Guidelines for RNA-seq projects: applications and opportunities in non-model decapod crustacean species

Tuan Viet Nguyen[*†a], Hyungtaek Jung[†b], Guiomar Rotllant[c], David Hurwood[d], Peter Mather[d], Tomer Ventura[a]

[a] GeneEcology Research Centre, Faculty of Science, Health, Education and Engineering, University of the Sunshine Coast, 4 Locked Bag, Maroochydore, QLD 4558, Australia

[b] Centre for Tropical Crops and Biocommodities, Institute for Future Environment, Queensland University of Technology, GPO Box 2434, Brisbane QLD 4001, Australia

[c] Institut de Ciències del Mar (CSIC), Passaig Marítim de la Barceloneta 37, 08003 Barcelona, Spain

[d] School of Earth, Environment and Biological Sciences, Science and Engineering Faculty, Queensland University of Technology, GPO Box 2434, Brisbane QLD 4001, Australia

[*]: Corresponding author: Tuan Viet Nguyen (tnguyen@usc.edu.au)

[†]: Equal contribution

Abstract (150-200 words)

Next Generation Sequencing (NGS) has dramatically changed the way biological research is being conducted in the post-genomic era and they have only been utilized widely over the recent decade for studies of non-model decapod crustacean species, predominantly by sequencing the transcriptome of various tissues across different life stages. NGS can now provide a rapid, cost-effective solution for discovery of genetic markers crucial in many applications that would previously have otherwise taken years to develop. Sequencing of the entire transcriptome (referred to as RNA sequencing; RNA-seq) is one of the most popular NGS tools. RNA-seq studies of non-model species in crustacean taxa however, have faced some problems, including a lack of "good" experimental study design, a relative paucity of gene annotations, combined with limited knowledge of genomic technologies and analyses. The aim of the current review is to assist crustacean biologists to develop a better appreciation of the applications and scope of RNA-seq analysis, understand the basic requirements for optimal RNA-seq studies and provide an overview of each step from RNA-seq experimental design to bioinformatics approaches to data analysis. Insights that have resulted from RNA-seq studies across a wide range of non-model decapod species are also summarized.

Keywords: RNA-sequencing, Next generation sequencing, Differential gene expression, shrimp, prawn, crab, lobster, crayfish, *in silico*

INTRODUCTION

Next generation sequencing (NGS) technologies have rapidly transitioned bioscience into the post-genomic era, resulting in easier, cheaper, and faster DNA sequencing. Application of advanced NGS platforms has allowed multiple techniques to be developed that address biological challenges. These include; RNA-sequencing (RNA-seq), whole-exome sequencing, chromatin immunoprecipitation sequencing (ChIP-seq), microRNA sequencing (miRNA-seq), restriction assisted DNA sequencing (RAD-seq), and small RNA sequencing. Among these, RNA-seq is a technique that has revolutionized gene expression studies and marker discovery (single sequence repeats [SSRs]/microsatellites and single nucleotide polymorphisms [SNPs]) (Das et al. 2016; Lister et al. 2009; Marguerat and Bähler 2010; Mykles et al. 2016; Ozsolak and Milos 2011; Wang et al. 2009; Wilhelm and Landry 2009). The RNA-seq platform is based on the analysis of the transcriptome - a small portion of the whole genome that is transcribed from chromosomal DNA into RNA molecules - a dynamic set of elements that change depending on developmental stages or physiological conditions. Also, by analysing the sequenced transcriptome, genetic polymorphisms including SNPs and SSRs can be mined and analysed with ease (Jaramillo et al. 2016; Jin et al. 2013; Jung et al. 2011; Jung et al. 2016; Lv et al. 2014; Meng et al. 2015; Nguyen et al. 2016).

While RNA-seq techniques have had a major impact on model species (which in this review is defined as a species with a well-characterized genome, e.g. *Daphnia pulex*), the application of RNA-seq approaches in non-model decapod crustacean taxa is still limited by the small size of the research community and the subsequent bottleneck of bioinformatics analysis capabilities. Many NGS analytical tools are available and by default, are developed for model species, making it difficult for researchers investigating non-model organisms to navigate through and identify appropriate tools. Designing and evaluating a pipeline for transcriptomics projects in non-model species therefore, can be considered a crucial step prior to project initiation. While the transcriptome can encompass many categories of different types of RNA (mi-RNA, small nuclear RNA, non-coding RNA etc.), this review will focus mainly on mRNA sequencing using Second Generation Sequencing (SGS) technology – we intend to use the same classification proposed by Schadt et al. (2010), that defined Sanger sequencing as First Generation, ”wash-and-scan” sequencing technology as Second Generation, and single molecule real time sequencing as Third Generation. Under this classification scheme, SGS includes a number of platforms, notably Illumina, Solid, Ion Torrent/Ion Proton, Roche 454; whereas PacBio and Oxford Nanopore are classified as Third Generation Sequencing (TGS) Technology. Here we will focus primarily on different strategies to initiate a transcriptome study, briefly addressing several platforms that currently are available, as well as recommending a number of experimental designs, bioinformatics software for *de novo* assembly and specific data analyses for decapod crustacean species. Finally, we review recent biological insights gained from application of SGS in crustacean transcriptomics and highlight opportunities as well as challenges for applied RNA-seq in the future.

OVERVIEW OF RNA-SEQ TECHNOLOGY

PRE-SEQUENCING

New sequencing technologies and new sequencing chemistries are being developed rapidly. The arrival of SGS, and more recently TGS, has completely changed the way researchers approach unanswered phenomena in basic, applied, and clinical research. Each sequencing platform is based on different proprietary chemistries and technologies and each has unique strengths and weaknesses. Details on sequencing chemistry have been summarized elsewhere (Goodwin et al. 2016; Koboldt et al. 2013; Metzker 2010; Reuter et al. 2015). Currently, Illumina is the most widely utilized SGS for RNA-seq, since the platform enables deep coverage of the transcriptome and provides long, low-error reads that are suitable for mapping to reference genomes and transcriptome assemblies (Goodwin et al. 2016; Metzker 2010; Niedringhaus et al. 2011). Performance benchmarking of many SGS platforms has been conducted for several years (Finseth and Harrison 2014; Glenn 2011; Goodwin et al. 2016; Lahens et al. 2017; Lam et al. 2012; Liu et al. 2012) and an online archive of sequencing platforms is available on the market and can be found at https://allseq.com/knowledgebank/. Given the popularity of Illumina Sequencers in general, it tends to be the technology most widely applied in crustacean transcriptome projects (Havird and Santos 2016).

In brief, RNA-seq includes the use of an SGS platform to generate a huge amount of sequence data. Due to technical constraints of the approach (most SGS platforms can only generate short to medium length reads, approximately 50-300 bp), RNA transcripts must be fragmented into shorter sequences. In the absence of a reference genome, short reads are then reconstructed to make a reference transcriptome, referred to as a *de novo* assembly. Following this, raw reads can be realigned (or mapped) to the previously generated reference sequence and counted, thus providing a digital measurement of specific transcript abundances that can facilitate biological interpretation. Where key genes are targeted (based on either high differential expression or previously identified in the literature), they can be validated by replicating samples across a range of experimental conditions (eg. in different tissue types, at different life history stages, between sexes, etc.). A popular approach for validation includes quantitative real-time PCR (qRT-PCR) where relative transcript abundance can be assessed under more strictly controlled conditions. Most RNA-seq strategies that utilize SGS can be summarized by a basic workflow (Figure 1).

Experimental design

Designing an RNA-seq experiment requires a solid biological understanding of the taxa under investigation and the question(s) to be addressed. Poor or inappropriate decisions at this stage can result in a large amount of unusable data. A good experimental design for every NGS–based experiment therefore, is a basic requirement that cannot be over-emphasized.

In general, several factors must be considered prior to the initiation of any well-designed sequencing project. Essentially, an appropriate experimental design is a balance between the level of biological versus technical replication (Figure 2) and the resulting depth of coverage for each tissue type, life stage, sex etc., within a framework of time and financial constraints. It is advisable that researchers without much prior experience should seek suggestions from professional service providers including bioinformaticians and biostatisticians, as well as the sequencing provider. This review highlights some of the pitfalls to be aware of and sets the scene for appropriate study design. Several studies provide direction on how to design a statistically valid RNA-seq experiment (Auer and Doerge 2010; Conesa et al. 2016; Fang and Cui 2011; Yang and Wei 2015). In general, a comprehensive transcriptome requires multiple tissues from multiple developmental stages while gene expression studies require samples that represent contrasting treatments (e.g. male vs female, control vs hormone treated, salinity vs freshwater acclimation, or different developmental/life history stages).

