For GigaScience (Technical Note):

<https://academic.oup.com/gigascience/pages/technical_note>

**Re-assembling and evaluating quality of 678 microbial eukaryotic reference transcriptomes**

Lisa K. Johnson1,2, Harriet Alexander1, C. Titus Brown1,2,3\*

1 Department of Population Health and Reproduction, School of Veterinary Medicine, University of California, Davis

2 Molecular, Cellular, and Integrative Physiology Graduate Group, University of California, Davis

3 Genome Center, University of California, Davis

\* Correspondence: [ctbrown@ucdavis.edu](mailto:ctbrown@ucdavis.edu)

**Abstract**

The Marine Microbial Eukaryotic Transcriptome Sequencing Project (MMETSP) is a unique and large set of RNA-seq data, allowing for evaluation with quantitative metrics. We present an automated and extensible *de novo* transcriptome pipeline used to re-assemble and evaluate 678 samples from the MMETSP data set. Re-assemblies show improved qualities and additional content compared to assemblies generated with a previous pipeline. Phylogenetic trends in assembly qualities were observed.

**Introduction**

Good quality transcriptome assemblies are necessary as a reference for studying the underlying contribution of gene expression to the biology of an organism. In the absence of a previously-existing transcriptome, *de novo* assemblies must be generated from raw RNAseq data.

Bioinformatics tools and workflows used to generate *de novo* transcriptome assemblies undergo a high turnover rate. At this time, the major *de novo* transcriptome assemblers include Trinity (Grabherr et al. 2011), SOAPdenovo-Trans (Xie et al. 2014), Trans-ABySS (Robertson et al. 2010), Oases (Schulz et al. 2012), IDBA-tran (Peng et al. 2013) and a new lineage of assemblers, including Shannon (Kannan et al. 2016). The development of new and improved tools is an active area of research and will continue to evolve in the future as algorithms continue to improve to deal with challenges associated with *de novo* transcriptome assembly, such as spliced isoforms and allelic variants (e.g. discussed in Chang et al. 2014, Grabherr 2011).

Best practices for the analysis of RNA-seq data are not well defined and lack a consensus workflow, especially for non-model species (Conesa et al. 2016).

We present an automated open-access, modularized and extensible pipeline developed for this study to address the following problems: future software version updates and new software programs, failure of individual processes during execution of the pipeline, and addition of more samples.

Eel pond protocol, automated.

In running a re-assembly of the MMETSP we sought to both improve the MMETSP reference transcriptomes and to create a platform to facilitate future re-assembly efforts. Moreover, the large number of mRNA sequence datasets combined with the vast genetic diversity encompassed by the MMETSP data make it possible to explore and statistically compare the results of two different assembly pipelines. Through this comparison we were able to examine a variety of transcriptome metrics to determine which of the assembly pipelines generated more unique and biologically relevant content. Moreover, we compared transcriptome quality metrics at the phylum-level to determine if any biases or inherent differences between broad taxonomic groups.

The Marine Microbial Eukaryotic Transcriptome Sequencing Project (MMETSP) created a catalogue of genes facilitating investigation of how organisms thrive in diverse marine habitats and how they influence marine ecosystems, biogeochemical cycles and the composition of the atmosphere (Keeling et al. 2014). This is one of the largest publicly-available mRNA sequencing data sets, purposefully-built from a diversity of non-model eukaryotic taxa with a standardized library preparation.

Reference transcriptome assemblies generated from these data have facilitated investigations in phylogenetic analyses (Durkin et al. 2016), phylogenetic tools (Grant and Katz 2014), differential gene expression analysis (Frishkorn et al. 2014), inter-group comparisons (Koid et al. 2014), meta-omic approaches (Alexander et al 2015) have been developed and interesting biological features discovered (Alkalaeva and Mikhailova 2016, Heaphy *et al*. 2016, Swart *et al*. 2016), and gene diversity and evolution (Grousman et al. 2015).

The original reference MMETSP transcriptomeswere generated by the National Center for Genome Research (NCGR). Software tools have been updated and new tools are available.

**Methods**

*Automated Pipeline*

An automated pipeline was developed to execute the steps of the Eel Pond mRNAseq Protocol (<http://eel-pond.readthedocs.io/en/latest/>), a lightweight protocol for assembling RNA-seq reads using the Trinity *de novo* transcriptome assembler (Brown et al. 2013). This protocol has been demonstrated to produce good quality *de novo* transcriptome assemblies (Lowe, Swalla, Brown, 2014). The pipeline was used to assemble all of the data from the MMETSP (Figure 1). The pipeline code and installation instructions are available: <https://doi.org/10.5281/zenodo.249982>.

