

## **Title page**

**Journal:** Genomics, Proteomics, and Bioinformatics

**Title:** Novel allergen discovery through comprehensive *de novo* transcriptomic analyses of 5 shrimp species

### **Authors:**

Shaymaviswanathan Karnaneedi<sup>1-6</sup>, Roger Huerlimann<sup>4-6</sup>, Elecia B. Johnston<sup>1-3,5</sup>, Roni Nugraha<sup>1,7</sup>, Thimo Ruethers<sup>1-3,5,6</sup>, Aya C. Taki<sup>1-3</sup>, Sandip D. Kamath<sup>1-3</sup>, Nicholas M. Wade<sup>4,8</sup>, Dean R. Jerry<sup>4,5,9</sup>, Andreas L. Lopata<sup>1-3,5\*</sup>

<sup>1</sup>Molecular Allergy Research Laboratory, College of Public Health, Medical and Veterinary Sciences, James Cook University, Townsville, QLD 4811, Australia

<sup>2</sup>Australian Institute of Tropical Health and Medicine, James Cook University, Townsville, QLD 4811, Australia

<sup>3</sup>Centre for Food and Allergy Research, Murdoch Children's Research Institute, The Royal Children's Hospital, 50 Flemington Road, Parkville, Victoria 3052, Australia

<sup>4</sup>ARC Research Hub for Advanced Prawn Breeding, Australia

<sup>5</sup>Centre for Sustainable Tropical Fisheries and Aquaculture, College of Science and Engineering, James Cook University, Townsville, QLD 4811, Australia

<sup>6</sup>Centre for Tropical Bioinformatics and Molecular Biology, James Cook University, Townsville, QLD 4811, Australia

<sup>7</sup>Department of Aquatic Product Technology, Bogor Agricultural University, Bogor, Indonesia

<sup>8</sup>CSIRO Agriculture and Food, Aquaculture Program, 306 Carmody Road, St Lucia, QLD 4067, Australia

<sup>9</sup>Tropical Futures Institute, James Cook University, 149 Sims Drive, Singapore 387380, Singapore

E-mail: [andreas.lopata@jcu.edu.au](mailto:andreas.lopata@jcu.edu.au) (Lopata AL)

**Running title:** *Karnaneedi et al / Transcriptomic analysis of shrimp allergens*

## **Abstract**

Shellfish allergy affects up to 2% of the world's population and persists for life in most patients. The diagnosis of a shellfish allergy, in particular shrimp, is however often challenging due to the similarity of allergenic proteins in other invertebrates. Despite the clinical importance, the complete allergen repertoire of allergy-causing shrimps remains unclear. Here we mine the complete transcriptome of five frequently consumed shrimp species to identify and compare allergens with all known allergen sources. The transcriptomes were assembled *de novo* from raw RNA-Seq data of the whiteleg shrimp (*Litopenaeus vannamei*), black tiger shrimp (*Penaeus monodon*), banana shrimp (*Fenneropenaeus merguensis*), king shrimp (*Melicertus latisulcatus*), and endeavour shrimp (*Metapenaeus endeavouri*). Trinity was used to assemble the transcriptome, and Transrate and BUSCO applied to verify the assembly. Blast search with the two major allergen databases, WHO/IUIS Allergen Nomenclature and AllergenOnline, successfully identified all seven known crustacean allergens. Salmon was utilised to measure their relative abundance, demonstrating sarcoplasmic calcium-binding protein, arginine kinase and myosin light chain as highly abundant allergens. In addition, the analyses revealed up to 40 unreported allergens in different shrimp species, including heat shock protein (HSP), alpha-tubulin, chymotrypsin, cyclophilin, beta-enolase, aldolase A, and glyceraldehyde-3-phosphate dehydrogenase (G3PD). Multiple sequence alignment, conducted in Jalview 2.1 with Clustal Omega, demonstrated high homology with allergens from other invertebrates including mites and cockroaches. This first transcriptomic analyses of allergens in a major food source provides a valuable genomic resource for investigating shellfish allergens, comparing invertebrate allergens and developing improved diagnostics and novel immunotherapeutics for food allergy.

**Keywords:** Allergy; Prawn; Tropomyosin; Allergen; RNA-Seq

## **Introduction**

Food allergy affects up to 10% of children and 10% of adults, and the prevalence is projected to be on the rise [1, 2]. Food allergy is caused through ingestion of food that contains allergenic proteins that triggers adverse reactions in sensitised individuals [3, 4]. The term “allergen” refers to a protein capable of inducing sensitisation and subsequent allergic immune responses through immunoglobulin E (IgE)-mediated type 1 hypersensitivity in patients [3-7].

Shellfish allergy is often lifelong, similar to peanut allergy, affects about 2% of the global population and appears to be highly prevalent in the Asia-Pacific region and other countries where seafood consumption is high [8-11]. A recent epidemiology study from Vietnam revealed that the prevalence of shellfish allergy is as high as 4.2% [11], while up to 3% of adults in the USA are sensitised to shellfish [1].

Among shellfish allergic individuals, shrimp allergy seems to be the most prominent crustacean allergy and remains to be difficult to diagnose and manage, for multiple reasons. Shrimp accounts for one of the most prevalent events of food derived anaphylactic reactions after peanuts and tree nuts [12-15].

The management of shrimp allergy is often challenging due to immunological cross-reactivity to molecularly similar allergens [12, 13, 16-22]. The similarity of shrimp allergens to proteins of other shellfish species, including crabs and lobsters, and other invertebrates such as house dust mites (HDM) and cockroaches, can induce unexpected allergic reactions [15-20, 23, 24]. Although this cross-reactivity has been observed in the clinical setting, the underlying molecular basis is not well understood.

Over the past decades, more than 2000 allergens are now well characterised and accessible via several databases, including the World Health Organization & International Union of Immunological Societies (WHO/IUIS) Allergen Nomenclature database ([www.allergen.org](http://www.allergen.org)) having the most stringent inclusion criteria [7, 25], and the highly peer-reviewed AllergenOnline: Home of the FARRP (Food Allergy Research and Resource Program) Allergen Protein database ([www.allergenonline.org](http://www.allergenonline.org)) [26, 27].

Allergen discovery is traditionally conducted using whole allergen sources and isolation of IgE antibody binding proteins [4, 28-31]. However, this approach has many

limitations, including low sensitivity and small patient cohorts that does not allow the detection of all possible allergenic proteins [32].

Here we report the first complete transcriptome analysis of shellfish food allergen source, with focus on shrimp allergens, the most common shellfish allergy. The transcriptomes of five most frequently consumed shrimp species were assembled *de novo* and screened for the presence of similar amino acid (AA) sequences to 2,172 allergens in the WHO/IUIS Allergen Nomenclature and AllergenOnline databases (Figure 1).

## **Results**

### **Assessment of 15 assembled transcriptomes**

Illumina HiSeq® 2500 (Illumina Australia and New Zealand, VIC, Australia) sequencing produced 125bp paired-end sequencing data with a total number of paired-end reads for each sample of approximately 20 million reads. The *de novo* assembly for 15 samples (three replicates each for five shrimp species) resulted in 28,101 to 42,510 contigs (Table 1). All 15 samples had more than 87% of read pairs that mapped back to the contigs within the assembled transcriptome. TransRate scores (assembly scores) for each of the 15 transcriptomes were approximately 0.4 (Table 1). BUSCO (Benchmarking Universal Single-Copy Orthologs) results, overall, had a complete genes (C) score ranging between 43% - 67%; fragmented genes (F) score ranging between 16% - 26%; and missing genes (M) score between 14% - 32% (Table 1). The transcriptomes of *P. monodon* and *F. merguensis* had the highest values for complete BUSCO's (C scores) (Table 1). An observable pattern here is that both these shrimp species also had the highest number of contigs and assembly size.

### **Large numbers of allergens identified within the transcriptomes**

After duplicate removal, the results yielded 40 unique allergen (AA sequences) identified in whiteleg shrimp (*L. vannamei*), 44 in black tiger shrimp (*P. monodon*), 42 in banana shrimp (*F. merguensis*), 44 in king shrimp (*M. latisulcatus*), and 50 in endeavour shrimp (*M. endeavouri*) (Figure 2). Approximately two thirds of allergen AA sequences that matched with all five shrimp species' transcriptomes, belonged to shellfish, mites, and fungi species (Figure 2). The remaining allergen AA sequences belonged to plants, insects, fish and other species.

### **Known crustacean allergens identified**

Contigs that matched with the major shrimp allergen tropomyosin (TM) were identified in all five species, with some species having more than one contig representing this allergen (Figure 3.A). *L. vannamei* shrimp's TM\_Contig\_1 has a 100% AA sequence identity with the previously recorded and IUIS registered Lit v 1 (ACB38288). This is a similar finding to *P. monodon*'s TM\_Contig\_1, which has a 100% sequence identity with Pen m 1 (AAX37288). Both TM\_Contig\_1's also match with a 100% similarity with each other (Figure 3.A). Overall, TM\_Contig\_1 of all five species showed a high

sequence similarity (pairwise identity of 99-100%) with Lit v 1 and Pen m 1, (Figure 3.A). However, TM\_Contig\_2 of *P. monodon* and *M. endeavouri* only showed a pairwise identity (PI) of 91% and 82%, respectively, with both Lit v 1 and Pen m 1 (Figure 3.A). The inclusion of HDM and cockroach tropomyosin allergens, Der p 10 (AAB69424), Bla g 7 (AAF72534), and Per a 7 (CAB38086), in the analyses of tropomyosin AA sequences revealed to have more than 70% PI with shrimp TM (Figure 3.A). Molecular phylogenetic tree analyses of previously published AA sequences of TM revealed that the crustacean TM's are not only very similar to each other, but also to insect and mite TM's. In comparison, molluscs, which are also grouped as "shellfish" with crustaceans, were found to be only distantly related in terms of TM AA sequence (Figure 3.B).

One contig in each of four shrimp species was identified as the arginine kinase (AK) allergen, while *M. endeavouri* had two contigs. In contrast to TM, all contigs were highly similar to each other and to the published AK allergens in *L. vannamei*, Lit v 2 (ABI98020), and *P. monodon*, Pen m 2 (AAO15713), with more than 95% PI (Figure 4.A). They were all also found to be more similar to the published cockroach AK allergens in *B. germanica*, Bla g 9 (ACM24358), and *P. americana*, Per a 9 (AAT77152), (83-84% identity) than the published HDM AK allergens in *D. pteronyssinus*, Der p 20 (ACD50950), and *D. farinae*, Der f 20 (AIO08850) (78-79% identity) (Figure 4.A). Similar to TM, published AA sequences of crustacean AK are more closely related to each other; and insects and mites as opposed to molluscs (Figure 4.B).

Only one contig each from all five shrimp species analysed matched myosin light chain (MLC), and demonstrated almost identical AA sequences. Interestingly, they were not at all similar to the published MLC allergens in *L. vannamei*, Lit v 3 (ACC76803), or *P. monodon*, Pen m 3 (ADV17342), with only 16-17% PI (Figure 5.A). Instead, they were found to be more similar to the *C. crangon* (North-sea shrimp) MLC allergen, Cra c 5 (ACR43477) with 86-87% PI (Figure 5.A). The contigs were also identified to be more closely related to the american HDM, *D. farinae*, MLC allergen, Der f 26, (51-54% identity) than the German cockroach, *B. germanica*, MLC allergen, Bla g 8 (18-19% identity) (Figure 5.A). Molecular phylogenetic tree analyses on the distance of MLC among edible crustaceans, molluscs and allergy causing mites confirmed that not all crustacean MLC are closely related to each other. For example, mud crab (*S.*

*paramamosain*) is more closely related to molluscs' MLC than shrimps and crayfish (Figure 5.B). Black tiger shrimp (*P. monodon*) and whiteleg shrimp (*L. vannamei*) contain MLC that are distantly related to kuruma shrimp (*M. japonicus*) and north-sea shrimp (*C. crangon*), but closely related to MLC from German cockroach (*B. germanica*) (Figure 5.B).

Four of the shrimp species had two contigs matching sarcoplasmic calcium-binding protein (SCBP), while *M. endeavouri*, only had one. SCBP\_Contig\_1 from all five shrimp species were highly similar to each other and also with the published SCBP allergen in *L. vannamei*, Lit v 4 (ACM89179) and *P. monodon*, Pen m 4 (ADV17343) with PI close to 100%, but only over 80% with the published SCBP allergen in *C. crangon*, Cra c 4 (ACR43475) (Figure 6.A). In contrast, SCBP\_Contig\_2 from the four species, except *M. endeavouri*, were only 82-84% identical to Lit v 4, Pen m 4 and Cra c 4, with the latter having a slightly higher match than the two former (Figure 6.A). Unlike MLC, but similar to TM and AK, the published AA sequences of SCBP in a phylogenetic tree analyses portrayed that all SCBP from edible crustaceans and molluscs are very closely related to other species within the same phylum, but distantly related between the phyla (Figure 6.B).