Biological and technical replicates

In the NGS context, technical replication refers to multiple libraries from the same biological sample (i.e. the technical steps are performed separately) (Figure 2). While potentially increasing the depth of reads, any variation recorded among technical replicates will also help identify inconsistencies associated with sampling techniques, PCR biases or sequencing errors. In some rare cases, sample collection, storage or processing can be a source of technical variance owing to the relative instability of RNA. It is advisable to employ several randomization techniques during sequencing, for example, multiplexing (mixing of different libraries, each tagged using a different barcode), splitting technical repeats between multiple lanes, or randomization of different libraries in the same lane (an excellent review on statistical randomization for RNA-seq can be found elsewhere (Auer and Doerge 2010)). Ultimately, this type of replication provides some measure of the quality and/or reliability of the analysis.

Biological replication alternatively, relates to different biological samples (e.g. same tissue type but from different individuals) that are processed separately (Figure 2). Biological replication is desirable since it quantifies natural variation among individuals within the experimental cohort. Furthermore, increasing the sample size (number of biological replicates) not only increases sequencing depth, but also provides greater statistical power to detect differences among treatments where they may exist.

Nevertheless, with a very large sample size, accommodating both technical and biological variation can become very costly and may also result in a complex assay to analyse. When sequencing individuals from a population with large levels of genetic variation, for example when dealing with wild-caught individuals, the more biological replicates, the more likely it is to capture genuine differential expression among groups. In general, most SGS experiments conducted on crustaceans tend to be under-replicated and while there is no gold standard for this matter, it is currently acceptable for RNA-seq experiments to consist of a minimum of three biological samples to provide adequate statistical power; a number of published studies have shown that the power to detect differential expressed genes improves from two samples to five samples per treatment (Dillies et al. 2013; Kvam et al. 2012). Similarly, other studies have proposed that sequencing fewer reads and including more biological replicates is an effective strategy to increase statistical power and accuracy in large-scale differential expression RNA-seq studies (Liu et al. 2014). More recently, results suggest that at least six biological replicates may be needed in more sophisticated RNA-seq experiments and up to 12 replicates per experimental group (Schurch et al. 2016). However, for samples that are very different from each other in terms of transcription level (for example, differential expression profiles between brain versus ovary), less replication may also be acceptable. It is also important to highlight the fact that replicates in an RNA-seq-based study are required for publication in some journals (e.g., refer to section 2.6.7 at https://www.frontiersin.org/about/author-guidelines). To conclude, we would recommend maximizing biological replicates to include at least three samples for each experimental condition in every non-model decapod crustacean RNA-seq study.

Choice of sequencing platforms

There are several sequencing methods that researchers can choose from, including single-end (SE) /paired-end (PE) reads, strand-specific, or non-strand-specific library preparation. The decision on which is selected will be based on the desired outcome of the study but will also depend on budget constraints. For experiments on crustacean species in general, PE sequencing is recommended to obtain a reliable *de novo* assembly where no reference genome is readily available. Long read sequencing (e.g. PacBio, Nanopore sequencing), proven to be suitable for enhancing continuity of *de novo* transcriptome assembly, is currently relatively expensive and its application has been described elsewhere (Cartolano et al. 2016; Chen et al. 2017; Kuo et al. 2017). Illumina short read sequencing however, is by far the most widely used platform for transcriptome sequencing in crustaceans due to its cost-effectiveness (unit price per nucleotide), fast sequencing times and higher raw read accuracy.

Another consideration is to choose whether stranded sequencing will be needed. In brief, a stranded-specific RNA-seq can retain the gene orientation (sense or antisense transcript). A number of studies have attempted to compare between stranded vs non-stranded approaches and most have shown that a stranded RNA-seq approach is more advantageous due to better assembly of unannotated genes, ability to detect genes on the antisense strand as well as improved continuity of transcripts. New *de novo* assembly programs like Trinity (Grabherr et al. 2011) have a special mode for strand-specific data analysis that has proven to be more effective than non-stranded data (Levin et al. 2010; Parkhomchuk et al. 2009; Sultan et al. 2012; Zhao et al. 2015; Zhong et al. 2011). We therefore recommend strand-specific RNA-seq if possible for non-model decapod crustacean studies (Havird and Santos 2016).

Depth of sequence (number of reads)

The amount of sequencing needed for a given sample is determined by the aims of the experiment, the number of transcribed transcripts and the nature of the species' RNA samples (this is due to the fact that crustacean genomes can be quite complex compared to other invertebrates). To our knowledge, there has been no attempt to investigate the depth required for effective RNA-seq studies in crustaceans. A study of chicken RNA-seq data revealed that approximately 30 million reads (Illumina-75 bp PE) covered all annotated genes, while 10 million reads detected only ~80% (Wang et al. 2011). Whereas RNA-seq samples from six different phyla (Annelida, Arthropoda, Chordata, Cnidaria, Ctenophora and Mollusca) has suggested that approximately 20 million reads for tissue samples and 30 million for whole-animal samples were required to provide a good balance between total coverage and noise (Francis et al. 2013). Based on these data, it is acceptable that 20 million PE reads per sample for a diploid crustacean organism is a reasonable target to aim for, although there is no specific benchmark for all sequencing experiments.

It is also important to note that in order to detect transcripts with low expression, a deeper sequencing strategy may be needed. In the guidelines for the ENCODE project (https://www.encodeproject.org/), an experiment to evaluate similarity between two transcriptional profiles, requires 30 million PE reads that must be mapped to the genome or known transcriptome. Guidelines to detect novel elements or quantification of known transcript isoforms requires deeper sequencing (Refer to the whole guideline at https://www.encodeproject.org/about/experiment-guidelines/#guideline). Another tool, Scotty, can be used to assist in the design phase of RNA-Seq experiments (Busby et al. 2013). This program can confirm if the design applied has sufficient statistical power to detect differentially expressed genes (DEGs) at the predetermined level required. The program is freely available online at http://bioinformatics.bc.edu/marthlab/scotty/scotty.php/.

An interim conclusion to be drawn from the above sections is that there exists a trade-off between depth of reads per sample and the number of samples (which include technical and biological repeats).

The technology employed and financial limitations usually dictate a fine balance between these factors.

Tissues RNA extraction and cDNA library preparation

Library preparation is a crucial step prior to sequencing. It consists of a number of stages including RNA extraction, proper storage of RNA, quality checking of RNA, mRNA isolation and finally cDNA library generation.

In brief, extraction of total RNA from target tissue can be undertaken immediately on-site or samples can be stored in RNA-later® solution for later extraction. It is important to note that RNA is extremely fragile and degrades readily if stored under inappropriate conditions. Additionally, ribonucleases (RNases) which enzymatically degrade RNA pose a constant threat of contamination and degradation of purified RNA. Traditionally, RNA can be stored at −20°C, −80°C (most desirable) or in liquid nitrogen (-196°C) to provide protection. RNA storage solutions that include chelating agents which inhibit RNase activity, can be used, although these might interfere with reverse transcription and should thus be removed prior to these steps. To our knowledge, there is no crustacean specific RNA extraction kit available on the market, however several commercial kits for RNA extraction are still usable for crustaceans, in addition to in-house (modified) versions of RNA extraction methods that use Beta-mercaptoethanol or phenol-based compounds, with the latter being more popular in recent publications. A detailed review on the effect of RNA extraction methods on RNA-seq can be found elsewhere (Sultan et al. 2014). RNA can then be assessed for quality and quantity using a Nanodrop® spectrophotometer or BioAnalyzer®. It is important to note that RNA integrity number (RIN) that has been used as a standardized metric of RNA quality for vertebrate species, is not usually valid for crustacean samples with non-typical RNA profiles. RIN is calculated based on the ratio between 18S ribosomal RNA (rRNA) and 28S rRNA band intensities, which are usually very conserved across eukaryotes. However, the 28S rRNA of arthropods tends to break down into two subunits, preventing a reliable RIN value calculation (Macharia et al. 2015; McCarthy et al. 2015; Winnebeck et al. 2010).

RNA can be stored and shipped in ambient conditions after desiccation with RNA-stable solution (Seelenfreund et al. 2014). An important consideration when it comes to RNA extraction in crustacean species is tissues with high pigment content (i.e. eyestalk). For these tissue, extra caution is suggested to avoid extracting pigment contamination that will affect the quality of library preparation. Currently, there is no threshold for deciding if a sample is too degraded for whole-transcriptome analysis. In most cases however, sequencing facilities provide users with specific guidelines and technical notes recommended for producing the best results. Moreover, depending on each sequencing platform, different cDNA library preparation protocols may be required.

