De novo transcriptome assembly and annotation of the isopod *Idotea balthica*

Keith Yamada
Master's Thesis
Master's Degree Programme in Bioinformatics
Department of Information Technology
University of Turku
March 2016

Supervisors:
Pierre de Wit
Martti Tolvanen
Veijo Jormalainen

UNIVERSITY OF TURKU
Department of Information Technology

YAMADA, KEITH: *De novo* transcriptome assembly and annotation of the isopod *Idotea balthica*

Master's Thesis, 32 p., 6 appendix pages
Master's Degree Programme in Bioinformatics
March 2016

The Baltic Sea is a unique environment that contains unique genetic populations. In order to study these populations on a genetic level basic molecular research is needed. The aim of this thesis was to provide a basic genetic resource for population genomic studies by *de novo* assembling a transcriptome for the Baltic Sea isopod *Idotea balthica*.

RNA was extracted from a whole single adult male isopod and sequenced using Illumina (125bp PE) RNA-Seq. The reads were preprocessed using FASTQC for quality control, TRIMMOMATIC for trimming, and RCORRECTOR for error correction. The preprocessed reads were then assembled with TRINITY, a de Bruijn graph-based assembler, using different k-mer sizes. The different assemblies were combined and clustered using CD-HIT. The assemblies were evaluated using TRANSRATE for quality and filtering, BUSCO for completeness, and TRANSDECODER for annotation potential. The 25-mer assembly was annotated using PANNZER (protein annotation with *z*-score) and BLASTX.

The 25-mer assembly represents the best first draft assembly since it contains the most information. However, this assembly shows high levels of polymorphism, which currently cannot be differentiated as paralogs or allelic variants. Furthermore, this assembly is incomplete, which could be improved by sampling additional developmental stages.

Keywords: *de novo* transcriptome assembly, *Idotea balthica*, isopod, population genomics, RNA-Seq, SNP discovery

Table of contents

Preface

The original unsuccessful thesis plan was to do an expressed sequence tag (EST) differential expression experiment with the Baltic Sea brown algae *Fucus vesiculosus*. Dr. Veijo Jormalainen supervised the biological part, Dr. Martti Tolvanen supervised the bioinformatics and PhD student Luca Rugiu worked with me in the laboratory.

The algae were collected from three different areas of the Baltic Sea and then kept in aquaria at the Archipelago Research Institute, part of the Centre of Environmental Research of the University of Turku. The algae were stressed with increased heat and decreased salinity both together and separately for a total of three different stress regimes. Unfortunately, after the stresses were applied no high quality RNA could be extracted. It was thought that *F. vesiculosus* produces excessive amounts of stress related chemicals, such as phlorotannins, making RNA extraction not possible with the methods that were used. After two months and various different RNA extraction methods it was decided that a new topic should be chosen.

The new thesis topic was also part of the Baltic Sea Marine Biodiversity Project (project BAMBI, part of the BONUS program for Baltic Sea research) but now focused on *de novo* assembling a transcriptome for the Baltic Sea isopod *Idotea balthica*. The RNA sequencing (RNA-Seq) data was obtained from Dr. Pierre De Wit at the University of Gothenburg, Sweden, who also supervised the bioinformatics work. The analysis and progress for this thesis was updated regularly on GitHub (https://github.com/The-Bioinformatics-Group/Idotea_balthica_transcriptome_project).

This thesis topic was presented, near its completion, to BAMBI collaborators at the University of Gothenburg and the Sven Lovén Centre for Marine Sciences via videoconference. This thesis is also planned to be part of a publication in the spring of 2016.

List of abbreviations

BAMBI:	Baltic Sea marine biodiversity - addressing the potential of adaptation to climate change
bp:	Base pair
BUSCO:	Benchmarking universal single-copy orthologs
cDNA:	Complementary DNA
DNA:	Deoxyribonucleic acid
GB:	Gigabyte
GO:	Gene ontology
M:	Million
mRNA:	Messenger RNA
NGS:	Next-generation sequencing
NR:	Non-redundant protein sequence database
ORF:	Open reading frame
PE:	Paired-end
RNA:	Ribonucleic acid
RNA-Seq:	RNA sequencing
SNP:	Single nucleotide polymorphism

List of figures

List of tables

1. Introduction

1.1 Data possession and usage

The analysis will be conducted within the BONUS program project: “Baltic Sea marine biodiversity – addressing the potential of adaptation to climate change (BAMBI)” (http://www.bonusportal.org/projects/research_projects/bambi). The data produced will be a part of a larger research agenda and property of the project. The student has a right to use the data in his M.Sc.-thesis, as well as a right to co-author the paper(s) arising from the data as recommended by the “Responsible conduct of research” -guidelines of the Finnish Advisory Board on Research Integrity.

1.2 The Baltic Sea ecosystem and climate change

The Baltic Sea is a unique environment. Compared to the Atlantic Ocean, the Baltic Sea has a much lower salinity and has no tides contributing to the uniqueness of the local species. This unique genetic composition contributes to the ecosystem’s vulnerability (Johannesson & Andre, 2006), for example to climate change.

Global climate change is predicted to increase temperature, decrease salinity and decrease sea ice extent in the Baltic Sea (HELCOM, 2013; Meier, 2006). These changes will have a greater impact on the northern parts of the Baltic Sea (HELCOM, 2013). Since the changes are not uniform over the entire Baltic Sea not all populations will be stressed to the same extent. In order to determine each population’s potential for adaptation, genetic variation needs to be mapped. Before this type of population genomic study can be done a reference transcriptome is needed, which will be assembled in this thesis.

1.3 *Idotea balthica*

The BAMBI project will target a model ecosystem, which will be represented by a seaweed, a grazer, and a predator. This model is ideal since herbivory is an especially important interaction in the marine littoral environment (Poore et al., 2012). The isopod *Idotea balthica* will serve as the grazer in this model ecosystem. *I. balthica* has been shown to have a controlling influence on the seaweed *Fucus vesiculosus* and is a key food source to fish, such as the commercially important cod (*Gadhus morhua*) and perch (*Perca fluviatilis*) (Haavisto & Jormalainen, 2014; Leidenberger, Harding, & Jonsson, 2012).

Furthermore, *I. balthica* was chosen to represent the grazer in this model ecosystem because of its short generation time and ease of culturing (Hemmi & Jormalainen, 2004; Jormalainen, Honkanen, & Vesakoski, 2008). Its natural environment is also easily reproduced in the laboratory making environmental experimentation simple (Leidenberger et al., 2012). *I. balthica* is also already adapted to low salinity (Leidenberger et al., 2012) making it especially interesting to study since salinity is predicted to decrease even more (Meier & Eilola, 2011).

I. balthica can also serve as a wider model crustacean since there are currently very limited genomic resources, despite arthropods having the most species of any phylum (Bron et al., 2011; Stillman et al., 2008). The only publicly available crustacean genome (the water flea, *Daphnia pulex*) has uncovered that about 20% of its 27,000 predicted genes are tandem duplications (Gilbert, 2007). It is also expected that I. balthica will contain high levels of polymorphism as with many other marine crustaceans with large population sizes.

1.4 RNA-Seq and *de novo* assembly

RNA sequencing (RNA-Seq) uses next-generation sequencing (NGS) technology to analyze the transcriptome of a cell or cells (Kukurba & Montgomery, 2015; Wang, Gerstein, & Snyder, 2009). RNA-Seq is an improvement over previous methods such as microarray-based methods in that novel and rare transcripts can be discovered as well as a lower background signal, which improves the signal to noise ratio (Kukurba & Montgomery, 2015; Ozsolak & Milos, 2011; Zhao, Fung-Leung, Bittner, Ngo, & Liu, 2014). However, microarrays are still useful in transcriptomics for their reliability and cost-effectiveness in model organisms (Stefano, 2014). For this transcriptome experiment of a non-model organism, *Idotea balthica*, in which little is known about its transcripts RNA-Seq is the only option for creating a *de novo* assembly.

A *de novo* assembly of a transcriptome is where short RNA-Seq reads are assembled without a reference transcriptome or genome, unlike a mapping assembly. A common approach to this *de novo* assembly problem is the use of de Bruijn graphs, originally developed for genome assembly and now adapted for transcriptome assembly (Compeau, Pevzner, & Tesler, 2011; Pevzner, Tang, & Waterman, 2001). In general, the de Bruijn graph approach works by breaking the reads into substrings of length k (k-mers) and then tries to find a Eulerian path through the graphs based on k-mer overlaps (Compeau et al., 2011).

A comparison of three commonly used *de novo* assemblers (VELVET/OASES (Schulz, Zerbino, Vingron, & Birney, 2012; Zerbino & Birney, 2008), TRANS-ABYSS (Robertson et al., 2010), and TRINITY (Grabherr et al., 2011)) shows they are all able to handle the data from this experiment but that the ideal assembler is one that was developed using similar data (Martin & Wang, 2011).

1.5 Thesis goals

The goal of this thesis was to *de novo* assemble a reference transcriptome for the isopod, *Idotea balthica*, as a molecular resource for future population genomic studies.

2. Materials and Methods

2.1 *In vitro* data collection

The *in vitro* data was collected and generated by collaborators at the University of Gothenburg, Sweden as part of the BAMBI project.

2.1.1 Sampling

This *Idotea balthica* transcriptome describes the expressed genes of all tissues, except the gut, of a single adult male isopod (*Idotea balthica*). This individual was collected in September of 2014 in Kristineberg, Sweden (Lat: 58°14.869'N; Long: 11°26.883'E) and kept in an aquarium at the University of Gothenburg, Sweden for approximately six months.

2.1.2 RNA extraction and sequencing

The individual was killed and placed in RNA*later* (Thermo Fisher Scientific) and the extraction took place five days later. Total RNA was extracted using a TRIzol (Invitrogen) protocol. The RNA was cleaned with a Zymo RNA Clean and Concentrator Kit (Zymo Research). The concentration was measured using a QuBit RNA assay (Thermo Fisher Scientific). The quality was assessed using a MOPS denaturing agarose gel.

The RNA sample was shipped to the Science for Life Laboratory (www.scilifelab.se) in Stockholm, Sweden for library preparation and sequencing. The RNA concentration and quality were reassessed using an Agilent Bioanalyzer (Agilent Technologies). The complementary DNA (cDNA) libraries were prepared using a TruSeq RNA Library Preparation Kit v2 (Illumina), which included a poly-A selection of total RNA using oligo-dT beads.

A single lane of an Illumina HiSeq 2500 sequencer was used to paired-end (PE) sequence 125 base pair (bp) long reads at the Science for Life Laboratory.

2.2 Preprocessing

The 'best practice' transcriptome assembly pipeline by De Wit et al. (2015) was followed. All analyses were run on CSC - IT Center for Science's (Espoo, Finland; www.csc.fi) Taito super-cluster.

2.2.1 Visualization using FASTQC

Raw reads were assessed for quality using FASTQC (v0.11.2; www.bioinformatics.babraham.ac.uk/projects/fastqc). The visualization of the raw sequence data shows potential problems such as low-quality reads or regions of reads and adapter sequence contamination.