Seven contigs matched with Troponin C (TNC) in all five shrimp species, with *M. latisulcatus* and *M. endeavouri* having two contigs each whilst the other three shrimp species having only one each. All seven contigs were moderate to highly similar to each other and also with the published TNC allergens in *P. monodon*, Pen m 6 (ADV17344) and *C. crangon*, Cra c 6 (ACR43478), with PI ranging between 81-100% (Supplementary Figure 1). The PI of shrimp TNC with cockroach and storage mite TNC allergens ranged between 57-65% (Supplementary Figure 1). Meanwhile, only one contig from each shrimp species matched with Troponin I (TNI) allergen, and they were all highly identical to each other (PI: 87-99%), but were only moderately identical to the published TNI allergen in the narrow-clawed crayfish *P. leptodactylus*, Pon I 7 (P05547) (PI: 78-88%) (Supplementary Figure 2). Similarly, only one contig matched with Triosephosphate isomerase (TIM) allergen in each shrimp species and were all highly identical to each other and also with the published TIM allergen in *C. crangon*, Cra c 8 (ACR43476), (PI: 87-99%) (Supplementary Figure 3). However, they had lower PI to American HDM TIM allergens, Der f 25.01 (AGC56216) and Der f 25.02 (AIO08860), with PI values ranging between 66-69% (Supplementary Figure 3).

### Abundance of known crustacean allergens varies between shrimp species

The average expression or mean abundance, measured in transcripts-per-million (TPM), of TM across all five species ranges from 10,000 – 15,000 TPM (Figure 7.A). Comparing the difference in abundance between the two tropomyosin contigs within the same species (*P. monodon* and *M. endeavouri*), TM\_Contig\_2 of *P. monodon* was found to be significantly lower than its counterpart, TM\_Contig\_1 (Figure 7.A). Meanwhile, there was no significant difference between TM\_Contig\_1 and TM\_Contig\_2 of *M. endeavouri* (Figure 7.A). With AK, the mean abundance was approximately 40,000 – 80,000 TPM in all five species (Figure 7.B). Comparing the abundance of the two AK contigs in *M. endeavouri*, AK\_Contig\_1 was significantly lower than AK\_Contig\_2 (Figure 7.B). The mean abundance of MLC was approximately 30,000 – 50,000 TPM in all species (Figure 7.C). Meanwhile, for SCBP, the mean abundance was between 40,000 and 90,000 TPM in all species (Figure 7.D). Interestingly, SCBP\_Contig\_1 of *L. vannamei*, *P. monodon*, and *F. merguensis* were all significantly higher than their respective SCBP\_Contig\_2 (Figure 7.D). The same pattern could be visually observed on *M. latisulcatus* too but unfortunately, due to the presence of only two replicates instead of three (refer to: *Removal of inconclusive dataset* in Materials and methods section), the significance of this difference could not be statistically confirmed by T-test (GraphPad Prism (v7.03)). Similarly, one could not predict the significance of differences in the two TNC contigs of *M. latisulcatus*. However, for *M. endeavouri*, TNC\_Contig\_1 was found to be significantly higher than its TNC\_Contig\_2 (Figure 7.E). Overall, the mean abundance value for TNC was around 4,000 – 10,000 TPM for all five shrimp species (Figure 7.E). As for TNI and TIM, the mean abundance values for all five shrimp species were approximately 16,000 – 20,000 TPM (Figure 7.F) and 2,000 – 6,000 TPM respectively (Figure 7.G).

We then examined the difference in abundance of each allergen within individual shrimp species. We only took into account the contig with the highest PI value when there were more than one contig for that allergen. In all species, the top three most highly expressed allergens were SCBP, AK, and MLC (Figure 8). In fact, SCBP was the most highly expressed allergen in all species except *P. monodon*, where AK was higher (Figure 8.B). In descending order of abundance, these three allergens are followed by TNI, TM, TNC, and TIM (Figure 8). However, in *F. merguensis*, TM was higher than TNI, TNC and TIM (Figure 8.C). In addition, only in *F. merguensis*, TM's



abundance was not significantly different from all the three highly abundant allergens, namely, SCBP, AK and MLC (Figure 8.C).

### Evolutionary relationship of shellfish allergens TM, AK, MLC, and SCBP

The evolutionary distance of shrimp TM, AK, MLC, and SCBP were analysed among other edible crustacean and mollusc species; and allergy causing mite and insect species. The generated molecular phylogenies of all four shrimp proteins showed close affinities to homologues of other crustaceans such as crab, lobster and crayfish. However, homologues of the other class of “shellfish”, molluscs, have a distant relationship to shrimps. Molecular phylogenetic analyses of TM and AK revealed that allergy inducing mite and insect homologues are closer to shrimp TM and AK than molluscs. This observation is supported by a recent study by Nugraha et al. where IgE antibody binding epitopes demonstrated shared protein regions of clinical importance [33]. MLC of German cockroach is found to have a closer evolutionary relationship to the black tiger shrimp and whiteleg shrimp, whilst the MLC belonging to American house dust mite is closely related to MLC of a different subset of crustaceans, including, the north-sea shrimp, kuruma shrimp, and red swamp crayfish. Another interesting finding is that the crustacean MLC of mud crab have a closer evolutionary distance with homologues from the mollusca phylum, especially the pacific oyster, but not to those of the other crustaceans. Molecular phylogenetic analysis of SCBP shows a demarcated distance between the crustacean SCBP and mollusc SCBP. No insect or mite SCBP was included in this analyses as there were no AA sequence data available for insect or mite SCBP on NCBI Genbank or UniProt databases.

### Discovery of unreported shrimp allergens

Apart from the previously established shellfish allergens that were confirmed in the five shrimps' transcriptomes, some of the allergens of non-shellfish species are identified to be standout candidates to be unreported allergens in shrimps due to their high % PI values (Table 2). These allergens are heat shock protein (HSP), alpha-tubulin, chymotrypsin, beta-enolase, glyceraldehyde-3-phosphate dehydrogenase (G3PD), cyclophilin and aldolase A (Table 2). The HSP70 (Tyr p 28, AOD75395) from the storage mite *T. putrescentiae* that matched with the shrimp transcriptomes has the highest PI values with all 5 shrimp species (>82%) (Table 2). Other allergen AA sequences that matched with a PI of more than 70% to all 5 shrimp species'

transcriptomes are alpha-tubulin (Der f 33, AIO08861) and chymotrypsin (Der f 6, AAP35065) of the american HDM *D. farinae*, beta-enolase (Sal s 2, ACH70932) of the atlantic salmon *S. salar*, and glyceraldehyde-3-phosphate dehydrogenase or G3PD (Tri a 34, CAZ76054) of wheat *T. aestivum* (Table 2). The allergen cyclophilin (Asp f 27, CAI78448) of the common mould *A. fumigatus*, only matched with a PI of more than 70% with the banana and king shrimps' transcriptomes. Meanwhile, the allergen Aldolase A (Thu a 3, CAX62602) of the yellowfin tuna *T. albacares* matched with a PI of more than 70% only with the banana and endeavour shrimps' transcriptomes (Table 2).

## **Discussion**

Allergen discovery using traditional protein isolation and immunological assay methods, have identified and characterised seven shellfish allergens including the major allergen TM, in addition to AK, MLC, SCBP, TNC, TNI, and TIM. All seven allergens have been identified in various shrimp species except TNI. The increased reporting of allergic cross-reactivity in shrimp-allergic patients to non-shrimp sources demands a full analysis of allergenic proteins. This study utilised an advanced transcriptomic approach to discover the whole repertoire of shrimp allergens, both reported and unreported. Using this comprehensive approach, combining the generation of transcriptomes of five shrimp species and BLAST searching the transcriptomes with all known allergen AA sequences, we identified up to 50 allergens. The majority of identified allergens (45%) belong to the group of shellfish and mite allergens. This is not surprising as the shellfish group consists of crustacean (shrimp) and molluscs, which are often combined when analysing related allergens [15, 33, 34].

In line with existing studies, we confirmed the presence of TM in all five investigated shrimp species, however, the AA sequence was not always similar and the abundance varied significantly. The major allergen TM, a rod-shaped muscle protein (33 – 39 kDa), demonstrates 100% PI between the whiteleg and black tiger shrimp, as recently reported by Ruethers et al. [15], validating the *in silico* approach used in this study. Furthermore, we demonstrated for the first time that TM from banana and king shrimp also exhibited 100% PI to whiteleg and black tiger shrimp's TM. In contrast, known TM allergen from the king shrimp (Mel I 1; AGF86397) shares only 95% of AA identity with the other shrimp species, which was previously demonstrated to result in different allergenicity in patients [35]. This is an important finding which needs to be followed up in clinical studies. However, the identified TM in all five shrimps are very similar and therefore termed isoallergens. The IUIS Allergen Nomenclature identifies an isoallergen to be two proteins with the same biological function with more than 67% AA sequence identity and similar molecular size [36]. The high AA sequence identity (PI: >70%) of the house dust mite (HDM) and cockroach TM allergens (Der p 10, Bla g 7 and Per a 7) with all the analysed shrimp TM's indicate a likelihood of all these invertebrate allergens of being cross-reactive. As previously established, an AA

sequence identity of more than 70% would demonstrate a highly-likely possibility of cross-reactive IgE antibody binding to these allergens [32, 37]. Clinical studies have previously demonstrated a phenomenon named ‘HDM-cockroach-shrimp’ cross-reactivity [23, 38-42], and we provide here definitive molecular data on the AA sequence similarity of a major shrimp allergen with other invertebrate species.

AK, an important enzymatic protein which regulates the cellular ATP levels of invertebrates, is a heat labile protein (38 – 45 kDa) and highly concentrated in muscle tissue [43]. All five analysed shrimp species demonstrate very high AA sequence similarity with each other, indicating that shrimp allergic patients reacting to AK would most likely react to all five species. AK is considered a major allergen amongst insects and mites and potentially a pan-allergen implicated in cross-reactivity between invertebrate species [44-46]. All five shrimp AK identified in this study are highly-likely allergens with high AA sequence identity (>70%) to AKs from mites (Der p 20; Der f 20) and cockroaches (Bla g 9; Per a 9). Furthermore, the two almost similar AK contigs (PI: 99%) found in endeavour shrimp indicate that they are potential variants instead of isoforms [36], however, the significantly low abundance of AK\_Contig\_1 than AK\_Contig\_2 of endeavour shrimp reduces the likelihood of AK\_Contig\_1 having a role as an allergen.

MLC is part of a large macromolecular complex in muscle tissue consisting of two heavy and four light chains. There are two crustacean MLC allergens, the essential MLC1 (~18kDa) and the regulatory MLC2 (~20kDa) [47, 48]. MLC allergens have been identified in the whiteleg shrimp (Lit v 3) [49] and black tiger shrimp (Pen m 3) [28] – both MLC2 – but not in other shrimp species investigated in this study. While the AA sequences of MLC1 and MLC2 are very different from each other [50], the MLC contigs found in all 5 analysed shrimps are most likely MLC1. Furthermore, this study also suggests that HDM MLC (Der f 26) and cockroach MLC (Bla g 8) are most likely MLC1 and MLC2, respectively, explaining the close molecular phylogenetic relationship to crustacean, but not molluscs.

Another crustacean allergen involved in invertebrate muscle contraction is SCBP (20 – 24 kDa), through binding of calcium ions [51, 52]. We identified two different SCBP contigs for each shrimp species (except endeavour shrimp), with PI values between 81-85%, implicating the presence of SCBP isoallergens. However, the significantly low abundance of SCBP\_Contig\_2 diminishes its role as an allergen compared to SCBP\_Contig\_1.

Other muscle regulatory protein identified include troponin. This protein is composed of three subunits, suffixed C, I, and T, with Troponin C and I being registered as allergens. TNC has been identified as an allergen in various crustaceans [53-55], cockroaches [56] and the storage mite [57]. Meanwhile, TNI has only been identified in narrow-clawed crayfish [58]. TIM, an enzyme that is involved in glucose metabolism, is also a registered allergen in north-sea shrimp [53], red swamp crayfish [59], american HDM [60], octopus [61] and wheat [62]. TNC, TNI, and TIM are highly conserved among shrimp species with sequence homology higher than 80%, 78%, and 87%, respectively.

Having considered the different shrimp allergens, it is important to note that apart from allergen presence, the abundance of isoforms also needs to be taken into consideration. Correlation between protein abundance and RNA-Seq data has been established, with some post-transcriptional cellular processes affecting this interpretation [63]. A study on European HDM allergen transcript levels using RNA-Seq data concluded that allergens have a higher level of abundance than non-allergens [64]; and their results were found to be relatively similar to homologues identified in american HDM from a different study [65]. In particular, these dust mite allergen studies indicate that there is substantial correlation between RNA-Seq dependent abundance levels and a protein's allergen status. Therefore, when there were more than one contig identified for any allergens in the five shrimps, we proceeded to analyse the isoform that had the highest abundance. Comparing known allergens within every shrimp species, SCBP, AK, and MLC were the most abundant in all five shrimps. Interestingly, TM was found to be significantly less abundant than SCBP (in king and endeavour shrimp); AK (in black tiger shrimp); and all three

allergens (in whiteleg shrimp). However, TM is the major and most recognised shrimp allergen, despite its relatively low abundance. The stronger allergenicity is possible due to being very heat stable and having linear IgE binding epitopes as compared to AK, SCBP, or MLC [8, 28, 33].

In addition to previously implicated crustacean allergens, this study also identified up to 38 previously unreported but likely allergens (>50% PI), including seven proteins which are very-likely allergens (>70% PI). Three of these proteins, HSP70, alpha-tubulin, and chymotrypsin, have very high matches to known mite allergens in this study. Additionally, these three proteins have been identified as allergens in different mite and insect species [25, 66-69]. Subsequently, clinical cross-reactivity has been reported as crustacean-mite-insect syndrome [20, 38, 40, 42, 70-72], and we report here the most likely allergens involved. Furthermore, this study identified for the first time very likely allergens, responsible for possible cross-reactivity between shrimp and fish. Beta-enolase and aldolase A, enzymatic proteins that play a role in the glycolytic pathway [15], have previously been identified as heat labile allergens in various fish species and also chicken [73, 74]. Our findings implicate the importance of both proteins as strong candidate allergens in shrimps. In addition, other proteins that are identified to be candidate allergens include cyclophilin and G3PD. Cyclophilin allergen is generally found in fungi, plants and dust mites, and has been shown to have high rates of IgE-binding [60, 75]. Meanwhile, G3PD, an enzymatic protein that is involved in the process of glycolysis similar to aldolase A and beta-enolase, has been identified as allergens in wheat [76], and recently in cockroach and fish [25].