POST-SEQUENCING

Post-sequencing analyses include quality checking of raw sequences, trimming, *de novo* assembly of trimmed reads, read mapping and quantification, DEG assessment and finally biological interpretation (Figure 1).

Quality control for SGS data

Current SGS runs generate millions, or even hundreds of millions of read sequences. Technologies advancement reduces the error rate; however, every platform still produces read errors that require the application of a quality control program post-sequencing. Read errors, while relatively negligible in number compared with the massive dataset generated, still pose a hurdle for downstream analysis. For instance, errors in base-calling cause improper connection of nodes in *de novo* assembly (thus expanding running time and increase memory needed to store the nodes). In addition, incorrect SNP detection can result from an inability to differentiate between a true polymorphism and a sequencing error (Kelley et al. 2010). Several quality control tools have been developed for NGS data (most popular tools are summarized in Table 1).

In general, quality control of raw reads from NGS sequencers can be completed in a few simple steps. Raw read statistics can then be checked with FASTQC software (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). A variety of parameters can be used to trim the data. The most important is the PHRED quality score (a base-calling score ranking system that allows users to judge the confidence of a nucleotide presumed to be correctly called (Ewing et al. 1998)). Some other considerations include reads average length, total number of base pairs and adapters contamination. In addition, reads generated on the Illumina platform are considered to have a relatively higher error rate towards the 3′-end of the read (Schirmer et al. 2015), so if a drop in quality is detected, it is acceptable to trim off a portion of the read from that end. Some commonly used criteria for trimming reads include; minimum read length, minimum quality score, and homopolymer trimming. Read duplication is also a factor to consider during the quality control step in NGS projects. In brief, read duplications are identical reads that map to the same genomic location (effects of PCR amplification bias, excess computational resources, and errors). Raw reads may also need to be cleaned from artificially introduced sequences - PCR primers or sequencing adapters; these are usually addressed in most quality control packages. In a benchmarking study, it was shown that trimming applied in every sequencing project will improve not only quality of the results, but also reduce analysis duration (Del Fabbro et al. 2013). In general, normal quality trimming with a PHRED score ranging from 20 to 30 is normal for most RNA-seq experiments, while a PHRED score threshold of 30 or above is usually required for variant calling experiments (Ledergerber and Dessimoz 2011). However, in one particular study, the authors highlighted that although strict trimming is usually applied, in some cases a more gentle trimming (PHRED score <2 or <5) might be more optimal (MacManes 2014). This is due to the fact that short and low expressed transcripts suffer from heavy negative bias when using harsh trimming (MacManes 2014). Therefore, lowering the PHRED score threshold in the quality control step can result in a greater transcript discovery rate. As a conclusion, we suggest gentle trimming initially as suggested in the above study. A list of some popular software packages for NGS quality control can be found in Table 1.

De novo assembly for non-model species and transcript clustering

For non-model decapod species, it is often difficult to align RNA-seq data to a reference genome from relatively recently diverged organisms (currently there are very few crustacean reference genomes available – see section 3.2). An alternative strategy therefore, is to construct a *de novo* assembly (new assembly) from high-quality reads. The primary aim is to extend the short reads from the sequencer into longer continuous sequences (contigs) that reflect the mRNAs transcribed in the cell without any chimeric/fusion events. A number of *de novo* transcriptome assemblers have been developed (initially they were simply modified genome assemblers), including the Velvet/Oases pipeline (Schulz et al. 2012; Zerbino and Birney 2008), SOAPdenovo (https://soap.genomics.org.cn/soapdenovo.html) and Trans-Abyss (Robertson et al. 2010). More recently, the Trinity software (Grabherr et al. 2011) has become available, developed specifically for *de novo* transcriptome assembly from short-read RNA-seq data. Since reads from SGS are short in length compared with pyrosequencing output (Liu et al. 2012), transcriptome *de novo* assemblers often employ a De Bruijn graph algorithm instead of the traditional **O**verlap **L**ayout **C**onsensus (OLC). This minimizes the amount of memory required to handle numerous parallel calculations. Further information on graph algorithms can be found elsewhere (Li et al. 2012; Miller et al. 2010).

Most *de novo* assemblers are freely distributed but usually required operating using command line, which deters many biologists without programming skills. To overcome this issue, bioinformatics platforms such as Galaxy (https://www.usegalaxy.org/) and CyVerse (https://www.cyverse.org/) embed command line packages into user-friendly interfaces. Yet, there is a limited flexibility in utilizing these tools. Learning how to use command line programming can be time consuming and potentially is out of reach for many non-model biology researchers and this can slow the pace at which NGS studies are performed on these species. To address this, users can use commercial products (usually with a “point-and click” user-friendly interface) that are available on the market. A summary of some notable *de novo* assemblers can be found in Table 2. A performance comparison of commonly used *de novo* transcriptome assemblers can be found elsewhere (Amin et al. 2014; Finseth and Harrison 2014; Ghangal et al. 2013; Surget-Groba and Montoya-Burgos 2010; Zhao et al. 2011).

One significant challenge associated with *de novo* assembly is the lack of software to identify the assembly that is most accurate. To address this challenge, **S**equence **C**omparative **A**nalysis using **N**etworks (SCAN) was created (Misner et al. 2013). SCAN uses a reference dataset (from a related genome) to identify the most accurate *de novo* assembly and to classify "good" transcripts in these assemblies (Misner et al. 2013). A similar program was generated for this purpose, named DETONATE (an abbreviation of ***DE*** *novo* **T**ranscript**O**me r**N**a-seq **A**ssembly with or without the **T**ruth **E**valuation) (Xie et al. 2014). This program combines multiple factors into a single evaluation score that then can be used to select the best assembler. The software is distributed freely at https://deweylab.biostat.wisc.edu/detonate. Another approach is to employ the CEGMA pipeline (**C**ore **E**ukaryotic **G**enes **M**apping **A**pproach) (Parra et al. 2007) or BUSCO (**B**enchmarking **U**niversal **S**ingle-**C**opy **O**rthologs) (Simão et al. 2015). These programs scan the *de novo* assembly against a dataset of core eukaryotic genes that are well conserved across several eukaryotic taxa, to calculate the coverage of protein-coding genes, thus estimating the degree of completeness of the reconstruction and the full-length complement of transcript sequences comprising the *de novo* transcriptome assembly. As a concluding remark, benchmarking assemblies are an option that can be trialled, but the practice is still in development.

Transcriptome mapping

Following *de novo* assembly, reads can be aligned against the *de novo* assembly (mapping). The mapping step can serve two purposes: i) a remapping step can be used to assess the assembly quality and ii) the alignment can then be quantified; gene expression levels can be inferred from the total counts of reads aligned to each contig. Furthermore, mapping also enables variant calling for transcripts of interest.

Stringent parameters may result in a small subset of reads mapped, while less stringent settings reduce mapped read specificity. To gain a balance between sensitivity and specificity, trials with different parameters can be performed. Popular aligners for RNA-seq include: Bowtie 1 (Langmead et al. 2009) and Bowtie 2 (Langmead and Salzberg 2012), BWA (Li and Durbin 2009), GSnap (Wu and Nacu 2010), and commercial programs including CLC Genomics Workbench®, DNA-STAR® or Partek Genomics®. A detailed list of available aligners can be found at https://www.ebi.ac.uk/~nf/hts_mappers/ (Fonseca et al. 2012). Comparisons of different aligners usually takes into consideration running time, accuracy, as well as the sensitivity of mapped reads (Baruzzo et al. 2017; Grant et al. 2011; Hatem et al. 2013; Li and Homer 2010). Critically, for non-model organisms where no genome sequence is available, it is hard to define which are the best mapping parameters to apply. This is due to the occurrence of isoforms and splice variants that cannot be accurately determined without access to a reference genome. Reads can be mapped randomly to shared exons between splice variants, biasing the resulting count and confounding the biological interpretation.

Quantifying transcript level and analysis of differential gene expression

To quantify gene expression, RNA-seq reads need to be aligned to a reference genome from model organisms or to the transcriptome sequences reconstructed using *de novo* assembly strategies for organisms without reference genome sequences. The number of mapped reads is calculated based on the outcome of the alignment and can be used to estimate the relative expression level of individual genes. Following this, statistical methods are applied to test for significant differences among experimental groups. The data however, first needs to be normalized since there are inherent differences in total reads per sample, resulting in over-represented long transcripts. With rapid development of RNA-seq technology, there are now numerous tools available to estimate gene expression levels, which vary in their efficiency. Popular RNA-seq quantification (reads counting) tools include: RSEM (Li and Dewey 2011), eXpress (Roberts and Pachter 2012), HTSeq (Anders et al. 2015), Salmon (Patro et al. 2017) and kallisto (Bray et al. 2016). Several studies have also been conducted to compare the pros and cons of each tool (Chandramohan et al. 2013; Li and Homer 2010; Teng et al. 2016).

