the percentage of each group's allergens that are similar to the analysed organism.
3. The pie charts can be generated in Microsoft Excel (https://www.microsoft.com/en-us/microsoft-365/excel) or any data analysis software such as GraphPad Prism available at (https://www.graphpad.com/scientific-software/prism/).
4. An example of these types of pie charts showing the comparative distribution (in percentages) of the identified potential allergens amongst different groups of allergen sources can be viewed in the reference study for this methods chapter, the "*Novel allergen discovery through comprehensive* de novo *transcriptomic analyses of five shrimp species*" study by Karnaneedi et al. (2021) [7].

3.11. Comparing putative new allergens to homologous allergens in other food sources

1. Each match in the BLAST search results would identify the contig from the analysed organism that has the highest match with a known allergen from another food source.
2. The AA sequence of this contig can be exported from the Geneious software into a FASTA format.
3. Both the AA sequence of the analysed organism's transcriptome's translated contig and the AA sequence of the known registered allergen can be copied and pasted into a multiple sequence alignment software. Additionally, other known homologous allergens from other species can also be included in this comparative analysis.
4. The multiple sequence alignment software used in this example is a free online software, the Clustal Omega program by the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI), which can be accessed using the URL: https://www.ebi.ac.uk/Tools/msa/clustalo/.

5. Select “Protein” as the input sequence, and insert the AA sequence of the translated contig and all the other homologous allergen’s AA sequences. Select “ClustalW with character counts” as the output format and submit the job.
6. Once the sequence alignment is complete, select the “Results Summary” tab and select the “Percent Identity Matrix”. This data can then be copied and pasted, or imported, into Microsoft Excel.
7. Select appropriate or desired colour shading for the values based on high and low percentage identity between the different organism’s homologous allergen.
8. This “Identity Matrix” analysis is an example method to compare new allergens identified in a food source with other known homologous allergens [7, 16].
9. For full visualisation of the identity matrix generated for the example used in this chapter, please refer to Fig. 2.

Tropomyosin

		Species	Contig ID/Accession ID	#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Transcriptome Contigs	Shrimp	*L._vannamei_TM_Contig_1*	TRINITY_DN8531_c1_g1_i23	1	100																
		P._monodon_TM_Contig_1	TRINITY_DN9650_c0_g5_i8	2	100	100															
		P._monodon_TM_Contig_2	TRINITY_DN10227_c1_g1_i21	3	91	91	100														
		F._merguiensis_TM_Contig_1	TRINITY_DN9629_c3_g1_i18	4	100	100	91	100													
		M._latisulcatus_TM_Contig_1	TRINITY_DN10005_c3_g3_i4	5	100	100	91	100	100												
		M._latisulcatus_TM_Contig_2	TRINITY_DN7339_c0_g1_i14	6	95	95	96	95	95	100											
		M._latisulcatus_TM_Contig_3	TRINITY_DN8176_c0_g1_i1	7	61	61	60	61	61	60	100										
		M._endeavouri_TM_Contig_1	TRINITY_DN9495_c0_g1_i3	8	99	99	91	99	99	96	61	100									
		M._endeavouri_TM_Contig_2	TRINITY_DN9495_c0_g1_i16	9	82	82	87	82	82	83	57	82	100								
		M._endeavouri_TM_Contig_3	TRINITY_DN146_c0_g1_i1	10	61	61	60	61	61	60	100	61	57	100							
		M._endeavouri_TM_Contig_4	TRINITY_DN5622_c0_g1_i1	11	57	57	53	57	57	55	59	56	52	59	100						
Allergen		*L._vannamei_TM_[Lit_v_1]*	ACB38288	12	100	100	91	100	100	95	61	99	82	61	57	100					
		P._monodon_TM_[Pen_m_1]	AAX37288	13	100	100	91	100	100	95	61	99	82	61	57	100	100				
		M._latisulcatus_TM_[Mel_l_1]	AGF86397	14	95	95	94	95	95	98	60	95	82	60	56	95	95	100			
	HDM	*D._pteronyssinus_TM_[Der_p_10]*	AAB69424	15	81	81	78	81	81	80	59	81	72	59	55	81	81	80	100		
	Cockroach	*B._germanica_TM_[Bla_g_7]*	AAF72534	16	82	82	81	82	82	83	56	83	74	57	52	82	82	82	80	100	
		P._americana_TM_[Per_a_7]	CAB38086	17	82	82	81	82	82	82	56	83	74	56	52	82	82	81	80	97	100