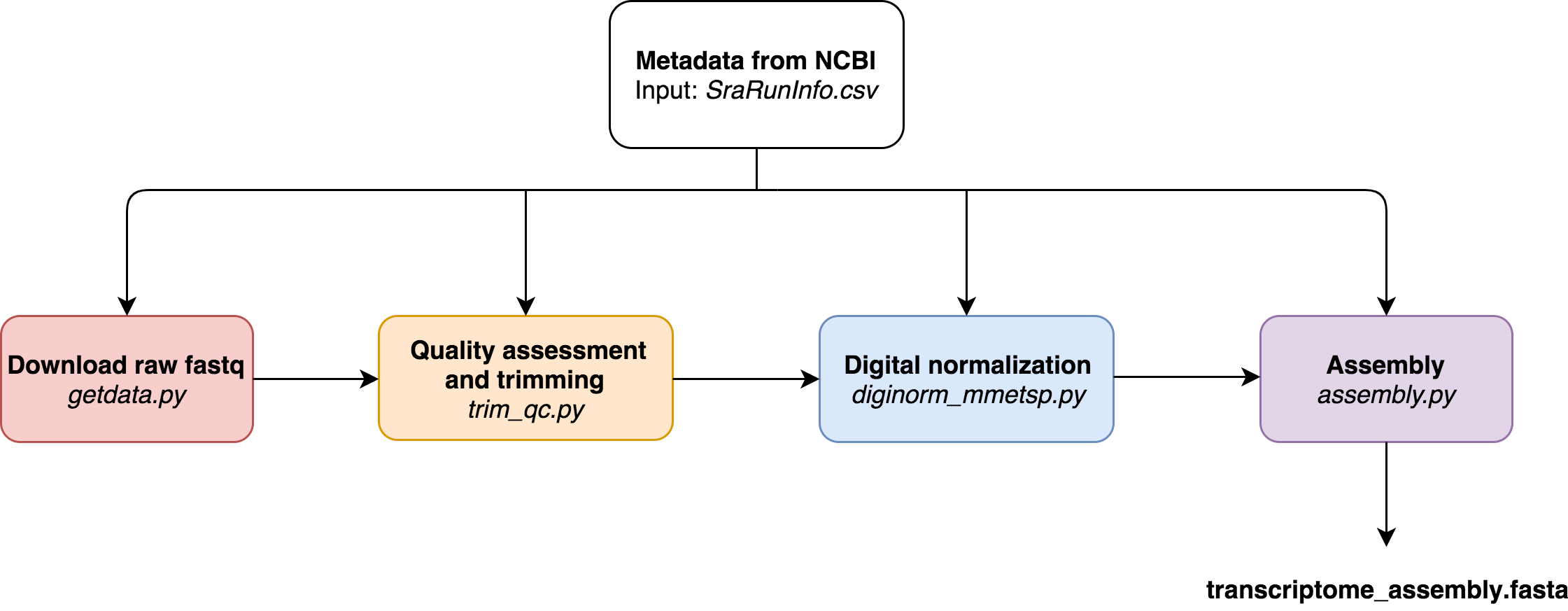


Figure 1. The modularized *de novo* transcriptome assembly pipeline workflow. Metadata in the SraRunInfo.csv are downloaded from NCBI: 1.) downloading data with the *fastq-dump* script from the *SRA Toolkit* (Leinonen et al 2014), 2.) quality assessment with *FastQC* (Andrews 2010) and trimming residual adapters and low quality bases (Q<2) with *Trimmomatic* (MacManes 2014; Bolger et al. 2014), 3.) digital normalization with *khmer* version 2.0 (Crusoe et al. 2015), 4.) assembly with *Trinity* (Grabherr et al. 2011). Each script is controlled by the metadata file, *SraRunInfo.csv* obtained from the SRA.

1. Download Data

Initial assemblies were run by the the National Center for Genome Resources (NCGR) using methods and data described in Keeling et al. (2014).

Raw RNA-seq data were obtained from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA): PRJNA231566. Data were paired-end (PE) reads with a length of 50 bases each. The metadata file (SraRunInfo.csv) was obtained from the SRA and 719 samples were downloaded and extracted using fastq-dump in the SRA Toolkit (version 2.5.4) (Leinonen et al. 2011). There were 18 MMETSP samples with more than one sequence record. In this case, digital normalization was performed on each sample record, and then sample records were pooled and assembled together for a total of 678 assemblies.

2. Quality Control

Reads were analyzed with the fastQC (version 0.11.5) program to confirm overall qualities before and after trimming. Per base qualities were consistently >Q30 across the lengths of all reads before trimming. Trimmomatic (version 0.33) was used to trim residual TruSeq adapters and bases <Q2 (Bolger et al. 2014), settings as recommended by MacManes (2014).

3. Digital Normalization

For memory-efficient assembly, reads were interleaved and normalized to a *k*-mer coverage of 20 using khmer software (version 2.0) (Crusoe et al. 2015, Zhang et al. 2014, Brown et al. 2012). Orphaned reads, where the pair was removed during normalization, were included in the assembly.

4. Assembly

Transcriptomes were assembled from normalized reads with Trinity 2.2.0. Other versions were tested, see Supplemental.