2.2.2 Trimming using TRIMMOMATIC

After the potential problems were found, TRIMMOMATIC (v0.33) was used to remove them (Bolger, Lohse, & Usadel, 2014). First, the adapters were trimmed using the included adapter sequence file, using a maximum of two mismatches, a palindrome clip threshold of 30 and a simple clip threshold of 10. Next, the leading ends were trimmed for low-quality (<3) or N bases followed by the same trimming for the trailing ends. Then, a sliding window of four bases was used for trimming if the average quality per base dropped below 15. Finally, sequences that were less than 36 bases were dropped.

The trimmed files were re-visualized using FASTQC to verify the potential problems had been removed.

2.2.3 Error correction using RCORRECTOR

In addition to the standard preprocessing strategy of trimming, error correcting was also done to reduce the number of substitution errors incorporated into the assembly (De Wit, Pespeni, & Palumbi, 2015). Error correction was done using RCORRECTOR (v2015-11-06; Song & Florea, 2015). The k-mer size was set to 25. Then, a k-mer size of 20 was used to test the sensitivity to k-mer size.

2.2.4 *In silico* normalization using TRINITY

Assembling hundreds of millions of reads requires a large amount of computational resources and run time. Normalization reduces the amount of resources and run time needed, and can also reduce the incorporation of errors (Brown, Howe, Zhang, Pyrkosz, & Brom, 2012). TRINITY (v2.1.0) was used for *in silico* normalization (Grabherr et al., 2011; Haas et al., 2013). A k-mer size of 25 and a coverage limit of 50 were used.

2.3 *De novo* Assembly using TRINITY

The approximately 10 million (M) preprocessed sequences were *de novo* assembled using TRINITY (v2.1.0). TRINITY was run three times using 20-mers, 25-mers and 30-mers to understand the effect of k-mer size on the assembly, all other parameters were left at default. As the size of the k-mers increase, specificity should improve but at the cost of sensitivity (Haas et al., 2013).

2.4 Post-processing

After the reads are assembled into contigs (putative transcripts) they need to be assessed for assembly errors, such as chimeras, and for contamination such as ribosomal RNA (rRNA) or bacterial transcripts (De Wit et al., 2015).

2.4.1 Clustering using CD-HIT

In order to maximize the number of potential transcripts found, all three assemblies (20, 25, and 30-mers) were combined and then clustered using CD-HIT (v4.6.1; Fu, Niu, Zhu, Wu, & Li, 2012; Li & Godzik, 2006). As a result of clustering, many potential paralogs were collapsed into a single contig. A sequence identity threshold of 0.95 and a word length of 10 were used.

2.4.2 Filtering using TRANSRATE

Post-assembly filtering is suggested to remove or reduce contamination (De Wit et al., 2015). Filtering was done using TRANSRATE (v1.0.1; Smith-Unna, Boursnell, Patro, Hibberd, & Kelly, 2015) and a custom script (prune_fasta.pl). TRANSRATE filtered the contigs by automatically determining the contig score cutoff that maximized the overall assembly score (Smith-Unna et al., 2015). The custom script was used to filter out contigs that were less than 600bp.

Another custom script (rename_fasta.py) was used to modify the clustered assembly's FASTA headers to work with TRANSRATE.

2.5 Evaluation

In order to chose the assembly that best represents the true expressed sequences each should be evaluated for supporting experimental evidence, completeness and containing an open reading frame (ORF) (De Wit et al., 2015).

2.5.1 Read-mapping using TRANSRATE

Assemblies were quality evaluated and compared using TRANSRATE. Experimental evidence for each contig can be found if the reads can be mapped to them (Smith-Unna et al., 2015). The normalized reads were used for read-mapping. All assemblies were compared to each other to determine which assembly was the most accurate.

2.5.2 Completeness assessment using BUSCO

Benchmarking universal single-copy orthologs (BUSCO; v1.1) was used to evaluate the completeness of the assemblies (Simão, Waterhouse, Ioannidis, Kriventseva, & Zdobnov, 2015). This software relied on BLAST+ v2.2.31 (Camacho et al., 2009), HMMER v3.1b2 (Eddy, 2011), EMBOSS v6.5.7 (Rice, Longden, & Bleasby, 2000) and PYTHON v3.4.0. BUSCO was run in transcriptome mode using the arthropod lineage dataset.

The exact number of times a duplicate BUSCO was found was determined using a custom script (BUSCOdup.py).

2.5.3 Predicting ORFs with TRANSDECODER

To filter out non-coding contigs, such as rRNA, and maximize the number of annotatable protein coding genes ORFs should be predicted (De Wit et al., 2015). Putative proteins were identified by predicting ORFs using TRANSDECODER (v2.0.1; Haas et al., 2013).

2.6 Annotation

Finally, to discover new or interesting genes and contaminant sequences annotation was done.

2.6.1 PANNZER

To maximize the reliability of the functional annotations PANNZER was used (Koskinen, Toronen, Nokso-Koivisto, & Holm, 2015). PANNZER searches the UniProtKB/Swiss (http://www.uniprot.org) database and uses a *Z*-score for a more reliable automated annotation (Koskinen et al., 2015). As a result PANNZER provides both free text descriptions and gene ontology (GO) terms. The query taxon id 82763 was used; the other settings were left as default.

2.6.2 BLASTX to nr

Additional annotations were found using a naïve BLAST approach. The raw 25-mer assembly was used by BLASTX (v2.2.31+; Camacho et al., 2009) to search the National Center for Biotechnology Information (NCBI) non-redundant protein sequence database (nr; v2015-12-01) and output in XML format. The result file was parsed using a script (parse_blast.py) from De Wit et al. (2012). The parsed results were reparsed (ReParseBlastbycutoffs.py; De Wit et al., 2012) using an e-value cutoff of 10^{-5}.

Next, the reparsed results were simplified into a more readable format using the custom script ReParseBlast.py. The simplified results were filtered to separate out uninformative terms (i.e. hypothetical, putative, predicted and unknown) using the custom script ReParseBlastByTerm.py. Then, the term-filtered results were analyzed to uncover the top taxon groups using the custom script GroupTaxon.py. Finally, a list was created from the top bacterial and protozoan genera and was used to separate out those results using the custom script FilterByTaxon.py. The custom scripts can be found on GitHub (https://github.com/The-Bioinformatics-Group/Idotea_balthica_transcriptome_project/tree/master/Annotation/blastx2nr/scripts).

The numbers of contigs annotated by PANNZER alone, BLASTX alone, and by both PANNZER and BLASTX were determined using a custom script (AddAno.py).

3. Results

3.1 Preprocessing

3.1.1 Visualization using FASTQC

The FASTQC basic statistics for the raw reads show the sequences have no major issues such as poor quality or a shorter than expected length (Table 1). The total number of sequences is more than adequate and the percent GC content is within reason (Table 1).

Table 1. FASTQC basic statistics for raw and trimmed reads (same for both pairs)

	Raw	Trimmed
Total sequences	296,173,922	262,020,548
Sequences flagged as poor quality	0	0
Sequence length	126	36-126
%GC	39	39

The FASTQC basic statistics for the trimmed reads show the sequences were trimmed and that approximately 34M sequences were dropped (11.53%) (Table 1). As expected no sequences were flagged as poor quality and the percent GC remained constant (Table 1). Figure A9 also shows the adapters were successfully removed.

The full FASTQC reports are available on GitHub (https://github.com/The-Bioinformatics-Group/Idotea_balthica_transcriptome_project/tree/master/Preprocessing/FASTQC).

3.1.2 Trimming using TRIMMOMATIC

The TRIMMOMATIC results show 88.74% of read pairs survived trimming, while 10.29% survived as unpaired reads and 1.24% were dropped (Table 2).

Table 2. TRIMMOMATIC results

Input read pairs	296,173,922
Both surviving	262,020,548 (88.47%)
Forward only surviving	28,871,039 (9.75%)
Reverse only surviving	1,602,357 (0.54%)
Dropped	3,679,978 (1.24%)

3.1.3 Error correction using RCORRECTOR

The RCORRECTOR results show using a k-mer size of 20 corrected approximately 1.66M more bases than using a k-mer size of 25 (Table 3).

Table 3. RCORRECTOR results

	20-mers	25-mers
Stored k-mers	128,297,607	139,073,753
Weak k-mer threshold rate	0.009025	0.006066
Bad quality threshold	;	;
Processed reads	524,041,096	524,041,096
Corrected bases	46,535,134	44,876,577

3.1.4 *In silico* normalization using TRINITY

The TRINITY *in silico* normalization resulted in 10,099,252 (3.85%) PE reads being selected based on 25-mers and 50x coverage.

3.2 *De novo* Assembly using TRINITY

The TRINITY *de novo* assembly statistics are within reason (Table 4). Each assembly has a percent GC content around 39 as expected since the raw sequences have a 39% GC content (Table 4). The 25-mer and 30-mer assemblies show very similar results with approximately 115,000 contigs (putative transcripts) while the 20-mer assembly only has approximately 78,000 (Table 4). The 25-mer and 30-mer assemblies also have a higher average contigs length and N50 than the 20-mer assembly (Table 4).

Table 4. TRINITY statistics

	20-mers	25-mers	30-mers
Total Trinity ‘genes’	77,981	90,663	90,309
Total Trinity transcripts	78,037	115,931	115,561
Percent GC	39.23	39.36	39.31
N50	826	1,154	1,154
Shortest contig length	201	201	201
Median contig length	342	358	359
Average contig length	577.05	683.11	683.20
Longest contig length	10,790	21,569	21,376
Total assembled bases	45,031,014	79,193,347	78,951,148

3.3 Post-processing

Only the 25-mer assembly was used for further analysis since it contained the most contigs and the most bases.

3.3.1 Clustering using CD-HIT

The CD-HIT results show 309,529 contigs (all three assemblies combined) were clustered into 115,917 contigs and 97,530 putative genes (Table 5). All other statistics remained approximately constant in relation to the 25-mer assembly (Table 5).

Table 5. CD-HIT statistics

	25-mers	Clustered
Total Trinity ‘genes’	90,663	97,530
Total Trinity transcripts	115,931	115,917
Percent GC	39.36	39.07
N50	1,154	1,121
Shortest contig length	201	201
Median contig length	358	370
Average contig length	683.11	682.59
Longest contig length	21,569	21,569
Total assembled bases	79,193,347	79,123,721

3.3.2 Filtering using TRANSRATE

The TRANSRATE filtering results show a loss of more than half of the contigs for both the 25-mer assembly and the clustered assembly (Table 6, K25 'good' and C. 'good' respectively). Filtering based on a minimum length of 600bp resulted in an even greater loss of contigs for both assemblies (Table 6, the L600 columns). All other statistics remained constant indicating the filtering was done correctly (Table 6).

Table 6. Filtering statistics (K25 = 25-mer assembly, 'good' = TRANSRATE filtering, L600 = minimum length of 600, C. = clustered assembly). The total TRINITY 'genes' and median contig length could not be calculated.

