Examining De Novo Transcriptome Assemblies via a Quality Assessment Pipeline

Noushin Ghaffari*+, Osama A. Arshad+, Hyundoo Jeong+, John Thiltges, Michael F. Criscitiello, Byung-Jun Yoon, Aniruddha Datta, Charles D. Johnson

Abstract—New *de novo* transcriptome assembly and annotation methods provide an incredible opportunity to study the transcriptome of organisms that lack an assembled and annotated genome. There are currently a number of *de novo* transcriptome assembly methods, but it has been difficult to evaluate the quality of these assemblies. In order to assess the quality of the transcriptome assemblies, we composed a workflow of multiple quality check measurements that in combination provide a clear evaluation of the assembly performance. We presented novel transcriptome assemblies and functional annotations for Pacific whiteleg shrimp (*Litopenaeus vannamei*), a mariculture species with great national and international interest, and no solid transcriptome/genome reference. We examined Pacific whiteleg transcriptome assemblies via multiple metrics, and provide an improved gene annotation. Our investigations show that assessing the quality of an assembly purely based on the assembler's statistical measurements can be misleading; we propose a hybrid approach that consists of statistical quality checks and further biological-based evaluations.

Index Terms— bioinformatics, Pacific whiteleg shrimp, Litopenaeus vannamei, transcriptome assembly, workflow evaluation

1 Introduction

In the past few years, the fast growth of the Next Generation Sequencing (NGS) technologies has enabled scientists to explore genomes and transcriptomes in depth. RNA-Seq experiments sequence expressed messenger RNA in high-throughput fashion. Historically a problem with all mRNA measurement methods (Northern blot, qPCR, and microarray) was the requirement of prior information about the species. There was a need to know all or a portion of the sequence for a given gene to develop a quantification method. With RNA-Seq this is no longer necessary to measure the mRNA transcripts, however, the transcript/gene information in the cases where there is not an established reference must be derived from the data, i. e. non-model species.

Pacific whiteleg shrimp, *Litopenaeus vannamei (L. vannamei)*, is a prawn native to the eastern Pacific Ocean from Sonoran Mexico south to northern Peru, and heavily farmed in the Untied States and Latin America. In 2009 the U.S. per capita consumption of shrimp was 4.1 pounds, a year when the U.S. had a four billion dollar shrimp trade deficit. *L. vannamei* is a decapod (e.g. crabs, lobster, shrimp) crustacean of great interest as the dominant shrimp species in the global aquaculture industry. Whiteleg shrimp has a great potential to provide food security, however, suffers from panademics caused by viruses. The most well known viruses that affect panaeid shrimp are Hematopoietic Necrosis Virus (IHHNV), Yellow Head Virus (YHV), Taura Syndrome Virus (TSV), and White Spot Syndrome Virus (WSSV) [1]. Shrimp are invertebrate arthropods and do not benefit from immunoglobulin superfamily-based adaptive immune system, unlike sharks and all other jawed vertebrates [2], [3], [4]. Shrimp only have innate mechanisms of immunity, which presumably lack the high specificity and memory of an adaptive system. Therefore, classical vaccination of shrimp is impossible (still not completely proven a successful methodology [5]). Some of the complex innate immune system components of shrimp have been characterized functionally and biochemically, including mechanisms of apoptosis, phagocytosis, Toll-like receptor signaling, anti-microbials, clotting cascades, and a prophenyloxidase activating system [6], [7], [8], [9], [10], [11],

- *N. G. is with AgriLife Genomics and Bioinformatics, Texas A&M AgriLife Research, Texas A&M University, College Station, TX 77845. E-mail: nghaffari@tamu.edu. The first three authors (+ sign indicated) contributed equally to this work. Asterisk indicates corresponding author.*
- *O. A. A. is with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77845. E-mail: oarshad@tamu.edu.*
- *H. J. is with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77845. E-mail: hyundoo@tamu.edu.*
- *J. T. is with AgriLife Genomics and Bioinformatics, Texas A&M AgriLife Research, Texas A&M University, College Station, TX 77845. E-mail: jthiltges@tamu.edu.*
- *M. F. C. is with the Comparative Immunogenetics Laboratory, Departments of Veterinary Pathobiology and Microbial Pathogenesis and Immunology, Texas A&M University, College Station, TX 77845. E-mail: mcriscitiello@cvm.tamu.edu.*
- *B.-J. Yoon is with the College of Science and Engineering, Hamad bin Khalifa University (HBKU), Doha, Qatar, and the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA. E-mail: byoon@qf.org.qa.*
- *A. D. is with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77845. E-mail: datta@ece.tamu.edu.*
- *C. D. J. is with AgriLife Genomics and Bioinformatics, Texas A&M AgriLife Research, Texas A&M University, College Station, TX 77845. E-mail: charlie@ag.tamu.edu.*

xxxx-xxxx/0x/$xx.00 © 200x IEEE Published by the IEEE Computer Society

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2446478, IEEE/ACM Transactions on Computational Biology and Bioinformatics

[12]. These devices are regulated by complex systems employing JAK-STAT signaling and RNAi pathways [13], [14]. However, a great deal yet remains to be discovered in the analysis of a complete genome and transcriptome of the shrimp.

In recent years, there have been studies on transcriptome assembly and annotation of the *L. vannamei* [15], SNP detection in the transcriptome [16], next-generation sequencing datasets and unigene assignments [17], transcriptomic response to pollutant exposure [18], multiple cDNA libraries [19], and responses to viral infections in shrimp [20], [21], [22], [23], [24], [25].

By advancement of the NGS techniques, RNA-Seq has facilitated the measurement of the gene-expression level of thousands of genes simultaneously [26]. RNA-Seq applications include discovery of new splice junctions [27], prediction of absolute copy-number variation (CNV) [28], detecting single-nucleotide polymorphisms (SNP) [29], [30], and transcriptome assembly [31]. The transcriptome assembly methods reconstruct the transcriptome, using the typically short RNA-Seq reads, based on 1- a reference genome (reference-based), 2- without a reference (*de novo*), or 3- through a combined approach. Transcriptome assembly is challenging compared to genome assembly because of the uneven coverage across the transcriptome and alternative isoforms (due to sharing exons), which both cause difficulties for the algorithms [31], [32]. Despite the challenges of transcriptome assembly, some algorithms have succeeded in overcoming a majority of these challenges: Trinity [32], Oases [33], Trans-ABySS [34], MIRA [35], Rnnotator [36], KISS-PLICE [37], SAT-Assembler [38], T-IDBA [39], STM [40], and EBARDenovo [41]. All of these methods are based on De Brujin graphs and take a greedy approach (for more information refer to Appendix), except EBARDenovo.

In this study we used three leading transcriptome assembly algorithms, 1- Trinity, 2- SOAPdenovo-Trans, and 3- Trans-ABySS to reconstruct the transcriptome of the *L. vannamei*. To examine the quality of each assembly, we evaluated the results via a Quality Control (QC) pipeline. Furthermore, we enriched the gene annotation for the Pacific whiteleg shrimp. The goal of this study was not to comprehensively compare the performance of the assembly methods, rather we have focused on evaluating the results from multiple perspectives. Our assessments illustrate different characteristics of the respective assembly methods and their usage in the annotation.

2 Material and Methods

2.1 Study Design

The main objective of this study is to improve the available transcriptomic resources for Pacific whiteleg shrimp. We propose a workflow to assemble, quality check and annotate a transcriptome. As Figure 1 shows, quality assessments play an important role and should be performed after the assembly. If the quality metrics are satisfactory, combining contigs of multiple assemblies can be the next step. And finally one needs to proceed with annotations. The workflow encompasses the cases that scientists choose not to perform quality check or contig intersection. However, it is not recommended to omit the quality assessments. Furthermore, contig intersection is suggested as an alternative to annotating all assembled transcripts. It also can be used to verify the compatibility of assembled contigs, using different assemblers. Finally, the contigs (or interested contigs) will go through selected annotation steps. Supplemental Document provides more details on the individual steps and contig intersections: https://repository.tamu.edu/handle/1969.1/154308.

The *de novo* assembly algorithms perform differently, and a great deal of attention has been paid to their comparisons [42], [43], [44], [45], [46], [47], [48], [49], [50], [51]. Those studies aimed to address transcriptome assembly and its associated computational challenges, to compare the performance of different assemblers, to use genomic assemblers on transcriptomic data, to use simulated datasets, or to complete the transcriptome for model species. The non-Illumina platform, e.g. 454 Roche long reads, which requires specific assembly approaches have also been examined in those studies.

