Improving reference genomes and transcriptomes towards gene expression profiling of sequencing data sets containing multiple eukaryotic species

PI: N. Tessa Pierce
Co-PI: Lisa Johnson
UC Davis Data Intensive Biology Lab
Startup Allocations: TG-MCB170160 and TG-BIO160028

Background and Support

High-throughput sequencing methods have revolutionized ecological and evolutionary studies by facilitating research at the whole-genome level for multiple species. However, generation of high-quality references remains a complex and computationally-intensive endeavor, complicated by the biological properties of the sequence (e.g. polymorphism, gene splicing and duplication, etc), technical limitations of the sequencing platforms and analysis tools, and the scale (often hundreds of samples) required for balanced experimental designs. Furthermore, we have shown that the goal to generate a single reference genome or transcriptome as a final product must be re-evaluated in light of the information gained via reanalysis with new tools (Johnson et al. 2018).

Our research aims to address these issues by systematically evaluating transcriptome and genome assembly, annotation and analysis methods to produce best practices, user-friendly, automated pipelines that facilitate future research. Here, we leverage biological datasets at two scales: first a broad-scale analysis of 678 species in the Marine Microbial Eukaryotic Transcriptome Sequencing Project (MMETSP), and second, finer-scale analysis on the evolution of salinity tolerance in sixteen species of a relatively well-studied species group, *Fundulus* killifish, exposed to acute hypersalinity challenge. The phylogenetic breadth of these projects enables us to build better bioinformatic pipelines agnostic of species that can be applied by investigators to sequencing data from multiple species as well as repeatedly to our own data as software tools and annotation databases improve.

These projects are supported through November 2019 by a Moore Foundation Data-Driven Science investigator award to the PI of our lab, Dr. C. Titus Brown. PI Dr. N. Tessa Pierce is funded by an NSF Postdoctoral Research Fellowship in Biology (#1711984; 2017-2020), entitled "Improving RNA-Seq Analysis through Graph-based Analysis and Computational Indexing," which uses the MMETSP data to test RNAseq analyses. Co-PI Lisa Johnson (Cohen) received an NSF Graduate Research Fellowship Honorable Mention in 2016 and has authored the first publication for this project leveraging the MMETSP dataset (Johnson et al. 2018). Results from these analyses are essential preliminary data for funding proposals to be submitted in the coming year.

Research Objectives

Project 1: RNA-Seq tool development using the MMETSP dataset

The MMETSP consists of RNAseq data from 678 cultured protist species spanning more than 40 eukaryotic phyla, generated to broaden the diversity of sequenced marine protists and facilitate our understanding of their evolution and roles in marine ecosystems and biogeochemical cycles (Keeling et al. 2014). This database forms a rich reference which can be queried to help functionally characterize genes from environmental samples containing species that cannot be cultured. However, the MMETSP data also provides a unique opportunity to assess RNAseq analysis methods using one of the largest and most diverse publicly-available RNAseq data sets at this time. MMETSP is also unusually well-suited for benchmarking analyses, as data submitted by a large consortium of international investigators was prepared with a single, standardized library preparation and sequenced at a single facility.

MMETSP was first analyzed in 2013, with *de novo* reference transcriptomes generated by the National Center for Genomic Resources (NCGR) using the Trans-ABySS assembler (Robertson et al. 2010). However, transcriptome assembly is an area of active research, with many new tools and improvements since this initial analysis. The first part of this project (now published) demonstrated that application of a new assembly software tool, Trinity (Grabherr et al. 2011) improved the quality of the assemblies and uncovered new content (Johnson et al. 2018). Comparing the original Trans-ABySS "NCGR" pipeline and our Trinity-based "DIB" assembly pipeline, we see (Figure 1) differences in the number of contigs (left), the assembly content measured by alignments (middle) and unique *k*-mers (*k*=25) (right). These assemblies are annotated with the dammit annotation program, developed by Camille Scott who is a graduate student in our lab.