*Annotation*

Transcriptomes were annotated using the dammit pipeline (Scott 2016), which relies on the following databases as evidence: Pfam-A (Finn et al. 2016), Rfam (Gardner et al. 2008), OrthoDB (Zdobnov et al. 2016).

*Evaluation*

All assemblies were evaluated using metrics generated by the TransRate program (Smith-Unna et al. 2016). Trimmed reads were used to calculate a TransRate score for each assembly, which represents the geometric mean of all contig scores multiplied by the proportion of input reads providing positive support for the assembly (Smith-Unna et al. 2016). New re-assemblies will hereafter be referred to as “DIB” since we generated them in the Data Intensive Biology lab at UC Davis. Previous assemblies generated by the National Center for Genome Research will be referred to as “NCGR”. Comparative metrics were calculated for each MMETSP sample between DIB and the NCGR assemblies using the Conditional Reciprocal Best BLAST hits (CRBB) algorithm (Aubry et al. 2014). A forward comparison was made with the NCGR assembly used as the reference and each DIB assembly as the query. Reverse comparative metrics were calculated with each Trinity assembly as the reference and the NCGR assembly as the query.

Benchmarking universal single-copy orthologs (BUSCO) software (version 2) was used to compare 234 genes from a database of Protist with open reading frames in assemblies. BUSCO scores have been suggested to be a measure of assembly completeness (Simão et al. 2015). The eukaryotic database with 303 genes was also compared (Supplemental).

Unique *k*-mers were calculated by digesting each assembly using a HyperLogLog function to count the number of distance *k*=25 elements in the assembly (Irber and Brown 2016). To compare the similarities between samples, Minhash sketches were calculated with sourmash for read one of the first one million reads of each sample (Brown and Irber 2016).

**Results**

"X % of the samples have higher ORF content in the DIB assembly than in the NCGR assembly"

“DIB” for Data Intensive Biology assemblies had more contigs (Figure 2 - left), 43,882 ± 26,116 contigs with some samples producing over 190,000 contigs, while the mean of NCGR was 30,179 ± 21,341 contigs. A two-sample Komogorov-Smirnov test comparing the two distributions containing 635 of the 678 samples (excluding redundancies) indicated p < 2.2e-16 (D = 0.29793). Only 17 samples had more contigs in the NCGR assemblies compared to the DIB re-assemblies. Moreover, the qualities of our assemblies are higher. The mean score of DIB re-assemblies was 0.31 ± 0.1 while the NCGR transrate score was 0.22 ± 0.09, p < 2.2e-16 (D = 0.48827) (Figure 2 - right).

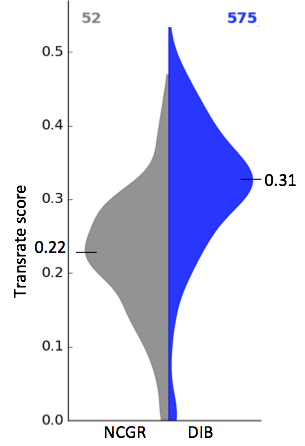
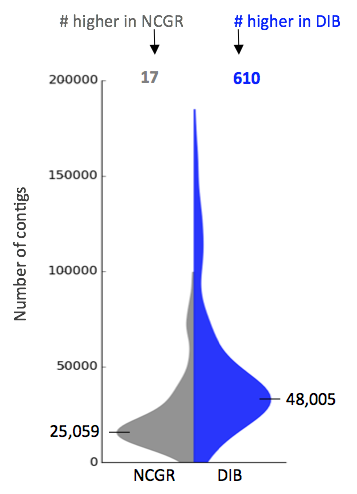


Figure 2. Split violin plots of frequency distributions of the number of assembled contigs (A) and TransRate quality scores (B) of each pipeline. In the blue (right side of each plot) are the “DIB” re-assemblies and in gray (left side of each plot) are the original assemblies from NCGR. The number on top in blue shows the numbers of assemblies where DIB has a higher value than NCGR or in gray where NCGR has a higher number.

The proportion of references with a conditional reciprocal best (CRB) BLAST hit was different when DIB assemblies were used as a reference compared to when NCGR assemblies were used as a reference (Figure 3). The mean proportion for DIB re-assemblies was 0.70 ± 0.28 while the mean proportion for NCBR assemblies was 0.49 ± 0.09, p < 2.2e-16 (D=0.7616). The difference between these indicates there is a ~30% difference in content between assemblies performed by NCGR with the ABySS assembler and the Trinity assemblies performed with this pipeline.

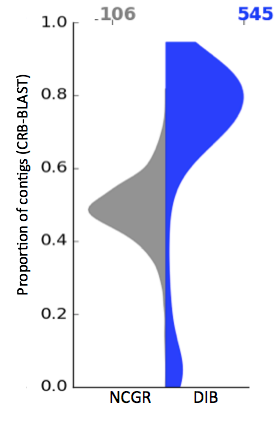


Figure 3. Split violin plot showing the proportion of references with a conditional reciprocal best (CRB) BLAST hit. The distribution plot on the left (grey) contains data were NCGR assemblies were aligned against the DIB assemblies as reference and on the left (blue), the DIB assemblies were aligned against the NCGR assembly as reference.