We used the cutting-edge transcriptome assembly algorithms on Illumina pair-end (PE) reads, and annotated the resulting transcripts. Throughout this manuscript, we will discuss the criteria that can be used to compare different assembly results. As it is not possible to pick one universal best transcriptome assembly program, it is important to understand the power and limitations of each algorithm, and choose the appropriate tool accordingly. Due to the availability of High Performance Computing (HPC) facilities and time usages optimizations on many of the assembly algorithms, the best practice would be assembling input data utilizing multiple assemblers simultaneously, selecting the results with maximum accuracy, and ultimately moving forward to annotation (if applicable).

2.2 Transcriptome Assemblies

In this study, we used shrimp RNA-Seq reads [15], and assembled it using three state-of-the-art transcriptome assemblers: 1- Trinity (release r2013-02-25, release r2014-04-13, and release r2014-07-17), 2- SOAPdenovo-Trans (release 1.03), 3- Trans-ABySS (version 1.5.1).

The assemblers have a large number of adjustable parameters, however, we ran majority of them using their default settings. The Supplemental Documents describes all the tools used in this study and their settings. All of our assembly, mapping and annotations results are publicly available. By exploring the wide-range of parameter combinations for the different assemblers, it might be possible to improve their qualities. On the other hand, these settings can be very subjective, and our goal is not to carry out a comprehensive evaluation of different programs for *de novo* transcriptome assembly. Our interest lies in finding transcriptome assemblies that are of high quality and accurate (and preferably similar), and can be used for annotation.

2.3 Transcriptome Annotations

The three transcriptome annotation approaches are used

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2446478, IEEE/ACM Transactions on Computational Biology and Bioinformatics

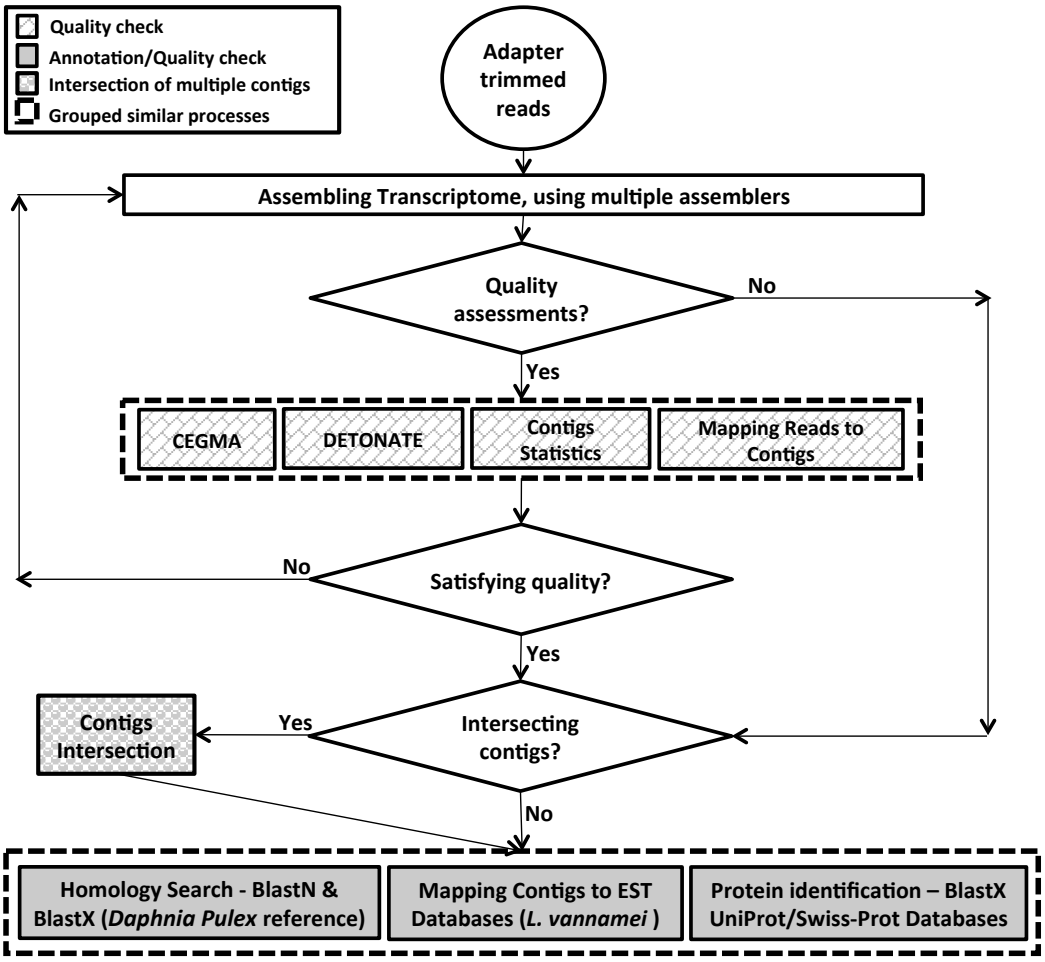


Fig. 1. Workflow for transcriptome assembly, quality assessments, and annotations. The quality assessments, and intersecting contigs are optional steps, however, quality checks are highly recommended. The workflow can be used for any transcriptome assembly/annotaion study. The databases that are used for annotaion of Pacific whiteleg shrimp are provided within the annotations steps.

for each assembly:

- BlastN/BlastX assembled contigs against *Daphnia pulex (D. pulex)* Transcripts and CDS/Proteins, respectively
- Mapping contigs against *L. vannamei* expressed sequence tag (EST) databases
- BlastX assembled contigs to UniProt/Swiss-Prot protein database

Taking all the three approaches for all the assembly results not only is helpful in the annotation, but is also used as a validation method. More details are provided in the RESULTS section.

We chose *Daphnia pulex* for annotating our contigs due to its phylogenetic relationship to panaeid shrimp and its completed genome. The subphylum Crustacean encompasses 67,000 species of arthropods but has few model organisms that have been subjected to comprehensive transcriptomic analyses. The first assembled crustacean genome was the water flea *D. pulex* [52]. It remains the lone crustacean with both well developed genomic and transcriptomic resources. As a cladoceran branchiopod, it is of a lineage that dates to the Permian, is most closely allied with insects, and may help distinguish fundamental genomic signatures of crustaceans from those of insects and arthropods in general [53]. Therefore, *Daphnia* is at present the obvious point of comparison for –omics in any crustacean. *D. pulex* and *L. vannamei* shared a common ancestor approximately 530 million years ago.

In the second approach, we mapped our contigs to *L. vannamei* expressed sequence tags (EST) available at NCBI dbEST [54] and the Penaeus Genome Database (PAGE) [55]. Before the advent of next-generation sequencing technologies, ESTs were the most powerful resources for collecting and curating what would now be considered transcriptomic datasets. EST's result from single Sanger sequencing runs from one end of cDNA clones of a library generated from a particular species, tissue, cell population or developmental state. These usually will not contain an entire coding sequence, but an incomplete "tag" with the sequence identifying a transcript that was expressed in the target cell population. This can be assessed in many ways, including being used as a handle to further characterize the full-length transcript, genomic locus, and expression patterns of the gene. Before our next-generation RNA-Seq based work and that of others in shrimp, EST resources were developed in *L. vannamei* that contributed over 165,000 ESTs to the public domain [22], [41], [42].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2446478, IEEE/ACM Transactions on Computational Biology and Bioinformatics

3 RESULTS

In this section, we present our transcriptome assemblies, quality assessment metrics, and annotation results. Three different releases of the Trinity algorithm were used to assemble the Pacific whiteleg shrimp. The Trinity contigs generated by release r2013-02-25 were originally reported, and thoroughly annotated by [15]. We referred to the Trinity run for releases r2013-02-25, release r2014-04-13, and release r2014-07-17 as Trinity_Run1_rFeb13, Trinity_Run2_rApr14 and Trinity_Run3_rJul14, respectively. All three runs were examined by our QC workflow and further annotated. The Trinity developing team had pointed that in the more recent releases the highest quality isoform reconstructions are reported in order to reduce the noise; and possibly fewer transcripts are reported in the 2014 versions, compared with the 2013 release. The three Trinity assemblies, using different releases, varied in the N50s, however all of them had high N50 values. As mentioned above, the 2014 Trinity releases report the most reliable isoforms, and attempt to report the "best" genes that are inferred by their Expectation Maximization (EM) algorithm. We observed these variations by fewer "total contigs" and lower "N50". However, the performance of three assemblies through the QC assessments was very similar. Most importantly, different releases did not affect the downstream annotations.