In conclusion, this study accomplished the comparative analyses of all known shrimp allergens derived from five different shrimp species' transcriptomes assembled *de novo* from raw RNA-Seq data. The identification of previously known shrimp allergens validated the comprehensive approach utilised in this study. Moreover, over 30 additional proteins known for their allergenic properties in mite, fungi, plants, insect and fish were identified as candidate shrimp allergens. These includes HSP70, alpha-tubulin, chymotrypsin, beta-enolase aldolase A, cyclophilin and G3PD, which were further identified as very-likely candidate allergens of shrimps. Further immunological

studies would be required to confirm clinical allergenicity in patients. The findings of this study will enable improved diagnostics for shrimp allergy and future therapeutics for this lifelong disease.

## **Materials and methods**

### **Sample selection**

Specimen of the five species of shrimps (*Litopenaeus vannamei*, *Penaeus monodon*, *Fenneropenaeus merguensis*, *Melicertus latisulcatus*, and *Metapenaeus endeavouri*) were supplied by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) based in Queensland, Australia. *L. vannamei* and *P. monodon* samples originated from aquaculture farms whilst the other three species were caught as part of the CSIRO Northern Prawn Fishery Surveys from the benthic trawls in the Gulf of Carpentaria, Australia [77]. The shrimps were immersed in an ice-seawater slurry for a few minutes immediately after being caught, to be euthanized. Species-specific reference material were utilised to identify the species of shrimps [78]. Muscle tissue was then removed and stored in RNeasy Lysis Buffer (Qiagen, Crawley, VIC, Australia) [79]. *P. monodon* samples were collected as described by Huerlimann et al (2018) [80]. Total RNA was extracted from the muscle tissue of three randomly selected adult shrimps of each of the five shrimp species (total of 15 samples) with an RNeasy Universal Extraction kit (QIAGEN) using manufacturer's instruction in an RNase-free laboratory [79]. RNA concentration, quality and purity was assessed using a Nanodrop UV spectrophotometer (Thermo Fisher Scientific) and Agilent Bioanalyzer (Agilent Technologies), before being selected for sequencing.

### **Illumina library preparation and RNA sequencing**

All 15 samples were sequenced via Illumina HiSeq® 2500 System (Illumina Australia and New Zealand, VIC, Australia). Before sequencing, samples were quality checked with the Bioanalyzer RNA 6000 nano reagent kit (Agilent); and Illumina libraries were prepared using the TruSeq Stranded mRNA Library Preparation Kit (Illumina) according to established protocols at the Australian Genome Research Facility (AGRF). The resulting libraries were checked again with the TapeStation DNA 1000 TapeScreen Assay (Agilent). Cluster generation was performed immediately before sequencing on a cBot with HiSeq® PE Cluster Kit v4 – cBot. The sequencing was conducted using a HiSeq® SBS Kit on a HiSeq® 2500, operating with HiSeq Control Software v2.2.68 and base-calling with RTA v1.18.66.3. Raw RNA-Seq short read data for all samples are freely available on NCBI under BioProject PRJNA482687.



### De novo transcriptome assembly and quality control

RNA-Seq reads for all 15 samples were corrected using the software Rcorrector (v1.0.2) [81]. Transcriptomes of all 15 samples were individually assembled from their RNA-Seq data, *de novo*. The assembly was carried out using Trinity (v2.4.0) [82, 83]. The quality of the *de novo* transcriptome assembly was assessed using TransRate (v1.0.3) [84] and BUSCO (Benchmarking Universal Single-Copy Orthologs) (v1.2) [85] using the arthropoda odb9 database [86]. The quality score, also known as the TransRate score, is a score between 0.0 – 1.0 that is obtained by multiplying the mean of individual contig scores by the proportion of read pairs (original sequencing reads) that supported the transcriptome [84, 87]. The results of BUSCO assessment are given in percentages of complete (C), fragmented (F) and missing (M) genes within the transcriptome [85]. Using *L. vannamei* as an example, stepwise methods of sample extraction, sequencing, *de novo* transcriptome assembly and quality check are summarised and schematically represented in Figure 1.A.

### Removal of inconclusive dataset

Using the Rcorrected reads in an Assembly and Alignment-Free (AAF) method to create a phylogeny [88], it was discovered that one replicate of *M. latisulcatus* grouped with *M. endeavouri* rather than with the other two replicates of *M. latisulcatus*. To confirm the potentially misidentified sample, the assembled transcriptome was BLAST searched against the other *M. latisulcatus* and *M. endeavouri* transcriptomes, where the potentially misidentified sample also showed more similarity to *M. endeavouri*. Lastly, the transcriptomes were compared to known sequences of Enolase [89], which also confirmed that the misidentified sample is not *M. latisulcatus*.

### Allergen reference database construction

Known allergen AA sequences were retrieved from two reputable and peer-reviewed online databases to construct a reference allergen database for this study. The first is the World Health Organization & International Union of Immunological Societies (WHO/IUIS) Allergen Nomenclature database ([www.allergen.org](http://www.allergen.org)) [25]. The second is the AllergenOnline: Home of the FARRP (Food Allergy Research and Resource Program) Allergen Protein database (v.17) ([www.allergenonline.org](http://www.allergenonline.org)) [26, 27]. At the time of retrieval, the WHO/IUIS Allergen Nomenclature database contained 875 allergen AA sequences while the AllergenOnline database contained 2,035 allergen

AA sequences [25-27]. After removing duplicates between the 2 databases, a total of 2,172 allergen AA sequences were compiled to form the reference allergen database for this study.

### BLAST search for allergens

The allergen database and the assembled transcripts for all 15 samples were imported into the Geneious™ software (v8.1.9, Biomatters Limited, USA) [90]. In order to compare and search for transcripts which contain similar sequences to the allergen sequences compiled in the allergen database, blastx searches were carried out using the BLAST (Basic Local Alignment Search Tool) module within the Geneious™ software. The criteria for the search conducted are shown in Supplementary Table 1.

### Refining the BLAST search results

The BLAST search results were filtered for matched sequences with a PI of 50% or more. Subject coverage (percentage of the allergen sequence that is covered by the matching transcript from the transcriptome) was manually calculated using the formula: Subject coverage = Sequence length / Subject length x 100%, where Sequence length is the length of the matched consensus sequence and the Subject length is the actual length of the allergen sequence from the constructed database. Results were then filtered again by selecting only sequences that have 90% or more subject coverage.

Duplicates of allergen sequences that aligned with contigs within the transcriptome were removed by keeping the top-matched allergen-transcript consensus sequence. The BLAST search results of 3 replicates of each species were then combined to form one list of allergens for every species and the duplicates (between replicates) were removed. Stepwise methods of allergen database construction and the processing of transcriptome data such as BLAST search, results refinement and removal of duplicates are schematically represented in Figure 1, using the three assembled transcriptome replicates of *L. vannamei* as an example.

### Analysing the BLAST search results

For each shrimp species, the matched allergen AA sequences were grouped into: 'Shellfish', 'Mites', 'Insects', 'Fungi', 'Plants', 'Fish', and 'Other', based on the organism that the allergen was documented in. The proportion of allergen sequences belonging

to each group were graphed into a pie chart using GraphPad Prism version 7.03 for Windows [91] to show their distribution amongst different groups of allergen sources.

Multiple sequence alignment was conducted on all the contigs/transcripts that matched tropomyosin allergen in all five transcriptomes with shellfish tropomyosin allergens' sequences (as reference). Mites' and cockroaches' tropomyosin allergen sequences were also included in the multiple sequence alignment that was conducted in Jalview2.1 using Clustal Omega [92]. Comparative AA sequence identities were carried out between the contigs from all five shrimp species that matched with tropomyosin, and previously reported crustacean, mites, and cockroach tropomyosin allergens using Clustal Omega, EMBL-EBI [93]. The multiple sequence alignment and comparative sequence identities were carried out for other documented crustacean allergens: arginine kinase, myosin light chain, sarcoplasmic calcium-binding protein, troponin C, troponin I, and triosephosphate isomerase.

Non-crustacean allergens that have a PI value of more than 70% were shortlisted as highly likely candidates of unreported allergens in shrimp species. These unreported allergens were selected based on their match with the transcriptome of a minimum of 70% PI in at least one of the 5 shrimp species.

#### Measuring the abundance of allergen sequences

Abundance of each transcript/contigs within the transcriptomes, in transcript-per-million (TPM) values, was quantified using Salmon software [94]. Briefly, Salmon is a software that estimates the abundance of each contig by measuring the number of reads from the RNA-Seq data that align to the contig being measured [94]. Abundance estimation values for all known crustacean allergens were retrieved from all 15 samples. For each allergen in each sample, the estimated abundance value is the sum of all TPM values of all the contigs that matched with that allergen. The mean TPM values with standard deviation error bars for each allergen of the three replicates for each shrimp species are graphically represented in Figure 7 and 8. Standard deviation error bars were omitted from *M. latisulcatus* samples as only 2 replicates were investigated in this study. We first analysed the difference in abundance of all contigs representing a specific allergen, between the 5 shrimp species (Figure 7). In order to look for significant differences between two contigs representing the same allergen, we used unpaired *T-test* using GraphPad Prism version 7.03 for Windows [91]. Next,

we analysed the difference in abundance of allergens within each shrimp species (Figure 8). For this analyses, we only took into account the contig with the highest abundance, when there are more than one contig representing one allergen. To analyse significant differences between the seven crustacean allergens' abundance, we used One-way ANOVA test using GraphPad Prism version 7.03 for Windows [91].

#### *Molecular phylogenetic tree building of TM, AK, MLC and SCBP*

Published AA sequences of the four widely studied crustacean allergens, TM, AK, MLC and SCBP belonging to edible crustacean and mollusc species; and allergy causing mite and insect species were mined from NCBI Genbank and UniProt databases. The proteins which are not registered as an allergen in WHO/IUIS or AllergenOnline databases were also included. To determine the evolutionary distance between the same proteins from different species, molecular phylogenetic trees for each protein was built using MEGA X software (v10.0.5). The trees were constructed using the neighbour-joining method with the Poisson correction model. Hence, the branch lengths are the proportion of AA substitutions per site. Bootstrap test was also included (10,000 replicates) and the percentages are shown next to the branches. The gaps which occurred in alignment were treated as pairwise deletion.

### **Authors' contributions**

EJ and AL conceptualized the research objectives. SK, RH, EJ, RN, AT, and AL designed the research methods. RH, NW, and DJ provided the shrimp samples, conducted the RNA extraction and funded the RNA sequencing. SK and RH carried out the *de novo* transcriptome assembly, transcriptome quality check, BLAST search and relative abundance estimation. SK and RN constructed the allergen reference database and the molecular phylogenetic tree of known shellfish allergens. SK, RN, EJ, TR, AT, SDK, and AL determined the relevant allergens identified within the shrimp transcriptomes and the downstream analyses of these potential allergens. SK, RN, TR, RH, and AL designed the figures and tables. SK and AL wrote the first draft. All authors contributed to manuscript editing and revision.

### **Competing interests**

The authors have declared no competing interests.