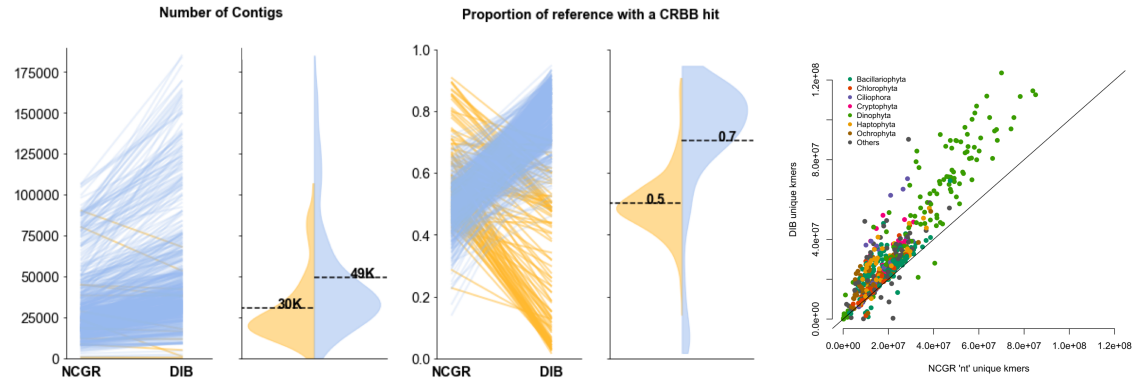

Figure 1. (left) Number of contigs compared between NCGR (yellow) and DIB (blue) *de novo* transcriptome assemblies. (middle) Proportion of contig sequence similarity as measured by a conditional reciprocal best BLAST (CRBB) algorithm between NCGR (yellow) and DIB (blue) assemblies, indicating that the two assemblies have different content. (right) Comparison of the unique *k*-mer (*k*=25) content between NCGR assemblies (x-axis) and DIB assemblies (y-axis) indicating that the DIB assemblies had more unique content.

The MMETSP data is publicly available from the NCBI Sequence Read Archive (PRJNA231566) from 678 species, 168,152,967,701 PE50 reads total, 1TB raw data storage with the following phyla (N=species): Bacillariophyta (N=173), Dinophyta (N=114), Ochrophyta (N=73), Haptophyta (N=61), Ciliophora (N=25), Chlorophyta (N=62), Cryptophyta (N=22), Others (N=13).

Improving Reference Generation with *eelpond*

Given these improvements to old data using new tools, in the second part of this project we are developing an automated best-practices pipeline, called eelpond (https://github.com/dib-lab/eelpond), for non-model species within a framework that facilitates software installation and new tool integration. Eelpond accepts a set of raw fastq RNA sequencing reads and an experimental design matrix, and generates annotated *de novo* transcriptome assemblies. It will also run preliminary differential expression analysis with the DESeq2 R package. A new multi-assembler protocol, the Oyster River Protocol (ORP; MacManes 2018) suggests that merging contigs from several assemblers (Spades; Bushmanova et. al 2018, Shannon, Trinity; Grabherr et. al 2011) and multiple *k*-mer lengths may outperform assemblies with the currently-favored assembler, Trinity. We are integrating each of these assemblers (including the full ORP pipeline) into eelpond for benchmarking and comparison with the MMETSP data.

Gene Pathway and Structure Discovery (*glympsed*) & Phylotranscriptomics

The final goal of our MMETSP project is to develop an unsupervised deep learning pipeline, called glympsed, using the "denoising autoencoder" algorithm (Vincent et al 2008) to discover common expression patterns in large RNA sequencing data sets (>100 samples and 30,000 genes/sample). This

method has been shown to successfully identify common metabolic patterns of gene expression where there are unknown relationships between samples (Tan et al. 2016). Code for the glympsed project was written as a collaborative group in 2016 using a a large test set from all publicly-available *Pseudomonas* bacteria data, with the aim of testing and optimizing the glympsed pipeline with MMETSP data. Now that the MMETSP re-assemblies and expression quantification results are complete (Johnson et al. 2018), we can use these results as input to run the glympsed pipeline.