DEG analysis programs perform statistical tests to determine if fold change results under different experimental conditions are significant (e.g. among tissues types, life stages etc.). Many programs have been developed for DEG analysis (a brief summary of popular DEG tools can be found in Table 3) and several comparative assessments are available (Khang and Lau 2015; Kvam et al. 2012; Rajkumar et al. 2015; Robles et al. 2012; Soneson and Delorenzi 2013; Zhang et al. 2014). Much like assembly and mapping, there is no guarantee as to which tool is the best, or which parameters will result in the highest accuracy or robustness of the results generated (Zhang et al. 2014). Most DEG call methods are designed to address analysis of RNA-seq experiments that have biological replicates. There are a few tools however, that can handle non-replicated experiments (e.g. GFOLD (Feng et al. 2012), EdgeR (Robinson et al. 2010), NOISeq (Tarazona et al. 2011). A recent study recommended using EdgeR (Robinson et al. 2010) or DESeq2 (Love et al. 2014) for experiments with less than 12 replicates per group, while they suggest studies with more than 12 replicates should use DESeq2 for the statistical analysis (Schurch et al. 2016). An alternative strategy is to employ several software packages and then compare the outcome of each approach, highlighting not only the similarity, but also differences among these analyses. Fold change is an important parameter to consider, but will depend on the number of reads that are assigned to a specific transcript. If the depth is low, yet with high fold change between groups, it should be considered as noise. For example, 10 X 100 base reads mapped onto a 1 Kb transcript per sample in one group (giving an average depth of 1) compared to 1 read on average per sample in the other group is a 10-fold change, yet the coverage is very low and should be validated using additional samples via qPCR.

Annotation of transcripts

After all reads have been assembled *de novo* into contigs, the next step is to annotate all the contigs based on the most up-to-date database (i.e. identify homology to previously characterized genes). The most common way to annotate a large number of transcripts is the Basic Local Alignment Search Tool (BLAST). As the number of contigs in every *de novo* assembly can be thousands to a few hundred thousand sequences, usage of an automated search tool, in particular BLAST+ (Camacho et al. 2009), is essential. For non-model species, many candidate protein databases are available including the non-redundant protein database (nr), UniProtKB/Swiss-Prot database, and the Reference Sequence database (RefSeq). RefSeq (nucleotide and protein) and UniProt/Swiss-Prot (protein) consist of curated, well annotated sequences, whereas the nr database includes both curated and non-curated databases. For most crustacean RNA-seq experiments, the nr database is considered to be the best choice due to the fact that very few crustacean genes have been properly annotated to date, a problem that has been highlighted (Clark and Greenwood 2016; Das and Mykles 2016).

After transcripts have been scanned against the protein database and assigned annotations, there is a variety of downstream packages that can further analyse a contig, including Gene Ontology (GO) term analysis, functional enrichment analysis, protein domain analysis (PFAM domain search - pfam.xfam.org), and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway maps (Kanehisa and Goto 2000). Each entry in the sequence database can be classified into a number of biologically relevant terms. In GO analysis, most genes can be assigned to one out of three basic ontologies: cellular component, biological process or molecular function. When comparing samples from two groups, differentially represented GO terms can help define the mechanism via which the groups differ from one another. Similarly, when using KEGG, contigs can be assigned with components of specific pathways. Differential expression can allow a detailed assessment of the changes in pathways between studied groups. PFAM shows domains within the open reading frames of contigs that enables characterization of the protein function, based on the architecture of domains in a polypeptide chain. A number of software packages are capable of extracting vast numbers of GO terms from public databases including Blast2GO (Conesa et al. 2005), DAVID - https://david.abcc.ncifcrf.gov/ (Huang da et al. 2009) or ermineJ (Lee et al. 2005). Among these programs, Blast2GO stands out as an easy to use, point-and-click program that has become very popular in the last few years.

In addition to Blast2GO, databases like PFAM (Finn et al. 2016), eggNOG (Huerta-Cepas et al. 2015) and InterProScan (Jones et al. 2014) can be employed to predict the function of unknown proteins. The Trinity RNA-seq package, Trinotate (https://trinotate.github.io/), uses UniProt, eggNOG and GO Pathway databases for annotating novel sequences and these have been widely used in recent years (Das and Mykles 2016; Das et al. 2016).

Validation of RNA-seq results

Validation is a very important step in every RNA-seq study. There is generally a very high correlation between RNA-seq and qPCR results with respect to relative gene expression. A significant point is that testing the same RNA samples used in the NGS platform for validation with techniques like qPCR or digital PCR only validates the sequencing accuracy result. Therefore, additional, independent biological replicates should be included to properly validate the biological interpretation from the RNA-seq experiment. Essentially, validation post sequencing is now mandatory for publication. An approach has been proposed to set the minimum acceptable standard for qPCR validation (Fang and Cui 2011) that takes a number of factors into consideration, including the number of genes tested and the number of isoform transcripts detected in the transcriptome. Nevertheless, up-to-date, qPCR techniques offer the easiest way to validate data in a transcriptomics study. One important note for researchers who are unfamiliar with the technology is that some RNA-seq pipelines allow RNA-seq analysis at the gene level (Trinity/RSEM for instance (Haas et al. 2013)). However, there is a deeper level of transcripts component (in which transcripts can be isoforms resulting from alternative splicing events or a single nucleotide variation). Therefore, researchers should design primers that are not included in these regions to avoid unreliable qPCR results between biological replicates. As a concluding remark for this section, several studies have compared RNA-seq results to qPCR data, and have found excellent correlations between the approaches (Everaert et al. 2017; Rajkumar et al. 2015; Wu et al. 2014).

PROMISES AND CHALLENGES OF RNA-SEQ BASED STUDIES IN CRUSTACEANS

Studied topics

SGS has revolutionized biological science, shifting it toward the post-genomics era. Transcriptomics studies in crustaceans include either:

1. Sequencing and annotation of the transcriptome of one (or several) tissue/s, or a whole individual of a particular taxa in a specific developmental stage or under specific experimental conditions.
2. Applying RNA-seq to identify DEGs among different physiological conditions, treatments, developmental stages and/or tissues.
3. Identification of novel transcripts – enzymes, receptors, hormones, neuropeptides.
4. Screening for variant mutations - SNPs, SSRs and/or microsatellites.
5. A combination of the above.

To date, several RNA-seq projects have been initiated on a variety of crustacean species. In Table 5, we have summarized several RNA-seq based studies on crustacean taxa that have been conducted over the last few years based on the following categories: *Aquatic toxicology*, *Reproduction & sexual differentiation*, *Disease resistance & immunology*, *Developmental biology*, and *Physiology*. This is however, by no means an exhaustive list as hundreds of applied RNA-seq studies have been undertaken in recent years, rather the list here illustrates several model comparative RNA-seq approaches.

Ongoing challenges for applied RNA-seq studies of crustaceans

– *Experimental limitation:* A good experimental design will have a major impact on data outcomes; it can prevent wasted resources and help avoid the generation of unpublishable results. The balance between sequencing cost and experimental design constraints is a major issue that has been highlighted in many review articles. Due to budgetary limitations, there will always be an incentive to cut costs by sequencing with higher depth but with little or no biological replication. Furthermore, where depth is added, a large number of reads will be also mapped to the already well-covered regions, while if additional replication was available, greater statistical power can be achieved resulting in better biological inference. To resolve this problem, optimal guidelines for the design of RNA-seq experiments are needed and should be applied accordingly. In parallel, biological replicates (at least 3 or greater) are required for an RNA-seq study to reach a basic publishable level. As another recommendation for best practice, is undertaking a pilot-sequencing project where a high number of libraries are run on one lane initially. This can be valuable in assessing the feasibility of the larger experiment, as well as providing a good indicator for how to address trade-offs between obtaining high quality output vs cost. Finally, although RNA-seq methods are becoming more robust and reliable and sometimes qPCR validations are proven to be unnecessary, a section for qPCR validation of selected genes/transcripts of interest may be beneficial to reveal the biological insights if the study has limited replications. Therefore, we recommend that for reliable biological interpretation and validation of RNA-seq analysis, the candidate genes themselves are tested for expression, rather than choosing random genes or genes showing high expression levels.