## References

- [1] Gupta RS, Warren CM, Smith BM, Jiang J, Blumenstock JA, Davis MM, et al. Prevalence and Severity of Food Allergies Among US Adults. *JAMA Network Open* 2019;2:e185630-e.
- [2] Tang MLK, Mullins RJ. Food allergy: is prevalence increasing? *Internal Medicine Journal* 2017;47:256-61.
- [3] Muraro A, Werfel T, Hoffmann-Sommergruber K, Roberts G, Beyer K, Bindslev-Jensen C, et al. EAACI Food Allergy and Anaphylaxis Guidelines: diagnosis and management of food allergy. *Allergy* 2014;69:1008-25.
- [4] Rahman AMA, Helleur RJ, Jeebhay MF, Lopata AL. Characterization of seafood proteins causing allergic diseases. InTech, 2012.
- [5] Knol EF. Requirements for effective IgE cross-linking on mast cells and basophils. *Molecular Nutrition and Food Research* 2006;50:620-4.
- [6] Stone KD. IgE, mast cells, basophils, and eosinophils. *J Allergy Clin Immunol* 2010;125:S73-S80.
- [7] Pomes A, Davies JM, Gadermaier G, Hilger C, Holzhauser T, Lidholm J, et al. WHO/IUIS Allergen Nomenclature: Providing a common language. *Mol Immunol* 2018;100:3-13.
- [8] Lopata AL, Kleine-Tebbe J, Kamath SD. Allergens and molecular diagnostics of shellfish allergy: Part 22 of the Series Molecular Allergology. *Allergo Journal* 2016;25:24-32.
- [9] Lee AJ, Gerez I, Shek LP-C, Lee BW. Shellfish allergy - an Asia-Pacific perspective. *Asian Pacific Journal of Allergy and Immunology* 2012;30:3-10.
- [10] Lee AJ, Thalayasingam M, Lee BW. Food allergy in Asia: how does it compare? *Asia Pacific allergy* 2013;3:3.
- [11] Le TTK, Tran TTB, Ho HTM, Vu ATL, Lopata AL. Prevalence of food allergy in Vietnam: comparison of web-based with traditional paper-based survey. *World Allergy Organization Journal* 2018;11:16.
- [12] Lopata AL, O'Hehir RE, Lehrer SB. Shellfish allergy. *Clinical & Experimental Allergy* 2010;40:850-8.
- [13] Sampson HA. Food anaphylaxis. *British Medical Bulletin* 2000;56:925-35.
- [14] Thalayasingam M, Gerez IFA, Yap GC, Llanora GV, Chia IP, Chua L, et al. Clinical and immunochemical profiles of food challenge proven or anaphylactic shrimp allergy in tropical Singapore. *Clinical & Experimental Allergy* 2015;45:687-97.
- [15] Ruethers T, Taki AC, Johnston EB, Nugraha R, Le TTK, Kalic T, et al. Seafood allergy: A comprehensive review of fish and shellfish allergens. *Mol Immunol* 2018;100:28-57.
- [16] Kamath SD, Johnston EB, Iyer S, Schaeffer PM, Koplin J, Allen K, et al. IgE reactivity to shrimp allergens in infants and their cross-reactivity to house dust mite. *Pediatr Allergy Immunol* 2017;28:703-7.
- [17] López-Matas MA, de Larramendi CH, Moya R, Sánchez-Guerrero I, Ferrer A, Huertas AJ, et al. In vivo diagnosis with purified tropomyosin in mite and shellfish allergic patients. *Annals of Allergy, Asthma & Immunology* 2016;116:538-43.
- [18] Popescu FD. Cross-reactivity between aeroallergens and food allergens. *World J Methodol* 2015;5:31-50.
- [19] Wang J, Calatroni A, Visness CM, Sampson HA. Correlation of specific IgE to shrimp with cockroach and dust mite exposure and sensitization in an inner-city population. *Journal of Allergy and Clinical Immunology* 2011;128:834-7.
- [20] Santos ABR, Chapman MD, Aalberse RC, Vailes LD, Ferriani VPL, Oliver C, et al. Cockroach allergens and asthma in Brazil: Identification of tropomyosin as a major allergen with potential cross-reactivity with mite and shrimp allergens. *The Journal of Allergy and Clinical Immunology* 1999;104:329-37.
- [21] Sampson HA. Food allergy. Part 2: Diagnosis and management. *Journal of Allergy and Clinical Immunology* 1999;103:981-9.
- [22] Lopata AL, Lehrer SB. New insights into seafood allergy. *Current Opinion in Allergy and Clinical Immunology* 2009;9:270-7.

- [23] Arlian LG, Morgan MS, Vyszynski-Moher DL, Sharra D. Cross-reactivity between storage and dust mites and between mites and shrimp. *Experimental and Applied Acarology* 2009;47:159-72.
- [24] Ayuso R, Lehrer SB, Reese G. Identification of Continuous, Allergenic Regions of the Major Shrimp Allergen Pen a 1 (Tropomyosin). *International Archives of Allergy and Immunology* 2002;127:27-37.
- [25] (WHO/IUIS) WHOaIUoIS. Allergen Nomenclature. <http://allergen.org/> last accessed).
- [26] Goodman RE, Ebisawa M, Ferreira F, Sampson HA, Ree R, Vieths S, et al. AllergenOnline: A peer-reviewed, curated allergen database to assess novel food proteins for potential cross-reactivity. *Molecular Nutrition & Food Research* 2016;60:1183-98.
- [27] (FARRP). Allergen Online: Home of the FARRP (Food Allergy Research and Resource Program) Allergen Protein Database. <http://www.allergenonline.org/databasebrowse.shtml> last accessed).
- [28] Kamath SD, Rahman AMA, Voskamp A, Komoda T, Rolland JM, O'Hehir RE, et al. Effect of heat processing on antibody reactivity to allergen variants and fragments of black tiger prawn: A comprehensive allergenomic approach. *Molecular Nutrition & Food Research* 2014;58:1144-55.
- [29] Abdel Rahman AM, Kamath SD, Gagné S, Lopata AL, Helleur R. Comprehensive Proteomics Approach in Characterizing and Quantifying Allergenic Proteins from Northern Shrimp: Toward Better Occupational Asthma Prevention. *Journal of Proteome Research* 2013;12:647-56.
- [30] Abdel Rahman AM, Kamath S, Lopata AL, Helleur RJ. Analysis of the allergenic proteins in black tiger prawn (*Penaeus monodon*) and characterization of the major allergen tropomyosin using mass spectrometry. *Rapid Communications in Mass Spectrometry* 2010;24:2462-70.
- [31] Kamath SD, Rahman AMA, Komoda T, Lopata AL. Impact of heat processing on the detection of the major shellfish allergen tropomyosin in crustaceans and molluscs using specific monoclonal antibodies. *Food Chemistry* 2013;141:4031-9.
- [32] Nugraha R, Kamath SD, Johnston E, Zenger KR, Rolland JM, O'Hehir RE, et al. Rapid and comprehensive discovery of unreported shellfish allergens using large-scale transcriptomic and proteomic resources. *Journal of Allergy and Clinical Immunology* 2017.
- [33] Nugraha R, Kamath SD, Johnston E, Karnaneedi S, Ruethers T, Lopata AL. Conservation Analysis of B-Cell Allergen Epitopes to Predict Clinical Cross-Reactivity Between Shellfish and Inhalant Invertebrate Allergens. *Frontiers in Immunology* 2019;10:2676.
- [34] Lopata AL, Kleine-Tebbe J, Kamath SD. Allergens and molecular diagnostics of shellfish allergy. *Allergo Journal International* 2016;25:210-8.
- [35] Koeberl M, Kamath SD, Saptarshi SR, Smout MJ, Rolland JM, O'Hehir RE, et al. Auto-induction for high yield expression of recombinant novel isoallergen tropomyosin from King prawn (*Melicertus latisulcatus*) for improved diagnostics and immunotherapeutics. *Journal of Immunological Methods* 2014;415:6-16.
- [36] Breiteneder H, Chapman MD. Allergen nomenclature. *Allergens and Allergen Immunotherapy* 2014:37-49.
- [37] Aalberse RC. Structural biology of allergens. *J Allergy Clin Immunol* 2000;106:228-38.
- [38] Ayuso R, Reese G, Leong-Kee S, Plante M, Lehrer SB. Molecular Basis of Arthropod Cross-Reactivity: IgE-Binding Cross-Reactive Epitopes of Shrimp, House Dust Mite and Cockroach Tropomyosins. *International Archives of Allergy and Immunology* 2002;129:38-48.
- [39] Faber MA, Pascal M, El Kharbouchi O, Sabato V, Hagendorens MM, Decuyper II, et al. Shellfish allergens: Tropomyosin and beyond. *Allergy: European Journal of Allergy and Clinical Immunology* 2017.
- [40] Gámez C, Zafra MP, Boquete M, Sanz V, Mazzeo C, Ibáñez MD, et al. New shrimp IgE-binding proteins involved in mite-seafood cross-reactivity. *Molecular Nutrition & Food Research* 2014;58:1915-25.
- [41] Reese G, Ayuso R, Lehrer SB. Tropomyosin: An Invertebrate Pan-Allergen. *International Archives of Allergy and Immunology* 1999;119:247-58.
- [42] Wong L, Huang CH, Lee BW. Shellfish and House Dust Mite Allergies: Is the Link Tropomyosin? *Allergy, Asthma & Immunology Research* 2016;8:101-6.
- [43] Abdel Rahman AM, Kamath SD, Lopata AL, Robinson JJ, Helleur RJ. Biomolecular characterization of allergenic proteins in snow crab (*Chionoecetes opilio*) and de novo sequencing of

- the second allergen arginine kinase using tandem mass spectrometry. *Journal of Proteomics* 2011;74:231-41.
- [44] Binder M, Mahler V, Hayek B, Sperr WR, Scholler M, Prozell S, et al. Molecular and immunological characterization of arginine kinase from the Indianmeal moth, *Plodia interpunctella*, a novel cross-reactive invertebrate pan-allergen. *J Immunol* 2001;167:5470-7.
- [45] Liu Z, Xia L, Wu Y, Xia Q, Chen J, Roux KH. Identification and Characterization of an Arginine Kinase as a Major Allergen from Silkworm (*Bombyx mori*) Larvae. *International Archives of Allergy and Immunology* 2009;150:8-14.
- [46] Bobolea I, Barranco P, Pastor-Vargas C, Iraola V, Vivanco F, Quirce S. Arginine Kinase from the Cellar Spider (*Holocnemus pluchei*): A New Asthma-Causing Allergen. *International Archives of Allergy and Immunology* 2011;155:180-6.
- [47] Liu J, Han LN, Zhang Q, Wang QL, Chang Q, Zhuang H, et al. Cloning and molecular characterization of a myosin light chain gene from *Puccinia striiformis* f. sp. *tritici*. *World J Microbiol Biotechnol* 2014;30:631-7.
- [48] Funkenstein B, Skopal T, Rapoport B, Rebhan Y, Du SJ, Radaelli G. Characterization and functional analysis of the 5' flanking region of myosin light chain-2 gene expressed in white muscle of the gilthead sea bream (*Sparus aurata*). *Comp Biochem Physiol Part D Genomics Proteomics* 2007;2:187-99.
- [49] Ayuso R. Myosin light chain is a novel shrimp allergen, Lit v 3. *J Allergy Clin Immunol* 2008;122:795-802.
- [50] Zhang Y-X, Chen H-L, Maleki SJ, Cao M-J, Zhang L-J, Su W-J, et al. Purification, Characterization, and Analysis of the Allergenic Properties of Myosin Light Chain in *Procambarus clarkii*. *Journal of Agricultural and Food Chemistry* 2015;63:6271-82.
- [51] Ayuso R. Sarcoplasmic calcium-binding protein is an EF-hand-type protein identified as a new shrimp allergen. *J Allergy Clin Immunol* 2009;124:114-20.
- [52] Johnston EB, Kamath SD, Iyer SP, Pratap K, Karnaneedi S, Taki AC, et al. Defining specific allergens for improved component-resolved diagnosis of shrimp allergy in adults. *Molecular Immunology* 2019;112:330-7.
- [53] Bauermeister K, Wangorsch A, Garoffo LP, Reuter A, Conti A, Taylor SL, et al. Generation of a comprehensive panel of crustacean allergens from the North Sea Shrimp *Crangon crangon*. *Molecular Immunology* 2011;48:1983-92.
- [54] Kalyanasundaram A, Santiago TC. Identification and characterization of new allergen troponin C (Pen m 6.0101) from Indian black tiger shrimp *Penaeus monodon*. *European Food Research and Technology* 2015;240:509-15.
- [55] Chao E, Kim H-W, Mykles DL. Cloning and tissue expression of eleven troponin-C isoforms in the American lobster, *Homarus americanus*. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology* 2010;157:88-101.
- [56] Hindley J, Wunschmann S, Satinover SM, Woodfolk JA, Chew FT, Chapman MD, et al. Blat g 6: a troponin C allergen from *Blattella germanica* with IgE binding calcium dependence. *J Allergy Clin Immunol* 2006;117:1389-95.
- [57] Jeong KY, Kim CR, Un S, Yi MH, Lee IY, Park JW, et al. Allergenicity of recombinant troponin C from *Tyrophagus putrescentiae*. *Int Arch Allergy Immunol* 2010;151:207-13.
- [58] Kobayashi T, Takagi T, Konishi K, Cox JA. Amino acid sequence of crayfish troponin I. *J Biol Chem* 1989;264:1551-7.
- [59] Yang Y, Zhang YX, Liu M, Maleki SJ, Zhang ML, Liu QM, et al. Triosephosphate Isomerase and Filamin C Share Common Epitopes as Novel Allergens of *Procambarus clarkii*. *J Agric Food Chem* 2017;65:950-63.
- [60] An S, Chen L, Long C, Liu X, Xu X, Lu X, et al. *Dermatophagoides farinae* allergens diversity identification by proteomics. *Mol Cell Proteomics* 2013;12:1818-28.
- [61] Yang Y, Chen Z-W, Hurlburt BK, Li G-L, Zhang Y-X, Fei D-X, et al. Identification of triosephosphate isomerase as a novel allergen in *Octopus fangsiao*. *Molecular Immunology* 2017;85:35-46.
- [62] Rozynek P, Sander I, Appenzeller U, Cramer R, Baur X, Clarke B, et al. TPIS – an IgE-binding wheat protein. *Allergy* 2002;57:463-.