Even though DIB re-assemblies had more contigs, the open reading frame (ORF) content is similar to the original assemblies, with a mean of 81.9% ± 9.68 of dib assemblies with ORF content and 76.79% ± 10.2 in the NCGR assemblies. ORF content in DIB assemblies were higher than NCGR assemblies, 95% of the samples (605 out of 635 non-redundant samples) (Figure 4-left). Complete Benchmarking Universal Single Copy Ortholog (BUSCO) percentages in the DIB re-assemblies (61.4% ± 0.20) were not significantly different compared to the original NCGR assemblies (60% ± 0.19) (p = 0.2096, D=0.058348) (Figure 4-right).

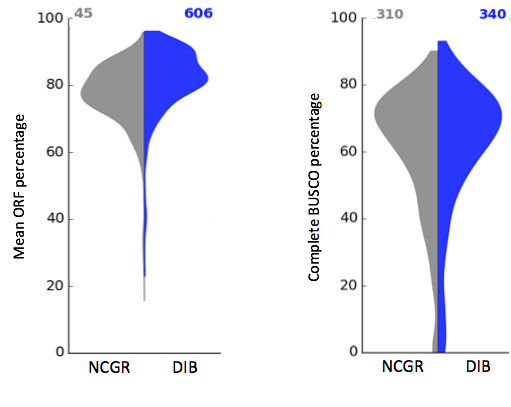


Figure 4. Split violin plots depicting the percentage of contigs with predicted open reading frame (ORF) in each assembly (left) and the percentage of complete benchmarking universal single-copy orthologs (BUSCO) recovered in each assembly(right). In the blue (right side of each plot) are the “DIB” re-assemblies and in gray (left side of each plot) are the original assemblies from NCGR.

Focusing on the contigs in each sample thatt were absent in NCGR but present in DIB (extra content seen in Figure 2), we use the dammit annotations (Scott 2016) to assess the percentage of thos that could be annotated with Pfam, Rfam, or OrthoDB databases (Figure 5). These annotations of contigs absent from NCGR but present in DIB have not been compared to the annotated gene names of contigs present in NCGR. Therefore, these are not necessarily unique annotations and could be allelic variants or alternatively spliced isoforms.

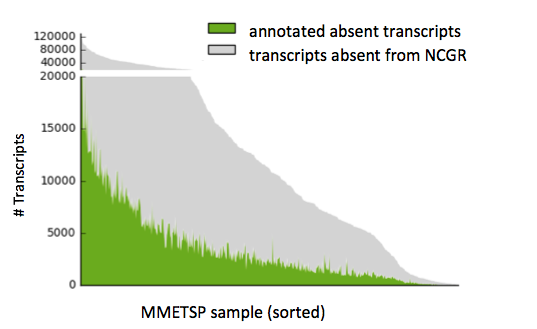


Figure 5. A histogram across the 678 samples depicting the number of contigs identified as present in DIB assembly but absent in the NCGR assembly. Highlighted in green are the contigs that could be annotated with a known gene name from the Pfam, Rfam, or OrthoDB databases.

To assess the unique nucleic acid sequence content within each of the assemblies, we quantified the total number of unique the *k*-mer. Comparing the raw sequence content regardless of annotation in terms of the number of unique *k*-mers (*k* = 25), X% of the samples fall above the 1:1 expectation indicating more unique *k*-mers in the DIB re-assemblies compared to NCGR assemblies (Figure 6).

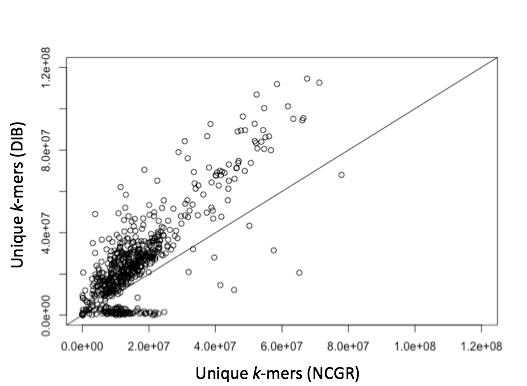


Figure 6. Unique *k*-mers (*k*=25) from the DIB re-assembly compared NCGR assembly for each of the 678 samples. The line indicates a 1:1 relationship between DIB and NCGR assemblies.

To examine phylogenetic differences in the assemblies, assembly metrics were averaged by taxa (Figure 7). The Dinoflagellates (Dinophyta) had more unique kmer content than other groups (can we do an ANOVA test to test variance?). Assemblies from Ciliates (Ciliophora) had lower open reading frame percentages. Assemblies from Dinoflagellates had higher ORF percentages and more contigs.