Percentage Identity

50 55 60 65 70 75 80 85 90 95 100

Fig. 2. Comparison of amino acid sequence identities of (1-11) contigs from five shrimp species that matched with tropomyosin (TM) allergen, (12-14) known shrimp TM allergen, and (15-17) house dust mite and cockroach TM allergen. The sequence identities were calculated

using multiple sequence alignment in Clustal Omega (EMBL-EBI). This figure is adapted from Karnaneedi et al. (2022).

3.12. Assessing the allergenicity potential of contig from transcriptome using AllerCatPro 2.0

1. An additional analysis can be conducted on the transcriptome contigs that were identified as new putative allergens by assessing the allergenicity potential by utilising protein allergenicity predicting web server known as AllerCatPro (https://allercatpro.bii.a-star.edu.sg/) [17, 18].
2. The current version of AllerCatPro, AllerCatPro 2.0 [17] assesses allergenicity potential by checking the similarity of the input contig sequence with 714 representatives in 3D model/structure database of known allergens as well as comprehensive dataset which includes known allergens (*n*=4979), low allergenic proteins (*n*=162), and autoimmune allergens (*n*=165) [17].
3. The input sequence into AllerCatPro 2.0 search engine can either be the AA or the mRNA sequence of the contig (from the transcriptome).
4. Export the AA or mRNA sequence of the new potential allergen identified in the transcriptome from Geneious, and input the sequence into the AllerCatPro2.0 web server's search tool (https://allercatpro.bii.a-star.edu.sg/).
5. Before submitting the sequence to the search tool, add a line above the sequence which corresponds to a FASTA format. The line should begin with ">" followed by any assigned name. For example: ">_Contig_06_Genus_Species".
6. After submitting the contig sequence, the search will generate results which will inform on the predicted most similar allergen, number of potential cross-reactive

proteins, protein family group, clinical relevance of the most similar allergen to the contig, and the similarity to an allergen, autoimmune allergen or low allergenic protein. Importantly, the results will indicate if there is strong, weak, or no evidence of allergenicity [17].

7. When applied to the transcriptome contigs which are found to be likely (>50% PI) and highly likely (>70% PI) new putative allergens, this analysis will add power to the potential allergenicity and indicate which new putative allergens should be studied further for clinical allergenicity using the traditional biomolecular and immunological allergen identification and characterisation methods.

4. Notes

1. If an assembled transcriptome of good quality metrics is already available, there will be no need to assemble the transcriptome *de novo* using RNA-Seq data.
2. If the transcriptomic analyses require a specific pool of organisms, another method to obtain the RNA-Seq data is by extracting total RNA from the samples and conducting RNA sequencing.
3. The BLAST tool can be utilised within the NCBI website, however, due to the large sizes of the transcriptome file and the allergen database, the search would either take a long time or fail. To circumvent this issue, the BLAST tool can be downloaded from the NCBI website (https://blast.ncbi.nlm.nih.gov/Blast.cgi) and run on the terminal.