We employed Trans-ABySS assembler using two different k-mer sizes 32 and 48 (default k-mer, and a larger k-mer to test the k-mer size effect). We refer to the Trans-ABySS runs as Trans-ABySS_Run1_kmer32 and Trans-ABySS_Run2_kmer48. The SOAPdenovo-Trans was run using the default k-mer value of 100.

3.1 Transcriptome Assembly Statistics

Principal metrics of an assembly are usually provided by the assembler or can be calculated independently. Those pertain to the size of the output along with statistics related to the length of the contigs [56]. The most commonly presented statistics of an assembly are total size in the base pairs (span) of the assembly, number of the assembled transcripts, length of the largest contig, and the mean and median of the contig length. These statistics for our different transcript assemblies are shown in Table 1A.

The default value of the minimum transcript length reported by Trinity is 200, thus, Trinity will only include transcripts longer than 200 base pairs in the final assembly. The default value of this parameter is 100 for SOAPdenovo-Trans, and equal to the k-mer size for Trans-ABySS. We ran Trans-ABySS using the default k-mer size of 32bp, and larger k-mer size of 48bp. The distribution of the lengths of the contigs for the various assemblies is shown in Figure 2. It can be observed that SOAPdenovo-Trans and Trans-ABySS have many contigs shorter than 200 base pairs. In order to compare the assemblies, and also because it is customary to discard the short contigs, the transcripts shorter than 200bp are filtered out from the SOAPdenovo-Trans and Trans-ABySS assemblies. The basic statistics of the assemblies, after removing the transcripts shorter than 200bp, are shown in Table 1B. In the subsequent analyses, we will use the assemblies in which transcripts shorter than 200bp have

TABLE 1: STANDARD ASSEMBLY METRICS.
(A) STATISTICS FOR ASSEMBLIES, NO FILTRATION

	SOAPdenovo-Trans	Trans-ABySS_Run1_kmer32	Trans-ABySS_Run2_kmer48	Trinity_Run1_rFeb13	Trinity_Run2_rApr14	Trinity_Run3_rJul14
Total number of contigs	147,493	512,210	255,886	110,474	102,093	103,773
Length of largest contig (bp)	30,864	17,067	22,752	31,344	21,010	20,992
Assembly size (bp)	85,069,279	134,839,057	127,101,139	125,657,935	94,245,425	93,511,053
Mean contig length (bp)	577	263	497	1,137	923	901
Median contig length (bp)	166	67	164	429	390	392
GC Content (%)	41.54	42.98	42.83	44.12	43.32	43.27

(B) STATISTICS FOR ASSEMBLIES AFTER FILTERING OUT THE TRANSCRIPTS SHORTER THAN 200 BP

	SOAPdenovo-Trans	Trans-ABySS_Run1_kmer32	Trans-ABySS_Run2_kmer48	Trinity_Run1_rFeb13	Trinity_Run2_rApr14	Trinity_Run3_rJul14
Total number of contigs	62,514	119,772	110,556	110,474	102,093	103,773
Length of largest contig (bp)	30,864	17,067	22,752	31,344	21,010	20,992
Assembly size (bp)	74,156,520	105,766,302	110,437,049	125,657,935	94,245,425	93,511,053
Mean contig length (bp)	1186	883	999	1,137	923	901
Median contig length (bp)	503	479	529	429	390	392
GC Content (%)	41.34	42.61	42.78	44.12	43.32	43.27

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2446478, IEEE/ACM Transactions on Computational Biology and Bioinformatics

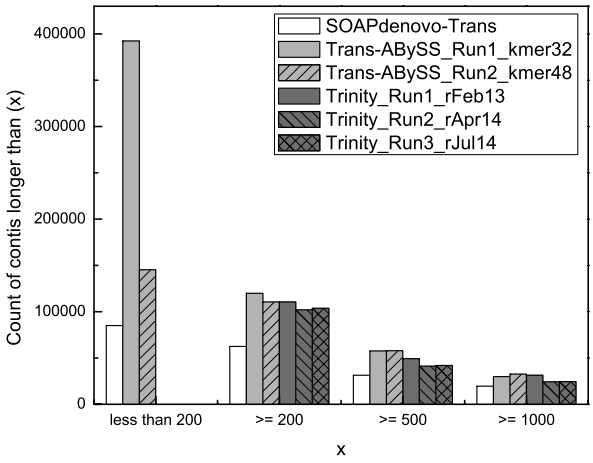


Fig. 2. Cumulative contig counts for various sizes.

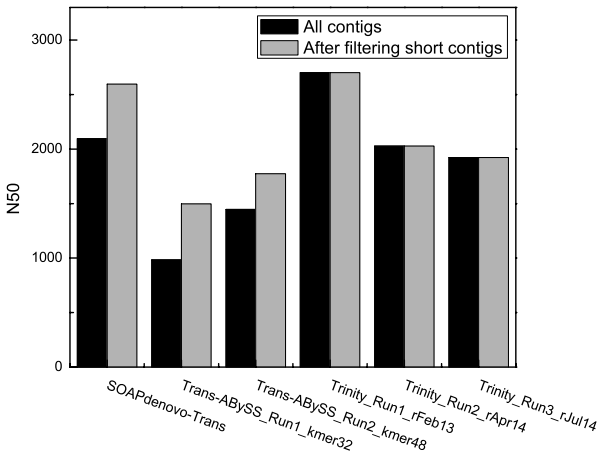


Fig. 3. N50 of the transcriptome assemblies including all the contigs and after removing the contigs shorter than 200bp.

been removed.

Another important metric of an assembly is N50. The N50 is defined as the contig size such that all the contigs equal to or greater than that size account for at least half of the total assembled bases [31]; it is also a weighted median of the lengths of the contigs [44]. Thus, the N50 value depends on the contigs' length; larger N50 indicates longer contigs, and possibly more continuous contigs. Figure 3 presents the N50 values for the assemblies. We filtered out contigs shorter than 200bp, and re-calculated the N50, which improved the results for SOAPdenovo-Trans and Trans-ABySS.

It should be noted that N50 is one of the quality check metrics of an assembly. A very low value of N50 can indicate a poor assembly (especially for a genome assembly), however, high N50 value may not be sufficient to pick one assembly result. The established transcriptome features, average gene and mRNA sizes for the species are essential in order to validate the contigs statistics of an assembly such as N50. There had been sturdies focusing on whiteleg shrimp transcriptome assembly, and we refer to their findings for acceptable N50s [57], [17], [15]; additionally, the N50 values of the tested assemblers were similar. Considering these two criteria, the runs had the N50 requirements and we validated them further through CEGMA, DETONATE, and Mapping Reads to the Assembled Contigs. As the following sections demonstrate, all assemblies passed all the QC metrics and were annotated.

3.2 CEGMA and DETONATE Validations

The transcript assemblies were compared in terms of their completeness, using Core Eukaryotic Genes Mapping Approach (CEGMA) pipeline (version 2.5) [58]. CEGMA defines a set of 248 highly conserved proteins that are present in a wide variety of eukaryotes. Completeness is defined as the number of these 248 Core Eukaryotic Genes (CEGs) that are present in the assemblies. This method is an important quality control check for eukaryotic assemblies, and we incorporated it in our pipeline. The number of CEGs represented in each assembly is shown in Table 2. All the assemblers preserved very large number of CEGs. Trinity runs, in particular, performed very well, conserving 245 and 246 CEGs.

Another recently developed assembly evaluation methodology and software package is DETONATE (DE novo TranscriptOme rNa-seq Assembly with or without the Truth Evaluation) [59]. DETONATE is based on a probabilistic model and evaluates assemblies with or without a reference. The two components of the package are RSEM-EVAL and REF-EVAL. RSEM-EVAL is a reference free approach that only relies on the input RNA-Seq and the resulting assembly. REF-EVAL needs a reference and provides more information about the assembly than currently available tools. In this study, we used RSEM-EVAL since the ground truth transcriptome of the *L. Vannamei* is being completed. The RSEM-EVAL is a model-based approach and provides a score to evaluate the assembly. The score is the log joint probability of the assembly and the reads, under the defined model. At the current stage the software handles single-end (SE) reads, however, DETONATE authors suggested the possibility of using the package for paired-end reads.