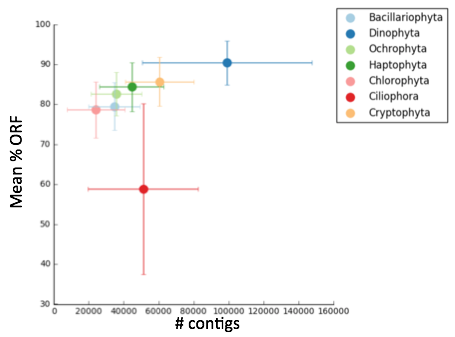
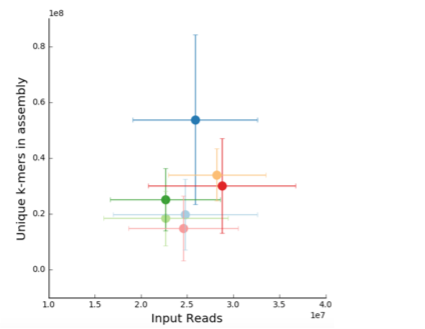


Figure 7. Averaged assembly metrics grouped by phylum for the the seven most common phyla in the MMETSP dataset, ranked according to number of samples. Taxonomic grouping are as follows, Bacillariophyta (light blue, N=193), Dinophyta (dark blue, N=128), Ochrophyta (light green, N=78), Haptophyta (dark green, N=63), Chlorophyta (pink, N=62), Ciliophora (red, N=31), Cryptophyta (orange, N=22). The mean for each phylum is plotted with bars indicating standard deviation across the phylum. Unique *k*-mers (*k*=25) in the assembly is plotted against vs. input reads (A left) and the mean percentage open reading frame (ORF) is plotted against vs. the number of contigs in the assembly (right).

**Discussion**

*Add some sort of lead in here reiterating why this is important.*

In Assemblathon2, Bradnam *et al*. (2013) compared different genome assembly pipelines with data from three vertebrate species. This study is the first to examine hundreds of transcriptome assemblies from different species.

*And then transition to what we did. Thinking about the overall structure we want to talk about:*

1. *Creating a modularized RNAseq assembly tool (granted-- this is largely covered in the oped*
2. *Why our assemblies >> theirs (or are they? Proof in 3 parts…)*
3. *Phylogenetic differences (aka cool biology that falls out of this work)*

*I would suggest trying to remind the reader of those three main points in the first paragraph and then maybe even break the discussion into subsections to make the content easier to break apart and take in. Suggested breaks are as follows:*

*\_\_\_\_\_\_\_\_\_\_\_\_\_*

*Modularized RNAseq assembly tool rationale (or some other phrase)*

The thought behind this project was two fold. First… The purpose of developing new algorithms and releasing new software or updates to existing programs are to improve sensitivity and specificity of the assemblies. Second… A benefit of a modularized pipeline includes the ease of running individual steps again if processes break or when updated software versions and/or additional samples become available. The pipeline used to generate the DIB assemblies was, by design, different than the pipeline used for the original assemblies. The DIB pipeline used Trinity (Grabherr et al. 2011) with adapter and light quality (Q<2) trimming in combination with digital normalization (Crusoe et al. 2015, Zhang et al. 2014, Brown et al. 2012). Digital normalization was used in the pipeline for this study to reduce the computational cost of assembly without negatively affecting the quality of the assembly. By contrast, the NCGR pipeline consisted of the Trans-ABySS assembler (Robertson et al 2010), contig scaffolding, and other processing steps (Keeling et al. 2014). As the two pipelines were so different, We did not expect the results to be similar between pipelines.

A benefit of having a pipeline documented as an available, open-source protocol is a set of standardized and documented methods for data analysis that anyone can follow and (in theory) arrive at similar results. Benefits of programmatically automating scripts include quick changes and fast execution of hundreds of samples. Examples of protocols and tutorials include the eel pond protocol (Brown et al. 2015), which was used as the basis for the automated protocol developed for this study, and the Oyster River Protocol (MacManes, 2015).

*Transcriptome assembly comparison*

*Here I would do an intro with some background on Transrate and why we chose this (basically paragraph 2). How we are using a combination of transcriptome quality metrics and more biological relevant metrics.*

*We looked at a variety of transcriptome quality* metrics, ranging from transcriptome metrics (not sure what to call it really) to more biologically relevant content-based metrics (e.g. BUSCO) to compare the DIB and NCGR assemblies . O'Neil and Emrich (2013) ran a meta-analysis of transcriptome evaluation metrics and determined that a combination of length and content-based metrics gives the most accurate assessment of assembly quality rather than length-based metrics alone such as N50, as transcripts, unlike genomic contigs, have natural length variation. Content-based metrics examined for this study included *k-*mer content, CRBB, BUSCO, ORF, and annotated gene information obtained from the dammit pipeline.

The two main software package tools available for evaluating transcriptome assemblies are TransRate and RSEM-EVAL (Li et al. 2014). RSEM-EVAL provides measurements for how compact the assembly is and the likelihood of support of the assembly from the RNA-Seq data. In the TransRate score calculation, each mapped read is fractionally assigned to each potential contig of origin using Salmon (Patro et al., 2015) in a process that is analogous to the proportional assignment of the EM procedure used in RSEM (Li et al., 2014). Therefore, it was not necessary to look at both RSEM-EVAL and the TransRate score metrics for this study.