- [63] Vogel C, Marcotte EM. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature reviews. Genetics* 2012;13:227-32.
- [64] Ogburn RN, Randall TA, Xu Y, Roberts JH, Mebrahtu B, Karnuta JM, et al. Are dust mite allergens more abundant and/or more stable than other *Dermatophagoides pteronyssinus* proteins? *The Journal of allergy and clinical immunology* 2017;139:1030-2.e1.
- [65] Chan T-F, Ji K-M, Yim AK-Y, Liu X-Y, Zhou J-W, Li R-Q, et al. The draft genome, transcriptome, and microbiome of *Dermatophagoides farinae* reveal a broad spectrum of dust mite allergens. *Journal of Allergy and Clinical Immunology* 2015;135:539-48.
- [66] Yusuf N, Nasti TH, Huang CM, Huber BS, Jaleel T, Lin HY, et al. Heat shock proteins HSP27 and HSP70 are present in the skin and are important mediators of allergic contact hypersensitivity. *J Immunol* 2009;182:675-83.
- [67] Radauer C, Bublin M, Wagner S, Mari A, Breiteneder H. Allergens are distributed into few protein families and possess a restricted number of biochemical functions. *J Allergy Clin Immunol* 2008;121:847-52.e7.
- [68] Wang H, Lin J, Liu X, Liang Z, Yang P, Ran P, et al. Identification of alpha-tubulin, Der f 33, as a novel allergen from *Dermatophagoides farinae*. *Immunobiology* 2016;221:911-7.
- [69] Jeong KY, Son M, Lee JH, Hong CS, Park JW. Allergenic Characterization of a Novel Allergen, Homologous to Chymotrypsin, from German Cockroach. *Allergy Asthma Immunol Res* 2015;7:283-9.
- [70] Rosenfield L, Tsoulis MW, Milio K, Schnittke M, Kim H. High rate of house dust mite sensitization in a shrimp allergic southern Ontario population. *Allergy, Asthma and Clinical Immunology* 2017;13.
- [71] Villalta D, Tonutti E, Visentini D, Bizzaro N, Roncarolo D, Amato S, et al. Detection of a novel 20 kDa shrimp allergen showing cross-reactivity to house dust mites. *European annals of allergy and clinical immunology* 2010;42:20.
- [72] Zhang H, Lu Y, Ushio H, Shiomi K. Development of sandwich ELISA for detection and quantification of invertebrate major allergen tropomyosin by a monoclonal antibody. *Food Chemistry* 2014;150:151-7.
- [73] Kuehn A, Codreanu-Morel F, Lehnert-Weber C, Doyen V, Gomez-Andre SA, Bienvenu F, et al. Cross-reactivity to fish and chicken meat - a new clinical syndrome. *Allergy* 2016;71:1772-81.
- [74] Kuehn A, Hilger C, Lehnert-Weber C, Codreanu-Morel F, Morisset M, Metz-Favre C, et al. Identification of enolases and aldolases as important fish allergens in cod, salmon and tuna: component resolved diagnosis using parvalbumin and the new allergens. *Clinical & Experimental Allergy* 2013;43:811-22.
- [75] Glaser AG, Limacher A, Fluckiger S, Scheynius A, Scapozza L, Cramer R. Analysis of the cross-reactivity and of the 1.5 A crystal structure of the *Malassezia sympodialis* Mala s 6 allergen, a member of the cyclophilin pan-allergen family. *Biochem J* 2006;396:41-9.
- [76] Sander I, Rozynek P, Rihs HP, van Kampen V, Chew FT, Lee WS, et al. Multiple wheat flour allergens and cross-reactive carbohydrate determinants bind IgE in baker's asthma. *Allergy* 2011;66:1208-15.
- [77] Kenyon RA, Ellis N, Donovan AG, van der Velde TD, Fry G, Tonks M, et al. (2015), 'An Integrated Monitoring Program for the Northern Prawn Fishery 2012–2015', *AFMA 2011/0811 Final Report*, CSIRO Oceans and Atmosphere, Brisbane, p. 200.
- [78] Grey (1983), 'A guide to the Australian penaeid prawns'.
- [79] Huerlimann R, Maes GE, Maxwell MJ, Mobli M, Launikonis BS, Jerry DR, et al. Multi-species transcriptomics reveals evolutionary diversity in the mechanisms regulating shrimp tail muscle excitation-contraction coupling. *Gene* 2020;752:144765.
- [80] Huerlimann R, Wade NM, Gordon L, Montenegro JD, Goodall J, McWilliam S, et al. De novo assembly, characterization, functional annotation and expression patterns of the black tiger shrimp (*Penaeus monodon*) transcriptome. *Scientific Reports* 2018;8:13553.
- [81] Song L, Florea L. Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. *GigaScience* 2015;4:48.
- [82] Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* 2013;8:1494-512.

- [83] Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 2011;29:644-52.
- [84] Smith-Unna RD, Bournsnel C, Patro R, Hibberd JM, Kelly S. TransRate: reference free quality assessment of de-novo transcriptome assemblies. *bioRxiv* 2015.
- [85] Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 2015;31:3210-2.
- [86] Zdobnov EM, Tegenfeldt F, Kuznetsov D, Waterhouse RM, Simão FA, Ioannidis P, et al. OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic acids research* 2017;45:D744-D9.
- [87] MacManes MD. The Oyster River Protocol: A Multi Assembler and Kmer Approach For de novo Transcriptome Assembly. *bioRxiv* 2017.
- [88] Fan H, Ives AR, Surget-Groba Y, Cannon CH. An assembly and alignment-free method of phylogeny reconstruction from next-generation sequencing data. *BMC Genomics* 2015;16:522.
- [89] Ma KY, Chan T-Y, Chu KH. Refuting the six-genus classification of *Penaeus* s.l. (Dendrobranchiata, Penaeidae): a combined analysis of mitochondrial and nuclear genes. *Zoologica Scripta* 2011;40:498-508.
- [90] 'Geneious', Biomatters Limited, USA.
- [91] 'GraphPad Prism', *GraphPad Software*, La Jolla California USA, USA.
- [92] Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 2009;25:1189-91.
- [93] 'Clustal Omega Multiple Sequence Alignment', EMBL-EBI, UK.
- [94] Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference. *Nature methods* 2017;14:417-9.

## **Figure legends**

**Figure 1: Schematic representation of (A) *de novo* transcriptome assembly and (B) transcriptomic analysis used in the identification of allergens in shrimps.** The example shown here is for *L. vannamei* species. LV1, LV2, and LV3 represents the 3 biological replicates of *L. vannamei* samples. 'n' value refer to the number of allergens identified in *L. vannamei*. A total of 40 allergens were identified.

**Figure 2: Total allergens identified from the transcriptomic analysis in each five shrimp species, distributed based on the matched allergen's source.** The distribution amongst different groups of allergen sources are shown in percentages and arranged in a descending order.

**Figure 3: A. Comparison of amino acid sequence identities of (1-7) contigs from five shrimp species that matched with tropomyosin (TM) allergen, (8-9) known shrimp TM allergen, and (10-12) house dust mite and cockroach TM allergen.** The sequence identities were calculated using multiple sequence alignment in Clustal Omega (EMBL-EBI). **B. Molecular phylogenetic tree based on published amino acid sequences of Tropomyosin (TM) from edible crustacean and mollusc species; and allergy causing mite and insect species.** The branches consist of UniProt ID/Genbank Accession ID, species name, and followed by common name in brackets. The numbers next to the branches indicate the bootstrap test percentage of 10,000 replicate trees.

**Figure 4: A. Comparison of amino acid sequence identities of (1-6) contigs from five shrimp species that matched with arginine kinase (AK) allergen (7-8) known shrimp AK allergen and (9-12) house dust mite and cockroach AK allergen.** The sequence identities were calculated using multiple sequence alignment in Clustal Omega (EMBL-EBI). **B. Molecular phylogenetic tree based on published amino acid sequences of Arginine kinase (AK) from edible crustacean and mollusc species; and allergy causing mite and insect species.** The branches consist of UniProt ID/Genbank Accession ID, species name, and followed by common name in brackets. The numbers next to the branches indicate the bootstrap test percentage of 10,000 replicate trees.

**Figure 5: A. Comparison of amino acid sequence identities of (1-5) contigs from five shrimp species that matched with myosin light chain (MLC) allergen, (6-8) known shrimp MLC allergen and (9-10) house dust mite and cockroach MLC allergen.** The sequence identities were calculated using multiple sequence alignment in Clustal Omega (EMBL-EBI). **B. Molecular phylogenetic tree based on published amino acid sequences of Myosin light chain (MLC) from edible crustacean and mollusc species; and allergy causing mite and insect species.** The branches consist of UniProt ID/Genbank Accession ID, species name, and followed by common name in brackets. The numbers next to the branches indicate the bootstrap test percentage of 10,000 replicate trees.

**Figure 6: A. Comparison of amino acid sequence identities of (1-9) contigs from five shrimp species that matched with sarcoplasmic calcium-binding protein (SCBP) allergen and (10-12) known shrimp SCBP allergen.** The sequence identities were calculated using multiple sequence alignment in Clustal Omega (EMBL-EBI). **B. Molecular phylogenetic tree based on published amino acid sequences of Sarcoplasmic calcium-binding protein (SCBP) from edible crustacean and mollusc species; and allergy causing mite and insect species.** The branches consist of UniProt ID/Genbank Accession ID, species name, and followed by common name in brackets. The numbers next to the branches indicate the bootstrap test percentage of 10,000 replicate trees.

**Figure 7: Abundance estimation values in transcript-per-million (TPM) for contigs in the 5 analysed shrimp species that matched with shrimp allergens.** **A:** tropomyosin, **B:** arginine kinase, **C:** myosin light chain, **D:** sarcoplasmic calcium-binding protein, **E:** troponin C, **F:** troponin I, and **G:** triosephosphate isomerase. *T*-test was employed to measure the significance of difference between two contigs from the same species, if present (\*:  $P \leq 0.05$ , \*\*:  $P \leq 0.01$ , \*\*\*:  $P \leq 0.001$ , \*\*\*\*:  $P \leq 0.0001$ )

**Figure 8: Abundance estimation values in transcript-per-million (TPM) for contigs that matched with shrimp allergens in the 5 analysed shrimp species.** **A:** *L. vannamei*, **B:** *P. monodon*, **C:** *F. merguensis*, **D:** *M. latisulcatus*, **E:** *M. endeavouri*. ANOVA test was employed to measure the significance of difference between the seven shrimp allergens. Only one contig with the highest Pairwise Identity with known shrimp allergens value was included where there was more than one contig for one allergen in each species. The contigs are arranged in descending order of on their abundance. Allergen abundance values with the same letter are not significantly different to each other.

**Table 1: Results of Trinity transcriptome assembly, TransRate, and BUSCO.** Shrimp species name (common name) and their 1-3 biological replicates are shown here with their transcriptomes' number of contigs and assembly size after assembly by Trinity. TransRate score and BUSCO scores (C: complete, F: fragmented, M: missing) of each transcriptome are also shown here.

**Table 2: List of unreported allergens identified that have a minimum of 70% pairwise identity value in at least one species.** List includes protein name, the common and scientific name of the allergen source, along with the allergen sequence's IUIS nomenclature. % Pairwise identity and E-values. Proteins with a % Pairwise identity of 70% or higher (highly likely to be allergenic) are highlighted in red.

## **Supplementary material**

**Supplementary Figure 1:** Comparison of amino acid sequence identities of (1-7) contigs from five shrimp species that matched with Troponin C (TNC) allergen, (8-9) known shrimp TNC allergen, and (10-14) cockroach and storage mite TNC allergen. The sequence identities were calculated using multiple sequence alignment in Clustal Omega (EMBL-EBI).

**Supplementary Figure 2:** Comparison of amino acid sequence identities of (1-5) contigs from five shrimp species that matched with Troponin I (TNI) allergen and (6) known crayfish TNI allergen. The sequence identities were calculated using multiple sequence alignment in Clustal Omega (EMBL-EBI).

**Supplementary Figure 3:** Comparison of amino acid sequence identities of (1-5) contigs from five shrimp species that matched with triosephosphate isomerase (TIM) allergen, (6) known shrimp TIM allergen, and (7-8) house dust mite TIM allergen. The sequence identities were calculated using multiple sequence alignment in Clustal Omega (EMBL-EBI).

**Supplementary Figure 4:** Multiple sequence alignment of (1-2) known shrimp tropomyosin (TM) allergen, (3-9) contigs from five shrimp species that matched with TM allergen and (10-12) TM allergen sequences from house dust mite and cockroaches. Multiple sequence alignment was conducted in Jalview 2.1 using Clustal Omega.

**Supplementary Figure 5:** Multiple sequence alignment of (1-2) known shrimp arginine kinase (AK) allergen, (3-8) contigs from five shrimp species that matched AK allergen and (9-12) AK allergen sequences from house dust mites and cockroaches. Multiple sequence alignment was conducted in Jalview 2.1 using Clustal Omega.

**Supplementary Figure 6:** Multiple sequence alignment of (1-3) known shrimp myosin light chain (MLC) allergen, (4-8) contigs from five shrimp species that matched with MLC allergen and (9-10) house dust mite and cockroach MLC allergen. Multiple sequence alignment was conducted in Jalview 2.1 using Clustal Omega.

**Supplementary Figure 7:** Multiple sequence alignment of (1-3) known shrimp sarcoplasmic calcium-binding protein (SCBP) allergen and (4-12) contigs from five shrimp species that matched with SCBP allergen. Multiple sequence alignment was conducted in Jalview 2.1 using Clustal Omega.