– *In silico annotation and functional annotation:* Annotation of RNA-seq data is based loosely on BLAST searches. In fact, many BLAST results produce “hypothetical”, “predicted”, “uncharacterized”, or “low-quality” assignments. This highlights the fact that gene databases for non-model species currently, are very limited. To add another layer of complexity, *Daphnia pulex*, the model species currently available for crustaceans, has a large number of genes that currently remain unannotated. Furthermore, when compared with Decapoda, it is very remotely related and in many cases, shares higher similarity with insects than with other crustacean taxa. Further downstream annotation is also a constraint for crustacean RNA-seq studies, as specific GO classification and KEGG pathways are still not available for these taxa. As a result, drawing biological interpretations from predicted results can be problematic. Moreover, similarities in structure do not necessarily correlate with equivalent functionalities. It is crucial therefore, to highlight that *in silico* prediction is only speculative and functional annotation is very important to validate any biological interpretations (in particular for novel genes). RNAi technology (gene silencing) is now the go-to method for gene functional studies in decapod crustaceans and it has been already applied in some cases (Sagi et al. 2013). Gene editing technologies, for example CRISPR/Cas9 technology, have emerged recently and hold great potential for functional annotation in decapod crustacean species (Mykles and Hui 2015). Employing RNAi and/or CRISPR-Cas9 in RNA-seq studies would be extremely helpful to highlight key genes and resolve functional roles of novel genes for crustacean species.

– *Combining transcriptomics/RNA-seq with other OMICS techniques:* In parallel with advances in RNA-seq technologies, other OMICS technologies including genomics, proteomics, metagenomics phylogenomics and phenomics have also developed rapidly. This highlights a challenge for RNA-seq studies, to make use of other OMICS approaches and to utilize them to create a multilayer outcome. One key reason why decapod crustacean genomes are not yet available is that they are often very large and complex which makes them hard to resolve. Nevertheless, draft genomes of a few crustacean species have been made publicly available recently including draft genomes for some decapods including: *N. denticulata* (Kenny et al. 2014), *P. vannamei* (Yu et al. 2015), *E. sinensis* (Song et al. 2016), *P. hawaiensis* (Kao et al. 2016), *P. monodon and M. japonicus (Yuan et al. 2017), and P. virginalis (Gutekunst et al. 2018)*. Utilizing these new genomic resources will allow better gene annotation and functional annotation of crustacean gene pathways. There is no doubt that in the near future, when the cost barrier for sequencing is essentially overcome, coupled with improved sequencing technologies, combining RNA-seq approaches with integrated OMICS will enable researchers to answer the most complex of biological questions.

CONCLUSIONS

To conclude, RNA-seq offers great promise for crustacean studies. It is a very powerful tool that can lead to developing a better understanding of underlying pathways and mechanisms that form the basis of many scientific questions. The guidelines offered here for future RNA-seq studies of crustaceans are an attempt to assist biologists who are not familiar with the complex and diverse array of bioinformatics software that are currently available. It is also important however, to highlight the gap between *in silico* prediction from RNA-seq analysis and *in vivo* results. This may be explained in general, by limitations on experimental designs in the past, the lack of annotation databases for crustacean species, as well as the need for question-driven research. In the future, we also suggest that RNA-seq should be integrated with other OMICs technologies to increase data output as well as improving biological insights.

Acknowledgement

The current work was supported through a Marie Curie International Research Staff Exchange Scheme Fellowship within the 7th European Community Framework Programme (612296-DeNuGReC). Tuan Viet Nguyen was supported through the Australian Research Council Discovery Project grant awarded to Dr Tomer Ventura (No. DP160103320) and a USC International PhD scholarship. The authors would like to acknowledge the precious help of Professor Abigail Elizur (University of the Sunshine Coast, Australia) and four anonymous reviewers for numerous feedback that improve the quality of this manuscript.

Species abbreviation

P. trituberculatus: Portunus trituberculatus

S. henanense: Sinopotamon henanense

P. vannamei: Penaeus vannamei

P. monodon: Penaeus monodon

M. japonicus: Marsupenaeus japonicus

P. virginalis: Procambarus virginalis or Procambarus fallax forma virginalis

S. olivacea: Scylla olivacea

S. paramamosain: Scylla paramamosain

E. sinensis: Eriocheir sinensis

N. norvegicus: Nephrops norvegicus

F. merguiensis: Fenneropenaeus merguiensis

P. clarkii: Procambrarus clarkii

M. rosenbergii: Macrobrachium rosenbergii

S. verreauxi : Sagmariasus verreauxi

N. denticulata: Neocaridina denticulata

P. hawaiensis: Parhyale hawaiensis

E. carinicauda: Exopalaemon carinicauda

M. olfersi: Macrobrachium olfersi

P. elegans: Palaemon elegans

P. australiensis: Paratya australiensis

E. carinicauda: Exopalaemon carinicauda

H. rubra: Halocaridina rubra

List of Figures

List of Tables

References

Amin, S., P. Prentis, E. Gilding & A. Pavasovic, 2014. Assembly and annotation of a non-model gastropod (Nerita melanotragus) transcriptome: a comparison of *De novo* assemblers. BMC Research Notes 7(1):488 doi:10.1186/1756-0500-7-488.

Anders, S., P. T. Pyl & W. Huber, 2015. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics 31(2):166-9 doi:10.1093/bioinformatics/btu638.

Auer, P. L. & R. W. Doerge, 2010. Statistical design and analysis of RNA sequencing data. Genetics 185(2):405-416 doi:10.1534/genetics.110.114983.

Baruzzo, G., K. E. Hayer, E. J. Kim, B. Di Camillo, G. A. FitzGerald & G. R. Grant, 2017. Simulation-based comprehensive benchmarking of RNA-seq aligners. Nature Methods 14(2):135-139 doi:10.1038/nmeth.4106.

Bray, N. L., H. Pimentel, P. Melsted & L. Pachter, 2016. Near-optimal probabilistic RNA-seq quantification. Nature Biotechnology 34(5):525-527 doi:10.1038/nbt.3519.

Busby, M. A., C. Stewart, C. A. Miller, K. R. Grzeda & G. T. Marth, 2013. Scotty: a web tool for designing RNA-Seq experiments to measure differential gene expression. Bioinformatics 29(5):656-657 doi:10.1093/bioinformatics/btt015.

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer & T. L. Madden, 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421 doi:10.1186/1471-2105-10-421.

Cartolano, M., B. Huettel, B. Hartwig, R. Reinhardt & K. Schneeberger, 2016. cDNA library enrichment of full length transcripts for SMRT long read sequencing. PLoS ONE 11(6):e0157779 doi:10.1371/journal.pone.0157779.

Chandramohan, R., P. Y. Wu, J. H. Phan & M. D. Wang, 2013. Systematic assessment of RNA-Seq quantification tools using simulated sequence data. ACM Conference on Bioinformatics, Computational Biology and Biomedicine 2013 doi:10.1145/2506583.2506648.

Chen, S.-Y., F. Deng, X. Jia, C. Li & S.-J. Lai, 2017. A transcriptome atlas of rabbit revealed by PacBio single-molecule long-read sequencing. Scientific Reports 7(1):7648 doi:10.1038/s41598-017-08138-z.

Clark, K. F. & S. J. Greenwood, 2016. Next-Generation Sequencing and the crustacean immune system: The need for alternatives in immune gene annotation. Integrative and Comparative Biology 56(6):1113-1130 doi:10.1093/icb/icw023.

Conesa, A., S. Gotz, J. M. Garcia-Gomez, J. Terol, M. Talon & M. Robles, 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21(18):3674-6 doi:10.1093/bioinformatics/bti610.

Conesa, A., P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szcześniak, D. J. Gaffney, L. L. Elo, X. Zhang & A. Mortazavi, 2016. A survey of best practices for RNA-seq data analysis. Genome Biology 17(1):13 doi:10.1186/s13059-016-0881-8.

Das, S. & D. L. Mykles, 2016. A comparison of resources for the annotation of a *de novo* assembled transcriptome in the molting gland (Y-Organ) of the Blackback Land Crab, *Gecarcinus lateralis*. Integrative and Comparative Biology 56(6):1103-1112 doi:10.1093/icb/icw107.

Das, S., S. Shyamal & D. S. Durica, 2016. Analysis of annotation and differential expression methods used in RNA-seq Studies in crustacean systems. Integrative and Comparative Biology 56(6):1067-1079 doi:10.1093/icb/icw117.

Del Fabbro, C., S. Scalabrin, M. Morgante & F. M. Giorgi, 2013. An extensive evaluation of read trimming effects on Illumina NGS data analysis. PLoS ONE 8(12):e85024 doi:10.1371/journal.pone.0085024.