With over 600 samples in the MMETSP dataset taken across a wide diversity of organisms, we had the statistical rigor to be able to compare the two assembly pipelines. [insert text here discussing the results of all the comparison figures] There were high quality and also low quality assemblies generated for this study. Some assemblies had as low as 3 contigs. [provide more discussion of quality etc. I think it is important to hammer home the idea this pipeline is better on many fronts.]

Thus, the evaluation metrics described here can serve as a framework for better contextualizing the quality of an externally generated protistan transcriptome. If an investigator generates a new transcriptome assembly, it may be useful to compare those evaluation metrics with the set of metrics collected from this study to see where it may fall along the distributions. While BUSCO scores from the core genes in the eukaryotic data set (Figure 4) or the protist data set (Supplemental) might not be the best absolute accurate measurements of the quality of a single transcriptome or genome, in combination with other metrics, and given BUSCO scores from transcriptomes from a range of species examined in this study, it may be useful to compare. For example, “transcriptome scores were all in the upper 80% percentile of all metrics, compared to other assemblies”.

*Phylogentic trends in transcriptome assembly metrics*

Assembly evaluation tools might yield results outside the range of what is normal for some organisms, in the case of low ORF predictions in Ciliates (Figure 7). It has recently been found that ciliates have an alternative triplet codon dictionary, with codons normally encoding STOP serving a different purpose (Alkalaeva and Mikhailova 2016, Heaphy et al. 2016, Swart et al. 2016). High unique *k*-mers and numbers of contigs in Dinoflagellates (Figure 7) make sense in in the context of evidence suggesting Dinoflagellates constitutively express genes, then regulate at the translational level (Aranda et al. 2016, Lin et al. 2011, Hou and Lin, 2009). It may be useful to incorporate strain-specific information like this into assembly software. The assemblies are not necessarily lower quality, but may be perceived as lower in quality because of unique features.

*Conclusion*

While contigs are linear predictions of full transcripts made by the assembly software, “final” assemblies are probabilistic approximations of graph structures traversed to generate a static set of transcripts that may change with updated software or changes in the pipeline. We realize that biologists generating data for new *de novo* transcriptomes care mostly about the biological questions that can be answered with these data, and often do not care about software details or the time needed to reassemble/re-analyze data. Yet, this study has demonstrated that there is value in re-analyzing data is that results could be different or better the second time around. Based on the comparisons of output from the DIB and NCGR pipelines (Figures 2,3,4) and between versions of Trinity (Supplement), we found that it may not be useful to spend time comparing transcriptomes generated by different software or versions to find the “optimal” pipeline or assembly software to use. However, it is worth spending time to evaluate an assembly to make sure the content is close to expected. While evaluation metrics can serve as a general diagnostic, domain-specific knowledge of the samples prior to assembly are important when evaluating whether a transcriptome is good quality or “finished”. Given a new set of data for a *de novo* transcriptome assembly, it is worth spending time to set up an assembly and analysis pipeline so that it can be run again if significant improvements to software are made. By using protocols, or the design from this pipeline, investigators have the capability of doing that.

**Acknowledgements**

Gordon and Betty Moore Foundation under award number GBMF4551. Scripts were run on high performance computing systems at iCER, MSU HPCC and NSF-XSEDE Jetstream with allocation TG-BIO160028.

**References**

Alkalaeva E and Mikhailova T. 2016. Reassigning stop codons via translation termination: How a few eukaryotes broke the dogma. DOI: 10.1002/bies.201600213

Aranda et al. 2016 [PMID: 28004835](http://www.nature.com/articles/srep39734)

Aubry S, Steven Kelly , Britta M. C. Kümpers, Richard D. Smith-Unna, Julian M. Hibberd. 2014. Deep Evolutionary Comparison of Gene Expression Identifies Parallel Recruitment of Trans-Factors in Two Independent Origins of C4 Photosynthesis. PLOS Genetics 12(5): e1006087. doi: 10.1371/journal.pgen.1006087

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Bradnam KR et al. 2013. Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species. *GigaScience*. DOI: 10.1186/2047-217X-2-10

Brown, C. Titus; Scott, Camille; Crusoe, Michael; Sheneman, Leigh;  
Rosenthal, Josh; Adina Howe (2013): khmer-protocols documentation.  
figshare. http://dx.doi.org/10.6084/m9.figshare.878460

The Eel Pond mRNAseq Protocol. <https://khmer-protocols.readthedocs.io/en/ctb/mrnaseq/>. Updated, 2016: <http://eel-pond.readthedocs.io/en/latest/>

Brown CT and Irber L. 2016. Sourmash: a library for MinHash sketching of DNA. The Journal of Open Source Software. <http://dx.doi.org/10.21105/joss.00027>

Brown CT, Howe A, Zhang Q, Pyrkosz AB, Brom TH. 2012. A reference-free algorithm for computational normalization of shotgun sequencing data. arXiv: 1203.4802.