To start, we will examine data from members of the dinoflagellata phylum in the MMETSP set, which exhibit two distinct lifestyles (endosymbiotic and free-living). With glympsed, we will investigate fundamental metabolic transcriptional differences between the two groups. We will then expand the this approach to additional taxonomic groups within the MMETSP. Following the development and successful analysis with the glympsed pipeline, the agalma phylotranscriptomics pipeline (Dunn et al. 2013) will be used to generate species trees based on pathway genes identified with glympsed.

Project 2: Comparative genomics and gene expression analysis of 16 Fundulus killifish species in response to osmotic stress

Killifish in the genus *Fundulus* have emerged as a comparative model system for studying the physiological and genetic mechanisms underlying osmotic tolerance (Whitehead 2010). Some *Fundulus* species can tolerate a range of environmental salinities (euryhaline) by switching osmoregulatory mechanisms while other *Fundulus* species require a narrower salinity range (stenohaline) in either fresh or marine waters. In collaboration with Andrew Whitehead's Environmental Genomics lab at UC Davis, we are studying the evolutionary history of osmotic adaptation using 16 estuarine and freshwater *Fundulus* species with representation from each of the three clades where species have independently radiated into freshwater environments. As these species are closely related and have a well-characterized phylogeny, this study presents a unique opportunity to study parallel evolution of adaptation of osmotic tolerance.

This project leverages both stress-response RNA-seq data and long-read genome data. RNAseq data were collected after exposure of 130 individuals to a common, controlled acute hypersalinity challenge experiment, and deeply sequenced on the Illumina sequencing platform, generating over 4 billion PE100 reads, 500GB raw data storage: *Adenia xenica* (N=9), *Fundulus catenatus* (N=7), *Fundulus chrysotus* (N=8), *Fundulus diaphanus* (N=7), *Fundulus grandis* (N=9), *Fundulus heteroclitus* (MDPL population) (N=9), *Fundulus heteroclitus* (MDPP population) (N=9), *Fundulus notatus* (N=9), *Fundulus nottii* (N=2), *Fundulus olivaceus* (N=8), *Fundulus parvapinis* (N=8), *Fundulus rathbuni* (N=9), *Fundulus sciadicus* (N=5), *Fundulus similis* (N=9), *Fundulus zebrinus* (N=4), *Lucania goodei* (N=9), *Lucania parva* (N=9) (Rodgers et al. 2018, Johnson and Whitehead 2018). After the experiment, a subset of four species (3 freshwater: *Fundulus nottii*, *Fundulus catenatus*, *Fundulus olivaceus*, and one marine: *Adinia xenica*) were selected for genome assembly via Oxford Nanopore Technologies (ONT) long-read sequencing.

Transcriptome Analysis with *eelpond*

Following the success of transcriptome assemblies of the MMETSP dataset, we are applying the same methods to assemble *de novo* transcriptomes for each of these 16 *Fundulus* species, which we can then use to analyze transcriptional responses to salinity change by clade and physiology. The end product for this project will be sets of candidate genes with common expression patterns and their corresponding physiological pathways. Based on previous work with microarrays and protein assays, we expect to see isoform switching in relevant genes (e.g. ATPase Na+/K+ transporting subunit alpha 1b vs. alpha 1a). These differences are hard to detect and require high quality reference transcriptomes. Improving the quality of the assemblies has included assembling and re-assembling with the new version of Trinity

(2.8.4), which improved the contiguity with fewer contigs (Figure 2, left) and benchmarking evaluation scores (BUSCO; Simão et. al 2015) (Figure 2, right). *De novo* transcriptome assemblies have been generated, twice with updated versions of Trinity. Now, they need to be improved with ORP and automated using eelpond.