Dillies, M.-A., A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, G. Marot, D. Castel, J. Estelle, G. Guernec, B. Jagla, L. Jouneau, D. Laloë, C. Le Gall, B. Schaëffer, S. Le Crom, M. Guedj & F. Jaffrézic, 2013. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Briefings in Bioinformatics 14(6):671-683 doi:10.1093/bib/bbs046.

Everaert, C., M. Luypaert, J. L. V. Maag, Q. X. Cheng, M. E. Dinger, J. Hellemans & P. Mestdagh, 2017. Benchmarking of RNA-sequencing analysis workflows using whole-transcriptome RT-qPCR expression data. Scientific Reports 7:1559 doi:10.1038/s41598-017-01617-3.

Ewing, B., L. Hillier, M. C. Wendl & P. Green, 1998. Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. Genome research 8(3):175-185.

Fang, Z. & X. Cui, 2011. Design and validation issues in RNA-seq experiments. Briefings in Bioinformatics doi:10.1093/bib/bbr004.

Feng, J., C. A. Meyer, Q. Wang, J. S. Liu, X. Shirley Liu & Y. Zhang, 2012. GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data. Bioinformatics 28(21):2782-8 doi:10.1093/bioinformatics/bts515.

Finn, R. D., P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, G. A. Salazar, J. Tate & A. Bateman, 2016. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Research 44(D1):D279-D285 doi:10.1093/nar/gkv1344.

Finseth, F. R. & R. G. Harrison, 2014. A comparison of Next Generation Sequencing technologies for transcriptome assembly and utility for RNA-Seq in a non-model bird. PLoS ONE 9(10):e108550 doi:10.1371/journal.pone.0108550.

Fonseca, N. A., J. Rung, A. Brazma & J. C. Marioni, 2012. Tools for mapping high-throughput sequencing data. Bioinformatics doi:10.1093/bioinformatics/bts605.

Francis, W. R., L. M. Christianson, R. Kiko, M. L. Powers, N. C. Shaner & S. H. Haddock, 2013. A comparison across non-model animals suggests an optimal sequencing depth for *de novo* transcriptome assembly. BMC Genomics 14(1):167.

Ghangal, R., S. Chaudhary, M. Jain, R. S. Purty & P. Chand Sharma, 2013. Optimization of *De Novo* short read assembly of Seabuckthorn *Hippophae rhamnoide*s L. transcriptome. PLoS ONE 8(8):e72516 doi:10.1371/journal.pone.0072516.

Glenn, T. C., 2011. Field guide to next-generation DNA sequencers. Molecular Ecology Resources 11(5):759-769 doi:10.1111/j.1755-0998.2011.03024.x.

Goodwin, S., J. D. McPherson & W. R. McCombie, 2016. Coming of age: ten years of Next Generation Sequencing technologies. Nature Reviews Genetics 17(6):333-351 doi:10.1038/nrg.2016.49.

Grabherr, M., B. Haas, M. Yassour, J. Levin, D. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman & A. Regev, 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology 29(7):644-652 doi:10.1038/nbt.1883.

Grant, G. R., M. H. Farkas, A. D. Pizarro, N. F. Lahens, J. Schug, B. P. Brunk, C. J. Stoeckert, J. B. Hogenesch & E. A. Pierce, 2011. Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). Bioinformatics 27(18):2518-2528 doi:10.1093/bioinformatics/btr427.

Gutekunst, J., R. Andriantsoa, C. Falckenhayn, K. Hanna, W. Stein, J. Rasamy & F. Lyko, 2018. Clonal genome evolution and rapid invasive spread of the marbled crayfish. Nature Ecology & Evolution 2(3):567-573 doi:10.1038/s41559-018-0467-9.

Haas, B. J., A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood, J. Bowden, M. B. Couger, D. Eccles, B. Li, M. Lieber, M. D. MacManes, M. Ott, J. Orvis, N. Pochet, F. Strozzi, N. Weeks, R. Westerman, T. William, C. N. Dewey, R. Henschel, R. D. LeDuc, N. Friedman & A. Regev, 2013. *De novo* transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. Nature Protocols 8(8):10.1038/nprot.2013.084 doi:10.1038/nprot.2013.084.

Hatem, A., D. Bozdağ, A. E. Toland & Ü. V. Çatalyürek, 2013. Benchmarking short sequence mapping tools. BMC Bioinformatics 14(1):184 doi:10.1186/1471-2105-14-184.

Havird, J. C. & S. R. Santos, 2016. Here we are, but where do we go? A systematic review of crustacean transcriptomic studies from 2014–2015. Integrative and Comparative Biology 56(6):1055-1066 doi:10.1093/icb/icw061.

Huang da, W., B. T. Sherman & R. A. Lempicki, 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Protocols 4(1):44-57 doi:10.1038/nprot.2008.211.

Huerta-Cepas, J., D. Szklarczyk, K. Forslund, H. Cook, D. Heller, M. C. Walter, T. Rattei, D. R. Mende, S. Sunagawa & M. Kuhn, 2015. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Research 44(D1):D286-D293 doi:10.1093/nar/gkv1248.

Jaramillo, M. L., F. Guzman, C. L. Paese, R. Margis, E. M. Nazari, D. Ammar & Y. M. R. Müller, 2016. Exploring developmental gene toolkit and associated pathways in a potential new model crustacean using transcriptomic analysis. Development Genes and Evolution 226(5):325-337 doi:10.1007/s00427-016-0551-6.

Jin, S., H. Fu, Q. Zhou, S. Sun, S. Jiang, Y. Xiong, Y. Gong, H. Qiao & W. Zhang, 2013. Transcriptome analysis of androgenic gland for discovery of novel genes from the oriental river prawn, *Macrobrachium nipponense*, using Illumina Hiseq 2000. PloS one 8(10):e76840 doi:10.1371/journal.pone.0076840.

Jones, P., D. Binns, H.-Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, J. Maslen, A. Mitchell, G. Nuka, S. Pesseat, A. F. Quinn, A. Sangrador-Vegas, M. Scheremetjew, S.-Y. Yong, R. Lopez & S. Hunter, 2014. InterProScan 5: Genome-scale protein function classification. Bioinformatics 30(9):1236-1240 doi:10.1093/bioinformatics/btu031.

Jung, H., R. E. Lyons, H. Dinh, D. A. Hurwood, S. McWilliam & P. B. Mather, 2011. Transcriptomics of a Giant Freshwater Prawn (*Macrobrachium rosenbergii*): *De novo* assembly, annotation and marker discovery. PLoS ONE 6(12):e27938 doi:10.1371/journal.pone.0027938.

Jung, H., B.-H. Yoon, W.-J. Kim, D.-W. Kim, D. Hurwood, R. Lyons, K. Salin, H.-S. Kim, I. Baek, V. Chand & P. Mather, 2016. Optimizing hybrid *de novo* transcriptome assembly and extending genomic resources for Giant Freshwater Prawns (*Macrobrachium rosenbergii*): The identification of genes and markers associated with reproduction. International Journal of Molecular Sciences 17(5):690 doi:10.3390/ijms17050690.

Kanehisa, M. & S. Goto, 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research 28(1):27-30.

Kao, D., A. G. Lai, E. Stamataki, S. Rosic, N. Konstantinides, E. Jarvis, A. Di Donfrancesco, N. Pouchkina-Stancheva, M. Sémon, M. Grillo, H. Bruce, S. Kumar, I. Siwanowicz, A. Le, A. Lemire, M. B. Eisen, C. Extavour, W. E. Browne, C. Wolff, M. Averof, N. H. Patel, P. Sarkies, A. Pavlopoulos & A. Aboobaker, 2016. The genome of the crustacean *Parhyale hawaiensis*, a model for animal development, regeneration, immunity and lignocellulose digestion. eLife 5:e20062 doi:10.7554/eLife.20062.

Kelley, D. R., M. C. Schatz & S. L. Salzberg, 2010. Quake: quality-aware detection and correction of sequencing errors. Genome Biology 11(11):R116 doi:10.1186/gb-2010-11-11-r116.

Kenny, N. J., Y. W. Sin, X. Shen, Q. Zhe, W. Wang, T. F. Chan, S. S. Tobe, S. M. Shimeld, K. H. Chu & J. H. Hui, 2014. Genomic sequence and experimental tractability of a new decapod shrimp model, *Neocaridina denticulata*. Marine Drugs 12(3):1419-37 doi:10.3390/md12031419.

Khang, T. F. & C. Y. Lau, 2015. Getting the most out of RNA-seq data analysis. PeerJ 3:e1360 doi:10.7717/peerj.1360.

Koboldt, Daniel C., Karyn M. Steinberg, David E. Larson, Richard K. Wilson & E. R. Mardis, 2013. The next-generation sequencing revolution and its impact on genomics. Cell 155(1):27-38 doi:10.1016/j.cell.2013.09.006.