	K25 'good'	K25 L600	C. 'good'	C. L600
Total Trinity 'genes'	30,147	21,395	-----	-----
Total Trinity transcripts	40,122	36,150	47,891	36,969
Percent GC	39.70	39.93	39.44	39.66
N50	1,710	1,778	1,618	1,715
Shortest contig length	201	600	201	600
Median contig length	900	1,170	-----	1,145
Average contig length	1,184.48	1,503.58	1,073.63	1,466.37
Longest contig length	21,522	21,569	21,569	21,569
Total assembled bases	47,523,534	54,354,429	51,417,044	54,210,125

Combining and clustering all three assemblies introduced non-unique FASTA headers. To fix this the headers were changed to be unique but in turn resulted in the loss of the TRINITY 'genes' information.

3.4 Evaluation

In order to maximize the potential for new information only the 25-mer, 'good' 25-mer, clustered, and 'good' clustered assemblies were evaluated.

3.4.1 Read-mapping using TRANSRATE

The TRANSRATE read-mapping results show the clustered assembly as the best in terms of having the highest percentage of reads mapped, the highest percentage of read pairs mapped back to the same contig ('good' mapped) and the highest optimal score (Table 7). The 25-mer 'good' assembly has the least number of potential bridges suggesting the least amount of misassembled contigs (Table 7). The clustered 'good' assembly has the highest TRANRATE score (Table 7).

The TRANSRATE 'good' filter reduced the percentage of mapped and 'good' mapped reads while also greatly reducing the number of potential bridges and increasing the score (Table 7).

Table 7. TRANSRATE read-mapping statistics. The best score for each metric is underlined.

	K25	K25 'good'	Clustered	C. 'good'
Percent mapped	80.04	66.75	84.44	73.83
Percent 'good' mapped	63.61	58.04	70.90	64.36
Potential bridges	42,305	7,925	34,351	9,761
Score	0.05841	0.3749	0.08598	0.39991
Optimal score	0.42087	0.4132	0.45763	0.44633

3.4.2 Completeness assessment using BUSCO

The BUSCO results show the 25-mer assembly as being the most complete (Table 8). However, the clustered 'good' assembly has a better ratio of single-copy BUSCOs to duplicated BUSCOs (Table 8). Each assembly has a greater than expected level of duplicated BUSCOs (Table 8).

The TRANSRATE 'good' filter improved the ratio of single-copy BUSCOs to duplicated BUSCOs but also reduced the overall number of BUSCOs found (Table 8).

Table 8. BUSCO results. The best score for each metric is underlined.

	K25	K25 'good'	Clustered	C. 'good'
Complete single-copy BUSCOs	759	762	824	884
Complete duplicated BUSCOs	423	378	367	310
Complete triplicated BUSCOs	244	185	199	148
Fragmented BUSCOs	386	320	372	333
Missing BUSCOs	863	1030	913	1000
Total BUSCO groups searched	2675	2675	2675	2675

Further analysis of the complete multiple-copy BUSCOs shows that most of the copies occur in pairs and the rest occur in a group of three (Table 8).

3.4.3 Predicting ORFs using TRANSDECODER

The TRANSDECODER results show the 25-mer assembly contains the most ORFs suggesting this assembly contains the most information (Table 9).

The TRANSRATE 'good' filter appears to have filtered out a large number of contigs containing ORFs, greatly reducing the amount of potentially interesting transcripts (Table 9).

Table 9. TRANSDECODER results. The best score is underlined.

	K25	K25 'good'	Clustered	C. 'good'
Total contigs	115,931	40,122	115,917	47,891
Putative peptides	<u>33,731</u>	21,439	32,253	22,635

3.5 Annotation

Only the 25-mer assembly was annotated as it represents the assembly with the most information.

3.5.1 PANNZER

PANNZER was able to assign 14,288 annotations to 13,807 putative peptides and assign 1,330,727 GO terms for the 25-mer assembly (https://github.com/The-Bioinformatics-Group/Idotea_balthica_transcriptome_project/tree/master/Draft_1; pannzer_K25.DE and pannzer_K25.GO respectively). These annotations are considered to be more reliable than a simple naïve BLAST approach.

3.5.2 BLASTX to nr

The BLASTX search using the nr database results show the initial number of contigs annotated as being much higher than the PANNZER results (Table 10).

However, after filtering out uninformative terms and bacterial and protozoan genera the total number of annotations decreased significantly (Table 10). The intermediate and final BLASTX annotations can be found on GitHub (https://github.com/The-Bioinformatics-Group/Idotea_balthica_transcriptome_project/tree/master/Annotation/blastx2nr).

Table 10. BLASTX annotation results

	Contigs Annotated	Total Annotations
Cuttoff e-5	38,872	54,465
Term-filtered	12,941	21,171
Genus-filtered	10,930	13,879

The BLASTX annotations were derived from a less reliable database and a simpler filtering algorithm than the PANNZER annotations and are therefore less reliable. However, the BLASTX method was able to annotate an additional 6,189 contigs (Figure 1). Overall 19,996 contigs were annotated.

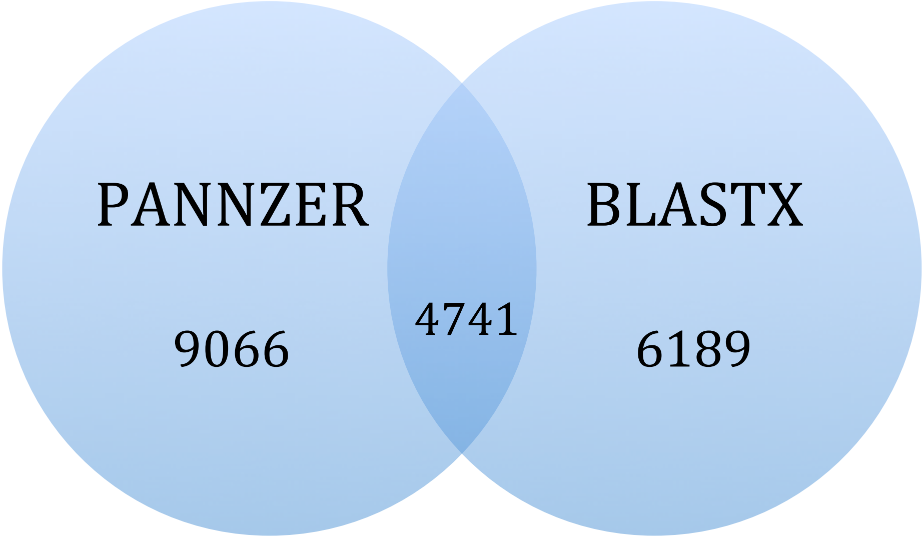

Figure 1. Annotation results. The numbers represent the number of contigs annotated by a given method.

4. Discussion

4.1 *In vitro* data collection

Current literature suggests sampling should include a range of developmental stages and tissues as well as genders (De Wit et al., 2015). In this experiment the maximum amount of tissue was sampled (i.e. the entire individual) but only the adult developmental stage and male gender were sampled. This could explain the estimated 68% level of transcriptome completeness result from BUSCO. If whole individuals of different developmental stages of both males and females were sampled then the transcriptome's completeness should increase.

The relatively low level of completeness could also be explained by the poly-A selection in conjunction with degraded RNA resulting in an increased number of fragmented or missing BUSCOs (Table 8).

Furthermore, current literature also states that the sequencing technology used should be as long as possible, paired and strand-specific (De Wit et al., 2015). Technologies such as PacBio are able sequence at the greatest lengths, often being able to sequence entire transcripts, however they are prone to a relatively high error rate and have relatively low output. In this experiment the sampled RNA was sequenced using Illumina's longest high output paired-end platform. Additional reliability could be gained by sequencing using a strand-specific protocol (Borodina, Adjaye, & Sultan, 2011). If new transcriptomes were assembled using the above recommendations they could be used to complement this initial *I. balthica* transcriptome even though different protocols would be used (Yu et al., 2015).

4.2 Preprocessing

Preprocessing was done to avoid incorporating sequencing errors and technical artifacts into the assembled transcriptome (De Wit et al., 2015).

4.2.1 Visualization using FASTQC

Visualization simplifies analyzing hundreds of millions of sequenced reads. FASTQC is a tool that provides simple graphs and tables to easily check for quality issues with the data. The FASTQC report provides a quick overview of potential problems with the data, however it should be noted that FASTQC was originally developed for genomic reads therefore the some of the potential problems may not apply to RNA-Seq data.

The Per base sequence quality graphs show the range qualities of each base across the reads. Figure A1 shows the quality tends to decrease as the length of the reads increase. The trimming of the ends is shown to be successful since the quality of the ends increased (Figure A1).

The Per tile sequence quality graphs show technical problems with the flowcell. Figure A2 shows a completely blue graph in all cases, which indicates a high quality run on a problem free flowcell.

The Per sequence quality scores graphs show if a subset of reads have a universally low score. Figure A3 shows no subset of reads that are of low quality. The trimming is also shown to have increased the minimum mean sequence quality score (Figure A3).

The Per base sequence content graphs show the proportion of each base for every position in the read. Figure A4 shows A/T consistently around 30% and G/C consistently around 20% as expected given the GC content is 39%. However, positions 1-12 show a clear bias most likely produced from priming random hexamers (Figure A4). This is a known true technical bias and cannot be corrected by trimming and does not seem to affect downstream analysis (Andrews, 2015).

The Per sequence GC content graphs show the GC content per read and compares them to a modeled normal distribution. This assumes that the reads are randomly sampled from a uniformly represented library such as genomic DNA. Figure A5 shows sharp spikes in the GC content indicating bias. In this case the spikes are likely overrepresented reads, which contain that specific GC content. Overrepresented reads are common in RNA-Seq data since the expression levels vary greatly. However, these spikes could also be caused by contamination (e.g. a different species).

The Per base N content graphs shows the percentage of N calls for each position in the read. If there is an N that means the sequencer was unable to confidently make a base call. Figure A6 shows there were almost no N calls at any position percentage wise.

The Sequence length distribution graphs show the distribution of fragment sizes. As expected from Illumina reads all fragment sizes are the same, in this case 125bp, for the raw data (Figure A7). After trimming the fragment size can be seen to range from 36bp to 125bp, with most of the fragments retaining near full length (Figure A7).

The Sequence duplication levels graphs show the degree of duplication for the first 100,000 sequences truncated to 50bp. Figure A8 shows high levels of duplication in all cases. In a genomic library it would be expected for most sequences to not be duplicated, otherwise that could indicate over-sequencing or some kind of bias. However, in a RNA-Seq library it is expected that the sequences in the library occur at very different levels therefore high levels of duplication are expected in highly expressed transcripts (Andrews, 2015).

The Adapter content graphs show the percentage of specific adapters for each position in the reads. Figure A9 shows the raw read pairs both contain adapter sequences towards the 3’-end of the reads and the trimmed reads no longer contain these adapter sequences.

The K-mer content graphs show the top 6 most position-biased 7-mer enrichments. Figure A10 shows high levels of enrichment at the beginning of the reads in all cases. In a genomic library it is assumed that a diverse library should not have a k-mer with positional bias. However, in an RNA-Seq library overrepresented sequences will cause enrichment to be shown. In this case the peaks at the beginning of the reads were probably caused by random primers (Hansen, Brenner, & Dudoit, 2010), also seen in Figure A4, given the fact that FASTQC only analyses the first 2% of the file causing an incomplete sampling of all possible random primers (Andrews, 2015). This is a known and common bias that is independent of organism and laboratory and appears to have a minimal downstream effect (Hansen et al., 2010). The other peaks were most likely caused overrepresented sequences.