Caron DA, Harriet Alexander, Andrew E. Allen, John M. Archibald, E. Virginia Armbrust, Charles Bachy, Callum J. Bell, Arvind Bharti, Sonya T. Dyhrman, Stephanie M. Guida, Karla B. Heidelberg, Jonathan Z. Kaye, Julia Metzner, Sarah R. Smith & Alexandra Z. Worden. 2017. Probing the evolution, ecology and physiology of marine protists using transcriptomics. Nature Reviews Microbiology 15, 6–20 (2017) doi:10.1038/nrmicro.2016.160

Chang Z. Wang Z, Li G. 2014. The impacts of read length and transcriptome complexity for de novo assembly: a simulation study. PLoS One. 2014; 9(4): e94825. doi: [10.1371/journal.pone.0094825](https://dx.doi.org/10.1371%2Fjournal.pone.0094825)

Cohen, Lisa; Alexander, Harriet; Brown, C. Titus (2017): Marine Microbial Eukaryotic Transcriptome Sequencing Project, re-assemblies. figshare.

<https://doi.org/10.6084/m9.figshare.3840153.v6>

Conesa A, Madrigal P. Tarazona S., Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak MW, Gaffney DJ, Elo LL, Zhang X, Mortazavi A. 2016. A survey of best practices for RNA-seq data analysis. Genome Biology. 17:13. DOI 10.1186/s13059-016-0881-8

Crusoe MR, Alameldin HF, Awad S *et al.* 2015. The khmer software package: enabling efficient nucleotide sequence analysis. *F1000 Research*. 4:900 doi: [10.12688/f1000research.6924.1](http://dx.doi.org/10.12688/f1000research.6924.1)

Finn RD, 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 (2016) Database Issue 44:D279-D285

Gardner PP, J. Daub, J.G. Tate, E.P. Nawrocki, D.L. Kolbe, S. Lindgreen, A.C. Wilkinson, R.D. Finn, S. Griffiths-Jones, S.R. Eddy and A. Bateman. 2008. Rfam: updates to the RNA families database. Nucleic Acids Research. doi: 10.1093/nar/gkn766

Gordon Robertson1, Jacqueline Schein1, Readman Chiu1, Richard Corbett1, Matthew Field1, Shaun D Jackman1, Karen Mungall1, Sam Lee2, Hisanaga Mark Okada1, Jenny Q Qian1, Malachi Griffith1, Anthony Raymond1, Nina Thiessen1, Timothee Cezard1,4, Yaron S Butterfield1, Richard Newsome1, Simon K Chan1, Rong She1, Richard Varhol1, Baljit Kamoh1, Anna-Liisa Prabhu1, Angela Tam1, YongJun Zhao1, Richard A Moore1, Martin Hirst1, Marco A Marra1,3, Steven J M Jones1,3, Pamela A Hoodless2,3 & Inanc Birol. 2010. *De novo* assembly and analysis of RNA-seq data. Nature Methods. doi:10.1038/nmeth.1517

Heaphy SM, Mariotti M, Gladyshev VN, Atkins JF, Baranov PV. 2016. Novel Ciliate Genetic Code Variants Including the Reassignment of All Three Stop Codons to Sense Codons in Condylostoma magnum. Mol Biol Evol (2016) 33 (11): 2885-2889. doi: 10.1093/molbev/msw166

Hou and Lin 2009. [PMID: 27426948](http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0006978)

Irber LC and Brown CT. 2016. Efficient cardinality estimation for *k*-mers in large DNA sequencing data sets. doi: <https://doi.org/10.1101/056846>

Kannan S, Hui J, Mazooji K, Pachter L, Tse D. 2016. Shannon: An information-optimal de novo RNA-seq assembler. doi: <https://doi.org/10.1101/039230>

Leinonen R, Sugawara H, Shumway M. 2011. The Sequence Read Archive. [Nucleic Acids Res](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3013647/). 2011 Jan; 39 (Database issue): D19–D21. doi: [10.1093/nar/gkq1019](https://dx.doi.org/10.1093%2Fnar%2Fgkq1019)

Li B, Fillmore N, Bai Y, Collins M, Thomson JA, Stewart R, Dewey CN. 2014. Evaluation of *de novo* transcriptome assemblies from RNA-Seq data. Genome Biology. DOI 10.1186/s13059-014-0553-5

Lin 2011 [PMID: 21514379](https://linkinghub.elsevier.com/retrieve/pii/S0923-2508(11)00068-4)

Lowe EK, Billie J Swalla, C Titus Brown. 2014. Evaluating a lightweight transcriptome assembly pipeline on two closely related ascidian species. PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.505v1 | CC-BY 4.0 Open Access | rec: 18 Sep 2014, publ: 18 Sep 2014

MacManes MD. 2014. On the optimal trimming of high-throughput mRNA sequence data. <https://doi.org/10.3389/fgene.2014.00013>