Kuo, R. I., E. Tseng, L. Eory, I. R. Paton, A. L. Archibald & D. W. Burt, 2017. Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human. BMC Genomics 18(1):323 doi:10.1186/s12864-017-3691-9.

Kvam, V. M., P. Liu & Y. Si, 2012. A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. American Journal of Botany 99(2):248-256 doi:10.3732/ajb.1100340.

Lahens, N. F., E. Ricciotti, O. Smirnova, E. Toorens, E. J. Kim, G. Baruzzo, K. E. Hayer, T. Ganguly, J. Schug & G. R. Grant, 2017. A comparison of Illumina and Ion Torrent sequencing platforms in the context of differential gene expression. BMC Genomics 18(1):602 doi:10.1186/s12864-017-4011-0.

Lam, H. Y. K., M. J. Clark, R. Chen, R. Chen, G. Natsoulis, M. O'Huallachain, F. E. Dewey & L. Habegger, 2012. Performance comparison of whole-genome sequencing platforms. Nature Biotechnology 30 doi:10.1038/nbt.2065.

Langmead, B. & S. L. Salzberg, 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods 9(4):357-359 doi:10.1038/nmeth.1923.

Langmead, B., C. Trapnell, M. Pop & S. L. Salzberg, 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10(3):R25 doi:10.1186/gb-2009-10-3-r25.

Ledergerber, C. & C. Dessimoz, 2011. Base-calling for next-generation sequencing platforms. Briefings in Bioinformatics 12(5):489-497 doi:10.1093/bib/bbq077.

Lee, H. K., W. Braynen, K. Keshav & P. Pavlidis, 2005. ErmineJ: tool for functional analysis of gene expression data sets. BMC Bioinformatics 6(1):269.

Levin, J. Z., M. Yassour, X. Adiconis, C. Nusbaum, D. A. Thompson, N. Friedman, A. Gnirke & A. Regev, 2010. Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nature Methods 7(9):709-15 doi:10.1038/nmeth.1491.

Li, B. & C. N. Dewey, 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12:323 doi:10.1186/1471-2105-12-323.

Li, H. & R. Durbin, 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25(14):1754-1760 doi:10.1093/bioinformatics/btp324.

Li, H. & N. Homer, 2010. A survey of sequence alignment algorithms for next-generation sequencing. Briefings in Bioinformatics 11(5):473-483 doi:10.1093/bib/bbq015.

Li, Z., Y. Chen, D. Mu, J. Yuan, Y. Shi, H. Zhang, J. Gan, N. Li, X. Hu, B. Liu, B. Yang & W. Fan, 2012. Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-Bruijn-graph. Briefings in Functional Genomics 11(1):25-37 doi:10.1093/bfgp/elr035.

Lister, R., B. D. Gregory & J. R. Ecker, 2009. Next is now: new technologies for sequencing of genomes, transcriptomes, and beyond. Current Opinion in Plant Biology 12(2):107-118 doi:10.1016/j.pbi.2008.11.004.

Liu, L., Y. Li, S. Li, N. Hu, Y. He, R. Pong, D. Lin, L. Lu & M. Law, 2012. Comparison of Next-Generation Sequencing systems. Journal of Biomedicine and Biotechnology 2012:11 doi:10.1155/2012/251364.

Liu, Y., J. Zhou & K. P. White, 2014. RNA-seq differential expression studies: more sequence or more replication? Bioinformatics 30(3):301-304 doi:10.1093/bioinformatics/btt688.

Love, M. I., W. Huber & S. Anders, 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology 15(12):550 doi:10.1186/s13059-014-0550-8.

Lv, J., P. Liu, B. Gao, Y. Wang, Z. Wang, P. Chen & J. Li, 2014. Transcriptome analysis of the *Portunus trituberculatus*: de novo assembly, growth-related gene identification and marker discovery. PLoS one 9(4):e94055.

Macharia, R. W., F. L. Ombura & E. O. Aroko, 2015. Insects RNA profiling reveals absence of hidden break in 28S Ribosomal RNA molecule of Onion Thrips, *Thrips tabaci*. Journal of Nucleic Acids 2015:8 doi:10.1155/2015/965294.

MacManes, M. D., 2014. On the optimal trimming of high-throughput mRNA sequence data. Frontiers in Genetics 5:13 doi:10.3389/fgene.2014.00013.

Marguerat, S. & J. Bähler, 2010. RNA-seq: from technology to biology. Cellular and Molecular Life Sciences 67(4):569-579 doi:10.1007/s00018-009-0180-6.

McCarthy, S. D., M. M. Dugon & A. M. Power, 2015. 'Degraded' RNA profiles in Arthropoda and beyond. PeerJ 3:e1436 doi:10.7717/peerj.1436.

Meng, X.-l., P. Liu, F.-l. Jia, J. Li & B.-Q. Gao, 2015. *De novo* transcriptome analysis of *Portunus trituberculatus* ovary and testis by RNA-Seq: Identification of genes involved in gonadal development. PLoS ONE 10(6):e0128659 doi:10.1371/journal.pone.0128659.

Metzker, M. L., 2010. Sequencing technologies—the next generation. Nature Reviews Genetics 11(1):31-46 doi:10.1038/nrg2626.

Miller, J. R., S. Koren & G. Sutton, 2010. Assembly algorithms for Next-Generation Sequencing data. Genomics 95(6):315-327 doi:10.1016/j.ygeno.2010.03.001.

Misner, I., C. Bicep, P. Lopez, S. Halary, E. Bapteste & C. E. Lane, 2013. Sequence comparative analysis using networks: Software for evaluating *de novo* transcript assembly from Next-Generation Sequencing. Molecular Biology and Evolution 30(8):1975-1986 doi:10.1093/molbev/mst087.

Mykles, D. L., K. G. Burnett, D. S. Durica, B. L. Joyce, F. M. McCarthy, C. J. Schmidt & J. H. Stillman, 2016. Resources and recommendations for using transcriptomics to address grand challenges in comparative biology. Integrative and Comparative Biology 56(6):1183-1191 doi:10.1093/icb/icw083.

Mykles, D. L. & J. H. Hui, 2015. *Neocaridina denticulata*: A decapod crustacean model for Functional Genomics. Integrative and Comparative Biology 55(5):891-7 doi:10.1093/icb/icv050.

Nguyen, C., T. G. Nguyen, L. Van Nguyen, H. Q. Pham, T. H. Nguyen, H. T. Pham, H. T. Nguyen, T. T. Ha, T. H. Dau & H. T. Vu, 2016. *De novo* assembly and transcriptome characterization of major growth-related genes in various tissues of *Penaeus monodon*. Aquaculture 464:545-553 doi:10.1016/j.aquaculture.2016.08.003.

Niedringhaus, T. P., D. Milanova, M. B. Kerby, M. P. Snyder & A. E. Barron, 2011. Landscape of Next-Generation Sequencing technologies. Analytical Chemistry 83(12):4327-4341 doi:10.1021/ac2010857.

Ozsolak, F. & P. M. Milos, 2011. RNA sequencing: advances, challenges and opportunities. Nature Reviews Genetics 12(2):87-98.

Parkhomchuk, D., T. Borodina, V. Amstislavskiy, M. Banaru, L. Hallen, S. Krobitsch, H. Lehrach & A. Soldatov, 2009. Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic Acids Research 37(18):e123-e123 doi:10.1093/nar/gkp596.

Parra, G., K. Bradnam & I. Korf, 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23(9):1061-1067 doi:10.1093/bioinformatics/btm071.

Patro, R., G. Duggal, M. I. Love, R. A. Irizarry & C. Kingsford, 2017. Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods doi:10.1038/nmeth.4197.

Rajkumar, A. P., P. Qvist, R. Lazarus, F. Lescai, J. Ju, M. Nyegaard, O. Mors, A. D. Børglum, Q. Li & J. H. Christensen, 2015. Experimental validation of methods for differential gene expression analysis and sample pooling in RNA-seq. BMC Genomics 16(1):548 doi:10.1186/s12864-015-1767-y.

Reuter, Jason A., D. V. Spacek & Michael P. Snyder, 2015. High-Throughput Sequencing Technologies. Molecular Cell 58(4):586-597 doi:10.1016/j.molcel.2015.05.004.

Roberts, A. & L. Pachter, 2012. Streaming fragment assignment for real-time analysis of sequencing experiments. Nature Methods 10:71 doi:10.1038/nmeth.2251.