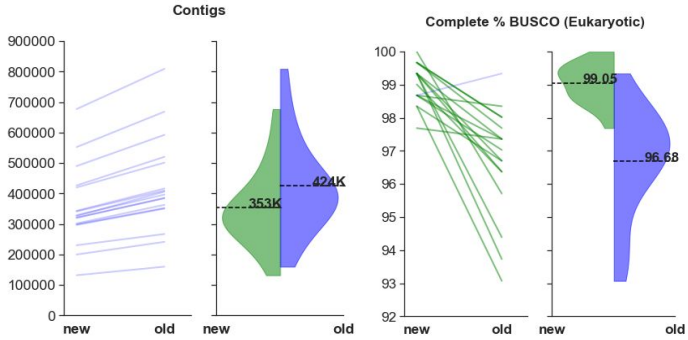

Figure 2. (left) Numbers of contigs decreased from old assemblies (purple) to new assemblies (green). (right) Benchmarking scores (BUSCO) increased with a tighter distribution in new assemblies (green) compared to the old assemblies (purple).

Generating High-Quality Genome References

Genomic sequence data from four species were generated (3 freshwater: *Fundulus nottii*, *Fundulus catenatus*, *Fundulus olivaceus*, and one marine: *Adinia xenica*) via Oxford Nanopore Technologies (ONT) PromethION desktop sequencer (approximately $3,500/species). Each genome is estimated to be around 1.1 Gb, based on the size of the sister species, *Fundulus heteroclitus*; therefore, we have collected between 30-50x coverage of ONT data. These genomes will better enable us to detect these small changes in the transcriptomes, as well as to discover changes in genomic architecture underlying adaptive responses to environmental change. Below is a table listing ONT data collected for each species:

ONT data collected thus far, which will contribute to *de novo* genomes of 4 *Fundulus* species:

Species	ONT long reads	N reads	Average len (bases)	reads N50
Adinia xenica	38.5 Gb	15,704,522	2,440	5,633; n=1,373,426
Fundulus nottii	33.4 Gb	5,160,367	6,480	12,995; n=700,534
Fundulus catenatus	40.3 Gb	23,701,206	1,699	3,439; n=2,687,295
Fundulus olivaceus	55 Gb	10,902,817+740,248	4,595	11,670; n=987,921

ONT is a new sequencing technology only available within the past 3 years, however it is showing promise improving genome assemblies because of the length of the reads (Ebbert et al. 2018). Because of the high error rate (~10-15%) associated with ONT data, we are in the process of generating 50x coverage of paired-end 150 bp short-read Illumina data, which has a low error rate (<1%) and will be used to correct the noisy long reads. The size of the Illumina short read data will be around 500 GB total.

Gene Pathway and Structure Discovery (*glympsed*) & Phylotranscriptomics

The killifish data also have more than 100 samples and will be run through the glympsed pipeline. We are looking for common patterns of gene expression across the clades, hypothesizing that there will be some genes with parallel avenues of evolution whereas some genes will have divergent modes of evolution.

B. Computational Methods

Transcriptome assemblies

We are working towards improving the quality of reference transcriptomes for these two projects: 16 *Fundulus* species + 678 MMETSP species = 694 species, around 700 transcriptomes. RNAseq samples from *Fundulus* species have approximately 10x more reads (average = 238,363,948) than samples from the MMETSP (average = 23,387,060 reads).

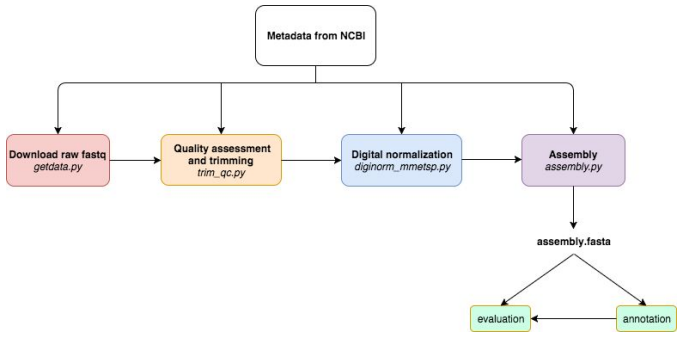

Figure 3. Workflow diagram for the eelpond RNAseq *de novo* transcriptome assembly pipeline, indicating milestones for each step: Download, quality assessment, digital normalization, assembly, annotation, and evaluation.