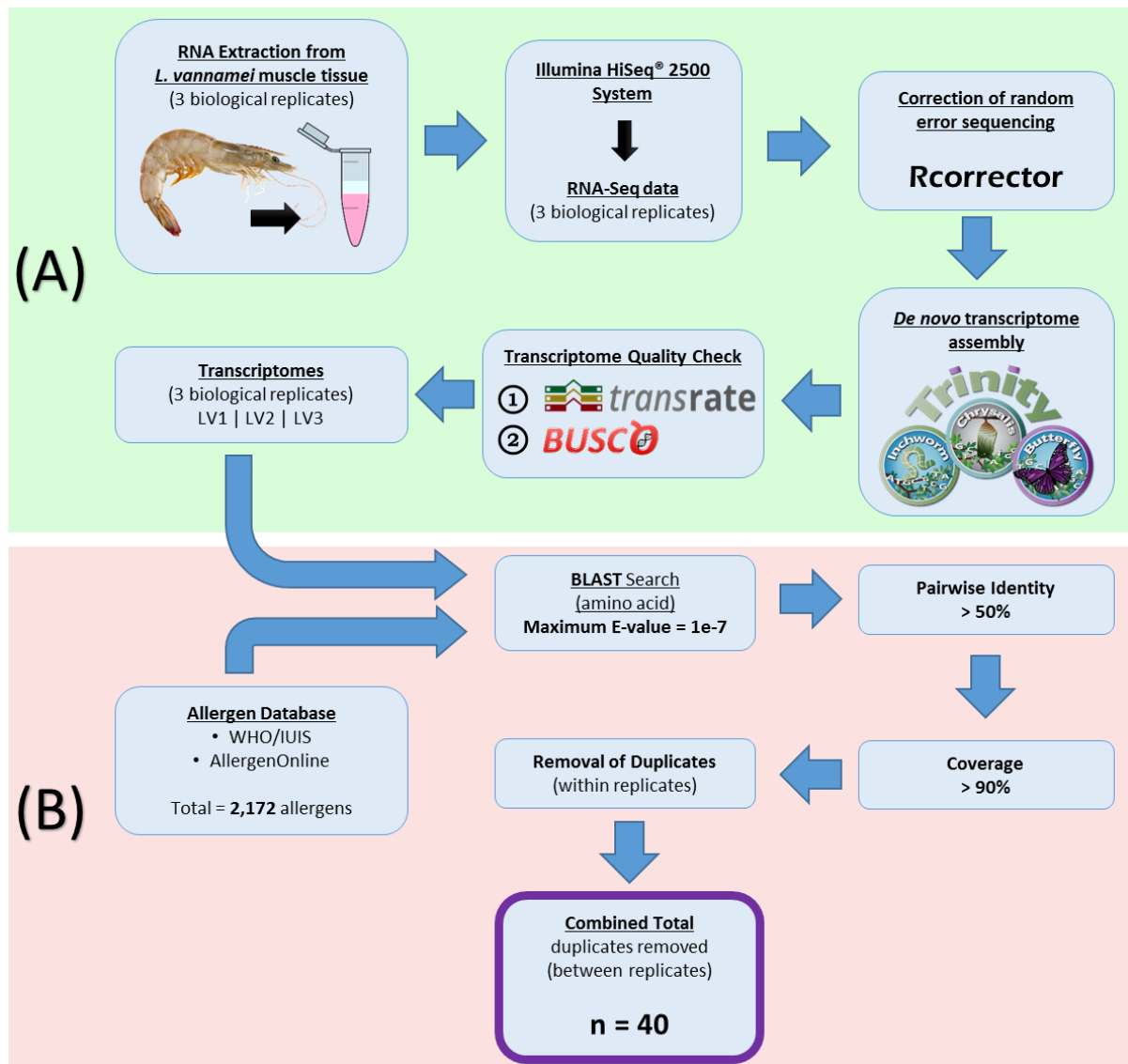
**Supplementary Figure 8:** Multiple sequence alignment of (1-2) known shrimp Troponin C (TNC) allergen, (3-9) contigs from five shrimp species that matched with TNC allergen and (10-14) TNC allergen sequences from house dust mites and cockroaches. Multiple sequence alignment was conducted in Jalview 2.1 using Clustal Omega.

**Supplementary Figure 9:** Multiple sequence alignment of (1) known crayfish Troponin I (TNI) allergen and (2-6) contigs from five shrimp species that matched with TNI allergen. Multiple sequence alignment was conducted in Jalview 2.1 using Clustal Omega.

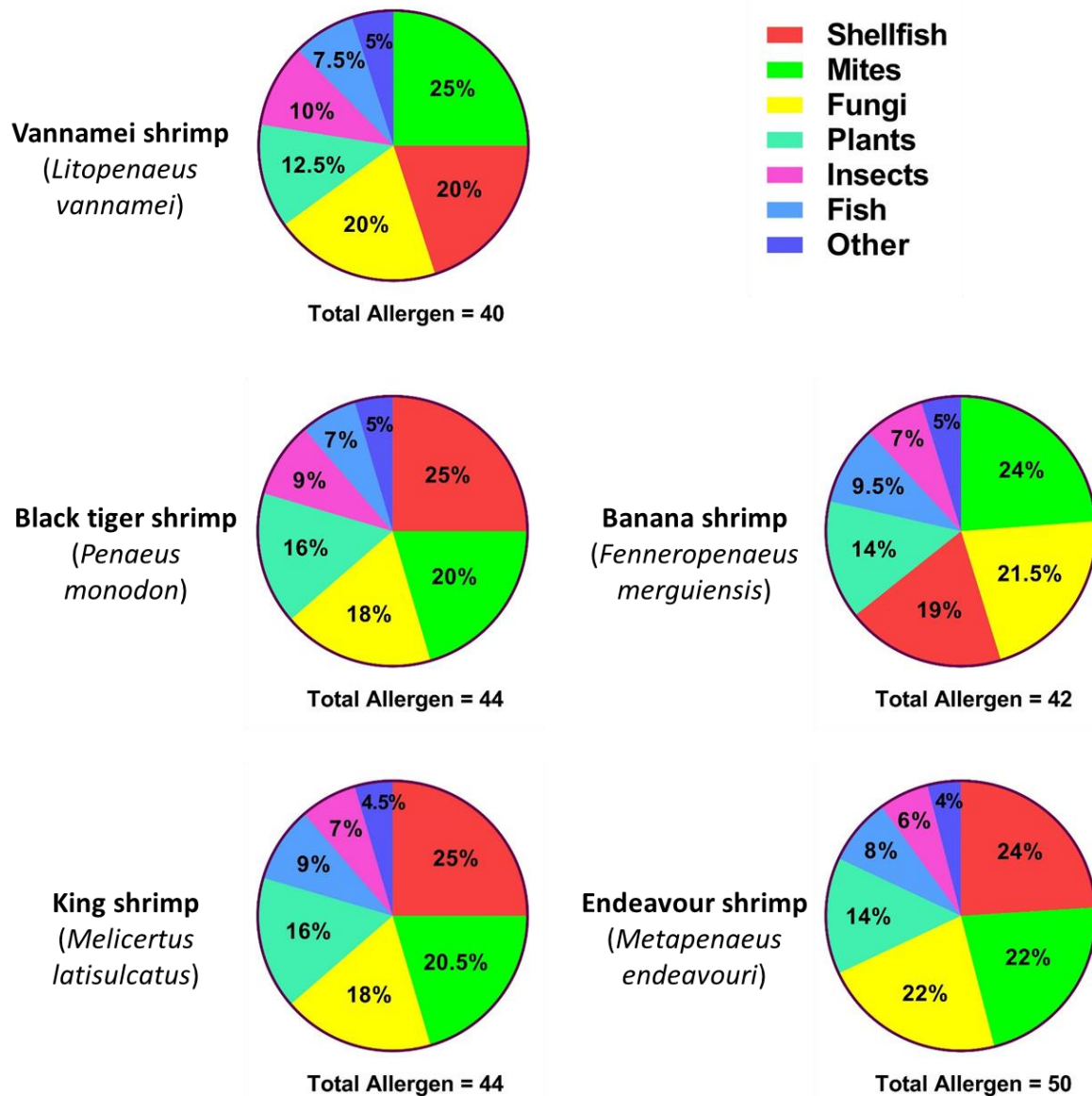
**Supplementary Figure 10:** Multiple sequence alignment of (1) known shrimp Triosephosphate isomerase (TIM) allergen, (2-6) contigs from five shrimp species that matched with TIM allergen and (7-8) TIM allergen sequences from house dust mites. Multiple sequence alignment was conducted in Jalview 2.1 using Clustal Omega.

**Supplementary Table 1: Criteria used in the BLAST search conducted.** The criteria shown here are only for the BLAST search utility within the Geneious™ software. Additional search criteria (for this project) was later used in the refining process of the search results, e.g. Minimum % Pairwise Identity of 50%.

## Figures



**Figure 1:** Schematic representation of (A) *de novo* transcriptome assembly and (B) transcriptomic analysis used in the identification of allergens in shrimps. The example shown here is for *L. vannamei* species. LV1, LV2, and LV3 represents the 3 biological replicates of *L. vannamei* samples. ‘n’ value refer to the number of allergens identified in *L. vannamei*. A total of 40 allergens were identified.



**Figure 2:** Total allergens identified from the transcriptomic analysis in each five shrimp species, distributed based on the matched allergen's source. The distribution amongst different groups of allergen sources are shown in percentages and arranged in a descending order.

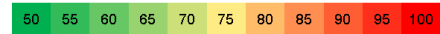


**A.**

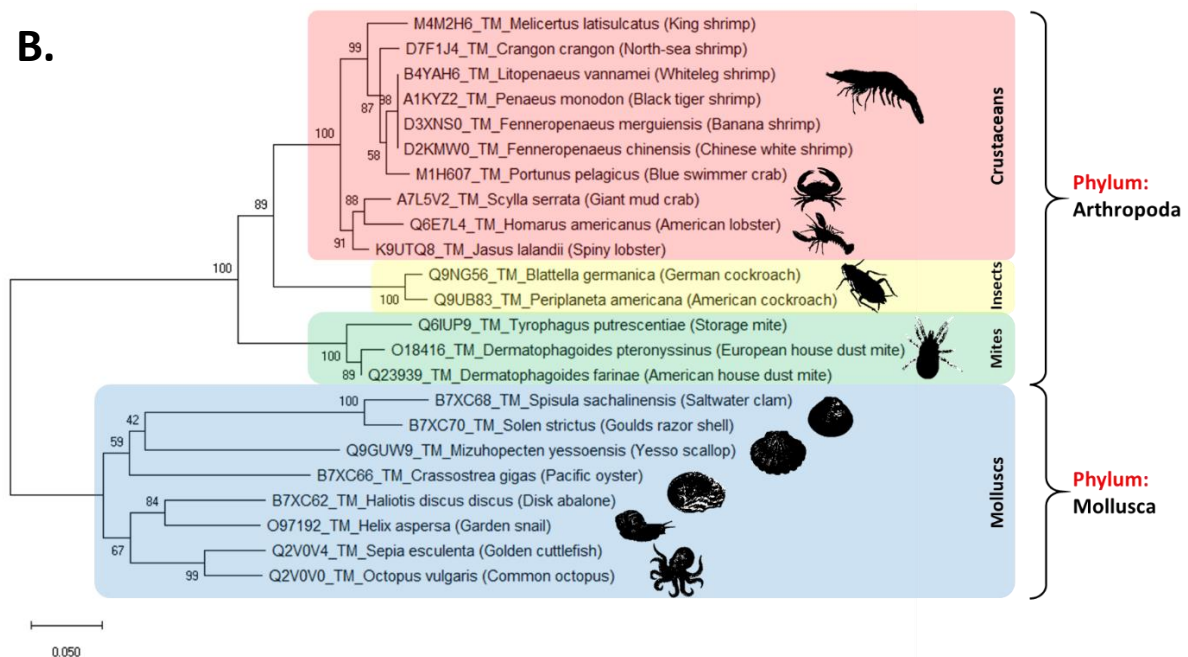
**Tropomyosin**

		Species	Contig ID/AccessionID	#	1	2	3	4	5	6	7	8	9	10	11	12
Contigs	Shrimp	<i>L_vannamei_TM_Contig_1</i>	TRINITY_DN8531_c1_g1_j23	1	100	100	91	100	100	99	82	100	100	81	82	82
		<i>P_monodon_TM_Contig_1</i>	TRINITY_DN9650_c0_g5_j8	2		100	91	100	100	99	82	100	100	81	82	82
		<i>P_monodon_TM_Contig_2</i>	TRINITY_DN10227_c1_g1_j21	3			100	91	91	91	87	91	91	78	81	81
		<i>F_merguensis_TM_Contig_1</i>	TRINITY_DN9629_c3_g1_j18	4				100	100	99	82	100	100	81	82	82
		<i>M_laticulatus_TM_Contig_1</i>	TRINITY_DN10005_c3_g3_j4	5					100	99	82	100	100	81	82	82
		<i>M_endeavouri_TM_Contig_1</i>	TRINITY_DN9495_c0_g1_j3	6						100	82	99	99	81	83	83
		<i>M_endeavouri_TM_Contig_2</i>	TRINITY_DN9495_c0_g1_j16	7							100	82	82	72	74	74
Allergen	Shrimp	<i>L_vannamei_TM_[Lit_v_1]</i>	ACB38288	8								100	100	81	82	82
		<i>P_monodon_TM_[Pen_m_1]</i>	AAX37288	9									100	81	82	82
	HDM	<i>D_pteronysinus_TM_[Der_p_10]</i>	AAB69424	10										100	80	80
	Cockroach	<i>B_germanica_TM_[Bla_g_7]</i>	AAF72534	11											100	97
		<i>P_americana_TM_[Per_a_7]</i>	CAB38086	12												