Robertson, G., J. Schein, R. Chiu, R. Corbett, M. Field, S. D. Jackman, K. Mungall, S. Lee, H. M. Okada, J. Q. Qian, M. Griffith, A. Raymond, N. Thiessen, T. Cezard, Y. S. Butterfield, R. Newsome, S. K. Chan, R. She, R. Varhol, B. Kamoh, A.-L. Prabhu, A. Tam, Y. Zhao, R. A. Moore, M. Hirst, M. A. Marra, S. J. M. Jones, P. A. Hoodless & I. Birol, 2010. *De novo* assembly and analysis of RNA-seq data. Nature Methods 7(11):909-912 doi:10.1038/nmeth.1517.

Robinson, M. D., D. J. McCarthy & G. K. Smyth, 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26(1):139-140 doi:10.1093/bioinformatics/btp616.

Robles, J. A., S. E. Qureshi, S. J. Stephen, S. R. Wilson, C. J. Burden & J. M. Taylor, 2012. Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. BMC Genomics 13(1):484 doi:10.1186/1471-2164-13-484.

Sagi, A., R. Manor & T. Ventura, 2013. Gene silencing in Crustaceans: From basic research to biotechnologies. Genes 4(4):620 doi:10.3390/genes4040620.

Schadt, E. E., S. Turner & A. Kasarskis, 2010. A window into third-generation sequencing. Human Molecular Genetics 19(R2):R227-R240 doi:10.1093/hmg/ddq416.

Schirmer, M., U. Z. Ijaz, R. D'Amore, N. Hall, W. T. Sloan & C. Quince, 2015. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Research 43(6):e37-e37 doi:10.1093/nar/gku1341.

Schulz, M. H., D. R. Zerbino, M. Vingron & E. Birney, 2012. Oases: robust *de novo* RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28(8):1086-1092 doi:10.1093/bioinformatics/bts094.

Schurch, N. J., P. Schofield, M. Gierliński, C. Cole, A. Sherstnev, V. Singh, N. Wrobel, K. Gharbi, G. G. Simpson, T. Owen-Hughes, M. Blaxter & G. J. Barton, 2016. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA 22(6):839-851 doi:10.1261/rna.053959.115.

Seelenfreund, E., W. A. Robinson, C. M. Amato, A.-C. Tan, J. Kim & S. E. Robinson, 2014. Long term storage of dry versus frozen RNA for next generation molecular studies. PLoS ONE 9(11):e111827 doi:10.1371/journal.pone.0111827.

Simão, F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva & E. M. Zdobnov, 2015. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31(19):3210-3212 doi:10.1093/bioinformatics/btv351.

Soneson, C. & M. Delorenzi, 2013. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics 14(1):91 doi:10.1186/1471-2105-14-91.

Song, L., C. Bian, Y. Luo, L. Wang, X. You, J. Li, Y. Qiu, X. Ma, Z. Zhu, L. Ma, Z. Wang, Y. Lei, J. Qiang, H. Li, J. Yu, A. Wong, J. Xu, Q. Shi & P. Xu, 2016. Draft genome of the Chinese mitten crab, *Eriocheir sinensis*. GigaScience 5:5 doi:10.1186/s13742-016-0112-y.

Sultan, M., V. Amstislavskiy, T. Risch, M. Schuette, S. Dökel, M. Ralser, D. Balzereit, H. Lehrach & M.-L. Yaspo, 2014. Influence of RNA extraction methods and library selection schemes on RNA-seq data. BMC Genomics 15(1):675 doi:10.1186/1471-2164-15-675.

Sultan, M., S. Dökel, V. Amstislavskiy, D. Wuttig, H. Sültmann, H. Lehrach & M.-L. Yaspo, 2012. A simple strand-specific RNA-Seq library preparation protocol combining the Illumina TruSeq RNA and the dUTP methods. Biochemical and Biophysical Research Communications 422(4):643-646 doi:10.1016/j.bbrc.2012.05.043.

Surget-Groba, Y. & J. I. Montoya-Burgos, 2010. Optimization of *de novo* transcriptome assembly from next-generation sequencing data. Genome Research 20(10):1432-1440 doi:10.1101/gr.103846.109.

Tarazona, S., F. García-Alcalde, J. Dopazo, A. Ferrer & A. Conesa, 2011. Differential expression in RNA-seq: A matter of depth. Genome Research 21(12):2213-2223 doi:10.1101/gr.124321.111.

Teng, M., M. I. Love, C. A. Davis, S. Djebali, A. Dobin, B. R. Graveley, S. Li, C. E. Mason, S. Olson, D. Pervouchine, C. A. Sloan, X. Wei, L. Zhan & R. A. Irizarry, 2016. A benchmark for RNA-seq quantification pipelines. Genome Biology 17(1):74 doi:10.1186/s13059-016-0940-1.

Wang, Y., N. Ghaffari, C. Johnson, U. Braga-Neto, H. Wang, R. Chen & H. Zhou, 2011. Evaluation of the coverage and depth of transcriptome by RNA-Seq in chickens. BMC Bioinformatics 12 doi:10.1186/1471-2105-12-S10-S5.

Wang, Z., M. Gerstein & M. Snyder, 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics 10(1):57-63 doi:10.1038/nrg2484.

Wilhelm, B. T. & J.-R. Landry, 2009. RNA-Seq—quantitative measurement of expression through massively parallel RNA-sequencing. Methods 48(3):249-257.

Winnebeck, E. C., C. D. Millar & G. R. Warman, 2010. Why Does Insect RNA Look Degraded? Journal of Insect Science 10:159 doi:10.1673/031.010.14119.

Wu, A. R., N. F. Neff, T. Kalisky, P. Dalerba, B. Treutlein, M. E. Rothenberg, F. M. Mburu, G. L. Mantalas, S. Sim, M. F. Clarke & S. R. Quake, 2014. Quantitative assessment of single-cell RNA-sequencing methods. Nature Methods 11(1):41-46 doi:10.1038/nmeth.2694.

Wu, T. D. & S. Nacu, 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics 26(7):873-881 doi:10.1093/bioinformatics/btq057.

Xie, Y., G. Wu, J. Tang, R. Luo, J. Patterson, S. Liu, W. Huang, G. He, S. Gu, S. Li, X. Zhou, T.-W. Lam, Y. Li, X. Xu, G. K.-S. Wong & J. Wang, 2014. SOAPdenovo-Trans: *De novo* transcriptome assembly with short RNA-Seq reads. Bioinformatics doi:10.1093/bioinformatics/btu077.

Yang, C. & H. Wei, 2015. Designing Microarray and RNA-Seq experiments for greater systems biology discovery in modern plant genomics. Molecular Plant 8(2):196-206 doi:10.1016/j.molp.2014.11.012.

Yu, Y., X. Zhang, J. Yuan, F. Li, X. Chen, Y. Zhao, L. Huang, H. Zheng & J. Xiang, 2015. Genome survey and high-density genetic map construction provide genomic and genetic resources for the Pacific White Shrimp *Litopenaeus vannamei*. Scientific Reports 5:15612 doi:10.1038/srep15612.

Yuan, J., X. Zhang, C. Liu, Y. Yu, J. Wei, F. Li & J. Xiang, 2017. Genomic resources and comparative analyses of two economical penaeid shrimp species, *Marsupenaeus japonicus* and *Penaeus monodon*. Marine Genomics doi:10.1016/j.margen.2017.12.006.

Zerbino, D. R. & E. Birney, 2008. Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Research 18(5):821-829 doi:10.1101/gr.074492.107.

Zhang, Z. H., D. J. Jhaveri, V. M. Marshall, D. C. Bauer, J. Edson, R. K. Narayanan, G. J. Robinson, A. E. Lundberg, P. F. Bartlett, N. R. Wray & Q.-Y. Zhao, 2014. A comparative study of techniques for differential expression analysis on RNA-Seq data. PLoS ONE 9(8):e103207 doi:10.1371/journal.pone.0103207.

Zhao, Q.-Y., Y. Wang, Y.-M. Kong, D. Luo, X. Li & P. Hao, 2011. Optimizing *de novo* transcriptome assembly from short-read RNA-Seq data: a comparative study. BMC Bioinformatics 12 doi:10.1186/1471-2105-12-S14-S2.

Zhao, S., Y. Zhang, W. Gordon, J. Quan, H. Xi, S. Du, D. von Schack & B. Zhang, 2015. Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap. BMC Genomics 16(1):675 doi:10.1186/s12864-015-1876-7.

Zhong, S., J. G. Joung, Y. Zheng, Y. R. Chen, B. Liu, Y. Shao, J. Z. Xiang, Z. Fei & J. J. Giovannoni, 2011. High-throughput illumina strand-specific RNA sequencing library preparation. Cold Spring Harbor Protocols 2011(8):940-9 doi:10.1101/pdb.prot5652.