We computed the RSEM-EVAL (version 1.6) scores for all the transcriptome assemblies. The higher scores correspond to better assemblies. As Table 3 presents, the Trinity_Run1_rFeb13 performs the best. The results for other two releases of Trinity are almost identical. The Trans-ABySS runs have alike scores. Also, the evaluation score for SOAPdenovo-Trans is comparable to Trinity and Trans-ABySS. The Trinity, Trans-ABySS and SOAPdenovo-Trans all have acceptable scores.

3.3 Mapping Reads to the Assembled Contigs

Another metric that can be used to evaluate the quality of an assembly is the input reads alignment rates against the assembled contigs. The higher number of the reads mapping to the contigs can be an indicator of a higher quality assembly. This approach is also reported in other studies as a measure of the "goodness" of assemblies [49].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2446478, IEEE/ACM Transactions on Computational Biology and Bioinformatics

TABLE 2: CEGMA EVALUATION

	SOAPdenovo-Trans	Trans-ABySS_Run1_kmer32	Trans-ABySS_Run2_kmer48	Trinity_Run1_rFeb13	Trinity_Run2_rApr14	Trinity_Run3_rJul14
Number of 248 ultra-conserved CEGs	239	241	246	246	245	245
Completeness (percentage of CEGs)	96.37	97.18	99.19	99.19	98.79	98.79

TABLE 3: RSEM-EVAL

	Using filtered contigs
SOAPdenovo-Trans	-19,921,043,419
Trans-ABySS_Run1_kmer32	-14,323,711,897
Trans-ABySS_Run2_kmer48	-13,014,100,387
Trinity_Run1_rFeb13	-4,295,090,084
Trinity_Run2_rApr14	-14,214,273,926
Trinity_Run3_rJul14	-14,508,408,664

The input reads were mapped back to the assembled transcriptome using the program bowtie2 (version 2.2.3) [60] with default parameters. The read-mapping rate (percentage of reads mapped to contigs) for each assembler is shown in Table 4. One can observe that Trinity and Trans-ABySS transcripts have higher read-mapping rates compared to SOAPdenovo-Trans. All three assemblers have higher than 75% mapping rates.

All the assembly outputs met the four QC requirements, and the next sections describe their annotation process.

3.4 Blast to *Daphnia Pulex* References

The assembled transcripts were analyzed for sequence conservation against the references of a related species *Daphnia pulex* (water flea) from Joint Genome Institute (JGI) [52] by using BLAST [61], a program designed to perform homology searches. The BlastX and BlastN tools were used to find similarities between our contigs and *D. pulex* proteins and transcripts/CDS, respectively. We ran the Blast tool of the CLC Genomics Workbench (version 6.0.1) [62] for this section.

An important parameter in a BLAST result is the Expected (E) value that defines the expected number of hits by chance. The E-values that are closer to zero depict more significant matches [61]. We filtered our BlastN and BlastX results for two significance levels of 1E-4 and 1E-10. The number of transcripts with a BlastX hit against the *D. pulex* protein data set is shown in Figure 4.

BlastN of the contigs from each assembly against the *D. pulex* transcript and CDS was also conducted, and the results are presented in Figure 5.

Trans-ABySS runs have the highest absolute number of hits against the reference proteins for BlastX. Trinity runs produce comparable number of hits to Trans-ABySS runs, especially in the 1E-10 filtrating threshold. SOAPdenovo-Trans had the minimum annotated contigs. The relative proportion of the contigs with a significant hit (ratio of contigs with a significant hit to the total number of contigs in the assembly) is fairly consistent across all three assemblers.

We also compared the protein homologies with *D. pulex* for transcriptome assembly results. The results from each BlastX search were intersected with the other corresponding sets to find the shared proteins. Supplemental Document provides the intersected protein hits, and also shows the results in Venn diagrams. There is a very high degree of concordance in the homologous proteins found in *D. pulex* from all the assemblies, with more than 7,200 proteins shared among all six assemblies (E-value < 1E-10). The Venn diagrams are drawn from the unique and identical protein significant hits using the JVENN program (version v.1.5) [63]. The JVENN program can find the overlaps among all six outputs of our study.

3.5 Mapping Reads to the EST References

The expressed sequence tags (EST) are short cDNA sequences and can be useful in annotating the assembled contigs. There are two major databases for Pacific whiteleg shrimp ESTs: 1- NCBI dbEST [54] and Penaeus Genome Database (PAGE) [55]. We used GMAP [64] (version 2014-08-04) for aligning transcriptome assemblies against aforementioned databases.

The GMAP results for all the assemblers are shown in Tables 5 and 6 for NCBI dbEST and PAGE, respectively. The Trans-ABySS runs, using k-mer sizes of 32 and 48, achieve the highest EST mapping rates, which we suspect is due to its shorter contigs (mean contig length, lower N50). Trinity and SOAPdenovo-Trans achieve closer EST mapping rates, and have closer contig metrics, i.e. mean length and N50. The three Trinity runs, using three different releases of the tool, have almost identical EST coverage rates, which indicates they share contigs that contain EST sequences. The similarities of Trinity runs for the annotaion steps assured us that the important contigs are preserved, regardless of the release date of the software. It is important to note that the EST sequences are very short (less than 2,200bp for all ~165K dbEST cases), and their quantity for *L. Vannamei* is very limited. Having high mapping results for them was not our expectation; rather, we present this approach as a possible annotation path, especially for assemblies with the short contigs such as Trans-ABySS.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2446478, IEEE/ACM Transactions on Computational Biology and Bioinformatics

TABLE 4: READ MAPPING RATE, MAPPING READS TO THE CONTIGS

	SOAPdenovo-Trans	Trans-ABySS_Run1_kmer32	Trans-ABySS_Run2_kmer48	Trinity_Run1_rFeb13	Trinity_Run2_rApr14	Trinity_Run3_rJul14
Number of reads mapped to contigs	151,233,987	181,585,589	186,695,843	179,204,712	181,184,808	179,944,005
Read mapping rate (%)	75.8	91.01	93.57	89.81	90.81	90.18

TABLE 5: GMAP RESULTS FOR NCBI EST DATABASE

	SOAPdenovo-Trans	Trans-ABySS_Run1_kmer32	Trans-ABySS_Run2_kmer48	Trinity_Run1_rFeb13	Trinity_Run2_rApr14	Trinity_Run3_rJul14
Mapping rate	7.47	18.90	22.02	8.53	8.50	8.66
% of EST covered by contigs	5.56	19.17	19.01	8.46	8.29	8.47

TABLE 6: GMAP RESULTS FOR PAGE DATABASE

	SOAPdenovo-Trans	Trans-ABySS_Run1_kmer32	Trans-ABySS_Run2_kmer48	Trinity_Run1_rFeb13	Trinity_Run2_rApr14	Trinity_Run3_rJul14
Mapping rate	7.32	18.66	21.78	8.36	8.33	8.49
% of EST covered by contigs	5.67	19.61	19.44	8.58	8.44	8.63

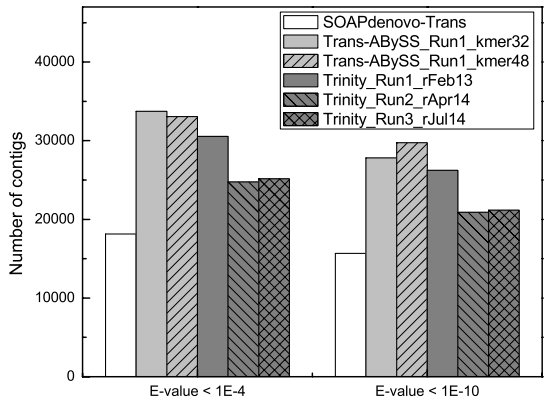


Fig. 4. Contig BLASTX hits against *Daphnia pulex* protein database.