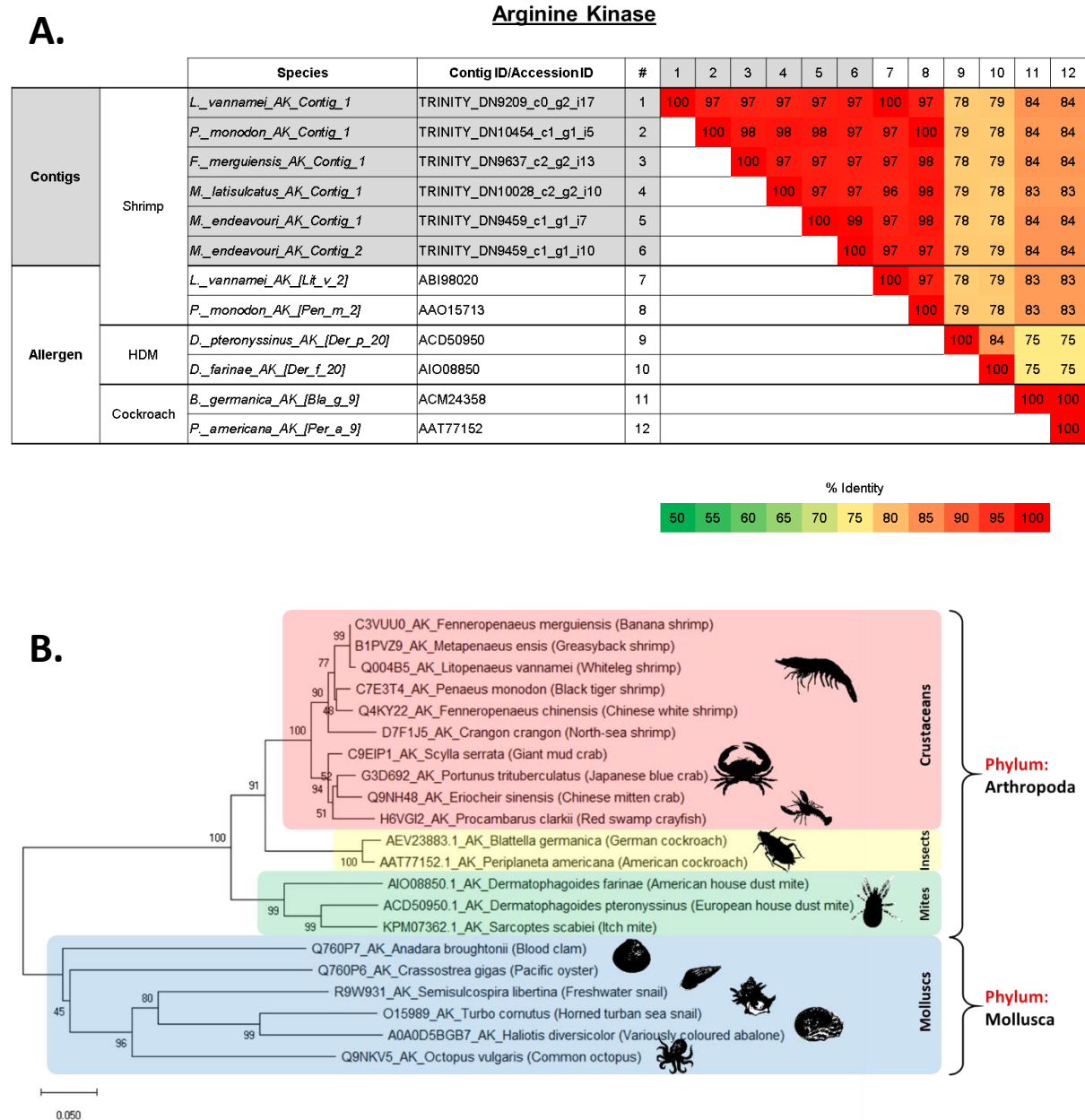
% Identity



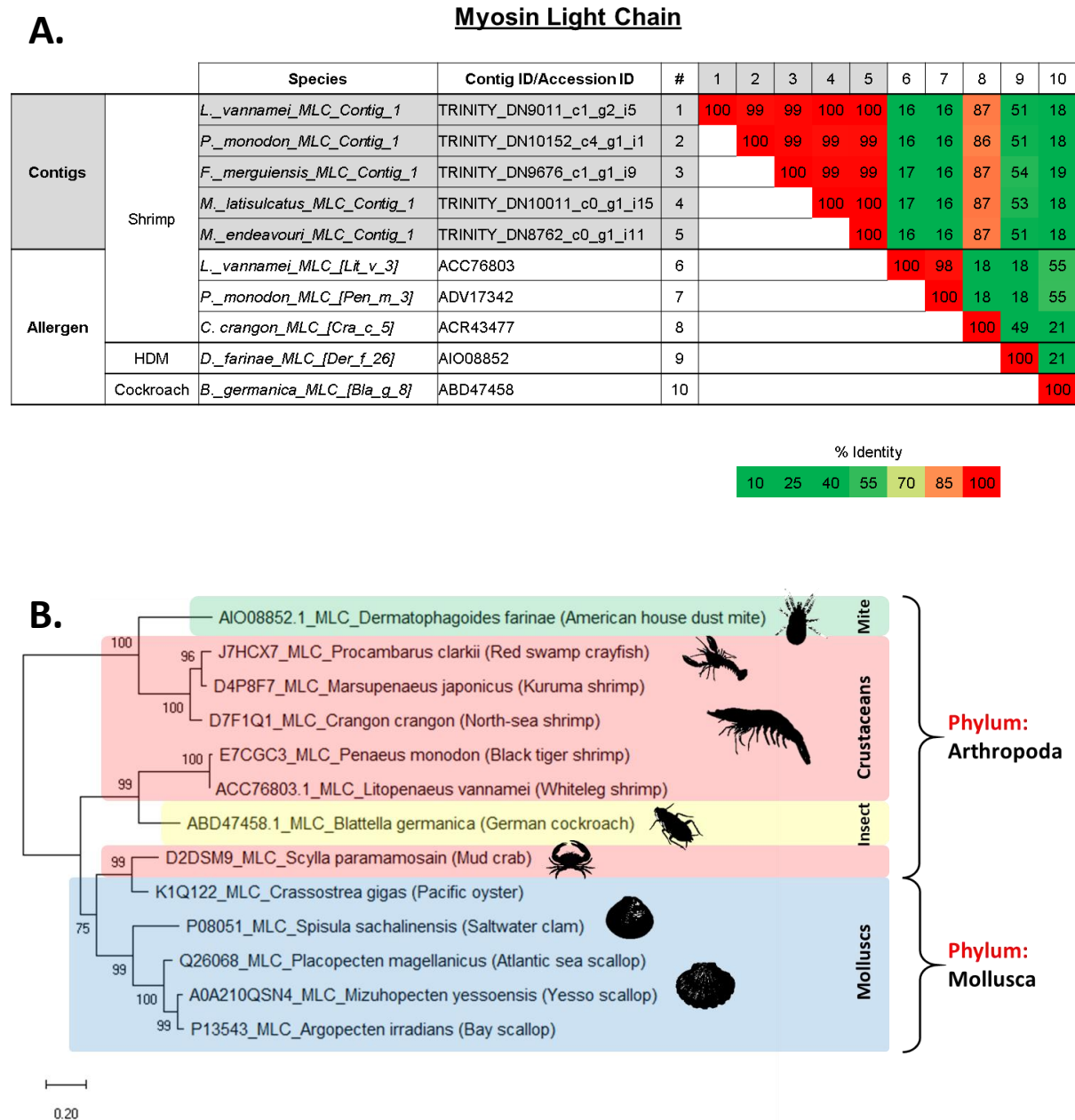
**B.**



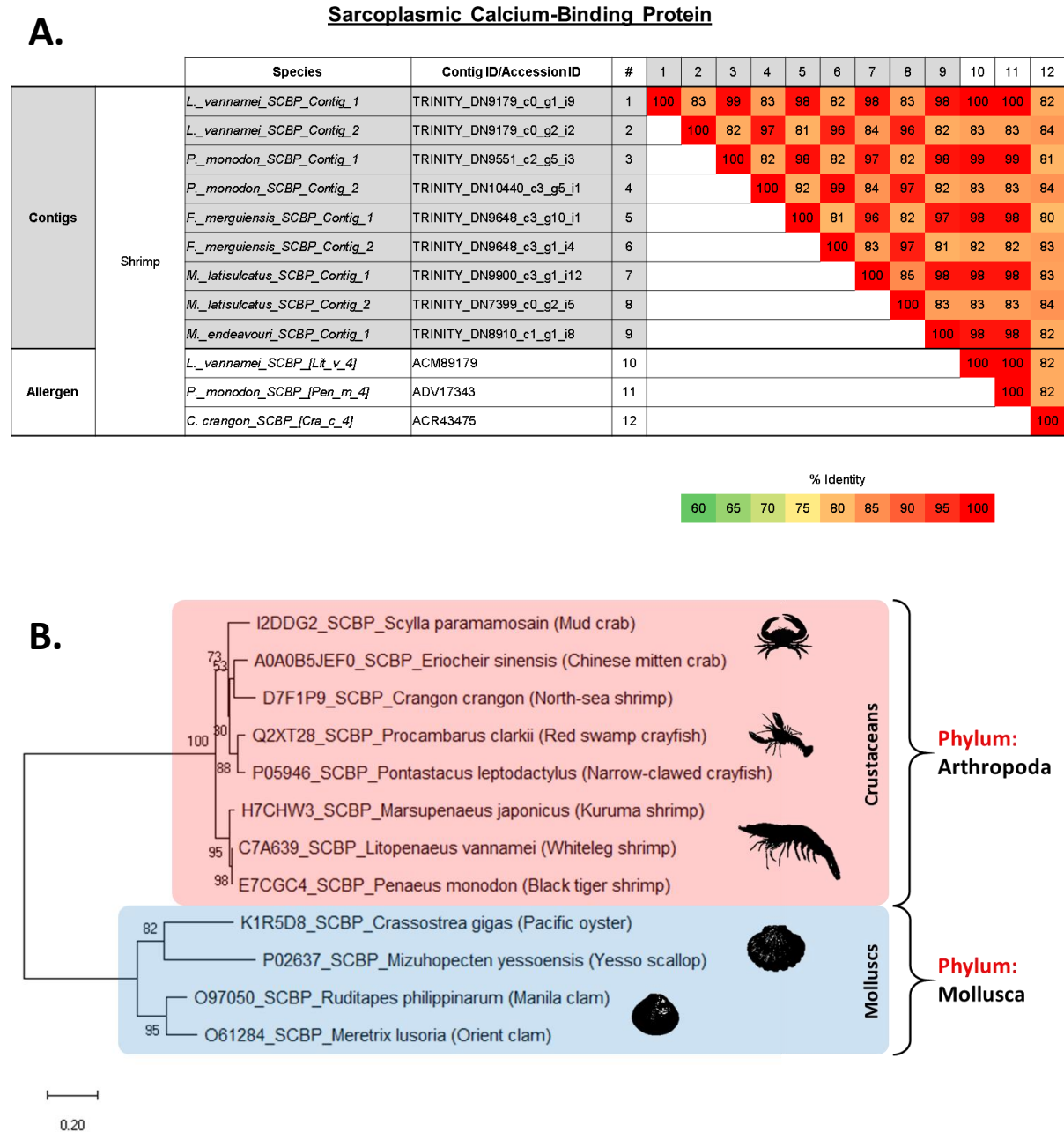
**Figure 3: A.** Comparison of amino acid sequence identities of (1-7) contigs from five shrimp species that matched with tropomyosin (TM) allergen, (8-9) known shrimp TM allergen, and (10-12) house dust mite and cockroach TM allergen. The sequence identities were calculated using multiple sequence alignment in Clustal Omega (EMBL-EBI). **B.** Molecular phylogenetic tree based on published amino acid sequences of Tropomyosin (TM) from edible crustacean and mollusc species; and allergy causing mite and insect species. The branches consist of UniProt ID/Genbank Accession ID, species name, and followed by common name in brackets. The numbers next to the branches indicate the bootstrap test percentage of 10,000 replicate trees.