MacManes MD. 2016. An opinionated guide to the proper care and feeding of your transcriptome. doi: http://biorxiv.org/content/early/2016/02/18/035642

Malik LI, Shravya Thatipally, Nikhil Junneti, Rob Patro. 2016 Graph regularized, semi-supervised (GRASS) learning improves annotation of de novo transcriptomes. doi: https://doi.org/10.1101/089417

Miriam L Land, Doug Hyatt, Se-Ran Jun, Guruprasad H Kora, Loren J Hauser, Oksana Lukjancenko and David W Ussery. 2014. Quality scores for 32,000 genomes. Standards in Genomic Sciences. DOI: 10.1186/1944-3277-9-20

Miriam L Land, Doug Hyatt, Se-Ran Jun, Guruprasad H Kora, Loren J Hauser, Oksana Lukjancenko and David W Ussery. 2014. Quality scores for 32,000 genomes. Standards in Genomic Sciences. 20149:20 DOI: 10.1186/1944-3277-9-20

O’Neil S and Emrich S. 2013. Assessing *de novo* transcriptome assembly metrics for consistency and utility. BMC Genomics. 14(1): 465. DOI: 10.1186/1471-2164-14-465

Patro R, Duggal G, Love M, Irizarry RA, Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods. DOI:10.1038/NMETH.4197

Peng Y, Leun HCM, Yiu, SM, Lv MJ, Zhu XG, Chin FYL. 2013. IDGA-tran: a more robust de novo de Bruijn graph assembler for transcriptomes with uneven expression levels. Bioinformatics (2013) 29 (13): i326-i334. <https://doi.org/10.1093/bioinformatics/btt219>

Schulz MH, Zerbino DR, Vingron M, Birney E. 2012. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics. 2012 Apr 15; 28(8): 1086–1092. doi: [10.1093/bioinformatics/bts094](https://dx.doi.org/10.1093%2Fbioinformatics%2Fbts094)

Scott, C. 2016. In prep. dammit: <https://github.com/camillescott/dammit/blob/master/CITATION>

Documentation: <http://www.camillescott.org/dammit/>

Smith-Unna R, Boursnell C, Patro R, Hibberd JM, Kelly S. 2016. TransRate: reference free quality assessment of de-novo transcriptome assemblies. Genome Research. DOI: 10.1101/gr.196469.115

Swart EC, Valentina Serra, Giulio Petroni2, Mariusz Nowacki. 2016. Genetic Codes with No Dedicated Stop Codon: Context-Dependent Translation Termination. Cell. Volume 166, Issue 3, 28 July 2016, Pages 691–702

Yinlong Xie Gengxiong Wu Jingbo Tang Ruibang Luo Jordan Patterson Shanlin Liu Weihua Huang Guangzhu He Shengchang Gu Shengkang Li Xin Zhou Tak-Wah Lam Yingrui Li Xun Xu Gane Ka-Shu Wong Jun Wang. 2014. SOAPdenovo-Trans: *de novo* transcriptome assembly with short RNA-Seq reads. Bioinformatics. 30 (12): 1660-1666. Doi: <https://doi.org/10.1093/bioinformatics/btu077>

Zdobnov EM et al. 2016. "OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs." [PMID:27899580](http://www.ncbi.nlm.nih.gov/pubmed?cmd=Search&term=27899580)

Zhang Q, Pell J, Canino-Koning R, Howe AC, Brown CT. 2014. These Are Not the K-mers You Are Looking For: Efficient Online K-mer Counting Using a Probabilistic Data Structure. PLOS One. 9(7): e101271.

**Supplemental Files**

Supplement 1 --- Matrix of all evaluation metrics: <https://github.com/ljcohen/MMETSP/blob/master/assembly_evaluation_data/MMETSP_all_evaluation_matrix.csv>

Supplement 2 --- Figures

S1. Comparison between Trinity versions (2 split-violin plots: 2014 vs. 2.2.0 and 2.2.0 vs. 2.3.2)

<https://github.com/ljcohen/MMETSP/blob/master/assembly_evaluation_data/transrate_reverse_trinity2014_v_trinity2.2.0.csv>

<https://github.com/ljcohen/MMETSP/blob/master/assembly_evaluation_data/transrate_reference_trinity2.2.0_v_trinity2014.csv>

<https://github.com/ljcohen/MMETSP/blob/master/assembly_evaluation_data/transrate_reference_trinity2.2.0_v_trinity2.3.2.csv>

(missing trinity2.3.2 vs. trinity2.2.0)

S2. Comparison of protist BUSCO scores, NCGR vs. DIB (need to generate this figure with these tables: <https://github.com/ljcohen/MMETSP/blob/master/assembly_evaluation_data/busco_scores_MMETSP_protist_trinity2.2.0.csv>

<https://github.com/ljcohen/MMETSP/blob/master/assembly_evaluation_data/busco_scores_imicrobe_protist.csv>

S3. All MMETSP sourmash assemblies

S4. All MMETSP reads vs. assembly