Eelpond To improve each MMETSP and *Fundulus* killifish transcriptome, we will run our automated eelpond pipeline (Figure 3; https://github.com/dib-lab/eelpond) to conduct read quality and error trimming, generate additional transcriptome assemblies, and perform assembly annotation and quality evaluation. Eelpond uses Python's snakemake automation tool to manage each run and uses conda for tool installation. While some of the components of eelpond are not compute intensive, the main assembler, Trinity, is *very* compute intensive (see "Code Performance and Resource Usage" document). To run this pipeline in parallel to process files from from start to finish, all tools must be run on the HPC. Since the pipeline is modularized, individual milestones can be adjusted (separate jobs submitted requesting different resources) so the whole pipeline does not require the same large amount of resources as Trinity. We are in the process of integrating the **Oyster River Protocol**, a multi-kmer, multi-assembler tool that may outperform Trinity assemblies (MacManes 2018). Since we already have Trinity assemblies, testing ORP will involve running the remaining assemblers (Spades x2 and Shannon) as well as the assembly merging script, Orthofuser. The resulting assemblies will also require annotations, which eelpond will perform with dammit (Scott *in prep*). Downstream analyses will be conducted with Salmon (Patro et al 2017) quantification and DESeq2 (Love et al. 2014), already integrated into the pipeline. While eelpond currently works on AWS and Jetstream, only minor modifications are needed to scale up to the Bridges HPC. Resource allocations requested by this proposal will allow us to develop the eelpond protocol for any hpc users with slurm job management, such as PSC Bridges.

Genome Assemblies

Four *Fundulus* killifish genomes will be assembled from Oxford Nanopore Technologies (ONT) long read data. New tools and methods for assembling and analyzing these noisy ONT long reads are being developed at a fast rate (de Lannoy et al. 2017). We will perform hybrid genome assemblies with ONT and Illumina data, which can produce high quality (>2 Mb contig N50) reference genomes (Tan *et al* 2018, Miller *et al*. 2018, Liu et al. 2018). We have found the best available assembler to be MaSuRCA, as evaluated by complete BUSCO identification (eukaryota database; Simão et. al 2015):

Seq data	Tool	bases	N contigs	Avg length	largest	N50	BUSCO
ONT	Canu	9,804,264	540	18,156	365,191	40,681; n = 43	0.7%
ONT	Miniasm	4,917,546	153	32,140	233,136	50,056; n = 25	0.0 %
Illumina	Megahit	1,183,861,293	1,038,799	1,139	88,218	3,846; n = 77,800	45.6 %
Illumina	ABySS	1,381,148,284	1,024,759	1,347	140,629	9,833; n = 37,013	77.9%
Hybrid	MaSuRCA	1,134,160,060	90,237	12,568	386,222	42,823, n = 7,616	86.2%

The job to run the **MaSuRCA** software tool required 1TB storage for intermediate files and >1 week to run for the 1.1 Gb genome assembly. Our measurements were validated by the manual and communications with the MaSuRCA authors. Following genome assembly, the intermediate files are deleted. We have successfully produced one MaSuRCA assembly on the MSU hpcc with an abbreviated set of ONT data. However, we were not able to benchmark this on PSC Bridges before this proposal, due to time and resource constraints on our startup allocations. After assembly with MaSuRCA, the quality of the reference genome assemblies will be improved by applying the **Pilon or Racon** scaffolding and consensus improvement tools, which are also resource intensive.