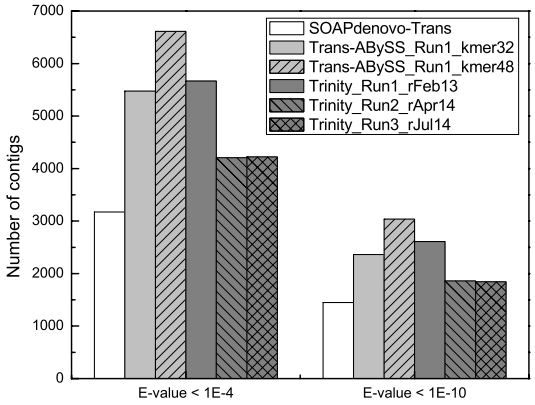


Fig. 5. Contig BLASTN hits against *Daphnia pulex* protein database.

3.6 Blast to UniProt/Swiss-Prot Databases

The protein sequences and functional annotations for the assemblies were assigned by employing the BlastX tool (BLAST+ version 2.2.29). BlastX queries all six open reading frames (ORF) of a sequence against the protein database. We used released July 2014 of UniProt/Swiss-Prot databases, installed on our local Red Hat Linux server, as the reference.

The BlastX search found numerous protein hits for each assembly. In order to select the most reliable hits, we filtered out any match with E-value greater than 1E-4. For the remaining hits of each assembly, we counted the number of times that each protein appeared for computing protein-hit frequencies. Finally, we categorized the protein-hit frequencies as X >= 200, 100 =< X < 200, 50 =< X < 100, 10 =< X < 50, and X < 10, where X is total number of protein hits. More details about the procedure can be found in the Supplemental Document. Figure 6 breaks down the appearance of protein hits for different assemblies. Comparing the protein hit frequencies shows that BlastX has annotated all three Trinity releases comparably. Trans-ABySS runs have close number of hits. SOAPdenovo-Trans generates the minimum number of protein hits, possibly due to its smaller total contigs. The variation of the protein hits is likely caused by the larger total contigs that Trans-ABySS and Trinity algorithms produced. The BlastX protein-hit frequencies are very similar for three Trinity releases, denoting that using different Trinity releases don't affect the annotations. BlastX was employed on contigs greater than 200bp. It should be noted that if the goal of the *de novo* transcriptome assembly study were to investigate the functional annotations

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2446478, IEEE/ACM Transactions on Computational Biology and Bioinformatics

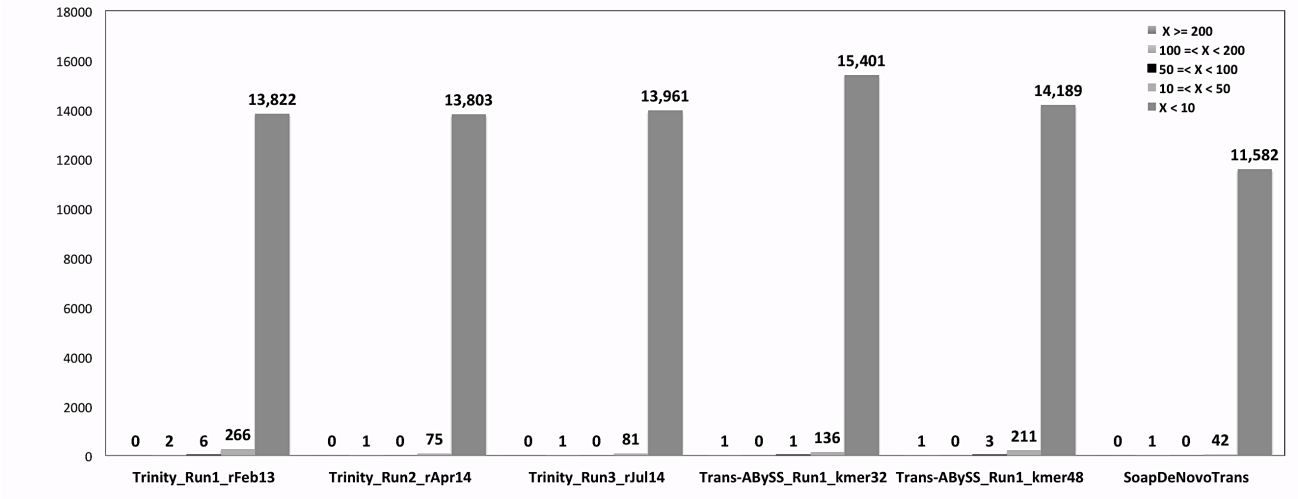


Fig. 6. BlastX search results for assemblies against UniProt/Swiss-Prot databases (X is defined as total number of hits).

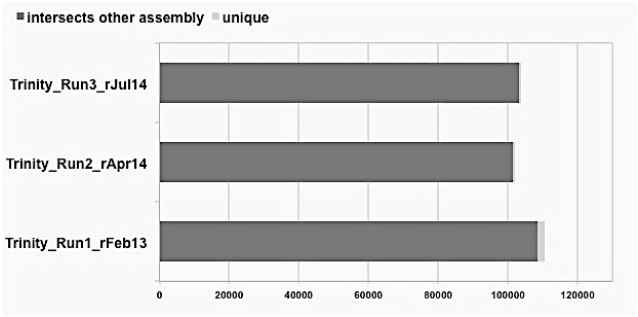


Fig. 7. Intersection of three Trinity runs

of the species, the total number of contigs would play an important role.

Finding the shared proteins among different assemblies can increase the confidence on the accuracy of the reported proteins. We selected the protein hits with frequency of at least 15 times in the SOAPdenovo-Trans results and matched them with Trinity and Trans-ABySS results. As Table 7 shows, most frequent protein hits appear in the annotations of all six runs, with minor exceptions. Our recommendation is to rely on the protein hits that are shared among at least three assembly BlastX results.

3.7 Contigs Intersection

The assembly methods differ in how they handle the challenges of reconstructing a complete transcriptome from short RNA-Seq reads. However, it is important to ensure that their final product is similar. To compare the assembler's outputs, we used BLAST (version 2.2.29+) [61] to match similar contigs between assemblies. A BLAST database was made from each assembly, and BlastN compared each database/assembly pair. Hits with >= 90% identity were considered matches, and were reported. We ran six assemblies, therefore, there were 36 Blast comparison pairs, and we provided the detailed comparison in the Supplemental Document. Overall, the intersection of contigs indicates great compatibility among the six assembled contig sets. As the Supplemental Table shows, in majority of comparisons, the two assemblers shared more than 75% of their contigs. Figure 7 depicts the intersection of contigs among three Trinity runs. The 2014 releases of the software have more similar Blast annotations. The 2013 release generated more contigs, and almost all of them are shared with 2014 releases. In summary, the three Trinity runs share more than 97% of their transcripts, demonstrating that different releases of the software produce very similar annotation results.

The main purpose of intersecting contigs is to determine the percentage of the overlap among different assembly algorithms' outputs. Our suggestion is to select the overlapping contigs for the annotations. In the current paper, we chose to include all the contigs in our annotations, since the main objective was to enhance the annotation knowledge for Pacific whiteleg shrimp. Furthermore, majority of the assemblers performed very similarly in quality metrics check steps, and overlapped well while intersecting the contigs. Therefore, we took advantage of all generated contigs, with the expense of more running time.

4 Conclusion

This study presented a workflow for transcriptome assemblies using RNA-Seq data. The purpose of this workflow is to check the quality metrics of each assembly and proceed to the annotations. We used the RNA-Seq data from Pacific whiteleg shrimp to assemble its transcriptome using well-known assembly methods. To ensure the quality of the assemblies, we examined each assembly via our workflow. We observed that the principal assembly metrics, such as mean contigs length and N50, are useful in initial judgments of the results; however, assemblies with the lower aforementioned values can perform well in downstream analysis. In our experiments, SOAPdenovo-Trans had lower total contig number, and Trans-ABySS (k-mer 32bp and 48bp) had relatively lower N50 and mean contig length compared to three Trinity runs. Nevertheless, SOAPdenovo-Trans and Trans-ABySS contigs had large

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2446478, IEEE/ACM Transactions on Computational Biology and Bioinformatics