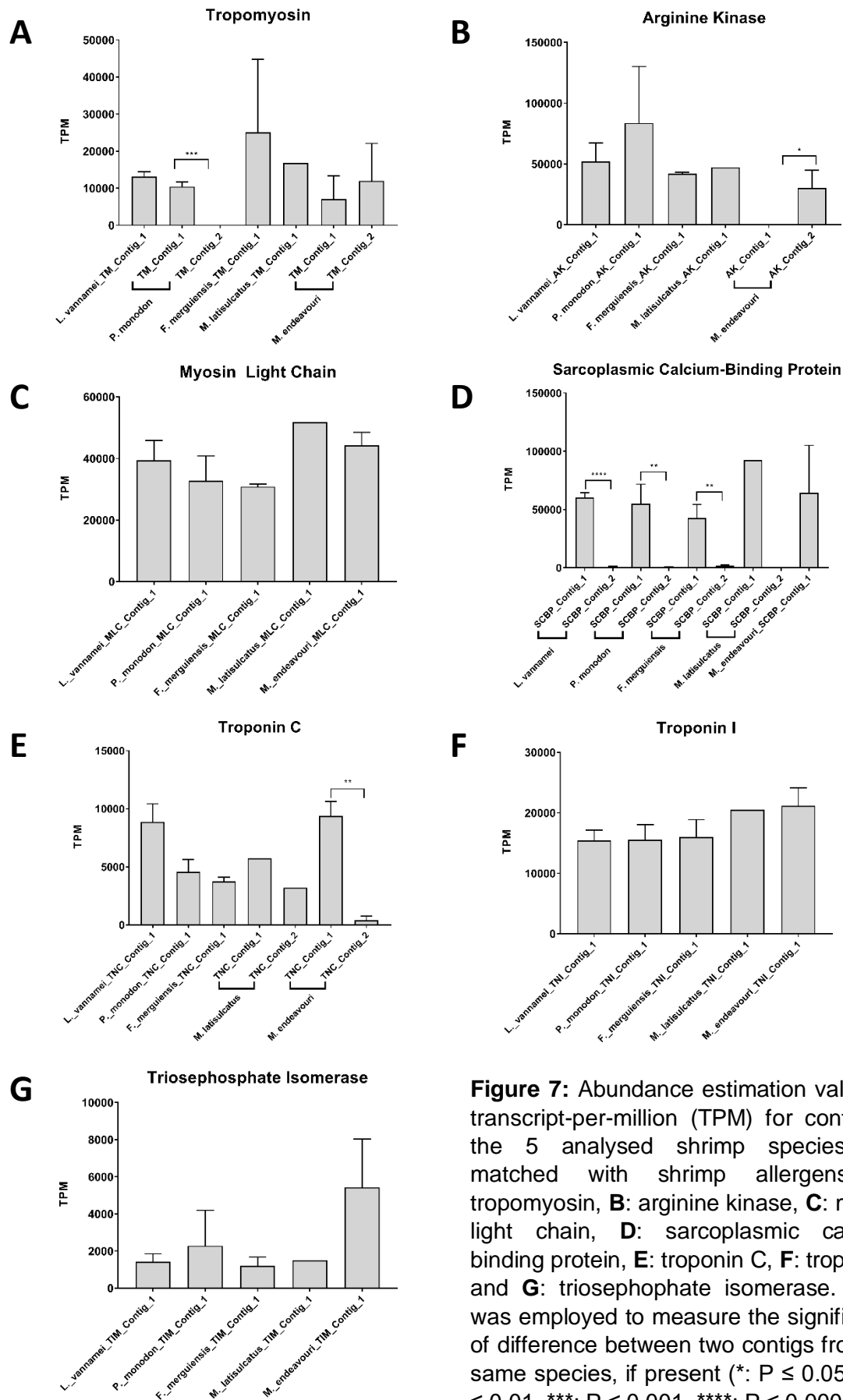
**Figure 4: A.** Comparison of amino acid sequence identities of (1-6) contigs from five shrimp species that matched with arginine kinase (AK) allergen (7-8) known shrimp AK allergen and (9-12) house dust mite and cockroach AK allergen. The sequence identities were calculated using multiple sequence alignment in Clustal Omega (EMBL-EBI). **B.** Molecular phylogenetic tree based on published amino acid sequences of Arginine kinase (AK) from edible crustacean and mollusc species; and allergy causing mite and insect species. The branches consist of UniProt ID/Genbank Accession ID, species name, and followed by common name in brackets. The numbers next to the branches indicate the bootstrap test percentage of 10,000 replicate trees.



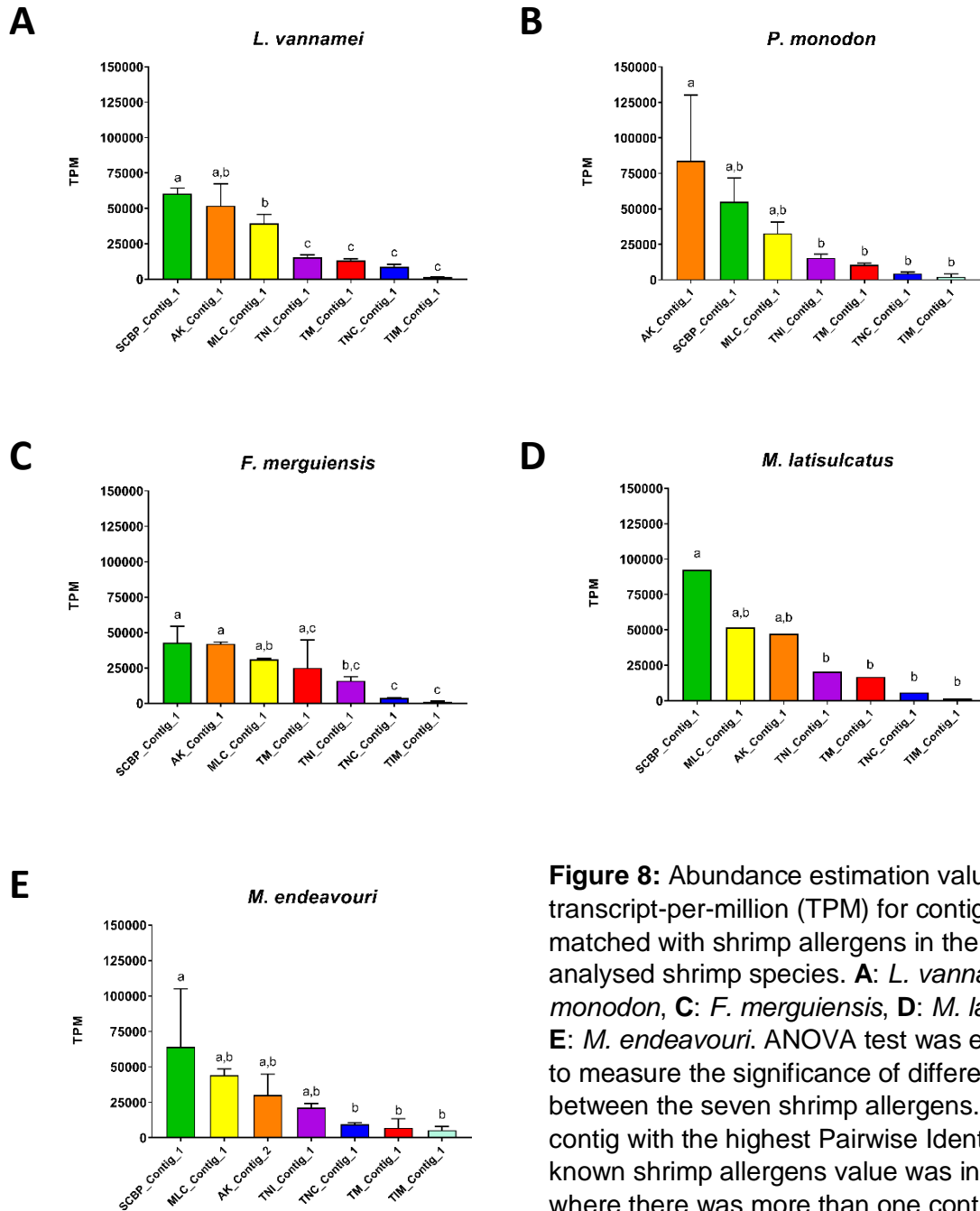
**Figure 5: A.** Comparison of amino acid sequence identities of (1-5) contigs from five shrimp species that matched with myosin light chain (MLC) allergen, (6-8) known shrimp MLC allergen and (9-10) house dust mite and cockroach MLC allergen. The sequence identities were calculated using multiple sequence alignment in Clustal Omega (EMBL-EBI). **B.** Molecular phylogenetic tree based on published amino acid sequences of Myosin light chain (MLC) from edible crustacean and mollusc species; and allergy causing mite and insect species. The branches consist of UniProt ID/Genbank Accession ID, species name, and followed by common name in brackets. The numbers next to the branches indicate the bootstrap test percentage of 10,000 replicate trees.



**Figure 6: A.** Comparison of amino acid sequence identities of (1-9) contigs from five shrimp species that matched with sarcoplasmic calcium-binding protein (SCBP) allergen and (10-12) known shrimp SCBP allergen. The sequence identities were calculated using multiple sequence alignment in Clustal Omega (EMBL-EBI). **B.** Molecular phylogenetic tree based on published amino acid sequences of Sarcoplasmic calcium-binding protein (SCBP) from edible crustacean and mollusc species; and allergy causing mite and insect species. The branches consist of UniProt ID/Genbank Accession ID, species name, and followed by common name in brackets. The numbers next to the branches indicate the bootstrap test percentage of 10,000 replicate trees.








**Figure 7:** Abundance estimation values in transcript-per-million (TPM) for contigs in the 5 analysed shrimp species that matched with shrimp allergens. **A:** tropomyosin, **B:** arginine kinase, **C:** myosin light chain, **D:** sarcoplasmic calcium-binding protein, **E:** troponin C, **F:** troponin I, and **G:** triosephosphate isomerase. *T*-test was employed to measure the significance of difference between two contigs from the same species, if present (\*:  $P \leq 0.05$ , \*\*:  $P \leq 0.01$ , \*\*\*:  $P \leq 0.001$ , \*\*\*\*:  $P \leq 0.0001$ )



**Figure 8:** Abundance estimation values in transcript-per-million (TPM) for contigs that matched with shrimp allergens in the 5 analysed shrimp species. **A:** *L. vannamei*, **B:** *P. monodon*, **C:** *F. merguensis*, **D:** *M. latisulcatus*, **E:** *M. endeavouri*. ANOVA test was employed to measure the significance of difference between the seven shrimp allergens. Only one contig with the highest Pairwise Identity with known shrimp allergens value was included where there was more than one contig for one allergen in each species. The contigs are arranged in descending order of on their abundance. Allergen abundance values with the same letter are not significantly different to each other.

## Tables

**Table 1:** Results of Trinity transcriptome assembly, TransRate, and BUSCO. Shrimp species name (common name) and their 1-3 biological replicates are shown here with their transcriptomes' number of contigs and assembly size after assembly by Trinity. TransRate score and BUSCO scores (C: complete, F: fragmented, M: missing) of each transcriptome are also shown here.

Shrimp species		Replicates	RNA-Seq	Transcriptome assembly metrics			Transrate quality assessment		BUSCO scores		
			Normalized read count	No. of contigs	Assembly size	GC content (%)	Proportion of read pairs mapped (%)	Assembly score	Complete (%)	Fragmented (%)	Missing (%)
	<b><i>L. vannamei</i></b> (Whiteleg shrimp)	1	1,412,587	32,302	28.6Mb	43.4	93.2	0.413	56	21	23
		2	1,412,010	33,574	29.4Mb	43.0	92.6	0.401	56	23	21
		3	1,070,376	28,101	22.7Mb	44.8	92.8	0.419	48	25	27
	<b><i>P. monodon</i></b> (Black Tiger shrimp)	1	1,609,374	41,971	37.9Mb	44.3	91.9	0.387	66	20	14
		2	1,443,066	40,927	36.5Mb	45.1	91.0	0.364	66	19	14
		3	1,643,259	42,510	38.1Mb	43.7	92.3	0.390	64	21	14
	<b><i>F. merguensis</i></b> (Banana shrimp)	1	1,486,264	37,572	31.4Mb	43.0	91.8	0.385	64	17	19
		2	1,657,940	41,336	34.8Mb	42.6	91.7	0.385	67	16	17
		3	1,602,775	38,638	33.5Mb	42.5	92.6	0.389	65	19	16
	<b><i>M. latisulactus</i></b> (King shrimp)	1	1,130,898	37,128	25.6Mb	42.9	90.7	0.410	46	26	27
		2	1,052,237	28,125	21.7Mb	42.8	92.2	0.411	43	25	32
	<b><i>M. endeavouri</i></b> (Endeavour shrimp)	1	1,142,169	35,407	25.9Mb	42.5	90.6	0.374	48	25	27
		2	1,035,324	30,879	23.2Mb	42.3	91.2	0.399	48	24	27
		3	1,081,301	38,204	25.5Mb	43.3	87.9	0.355	49	26	25

**Table 2:** List of unreported allergens identified that have a minimum of 70% pairwise identity value in at least one species. List includes protein name, the common and scientific name of the allergen source, along with the allergen sequence's IUIS nomenclature. % Pairwise identity and E-values. Proteins with a % Pairwise identity of 70% or higher (highly likely to be allergenic) are highlighted in red.

Allergens				LV Whiteleg shrimp (E Value)	PM Black tiger shrimp (E Value)	FM Banana shrimp (E Value)	ML King shrimp (E Value)	ME Endeavour shrimp (E Value)
Protein name	Source name		IUIS nomen- clature					
	Common	Scientific						
Heat shock-like protein	Storage mite	<i>Tyrophagus putrescentiae</i>	Tyr p 28	85.1% (0)	82.7% (0)	83.3% (0)	82.7% (0)	84.3% (0)
Alpha-tubulin	American house dust mite	<i>Dermatophagoides farinae</i>	Der f 33	81.8% (0)	81.7% (0)	81.6% (0)	81.6% (0)	81.6% (0)
Chymotrypsin	American house dust mite	<i>Dermatophagoides farinae</i>	Der f 6	78.7% (4.3E-94)	78.7% (2.13E-94)	79.3% (1.45E-94)	79.9% (3.71E-97)	80.5% (3.93E-95)
Enolase 3-2	Atlantic salmon	<i>Salmo salar</i>	Sal s 2	74.8% (0)	74.6% (0)	74.1% (0)	74.6% (0)	74.5% (0)
Glyceral-dehyde-3-phosphate dehydrogenase	Wheat	<i>Triticum aestivum</i>	Tri a 34	72.3% (1.25E-168)	72.0% (2.87E-172)	71.7% (1.86E-170)	72.0% (1.21E-174)	72.3% (4.91E-175)
Cyclophilin	Common mould	<i>Aspergillus fumigatus</i>	Asp f 27	61.9% (8.45E-65)	62.5% (1.87E-67)	70.3% (2.63E-75)	70.7% (4.38E-75)	69.3% (5.53E-76)
Aldolase A	Yellowfin tuna	<i>Thunnus albacares</i>	Thu a 3	66.0% (2.29E-164)	64.9% (2.25E-164)	70.1% (4.09E-169)	69.6% (1.47E-166)	70.1% (3.65E-167)