These four genomes will be annotated using the **MAKER** pipeline (Cantarel 2008; latest version v2.31.10 May 4, 2018), used to align RNAseq reads and to perform comparative genomic analysis to test for parallel loss of the genomic architecture that supports plasticity to salinity environments. Data have been collected and now need to be assembled and annotated. Genomes will then be used for comparative genomics analysis.

Gene Pathway and Structure Discovery (glympsed) & Phylotranscriptomics

To discover common expression patterns in large RNA sequencing data sets (>100 samples and 30,000 genes/sample), we are in the process of developing a pipeline using the unsupervised deep learning library, **Keras** (Chollet 2018). This pipeline, called "**glympsed**" (https://github.com/glympsed/glympsed) uses the "denoising autoencoder" algorithm (Vincent et al 2008) to extract features or patterns from the noisy dataset. This method has been shown to successfully identify common metabolic patterns of gene expression where there are unknown relationships between samples (Tan et al. 2016). The results of this pipeline will be used to generate hypotheses about pathway "features" that are present or absent in MMETSP and killifish samples.

Preliminary code and notebooks are available on GitHub, which were developed on AWS. Simulated data were created to test the denoising autoencoder model by randomly sampling beta values, assuming gene expression data follows negative binomial distribution. Jetstream resources will be used to develop code and run Jupyter notebooks for visualizations. Scripts will be written generate output tables containing gene networks and connect to gene regulatory network and pathway databases, such as **KEGG** (Kanehisa

et. al 2000). Tutorials will be developed and Jetstream images with required software installed will be made public.

Following the development and successful completion of results produced by the **glympsed** pipeline, the **agalma** phylotranscriptomics pipeline (Dunn et al. 2013) will be used to generate species trees based on pathway genes identified with glympsed.

Program	Resources	Total SUs
Development of glympsed and tutorials for phylotranscriptomics pipelines	Jetstream	110,000

C. Computational Research Plan: Resource usage with accompanying application efficiencies to achieve the research objectives.

We are requesting PSC Bridges (LM, RM and Pylon storage) in addition to Jetstream cloud computing resources. Jetstream is great for troubleshooting custom software, developing pipeline scripts and performing calculations and visualizations with data table inputs. PSC Bridges is great for storing raw sequencing data files, scaling up assembly pipelines with high memory processes in parallel, then evaluating and generating data tables.

Jetstream, like AWS elastic cloud computing where we have done some of our initial pipeline development and software installation, allows the freedom of installing with root privileges and frequently updating custom software, as well as hosting open data analysis collaboration with tools such as Jupyter notebooks. For students in a classroom setting and in order to provide persistent documentation, Jetstream resources are key to loading pre-installed software from images and install new software to test. Since we do not need the intermediate files beyond the time during computation, the *de novo* assembly products and gene expression quantification tables can be downloaded to local computers and the instances can be deleted.

We need access to an HPC with high memory nodes, like PSC Bridges to launch high memory jobs because the maximum Jetstream instance (s1.xxlarge: CPU 44, Mem 120 GB, Disk 480 GB) is too small to accommodate some of our assemblies and is also incapable of scaling to running hundreds of assemblies in parallel. We will delete intermediate files after assemblies have been generated and downloaded. Persistent storage is also a feature of Bridges that is not possible with Jetstream. One of the more frustrating practical aspects of using Jetstream is having to store large raw files for a data set that is around 1TB in volumes and keep the volume accessible each time you want to work with them from a different instance. Our experience and SUs measured during resources are explained in the attached "*Code Performance and Resource Usage*" document.