TABLE 7
INTERSECTING SOAPDENOVO-TRANS BLASTX HITS THAT APPEAR MORE THAN 15 TIMES WITH TRINITY AND TRANS-ABYSS RESULTS

Protein ID	SOAPdenovo-Trans	Trinity_Run1_rFeb13	Trinity_Run2_rApr14	Trinity_Run3_rJul14	Trans-ABySS_Run1_kmer32	Trans-ABySS_Run2_kmer48
sp|P05661|MYSA_DROME	122	133	108	111	220	323
sp|Q7KRI2|LOLAL_DROME	25	36	32	32	45	76
sp|Q05319|STUB_DROME	21	23	25	24	28	27
sp|Q24174|ABRU_DROME	21	40	19	19	28	24
sp|P04323|POL3_DROME	20	24	22	26	26	27
sp|P42283|LOLA1_DROME	20	23	15	16	16	25
sp|O15090|ZN536_HUMAN	19	24	22	23	42	29
sp|Q27712|CP2L1_PANAR	19	36	27	25	50	50
sp|Q9VCA2|ORCT_DROME	18	22	20	21	29	27
sp|P00765|TRYP_ASTAS	17	25	24	24	30	44
sp|P21902|PCE_TACTR	17	26	26	25	26	43
sp|Q9V7U0|RESIL_DROME	17	22	20	21	19	21
sp|P83088|FUCTC_DROME	16	22	23	22	22	28
sp|P36362|CHIT_MANSE	15	25	16	17	23	20
sp|Q96DM1|PGBD4_HUMAN	15	26	24	25	19	21
sp|Q9U572|CLOT_PENMO	15	NA	NA	NA	33	48
sp|Q9VS29|DSCL_DROME	15	50	24	22	NA	18

number of significant BlastN/BlastX hits, mapping rates, and similar performance in other metrics compared to Trinity. Therefore, their assembled transcripts and annotation are reported. Studying the similarities among the annotated contigs is essential for reporting the most reliable gene/protein hits. The intersection of contigs is proposed as

a method for annotating only the shared portion of the assembled contigs. The transcriptome assembly of non-model species, e.g. the Pacific whiteleg shrimp, are progressing and ensuring the validity of the generated contigs and their annotation is an important task. This work aimed to pave the way toward this goal. If the computational power is not a barrier, the superior solution is to perform multiple assemblies, assess the outputs, and eliminate the low quality results, and ultimately annotate the contigs via various annotation tools.

5 APPENDIX

5.1 RNA-Seq Data

In this study, we used Illumina Hiseq for RNA-Seq experiments that were initially reported in [15]. The samples are from four different shrimp tissues: hepatopancreas, gill, pleopod, and abdominal muscle. The RNA-Seq reads are publicly available using BioSample accessions: SAMN02918336, SAMN02918337, SAMN02918338, and SAMN02918339 at NCBI. Reads have no sequencing adapters attached (adapters are oligonucleotide sequences that are ligated to the cDNA fragments to facilitate the sequencing), and the total number of reads is 399,056,712.

5.2 De Bruijn Graph

De Bruijn graphs are directed mathematical graphs that are used for modeling overlapping sequences of symbols. Many transcriptome assembly algorithms utilize De Bruijn graphs by showing each node as a k-mer (sequence of letters of length k). Edges of the graphs connect nodes that differ in only one base [31].

ACKNOWLEDGMENT

The authors wish to thank Dr. Philip Blood and the support staff at Pittsburgh Supercomputing Center (PSC). This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number OCI-1053575. We also thank Dr. Spiros Vellas and the support staff at Texas A&M Supercomputing Facility. Finally, we appreciate all the support and help from The Brazos Cluster at Texas A&M University, Dr. Guy Almes (chair) and the support staff.

REFERENCES

[1] D. V. Lightner, "The Penaeid Shrimp Viral Pandemics due to IHHNV, WSSV, TSV and YHV: History in the Americas and Current Status," *Proceedings of the Thirty-second US Japan Symposium on Aquaculture,* US-

japan Cooperative Program in Natural Resources (UJNR). U.S. Department of Commerce, N.O.A.A., Silver Spring, MD, USA, pp. 6-24, 2003.

[2] M. F. Criscitiello, et al., "Evolutionarily conserved TCR binding sites, identification of T cells in primary lymphoid tissues, and surprising trans-rearrangements in nurse shark," *J. Immunology*, vol. 184, no. 12, pp. 6950-60, 2010.

[3] M. F. Criscitiello, et al., "Shark class II invariant chain reveals ancient conserved relationships with cathepsins and MHC class II," *Dev. Comp. Immunology*, vol. 36, no. 3. pp. 521-533, 2012.

[4] M. F. Criscitiello, and M.F. Flajnik, "Four primordial immunoglobulin light chain isotypes, including lambda and kappa, identified in the most primitive living jawed vertebrates," *Eur. J. Immunology*, vol. 37, no. 10, pp. 2683-2694, 2007

[5] S. S. Musthaq, and J. Kwang, "Oral Vaccination of Baculovirus-Expressed VP28 Displays Enhanced Protection against White Spot Syndrome Virus in Penaeus monodon," *PLoS ONE*, vol. 6, no. 11, 2011.

[6] L. S. Yang, et al., "A Toll receptor in shrimp," *Molecular Immunology*, vol. 44, no. 8, pp. 1999-2008, 2007.

[7] C. Y. Lai, W. Cheng, and C.M. Kuo, "Molecular cloning and characterisation of prophenoloxidase from haemocytes of the white shrimp, Litopenaeus vannamei," *Fish and Shellfish Immunology*, vol. 18, no. 5, pp. 417-30, 2005.

[8] X. D. Huang, X.D., et al., "Identification and functional study of a shrimp Relish homologue," *Fish and Shellfish Immunology*, vol. 27, no. 2, pp. 230-238, 2009.

[9] Huang, X.D., et al., "Identification and functional study of a shrimp Dorsal homologue," *Dev. Comp. Immunology*, vol. 34, no. 2, pp. 107-113, 2010.

[10] P. H. Wang, et al., "An immune deficiency homolog from the white shrimp, Litopenaeus vannamei, activates antimicrobial peptide genes," *Molecular Immunology*, vol. 46, no. 8-9, pp. 1897-1904, 2009.

[11] H. S. Ai, et al., "A novel prophenoloxidase 2 exists in shrimp hemocytes," *Dev. Comp. Immunology*, vol. 33, no. 1, pp. 59-68, 2009.

[12] H. S. Ai, et al., "Characterization of a prophenoloxidase from hemocytes of the shrimp Litopenaeus vannamei that is down-regulated by white spot syndrome virus," *Fish and Shellfish Immunology*, vol. 25, no. 1-2, pp. 28-39, 2008.

[13] W. Y. Chen, et al., "WSSV infection activates STAT in shrimp," *Dev. Comp. Immunology*, vol. 32, no. 10, pp. 1142-1150, 2008.

[14] J. Robalino, et al., "Double-stranded RNA induces sequence-specific antiviral silencing in addition to nonspecific immunity in a marine shrimp: convergence of RNA interference and innate immunity in the invertebrate antiviral response," *J. Virology*, vol. 79, no. 21, pp. 13561-13571, 2005.

[15] N. Ghaffari, A. Sanchez-Flores, D. Ryan, K. D. Garcia-Orozco, P. L. Chen, A. Ochoa-Leyva, A. A. Lopez-Zavala, J. S. Carrasco, C. Hong, L. G. Brieba, E. Rudino-Pinera, P. D. Blood, J. A. Sawyer, C. D. Johnson, S. V. Dindot, R. R. Sotelo-Mundo, and M. F. Criscitiello, "Novel Transcriptome Assembly and Improved Annotation of the Whiteleg Shrimp (Litopenaeus vannamei), a Dominant Crustacean in Global Seafood Mariculture", *Nature Scientific Reports*, vol. 4, DOI: 10.1038/srep07081, 2014.

[16] Y. Yang, et. al, "SNP Discovery in the Transcriptome of White Pacific Shrimp Litopenaeus vannamei by Next Generation Sequencing," *PLoS ONE*, vol. 9, no. 1, 2014.

[17] C. Li, et al., "Analysis of Litopenaeus vannamei transcriptome using the next-generation DNA sequencing technique," *PLoS ONE*, vol. 7, e47442, 2012.

[18] H. Guo, et al., "Trascriptome analysis of the Pacific white shrimp Litopenaeus vannamei exposed to nitrite by RNA-seq," *Fish and Shellfish Immunology*, vol. 35, pp. 2008-2016, 2013.

[19] N. A. O'Leary, et al., "Analysis of multiple tissue-specific cDNA libraries from the Pacific whiteleg shrimp Litopenaeus vannamei," *Integrative and Comparative Biology*, vol. 46, pp. 931-939, 2006.