4.2.2 Trimming using TRIMMOMATIC

Trimming is important for removing technical artifacts, such as adapters, and low quality regions. By trimming this non-biological information the assembly should be more accurate.

TRIMMOMATIC was designed to work well with paired-end Illumina reads. It works by executing the commands in the order they are given, which is important for trimming adapter sequences. The adapter sequences should be removed first since other processing can alter the adapter sequences making them more difficult to identify.

The trimming results were as expected based on the FASTQC report. The raw data was high quality so only approximately 3.7M (1.24%) reads were dropped. Approximately 262M (88.47%) reads survived as pairs, which are used for the assembly. An Additional 30M (10.29%) reads survived as unpaired reads that could also be used for assembling, albeit with less reliability since the other pair was dropped. The paired trimmed files were re-visualized with FASTQC to verify the results.

4.2.3 Error correction using RCORRECTOR

Error correction is especially important for accurate identification of SNPs (De Wit et al., 2015). Error correction can also improve the accuracy and quality of assembled transcripts (Macmanes & Eisen, 2013).

RCORRECTOR was specifically designed for correcting RNA-Seq reads and was optimized with Illumina reads (Song & Florea, 2015). Unlike many other programs RCORRECTOR is flexible enough to account for different expression levels and isoforms (Song & Florea, 2015). The fact that RCORRECTOR can explore multiple correction paths is especially important for *de novo* assembling an *I. balthica* transcriptome since it is thought to have highly polymorphic sequences.

The error correcting results show a small difference between using a k-mer size of 20 and 25 in terms of total corrected bases. Only about 2M more bases were corrected when using a k-mer size of 20, which accounts for less than 0.0033% of the total number of bases. A k-mer size of 25 was chosen since it was the average of the planned k-mer sizes to be used by TRINITY.

4.2.4 *In silico* normalization using TRINITY

In silico normalization or digital normalization has been shown to greatly reduce the number of sequences that need to be assembled, which greatly reduces the runtime and can improve the assembly's quality (Brown et al., 2012).

The normalization algorithm used by TRINITY is based on DIGINORM (Brown et al., 2012). However, the normalization algorithm used by TRINITY is optimized for TRINITY and includes an extra stringent filter to decrease the incorporation of errors (Brown, 2012).

The normalization results show that only 3.85% of the reads were selected, which agrees with the FASTQC Sequence duplication levels graphs (Figure A8).

4.3 *De novo* Assembly using TRINITY

TRINITY is made up of three parts that work sequentially to assembly a transcriptome (Grabherr et al., 2011; Haas et al., 2013). First, TRINITY actually uses another program, JELLYFISH (Marcais & Kingsford, 2011), to extract a k-mer abundance catalog. Next, INCHWORM reads the k-mer abundance catalog to assemble contigs. CHRYSALIS then clusters the contigs and constructs de Bruijn graphs. Finally, BUTTERFLY traces the paths of the graphs to report full-length putative transcripts.

Different k-mer sizes were used to find a balance between sensitivity and specificity. By default TRINITY uses a k-mer size of 25 so a larger (30-mer) and smaller (20-mer) k-mer size were chosen. The 30-mer assembly was very similar in every metric measured, while the 20-mer assembly produced fewer and shorter contigs (Table 4).

A common way to quickly inspect the assemblies for major problems is to calculate contiguity metrics such as N50. N50 is the minimum contig length that contains at least half of the assembled bases. The 25 and 30-mer assemblies had an N50 value of 1,154 meaning that most of the information in the assembly was contained in contigs longer than 1,154bp. While there is no universally ideal N50 value intuitively a reasonably higher value is more likely to contain more useable information in a *de novo* assembly. Assuming previously assembled crustacean transcriptomes are biologically accurate an N50 value of 1,154 and average length of 683bp would suggest this assembly produced near full length transcripts (Table A1). However, the N50 statistic can often be exaggerated due to an assembler generating chimeric contigs or too many isoforms especially for longer transcripts (Haas et al., 2013).

The 25-mer assembly was chosen as the best assembly because it contained the most contigs and the most assembled bases therefore was most likely to contain the most information. It should be noted that these contiguity metrics do not correlate well with the correctness of an assembly (Salzberg et al., 2012; Simão et al., 2015). Also, although the 25-mer assembly contains more than 100,000 contigs it unlikely that more than 30,000 are actually protein coding transcripts based on assumed homology with *D. pulex* (Gilbert, 2007).

4.4 Post-processing

Post assembly processing is import to reduce contamination such as bacterial or protozoan transcripts, noncoding RNAs and potential technical assembly software artifacts (De Wit et al., 2015).

4.4.1 Clustering using CD-HIT

Cluster was done to maximize the potential of discovering novel transcripts by combing all three assemblies. At the same time, however, clustering collapsed potential paralogs.

The clustering metrics were very similar to the 25-mer assembly, but the clustered assemble contained more TRINITY-assigned 'genes' (Table 6). These results are consistent with adding novel transcripts and collapsing paralogs when compared to the 25-mer assembly.

4.4.2 Filtering using TRANSRATE

Filter was done in an attempt to select contigs that are accurate and biologically relevant. TRANSRATE filters for accuracy by selecting contigs that are supported by the reads used for the assembly. A contig can be considered accurate, complete and non-redundant if the read-pairs map back correctly (Smith-Unna et al., 2015). To filter out biologically irrelevant contigs, such as those that are too short to be useful, a minimum length of 600bp was used (Cahais et al., 2012).

The metrics of the filtered assemblies show that more than half of the total number of contigs was removed, with the greatest loss resulting from the length-filter (Table 6). Even though the length-filter filtered out more contigs those assemblies still retained a greater number of bases when compared to their respective TRANSRATE 'good' filtered assemblies (Table 6). In order to maximize the number of possible annotations only the 'good' filtered and unfiltered assemblies were evaluated.

4.5 Evaluation

In addition to the basic contiguity metrics (i.e. number of contigs, median contig length, N50, etc.) assemblies should be evaluated for accuracy, completeness and annotatable content (De Wit et al., 2015).

4.5.1 Read-mapping using TRANSRATE

Read-mapping is a way to determine how accurate a contigs is by providing supporting evidence from the reads. In general most of the reads (~70%-80%) should map back to the assembly (Haas et al., 2013).

TRANSRATE maps the reads to the assembly and also determines how well the reads are mapped for every contig. First, SNAP (Zaharia et al., 2011), a faster BOWTIE2-like tool, is used to align the reads to the assembly. Then, SALMON (https://github.com/kingsfordgroup/sailfish/releases/tag/v0.3.0) is used to infer the most likely contig of origin for reads that map to multiple

contigs. Finally, TRANSRATE uses its own tools to evaluate how well the mappings are.
TRANSRATE calls a mapping ‘good’ if a mapping is consistent with a perfectly assembled contig. Here a perfectly assembled contig is defined as one that has both members of a pair aligned on the same contig, in the correct orientation and without overlapping either end.

The mapped reads also provide evidence for theoretical links (i.e. bridges) that suggest different contigs originate from the same transcript. Ideally no bridges should exist if each actual transcript is represented by a single contig.

Each contig can also be evaluated based on read evidence. TRANSRATE assigns each contig a score as a way to measure how accurate, complete and non-redundant a representation of an actual transcript it is. These contig scores are used to determine an overall score of the assembly that can be used to compare different assemblies made with the same reads. The higher the score the more biologically accurate the assembly should be. TRANSRATE also produces an optimized assembly (i.e. ‘good’ filtered assembly) based on the optimal assembly score.

TRANSRATE can also compare an assembly to a related reference species to provide external validation. However, this was not done since no closely related reference species exists.

Each assembly evaluated had most of the reads mapped, even using the ‘good’ criteria (Table 7). The number of potential bridges in the unfiltered assemblies was much higher than the filtered ones suggesting many contigs were incorrectly assembled as isomers or paralogs of a single transcript (Table 7). However, these bridges could also be explained by assuming the contigs actually represent highly polymorphic transcripts. The scores suggest the clustered ‘good’ assembly is the most biologically accurate.

4.5.2 Completeness assessment using BUSCO

Benchmarking universal single-copy orthologs (BUSCOs) are ideal for evaluating completeness since it makes sense evolutionarily to find only single-copies of these genes in the genome (Waterhouse, Tegenfeldt, Li, Zdobnov, & Kriventseva, 2013).

BUSCO identifies the longest open reading frame for each contig and compares it to an arthropod-specific set of BUSCOs using a hidden Markov model.

Each assembly contained approximately half of the total number of BUSCOs in their complete form and approximately 68% that were at least partially found (Table 8). This suggests all of the assemblies could be improved, for example by sampling a female and at multiple life stages. The number of multi-copy BUSCOs is also unexpectedly high (~15% duplicated and ~8% triplicated), for each assembly, because it is assumed these genes are evolving under single-copy control (Waterhouse, Zdobnov, & Kriventseva, 2011). However, this level of duplication is reasonable when compared to that of *D. pulex* (Gilbert, 2007). The high level of duplication is also consistent with the highly polymorphic nature of *I. balthica* causing the assembler to create multiple contigs for a single transcript.

The clustered assembly also appears to have failed at adding new information with respect to BUSCOs since it actually contains fewer BUSCOs than the 25-mer assembly (Table 8). Furthermore, the clustered assembly did not significantly reduce the number of duplicated BUSCOs suggesting many of the duplicated BUSCOs are less than 95% similar (Table 8).

4.5.3 Predicting ORFs using TRANSDECODER

Predicting ORFs was done to try and remove noncoding RNA, DNA contamination, chimeras and gene fragments (De Wit et al., 2015).

TRANSDECODER predicts likely coding sequences by identifying ORFs that are at least 100 amino acids longs. The ORF also has to have a log-likelihood score greater than zero.

The ORF prediction results show a greater number of putative peptides were found in the unfiltered assemblies compared to the filtered assemblies (Table 9). This indicates that the filter may have been too strict, excluding more than 10,000 ORF containing contigs. To maximize the number of possible annotations the 25-mer assembly was chosen since it contained the most putative peptides.

4.6 Annotation

4.6.1 PANNZER

PANNZER is a state of the art fully automated protein annotation tool (Koskinen et al., 2015).

PANNZER works by searching the manually curated UniProtKB/Swiss database and clustering the hits to more reliably predict a functional description and GO classes.

The PANNZER annotations show many duplicate annotations. This is consistent with the high level of duplicated BUSCOs found suggesting the transcriptome may contain highly polymorphic alleles that were assembled incorrectly as two paralogs.

4.6.2 BLASTX to nr

Additional annotations were found using BLASTX to search the nr database. Unlike the PANNZER pipeline, which used the TRANSDECODER predicted peptides to search the small manually curated database, this BLASTX pipeline used the nucleotide contig sequences to search the larger automated database. By using the nucleotide sequences to query the protein database the potential for finding different reading frame proteins increases. Also searching a larger database increases the likelihood of finding more annotations. However, since these annotations are derived from an automated database and only use a simple filtering strategy they should be used with caution.