5. References

1. Kamath SD, Rahman AMA, Voskamp A, Komoda T, Rolland JM, O'Hehir RE and Lopata AL (2014) Effect of heat processing on antibody reactivity to allergen variants and fragments of black tiger prawn: A comprehensive allergenomic approach. Molecular Nutrition & Food Research 58:1144-1155. doi: 10.1002/mnfr.201300584
2. Rahman AMA, Helleur RJ, Jeebhay MF and Lopata AL (2012) Characterization of seafood proteins causing allergic diseases. Allergic diseases–highlights in the clinic, mechanisms, and treatment. InTech:107-40.
3. Nugraha R, Kamath SD, Johnston E, Zenger KR, Rolland JM, O'Hehir RE and Lopata AL (2018) Rapid and comprehensive discovery of unreported shellfish allergens using large-scale transcriptomic and proteomic resources. Journal of Allergy and Clinical Immunology 141:1501-1504.e8. doi: 10.1016/j.jaci.2017.11.028
4. Nugraha R, Ruethers T, Johnston EB, Rolland JM, O'Hehir RE, Kamath SD and Lopata AL (2021) Effects of Extraction Buffer on the Solubility and Immunoreactivity of the Pacific Oyster Allergens. Foods 10. doi: 10.3390/foods10020409
5. Aalberse RC (2000) Structural biology of allergens. J Allergy Clin Immunol 106:228-38. doi: 10.1067/mai.2000.108434
6. Goodman RE, Ebisawa M, Ferreira F, Sampson HA, Ree R, Vieths S, Baumert JL, Bohle B, Lalithambika S, Wise J and Taylor SL (2016) AllergenOnline: A peer-reviewed, curated allergen database to assess novel food proteins for potential cross-reactivity. Molecular Nutrition & Food Research 60:1183-1198. doi: 10.1002/mnfr.201500769
7. Karnaneedi S, Huerlimann R, Johnston EB, Nugraha R, Ruethers T, Taki AC, Kamath SD, Wade NM, Jerry DR and Lopata AL (2021) Novel Allergen Discovery through Comprehensive De Novo Transcriptomic Analyses of Five Shrimp Species. International Journal of Molecular Sciences 22. doi: 10.3390/ijms22010032
8. (WHO/IUIS) WHOaIUoIS (2019) Allergen Nomenclature.
9. Song L and Florea L (2015) Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. GigaScience 4:48. doi: 10.1186/s13742-015-0089-y
10. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N and Regev A (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29:644-52. doi: 10.1038/nbt.1883
11. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, MacManes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, LeDuc RD, Friedman N and Regev A (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc 8:1494-512. doi: 10.1038/nprot.2013.084
12. Smith-Unna R, Boursnell C, Patro R, Hibberd JM and Kelly S (2016) TransRate: reference-free quality assessment of de novo transcriptome assemblies. Genome Res 26:1134-44. doi: 10.1101/gr.196469.115
13. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV and Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210-2. doi: 10.1093/bioinformatics/btv351

14. (FARRP) (2019) Allergen Online: Home of the FARRP (Food Allergy Research and Resource Program) Allergen Protein Database. In: Nebraska-Lincoln Uo (ed), 17 edn., University of Nebraska-Lincoln
15. Hileman RE, Silvanovich A, Goodman RE, Rice EA, Holleschak G, Astwood JD and Hefle SL (2002) Bioinformatic Methods for Allergenicity Assessment Using a Comprehensive Allergen Database. International Archives of Allergy and Immunology 128:280-291. doi: 10.1159/000063861
16. Ruethers T, Taki AC, Johnston EB, Nugraha R, Le TTK, Kalic T, McLean TR, Kamath SD and Lopata AL (2018) Seafood allergy: A comprehensive review of fish and shellfish allergens. Mol Immunol 100:28-57. doi: 10.1016/j.molimm.2018.04.008
17. Nguyen MN, Krutz NL, Limviphuvadh V, Lopata AL, Gerberick G F and Maurer-Stroh S (2022) AllerCatPro 2.0: a web server for predicting protein allergenicity potential. Nucleic Acids Research 50:W36-W43. doi: 10.1093/nar/gkac446
18. Maurer-Stroh S, Krutz NL, Kern PS, Gunalan V, Nguyen MN, Limviphuvadh V, Eisenhaber F and Gerberick GF (2019) AllerCatPro—prediction of protein allergenicity potential from the protein sequence. Bioinformatics 35:3020-3027. doi: 10.1093/bioinformatics/btz029