[20] X. Chen, et al., "Transcriptome analysis of Litopenaeus vannamei in response to white spot syndrome virus infection," *PLoS ONE*, vol. 8, no. 8, e73218, 2013.

[21] A. Clavero-Salas, et al. "Transcriptome analysis of gills from the white shrimp Litopenaeus vannamei infected with White Spot Syndrome Virus," *Fish and Shellfish Immunology*, vol. 23, pp. 459-472, doi:10.1016/j.fsi.2007.01.010, 2007.

[22] J. Robalino, et al., "Insights into the immune transcriptome of the shrimp Litopenaeus vannamei: tissue-specific expression profiles and transcriptomic responses to immune challenge," *Physiological Genomics*, vol. 29, pp. 44-56, doi:10.1152/physiolgenomics.00165.2006, 2007.

[23] S. Sookruksawong, F. Sun, Z. Liu, and A. Tassanakajon, "RNA-Seq analysis reveals genes associated with resistance to Taura syndrome virus (TSV) in the Pacific white shrimp Litopenaeus vannamei," *Dev. Comp. Immunology*, vol. 41, pp. 523-533, doi:10.1016/j.dci.2013.07.020, 2013.

[24] A. Veloso, et. al., "The transcriptomic response to viral infection of two strains of shrimp (Litopenaeus vannamei)," *Dev. Comp. Immunology*, vol. 35, pp. 241-246, doi:10.1016/j.dci.2010.10.001 2011.

[25] D. Zeng, et al., "Transcriptome analysis of Pacific white shrimp (Litopenaeus vannamei) hepatopancreas in response to Taura syndrome Virus (TSV) experimental infection," *PLoS ONE*, vol. 8, no. e57515, doi:10.1371/journal.pone.0057515, 2013.

[26] A. Mortazavi, "Mapping and quantifying mammalian transcriptomes by RNA-Seq," *Nature Methods*, vol. 5, pp. 621-628, 2008.

[27] M. Sultan, et. al, "A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome," *Science*, vol. 321, pp. 956-960, 2008.

[28] C. Alkan, et. al., "Personalized copy number and segmental duplication maps using next-generation sequencing," *Nature Genetics*, vol. 41, pp. 1061-1067, 2009.

[29] D. W. Craig, et. al., "Identication of genetic variants using barcoded multiplexed sequencing," *Nature Methods*, vol. 5, pp. 887-893, 2008.

[30] D. R. Bentley, et. al., "Accurate whole human genome sequencing using reversible terminator chemistry," *Nature*, vol. 456, no. 6, pp. 53-59, 2008.

[31] J.A. Martin and Z. Wang, "Next-generation transcriptome assembly," *Nature Reviews Genetics*, vol. 12, no. 10, pp. 671-682. doi: 10.1038/nrg3068, 2011.

[32] M. G. Grabherr, et. al, "Full-length transcriptome assembly from RNA-seq data without a reference genome," *Nature Biotechnology*, vol. 29, no. 7, pp. 644-652. doi: 10.1038/nbt.1883, 2011.

[33] M. H. Schulz, et. al., "Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels," *Bioinformatics*, vol. 28, pp. 1086–1092, 2012.

[34] G. Robertson, et. al, "De novo assembly and analysis of RNA-seq data," *Nature Methods*, vol. 7, no. 11, pp. 909-915, 2010.

[35] B. Chevreux B, et al., "Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs," *Genome Research*, vol. 14, pp. 1147–1159, 2004.

[36] J. Martin, "Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads," *BMC Genomics*, vol. 11, no. 663, 2010.

[37] G. A. T. Sacomoto, J. Kielbassa , R. Chikhi, R. Uricaru, P. Antoniou, M. F. Sagot, P. Peterlongo, and V. Lacroix, "KISS-PLICE: de-novo calling alternative splicing events from RNA-seq data," *BMC Bioinformatics*, vol. 13, no. Suppl 6:S5, 2012.

[38] Y. Zhang, Y. Sun, and J. R. Cole, "A Scalable and Accurate Targeted Gene Assembly Tool (SAT-Assembler) for Next-Generation Sequencing Data," *PLoS Computational Biology*, vol. 10, no. 8: e1003737. doi:10.1371/journal.pcbi.1003737, 2014.

[39] Y. Peng, H. Leung, S. M. Yiu, and F. Y. L. Chin, "T-IDBA: A de novo Iterative de Bruijn Graph Assembler for Transcriptome," *Proceedings of the 15th Annual International Conference on Research in Computational Molecular Biology (RECOMB),* Vancouver, Canada, 2011.

[40] Y. Surget-Groba and J. I.Montoya-Burgos, "Optimization of de novo transcriptome assembly from next-generation sequencing data," *Genome Research*, vol. 20, no. 10, pp. 1432–1440, 2010.

[41] H. T. Chu, W. W. Hsiao, J. C. Chen, T. J. Yeh, M. H. Tsai, H. Lin, Y. W. Liu, S. A. Lee, C. C. Chen, T. T. Tsao, and C. Y. Kao, "EBARDenovo: highly accurate de novo assembly of RNA-Seq with efficient chimera-detection," *Bioinformatics*, vol. 15, no. 8, pp.1004-1010, 2013.

[42] M. Mundry, E. Bornberg-Bauer, M. Sammeth, and P. G. Feulner, "Evaluating characteristics of de novo assembly software on 454 transcriptome data: a simulation approach," *PloS ONE*, vol. 7, no. 2, p. e31410, 2012.

[43] Y. Yang and S. A. Smith, "Optimizing de novo assembly of shortread rna-seq data for phylogenomics," *BMC Genomics*, 14 (1), 2013.

[44] K. Clarke, Y. Yang, R. Marsh, L. Xie, and K. K. Zhang, "Comparative analysis of de novo transcriptome assembly," *Science China Life Sciences*, vol. 56, no. 2, pp. 156–162, 2013.

[45] X. Ren, T. Liu, J. Dong, L. Sun, J. Yang, Y. Zhu, and Q. Jin, "Evaluating de bruijn graph assemblers on 454 transcriptomic data," *PloS ONE*, vol. 7, no. 12, e51188, 2012.

[46] S. Kumar and M. L. Blaxter, "Comparing de novo assemblers for 454 transcriptome data," *BMC Genomics*, vol. 11, no. 1, pp. 571, 2010.

[47] B. Feldmeyer, C. W. Wheat, N. Krezdorn, B. Rotter, and M. Pfenninger, "Short read illumina data for the de novo assembly of a non-model snail species transcriptome (radix balthica, basommatophora, pulmonata), and a comparison of assembler performance," *BMC Genomics*, vol. 12, no. 1, pp. 317, 2011.

[48] R. Garg, R. K. Patel, A. K. Tyagi, and M. Jain, "De novo assembly of chickpea transcriptome using short reads for gene discovery and marker identification," *DNA Research*, vol. 18, no. 1, pp. 53– 63, 2011.

[49] Q.-Y. Zhao, Y. Wang, Y.-M. Kong, D. Luo, X. Li, and P. Hao, "Optimizing de novo transcriptome assembly from short-read rna-seq data: a comparative study," *BMC Bioinformatics*, vol. 12, no. Suppl 14, pp. S2, 2011.

[50] B. Lu, Z. Zeng, and T. Shi, "Comparative study of de novo assembly and genome-guided assembly strategies for transcriptome reconstruction based on rna-seq," *Science China Life Sciences*, vol. 56, no. 2, pp. 143–155, 2013.

[51] M. B. Cougar, "Enabling large-scale next-generation sequence assembly with Blacklight," *Cocurrency and Computation*, vol. 26, pp. 2157-2166, 2014.

[52] J. K. Colbourne, et al. "The ecoresponsive genome of Daphnia pulex," *Science*, vol. 331, pp. 555-561, doi:10.1126/science.1197761, 2011.

[53] D. J. Taylor, et. al, "Phylogenetic evidence for a single long-lived clade of crustacean cyclic parthenogens and its implications far the evolution of sex," *Proceedings of Biological Sciences*, vol. 266, pp. 791-797, 1999.

[54] M. S. Boguski, T. M. Lowe and C. M. Tolstoshev, "dbEST--database for "expressed sequence tags," *Nature Genetics*, vol. 4, no. 4, pp. 332-333, 1993.