The BLASTX to nr annotations show many clusters of contigs with more than two copies. This could suggest a very high level of polymorphism with genes that have more than two different alleles, or that many are paralogs, or a combination of the two.

4.7 Conclusion

This first draft of the *Idotea balthica* transcriptome was created from a single adult male isopod. The RNA was sequenced using an Illumina HiSeq 2500 sequencer. Approximately 300M 125bp PE reads were produced. The reads were preprocessed (i.e. trimmed, error corrected and normalized) before being *de novo* assembled. In an attempt to improve the assembly, a combined and clustered assembly was created as well as filtered assemblies. The assemblies were evaluated by measuring accuracy, completeness and potential for annotations.

The unfiltered 25-mer assembly represents the best first draft assembly. The filtered assemblies appear to have removed a significant number of coding contigs and the clustered assembly appears to have collapsed paralogs.

The nature of this first draft assembly shows high levels contig duplication. One possible explanation is that the duplicates are paralogs, consistent with what is known about *D. pulex* (Gilbert, 2007) and *C. finmarchicus* (Lenz et al., 2014). Another explanation is that the duplicates are highly polymorphic allelic variants. Of course a combination of both of these explanations could also be possible. Without a high-quality reference transcriptome or genome of a closely related species it is not currently possible to determine if the paralogs should be collapsed. This first draft assembly also lacks completeness. The assembly can be found online (https://github.com/The-Bioinformatics-Group/Idotea_balthica_transcriptome_project/tree/master/Draft_1) and includes the assembly, PANNZER annotations, BLASTX to nr annotations, and contig scores.

The *Idotea balthica* transcriptome could be improved by assembling additional sequences from individuals of different ages as well as genders. The apparent paralogs could be resolved as true paralogs or allelic variants when the planned *Idotea balthica* genome project is completed. Furthermore, adding longer sequences (e.g. PacBio) could be used as a scaffold to improve assembly accuracy.

5. References

Andrews, S. (2015). FASTQC A Quality Control tool for High Throughput Sequence Data. Retrieved February 9, 2016, from http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3 Analysis Modules/

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. doi:10.1093/bioinformatics/btu170

Borodina, T., Adjaye, J., & Sultan, M. (2011). A Strand-Specific Library Preparation Protocol for RNA Sequencing. In *Methods in enzymology* (1st ed., Vol. 500, pp. 79–98). Elsevier Inc. doi:10.1016/B978-0-12-385118-5.00005-0

Bron, J. E., Frisch, D., Goetze, E., Johnson, S. C., Lee, C., & Wyngaard, G. A. (2011). Observing copepods through a genomic lens. *Frontiers in Zoology*, *8*(1), 22. doi:10.1186/1742-9994-8-22

Brown, T. C. (2012). What does Trinity's in silico normalization do? doi:http://dx.doi.org/10.6084/m9.figshare.98198

Brown, T. C., Howe, A., Zhang, Q., Pyrkosz, A. B., & Brom, T. H. (2012). A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data. Retrieved from http://arxiv.org/abs/1203.4802

Cahais, V., P., G., Tsagkogeorga, G., MELO-FERREIRA, J., BALLENGHIEN, M., WEINERT, L., ... GALTIER, N. (2012). Reference-free transcriptome assembly in non-model animals from next-generation sequencing data. *Molecular Ecology Resources*, *12*(5), 834–845. doi:10.1111/j.1755-0998.2012.03148.x

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, *10*(1), 421. doi:10.1186/1471-2105-10-421

Compeau, P. E. C., Pevzner, P. A., & Tesler, G. (2011). How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology*, *29*(11), 987–991. doi:10.1038/nbt.2023

De Wit, P., Pespeni, M. H., Ladner, J. T., Barshis, D. J., Seneca, F., Jaris, H., ... Palumbi, S. R. (2012). The simple fool's guide to population genomics via RNA-Seq: an introduction to high-throughput sequencing data analysis. *Molecular Ecology Resources*, *12*(6), 1058–67. doi:10.1111/1755-0998.12003

De Wit, P., Pespeni, M. H., & Palumbi, S. R. (2015). SNP genotyping and population genomics from expressed sequences -current advances and future possibilities. *Molecular Ecology, 24*, 2310–2323. doi:10.1111/mec.13165

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Computational Biology*, *7*(10), e1002195. doi:10.1371/journal.pcbi.1002195

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, *28*(23), 3150–3152. doi:10.1093/bioinformatics/bts565

Ghaffari, N., Sanchez-Flores, A., Doan, R., Garcia-Orozco, K. D., Chen, P. L., Ochoa-Leyva, A., ... Criscitiello, M. F. (2014). Novel transcriptome assembly and improved annotation of the whiteleg shrimp (Litopenaeus vannamei), a dominant crustacean in global seafood mariculture. *Scientific Reports*, *4*, 7081. doi:10.1038/srep07081

Gilbert, D. (2007). Daphnia pulex: Rich in tandem gene duplications. Retrieved February 8, 2016, from http://wfleabase.org/genome-summaries/gene-duplicates/daphnia-gene-tandems.html

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., ... Regev, A. (2011). Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature Biotechnology*, *29*(7), 644–652. doi:10.1038/nbt.1883.Trinity

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., ... Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, *8*(8), 1494–1512. doi:10.1038/nprot.2013.084

Haavisto, F., & Jormalainen, V. (2014). Seasonality elicits herbivores' escape from trophic control and favors induced resistance in a temperate macroalga. *Ecology*, *95*(11), 3035–3045. doi:10.1890/13-2387.1

Hansen, K. D., Brenner, S. E., & Dudoit, S. (2010). Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Research*, *38*(12), e131–e131. doi:10.1093/nar/gkq224

HELCOM. (2013). Climate change in the Baltic Sea Area: HELCOM thematic assessment in 2013. In *Baltic Sea Environment Proceedings No. 137*. Helsinki: Helsinki Commission. Retrieved from http://www.helcom.fi/Lists/Publications/BSEP137.pdf

Hemmi, A., & Jormalainen, V. (2004). Genetic and environmental variation in performance of a marine isopod: effects of eutrophication. *Oecologia*, *140*(2), 302–311. doi:10.1007/s00442-004-1574-7

Johannesson, K., & Andre, C. (2006). Life on the margin: genetic isolation and diversity loss in a peripheral marine ecosystem, the Baltic Sea. *Molecular Ecology*, *15*(8), 2013–2029. doi:10.1111/j.1365-294X.2006.02919.x

Jormalainen, V., Honkanen, T., & Vesakoski, O. (2008). Geographical divergence in host use ability of a marine herbivore in alga–grazer interaction. *Evolutionary Ecology*, *22*(4), 545–559. doi:10.1007/s10682-007-9181-9

Jung, H., Lyons, R. E., Dinh, H., Hurwood, D. A., McWilliam, S., & Mather, P. B. (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

Koskinen, P., Toronen, P., Nokso-Koivisto, J., & Holm, L. (2015). PANNZER: high-throughput functional annotation of uncharacterized proteins in an error-prone environment. *Bioinformatics*, *31*(10), 1544–1552. doi:10.1093/bioinformatics/btu851

Kukurba, K. R., & Montgomery, S. B. (2015). RNA Sequencing and Analysis. *Cold Spring Harbor Protocols*, *2015*(11), pdb.top084970. doi:10.1101/pdb.top084970

Leidenberger, S., Harding, K., & Jonsson, P. R. (2012). Ecology and distribution of the isopod genus in the Baltic Sea: key species in a changing environment. *Journal of Crustacean Biology*, *32*(3), 359–381. doi:http://dx.doi.org/10.1163/193724012X626485

Lenz, P. H., Roncalli, V., Hassett, R. P., Wu, L.-S., Cieslak, M. C., Hartline, D. K., & Christie, A. E. (2014). De Novo Assembly of a Transcriptome for Calanus finmarchicus (Crustacea, Copepoda) – The Dominant Zooplankter of the North Atlantic Ocean. *PLoS ONE*, *9*(2), e88589. doi:10.1371/journal.pone.0088589

Li, W., & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), 1658–1659. doi:10.1093/bioinformatics/btl158

Macmanes, M. D., & Eisen, M. B. (2013). Improving transcriptome assembly through error correction of high-throughput sequence reads. *PeerJ*, *1*, e113. doi:10.7717/peerj.113

Marcais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, *27*(6), 764–770. doi:10.1093/bioinformatics/btr011

Martin, J. a, & Wang, Z. (2011). Next-generation transcriptome assembly. *Nature Reviews. Genetics*, *12*(10), 671–682. doi:10.1038/nrg3068

Meier, M. H. (2006). Baltic Sea climate in the late twenty-first century: a dynamical downscaling approach using two global models and two emission scenarios. *Climate Dynamics*, *27*(1), 39–68. doi:10.1007/s00382-006-0124-x

Meier, M. H., & Eilola, K. (2011). Future projections of ecological patterns in the Baltic Sea. *Swedish Meteorological and Hydrological Institute*, (107). Retrieved from http://www.smhi.se/sgn0106/if/biblioteket/rapporter_pdf/Oceanografi_107.pdf

Ning, J., Wang, M., Li, C., & Sun, S. (2013). Transcriptome Sequencing and De Novo Analysis of the Copepod Calanus sinicus Using 454 GS FLX. *PLoS ONE*, *8*(5), e63741. doi:10.1371/journal.pone.0063741

Ozsolak, F., & Milos, P. M. (2011). RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics*, *12*(2), 87–98. doi:10.1038/nrg2934.RNA

Pevzner, P. A., Tang, H., & Waterman, M. S. (2001). An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences*, *98*(17), 9748–9753. doi:10.1073/pnas.171285098

Poore, A. G. B., Campbell, A. H., Coleman, R. A., Edgar, G. J., Jormalainen, V., Reynolds, P. L., … Emmett Duffy, J. (2012). Global patterns in the impact of marine herbivores on benthic primary producers. *Ecology Letters*, *15*(8), 912–922. doi:10.1111/j.1461-0248.2012.01804.x

Rice, P., Longden, I., & Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, *16*(6), 276–277. doi:10.1016/S0168-9525(00)02024-2

Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., … Birol, I. (2010). De novo assembly and analysis of RNA-seq data. *Nat Meth*, *7*(11), 909–912. Retrieved from http://dx.doi.org/10.1038/nmeth.1517

Salzberg, S. L., Phillippy, A. M., Zimin, A., Puiu, D., Magoc, T., Koren, S., … Yorke, J. A. (2012). GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Research*, *22*(3), 557–567. doi:10.1101/gr.131383.111

Schulz, M. H., Zerbino, D. R., Vingron, M., & Birney, E. (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

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

Smith-Unna, R., Boursnell, C., Patro, R., Hibberd, J. M., & Kelly, S. (2015). TransRate: reference free quality assessment of de-novo transcriptome assemblies. *bioRxiv*, 1–25. doi:10.1101/021626

Song, L., & Florea, L. (2015). Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. *GigaScience*, *4*(1), 48. doi:10.1186/s13742-015-0089-y

Stahl, B. A., Gross, J. B., Speiser, D. I., Oakley, T. H., Patel, N. H., Gould, D. B., & Protas, M. E. (2015). A Transcriptomic Analysis of Cave, Surface, and Hybrid Isopod Crustaceans of the Species Asellus aquaticus. *PLOS ONE*, *10*(10), e0140484. doi:10.1371/journal.pone.0140484

Stefano, G. B. (2014). Comparing Bioinformatic Gene Expression Profiling Methods: Microarray and RNA-Seq. *Medical Science Monitor Basic Research*, *20*, 138–142. doi:10.12659/MSMBR.892101

Stillman, J. H., Colbourne, J. K., Lee, C. E., Patel, N. H., Phillips, M. R., Towle, D. W., … Terwilliger, N. B. (2008). Recent advances in crustacean genomics. *Integrative and Comparative Biology*, *48*(6), 852–868. doi:10.1093/icb/icn096

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

Waterhouse, R. M., Tegenfeldt, F., Li, J., Zdobnov, E. M., & Kriventseva, E. V. (2013). OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Research*, *41*(D1), D358–D365. doi:10.1093/nar/gks1116

Waterhouse, R. M., Zdobnov, E. M., & Kriventseva, E. V. (2011). Correlating Traits of Gene Retention, Sequence Divergence, Duplicability and Essentiality in Vertebrates, Arthropods, and Fungi. *Genome Biology and Evolution*, *3*(0), 75–86. doi:10.1093/gbe/evq083

Yu, N. Y., Hallström, B. M., Fagerberg, L., Ponten, F., Kawaji, H., Carninci, P., … Daub, C. O. (2015). Complementing tissue characterization by integrating transcriptome profiling from the Human Protein Atlas and from the FANTOM5 consortium. *Nucleic Acids Research*, *43*(14), 6787–6798. doi:10.1093/nar/gkv608

Zaharia, M., Bolosky, W. J., Curtis, K., Fox, A., Patterson, D., Shenker, S., … Sittler, T. (2011). Faster and More Accurate Sequence Alignment with SNAP, 1–10. Retrieved from http://arxiv.org/abs/1111.5572

Zeng, V., Villanueva, K. E., Ewen-Campen, B. S., Alwes, F., Browne, W. E., & Extavour, C. G. (2011). De novo assembly and characterization of a maternal and developmental transcriptome for the emerging model crustacean Parhyale hawaiensis. *BMC Genomics*, *12*(1), 581. doi:10.1186/1471-2164-12-581

Zerbino, D. R., & Birney, E. (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

Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K., & Liu, X. (2014). Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. *PLoS ONE*, *9*(1), e78644. doi:10.1371/journal.pone.0078644