[55] J. H. Leu, et. al, "A review of the major penaeid shrimp EST studies and the construction of a shrimp transcriptome database based on the ESTs from four penaeid shrimp," *Marine Biotechnology*, vol. 13, pp. 608–621 2011.

[56] S. T O'Neil and S. J. Emrich, "Assessing de novo transcriptome assembly metrics for consistency and utility," *BMC Genomics*, vol. 14, no. 1, pp. 465, 2013.

[57] J. Wei, X. Zhang, Y. Yu, H. Huang, F. Li, and J. Xiang, "Comparative Transcriptomic Characterization of the Early Development in Pacific White Shrimp Litopenaeus vannamei," *PLoS One*, vol. 9, no. 9: e106201, 2014.

[58] G. Parra, K. Bradnam and I Korf, "CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes," *Bioinformatics*, vol. 23, no. 9, pp. 1061-1067, 2007.

[59] B. Li, et. al, "Evaluation of de novo transcriptome assemblies from RNA-Seq data," *bioRXiv*, doi: http://dx.doi.org/10.1101/006338, 2014.

[60] B. Langmead, and S. Salzberg, "Fast gapped-read alignment with Bowtie 2," *Nature Methods*, vol. 9, pp. 357-359, 2012.

[61] S. McGinnis and T. L. Madden, "BLAST: at the core of a powerful and diverse set of sequence analysis tools," *Nucleic Acids Research*, vol. 32, no. 2, pp. W20-W25, 2004.

[62] CLC Genomics Workbench 6.0.1 (http://www.clcbio.com)

[63] P. Bardou, J. Mariette, F. Escudié, C. Djemiel, and C Klopp, "jvenn: an interactive Venn diagram viewer," *BMC Bioinformatics,* vol 15, no. 293, doi:10.1186/1471-2105-15-293 , 2014.

[64] T. D. Wu and C. K. Watanabe, "GMAP: a genomic mapping and alignment program for mRNA and EST sequences," *Bioinformatics*, vol. 21, pp. 1859-1875, 2005.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2446478, IEEE/ACM Transactions on Computational Biology and Bioinformatics

N. Ghaffari received her Ph.D. degree in Electrical and Computer Engineering from Texas A&M University (TAMU), and her M.S. degree in Computer Information Systems from University of Houston - Clear Lake in 2011 and 2006, respectively. During her Ph.D. studies, she focused on complexity reduction of Gene Regulatory Networks. She is currently a bioinformatics scientist at AgriLife Genomics and Bioinformatics, where she regularly provides statistical and bioinformatics trainings for faculty/students across TAMU. She is the author/co-author in many peer-reviewed journal and conference papers. Few of her recent projects include discovering SNPs and CNVs for the first quarter horse mare sequenced, developing a new gene filtering method for RNA-Seq studies, and assembling/annotating transcriptome of the Pacific whiteleg shrimp. Her major research interests are computational biology, next-generation sequencing methods, genome and transcriptome assemblies and their evaluation, statistical design and modeling.

O. A. Arshad is a Ph.D student at the department of Electrical and Computer Engineering at Texas A&M University, College Station. Before starting his Ph.D, he worked as a lecturer at the School of Electrical Engineering and Computer Science at the National University of Sciences and Technology, Islamabad, Pakistan after completing a Masters degree in electrical engineering from Texas A&M and a bachelors in Electronics Engineering from Ghulam Ishaque Khan Institute of Engineering Sciences and Technology, Pakistan.

H. Jeong received the B.S. degree in electrical engineering from Inha University, Incheon, Korea, in 2007, and the M.S. degree in electrical engineering from Seoul National University, Seoul, Korea, in 2009. He is working toward the Ph. D. degree in the Department of Electrical and Computer Engineering at Texas A&M University. From 2009 to 2011, he was a software engineer with Samsung Electronics, Suwon, Korea where he worked on developing system software for smartphone. His research interest includes signal processing, computational network biology, and bioinformatics.

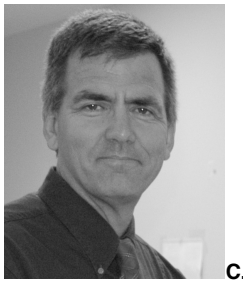
J. Thiltges received the B.S. (2004) in computer engineering and B.S. (2004) in chemical engineering from the Univ. of Nebraska-Lincoln. In 2013, he joined Texas A&M Genomics and Bioinformatics Services.

M. F. Criscitiello is an Assistant Professor in the Department of Veterinary Pathobiology of the College of Veterinary Medicine and the Department of Microbial Pathogenesis and Immunology of the College of Medicine at Texas A&M University. He received degrees from the University of North Carolina (B.S. Biology 1993), East Carolina University (M.S. Molecular Biology and Biotechnology 1997) and the University of Miami (Ph.D. Microbiology and Immunology 2003), and was a post-doctoral fellow at the University of Maryland studying comparative immunology. NIH and NSF grants have funded this work in shark and frog, but the Criscitiello lab also studies immunogenetics in agriculturally relevant species such as shrimp, tuna and cattle. Dr. Criscitiello's research merges immunology, genetics and evolution. A focus is the early natural history of the vertebrate adaptive immune system, with particular attention given to the genetic tricks of lymphocyte antigen receptor genes (e.g., antibodies and T cell receptors), mucosal immune mechanisms in the gut, and antigen presentation. He has received the Montague Center of Teaching Excellence Award in 2012 and the College of Veterinary Medicine Outstanding Research Achievement Award in 2014. Dr. Criscitiello is author of 28 peer reviewed publications, including a 2013 Cell cover article, and has an H-index of nine. He is a member of the American Association of Immunologists, Society for Experimental Biology and Medicine, International Society of Fish and Shellfish Immunology and the International Society of Comparative and Developmental Immunology. Dr. Criscitiello serves on the editorial boards of Experimental Biology and Medicine and Immunogenetics.

B.-J. Yoon received the B.S.E. (summa cum laude) degree from the Seoul National University, Seoul, Korea, in 1998, and the M.S. and Ph.D. degrees from the California Institute of Technology, Pasadena, in 2002 and 2007, respectively, all in Electrical Engineering. In 2008, he joined the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, where he was an Assistant Professor during 2008-2014 and an Associate Professor since 2014. Recently, Dr. Yoon joined the College of Science and Engineering at the Hamad bin Khalifa University (HBKU), Doha, Qatar, as a founding faculty member, where he is currently an Associate Professor. His recent honors include the Na-

tional Science Foundation (NSF) CAREER Award and the Best Paper Award at the 9th Asia Pacific Bioinformatics Conference (APBC). His main research interests include genomic signal processing (GSP), bioinformatics, and computational network biology.

A. Datta received the B. Tech degree in Electrical Engineering from the Indian Institute of Technology, Kharagpur in 1985, the M.S.E.E. degree from Southern Illinois University, Carbondale in 1987 and the M.S. (Applied Mathematics) and Ph.D. degrees from the University of Southern California in 1991. In August 1991, he joined the Department of Electrical and Computer Engineering at Texas A&M University where he is currently the J. W. Runyon, Jr. '35 Professor II and Director for the Center for Bioinformatics and Genomic Systems Engineering (CBGSE). His areas of interest include adaptive control, robust control, PID control and Genomic Signal Processing. He has authored or coauthored 5 books and over 100 journal and conference papers on these topics. He is a Fellow of IEEE, has served as an Associate Editor of the IEEE Transactions on Automatic Control (2001-2003), the IEEE Transactions on Systems, Man and Cybernetics-Part B (2005-2006) and is currently serving as an Associate Editor of the EURASIP Journal on Bioinformatics and Systems Biology, the IEEE Transactions on Biomedical Engineering, the IEEE Journal of Biomedical and Health Informatics, and IEEE Access.



C. D. Johnson received his BS and PhD from Texas A&M University from the Department of Soil and Crop Science; in between he received his MS from Clemson University in Horticulture. His postdoc was at University of Louisville, in the Center for Genetics and Molecular Medicine. Dr. Johnson is Director of Genomics and Bioinformatics at Texas A&M AgriLife in College Station, TX and leads the next generation sequencing program at A&M. Over 60 publications/patents (>3000 citations), two book chapters, and bioinformatics analysis tools for the study of miRNA which are in use throughout the world being distributed by several of the leading genomic technology companies.