6. Appendix

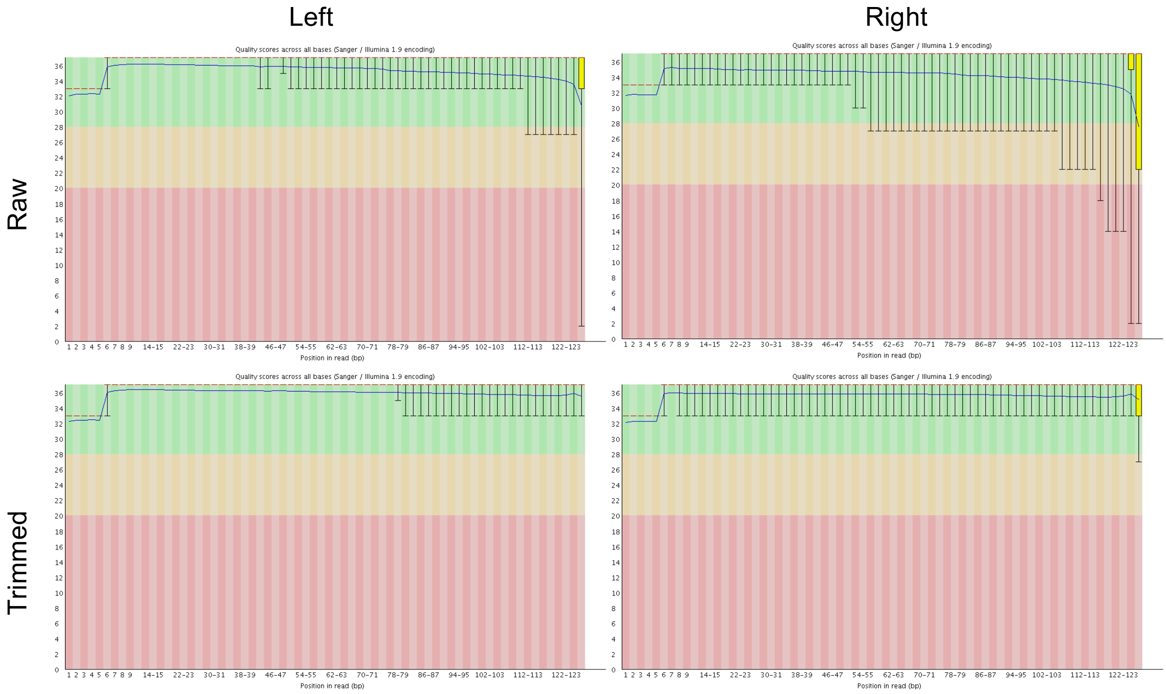

Figure A1. FASTQC: Per base sequence quality

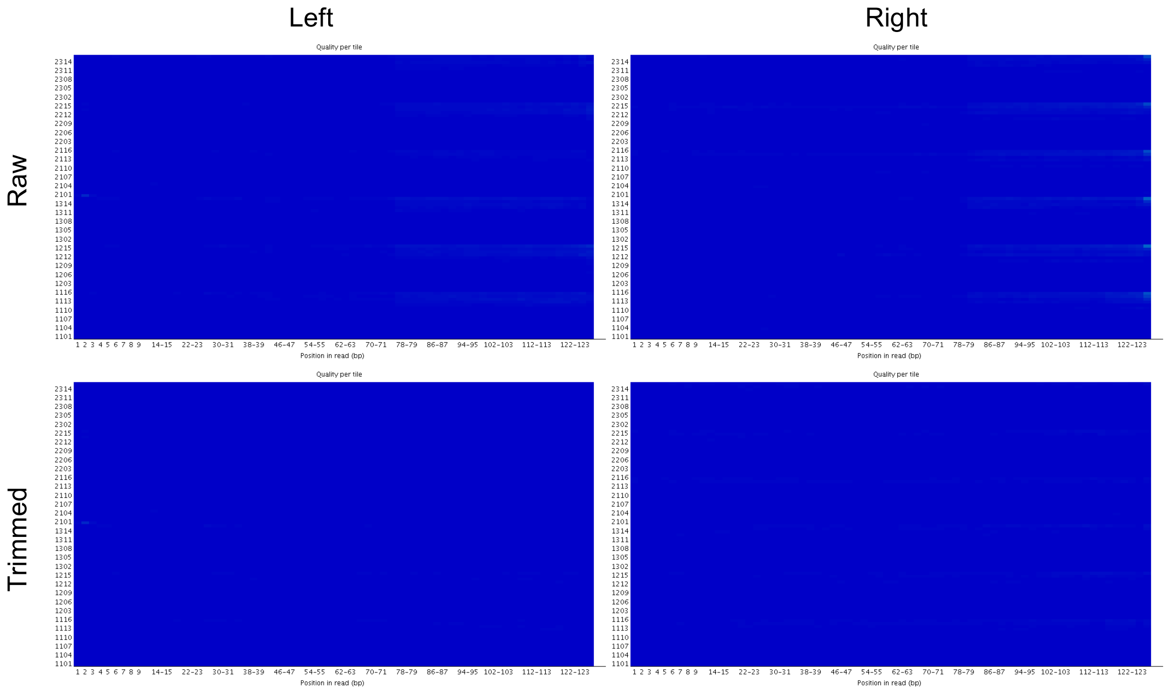

Figure A2. FASTQC: Per tile sequence quality

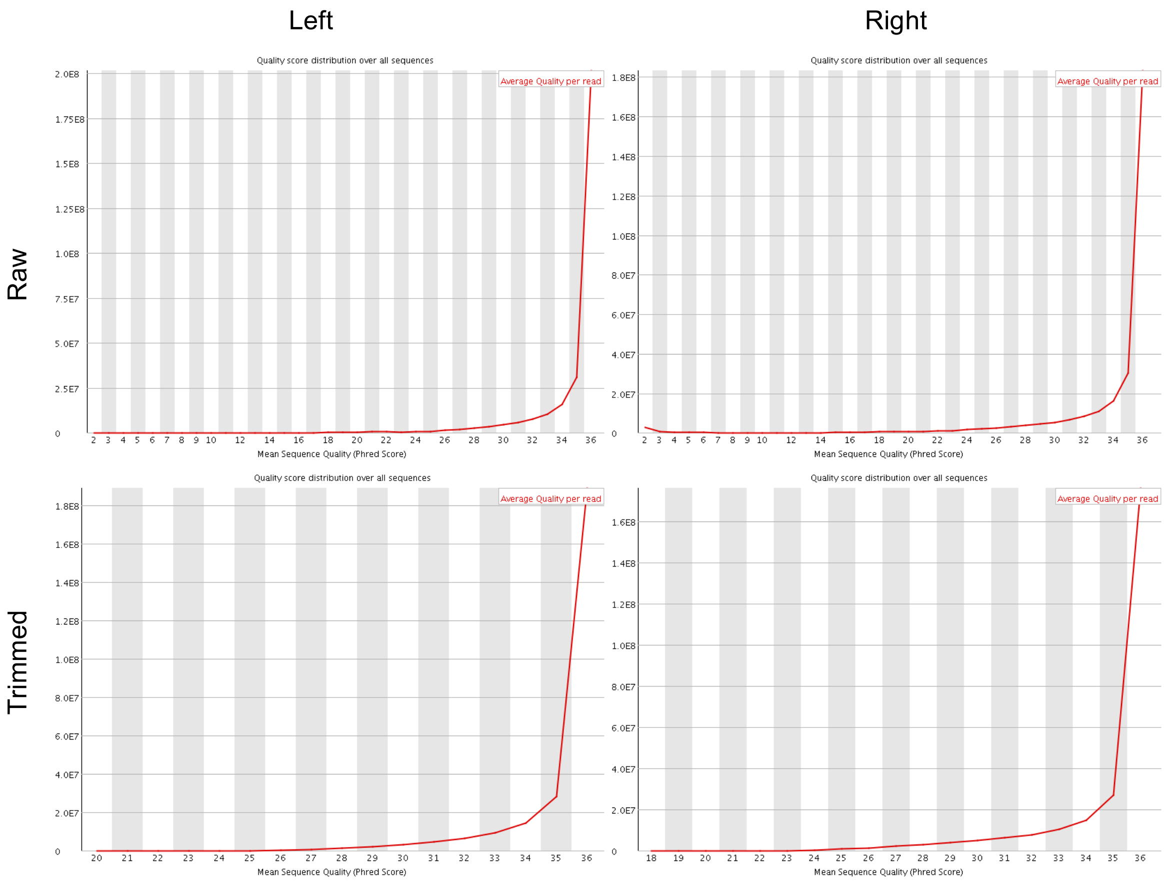

Figure A3. FASTQC: Per sequence quality scores

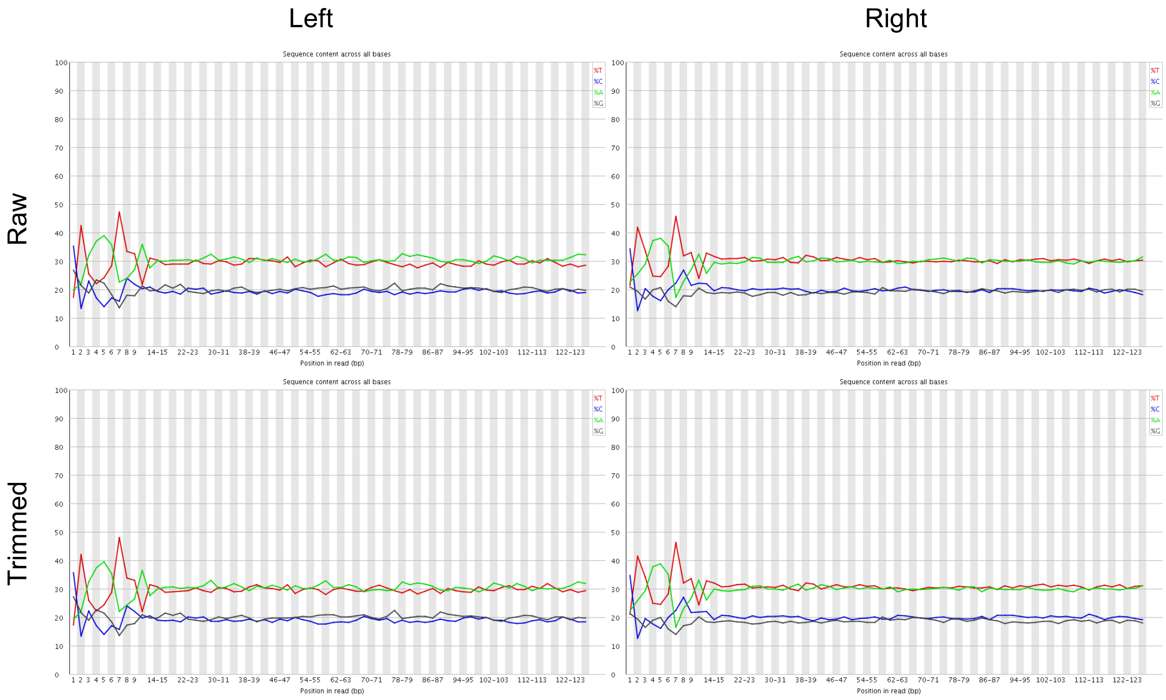

Figure A4. FASTQC: Per base sequence content

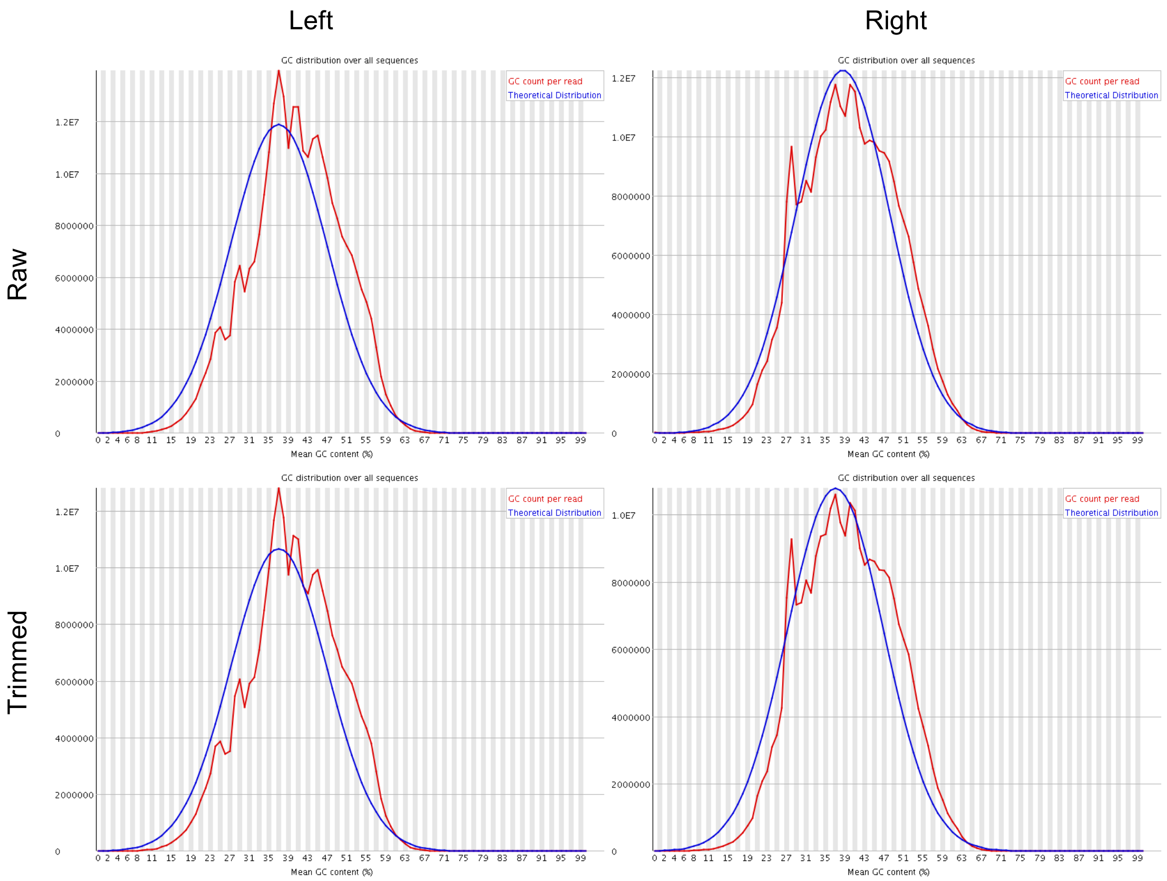

Figure A5. FASTQC: Per sequence GC content

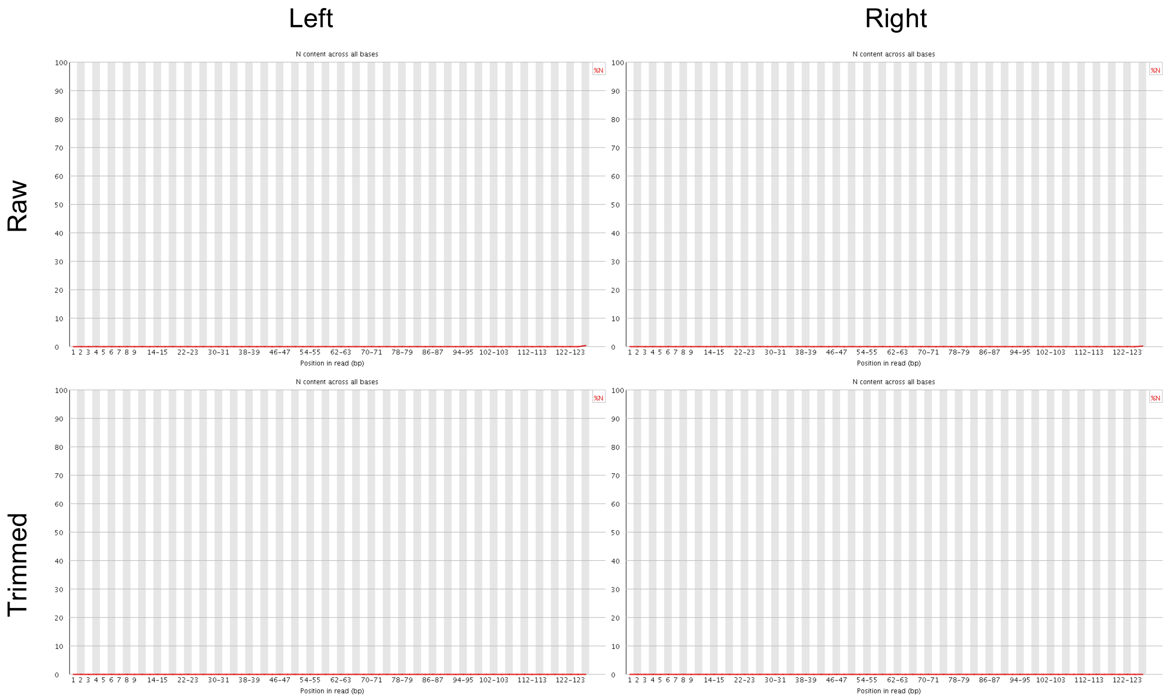

Figure A6. FASTQC: Per base N content

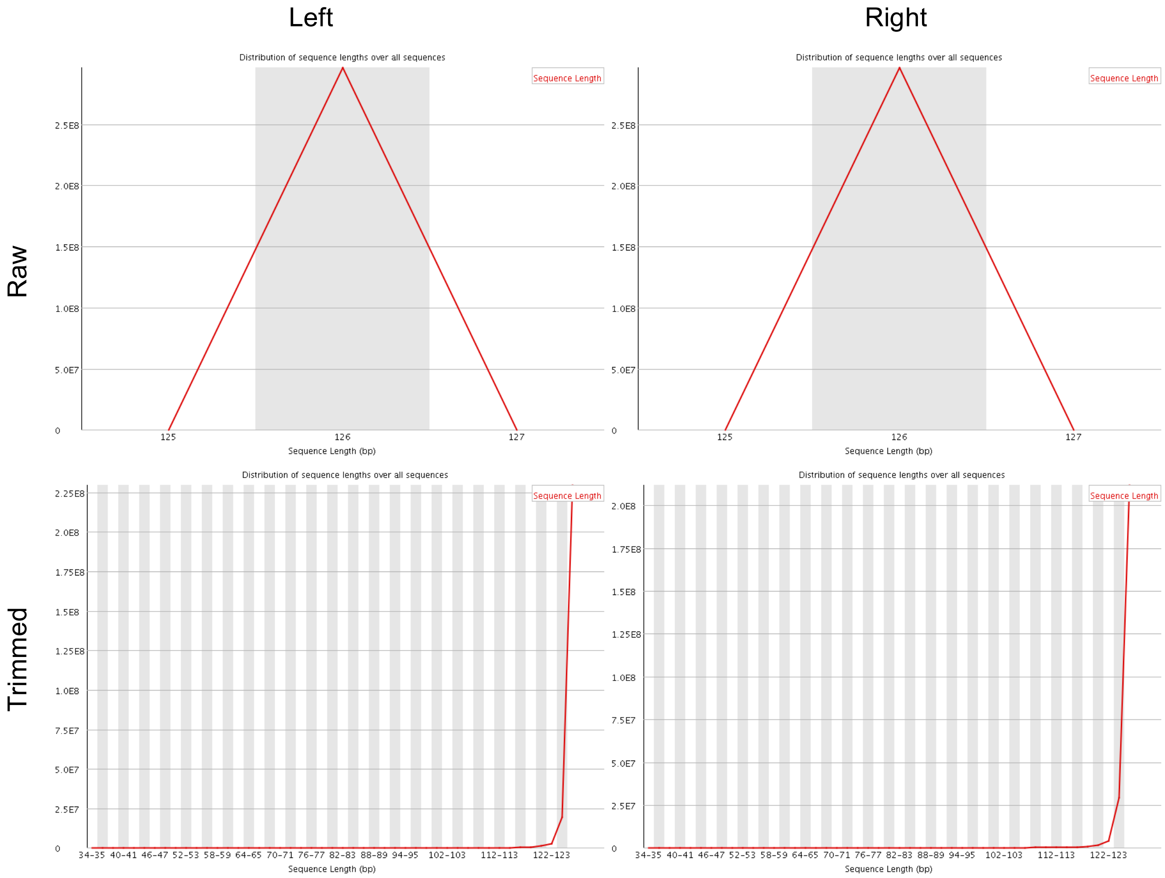

Figure A7. FASTQC: Sequence length distribution

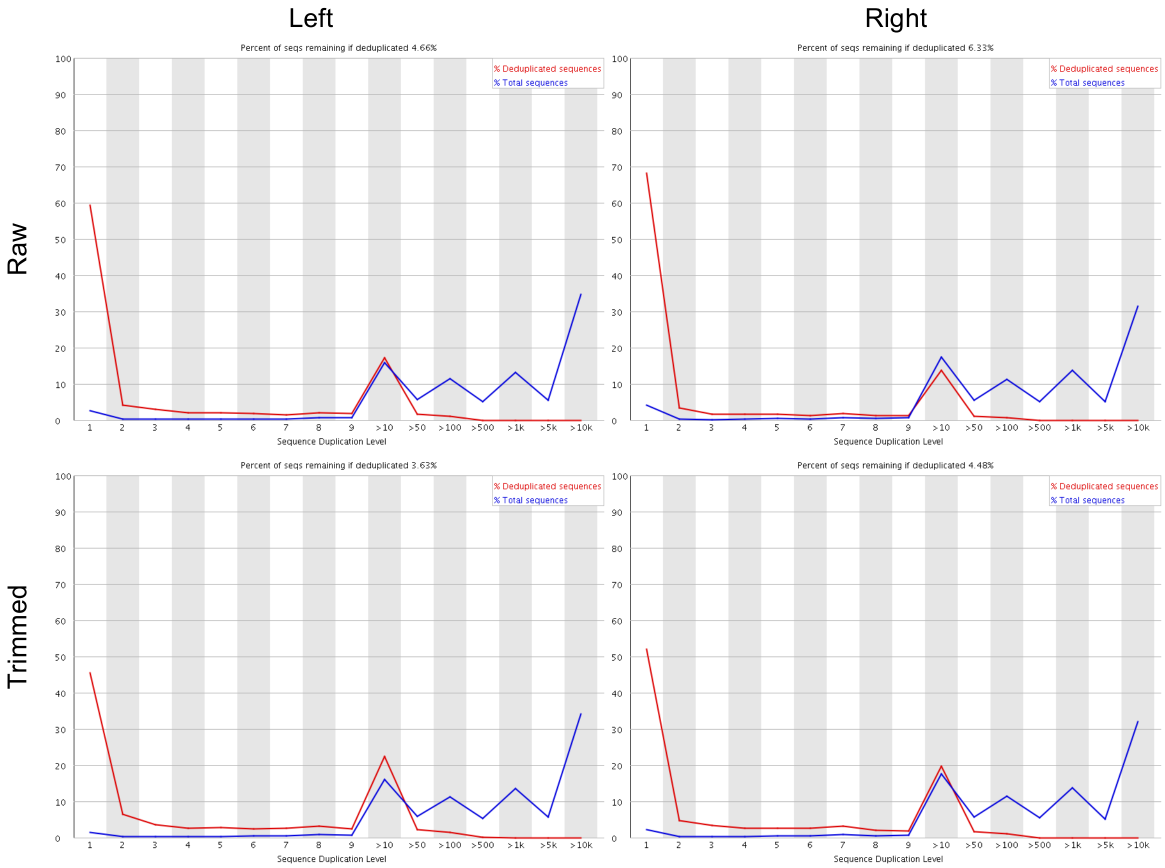

Figure A8. FASTQC: Sequence duplication levels

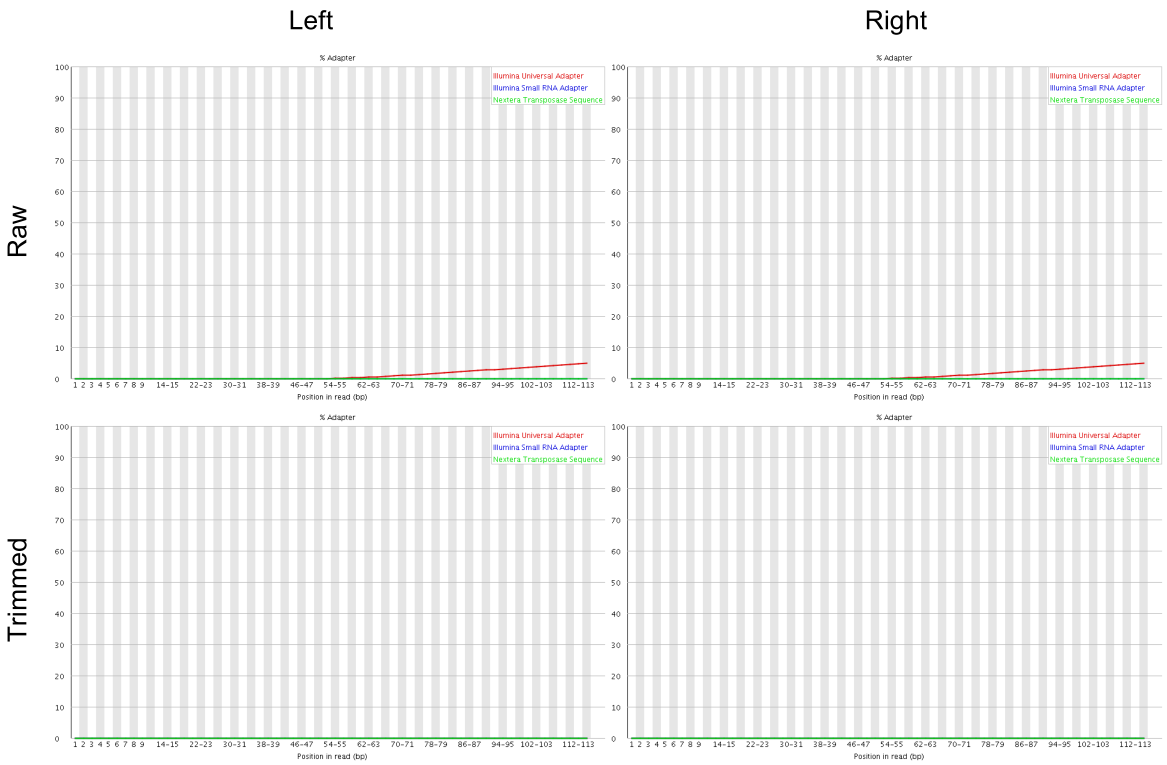

Figure A9. FASTQC: Adapter content

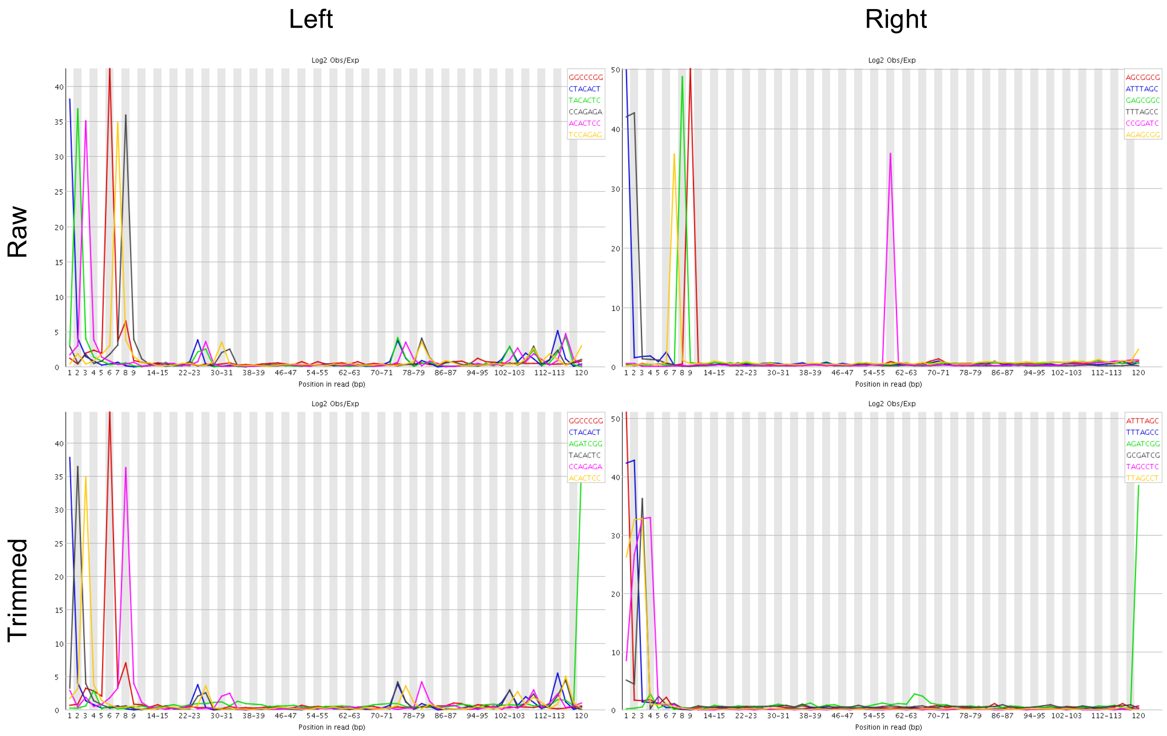

Figure A10. FASTQC: K-mer content

Table A1. Comparison of crustacean transcriptomes
Calanus finmarchicus (Lenz et al., 2014)
Litopenaeus vannamei (Ghaffari et al., 2014)
Asellus aquaticus (Stahl et al., 2015)
Macrobrachium rosenbergii (Jung et al., 2011)
Calanus sinicus (Ning, Wang, Li, & Sun, 2013)
Parhyale hawaiensis (Zeng et al., 2011)

	I. balthica	***C. finmarchicus***	***L. vannamei***	***A. aquaticus***	***M. rosenbergii***	***C. sinicus***	***P. hawaiensis***
Contigs	115,931	206,041	110,474	23,984	8,411	56,809	89,664
Shortest	201	301	201	21	40	?	?
Average	683.11	997	1,137.44	500	845	?	?
Longest	21,569	23,068	31,344	3,490	7,531	?	?
N50	1,154	1,418	2,701	800	?	873	1,510
GC	39	43	44	?	?	?	?
Total bases	79M	205M	126M	?	?	?	?
Technology	Illumina	Illumina	Illumnia	454	454	454	454
Assembler	Trinity	Trinity	Trinity	NGen	?	Newbler	Newbler
Common name	isopod	copepod	shrimp	isopod	prawn	copepod	amphipod

7. Acknowledgments

This research was partially supported by the University of Turku Joint Student Scholarship Fund.

The author wishes to acknowledge CSC – IT Center for Science, Finland, for computational resources.