D. Justification of the SU allocation amounts for all resources and resource types.

Based on the "Code Performance and Resource Usage" document, the total amount of SUs requested for the following year of research are outlined as follows:

Resource	SU's requested
Jetstream	110,000
Bridges LM	44,876
Bridges RM	139,750
Bridges Pylon storage	25 TB

D. Resource Appropriateness

We've used two XSEDE startups of PSC Bridges allocations (LM, RM, and Pylon storage) and two startups of Jetstream allocations generating *de novo* transcriptome assemblies and annotations for the MMETSP dataset and 17 killifish as well as a separate project on *Doryteuthis opalescens* squid, which is being prepared for publication by PI Pierce. As a result, we are experienced Jetstream and Bridges users. These XSEDE startup allocations on Jetstream have contributed to one publication thus far (Johnson et al. 2018) and will contribute to another two manuscripts before the conclusion of Ms. Johnson's PhD research. However, additional compute time is essential to complete this work and prepare results for publication. Completion of Lisa Johnson's PhD dissertation research depends on these resources.

We have participated as active members of the XSEDE community both on Jetstream and PSC Bridges, communicating with the help team with questions and reporting problems with jobs and compute nodes. We have served as PIs, Co-PIs and users of 5 education allocations where we have taught researchers (>300 total) in workshop settings how to use Jetstream resources. We have created two public images as a result of these allocations to make future software installation on new instances easier: "RNASeq_1DayWorkshop" and "dammit_annotation_v1.0". The following are workshops delivered with tutorials using Jetstream materials: Society for Environmental Toxicology And Chemistry meeting, November 2018 (TG-MCB180142), Global Invertebrate Genomics Alliance meeting, October 2018 (TG-MCB180140), Scripps Institution of Oceanography, Bioinformatics User's Group workshop, October 2017 (TG-BIO170083), Data Intensive Biology Summer Institute (DIBSI), non-model RNAseq topic workshop, July 2017 (TG-BIO170017), 2-week DIBSI beginner's workshop (ANGUS), June-July 2018, ANGUS 2017 (TG-BIO170017). The following is an example website for a workshop using Jetstream materials that is available to the public and persistent for users beyond the time of the workshop: https://setac-omics.readthedocs.io/en/latest/.

Local computing environment

As a members of Dr. C. Titus Brown's lab at UC Davis, we have access to two high performance computing clusters at Michigan State University (iCER's hpcc) and UC Davis (farm cluster). However, there are several issues with these resources. Our time on the MSU hpcc is expiring this fall 2019, following Dr. Brown's transition from MSU to UCD faculty in 2015. In addition, there has been limited capacity on the MSU hpcc for supporting the scale of these multi-species pipelines as we only

have access to 4 nodes with 1TB of resources and the queue time is long. Moving 700 transcriptome through this queue takes more than one month and our storage capacity is limited. With the PSC Bridges startup, we have been extremely pleased with the ease of submitting jobs and having them run smoothly on both the LM and RM partitions.

Research funds for AWS compute time are limited, as the Moore Foundation grant that has contributed to funding these projects ends in November 2019. On the UCD farm cluster, we do not have access to the high memory node without buying in to the high memory partition of the system. As our current account stands, we are limited to 12 jobs running at a time with 75 GB RAM each and less than 1TB of total storage.

Following the computing processes described in this proposal, raw sequence files will be stored locally on storage hard-drives and deposited in NCBI's Sequence Read Archive public repository. Data products, e.g. assemblies will be stored in public repositories such as Zenodo and the Open Science Framework (Foster and Deardorff, 2017).

E. Additional Considerations

Research outreach and training

Graduate students at UC Davis are required to rotate through labs during their first year prior to choosing a primary advisor. During rotations, small projects are completed. This allocation will contribute to computational training of several rotation students in our lab who are working on the MMETSP data sets using the eelpond protocol. New protocols developed, e.g. glympsed and eelpond-integrated Oyster River Protocol will contribute to tutorials taught during our Data Intensive Biology Summer Institute at UC Davis.

Funding

Analyses for the above-mentioned projects are funded until November 2019 by the Moore Foundation Data-Driven-Science investigator award to the PI of our lab, Dr. C. Titus Brown. Lisa Johnson (Cohen) received an NSF Graduate Research Fellowship Honorable Mention in 2016, which enabled her to serve as PI for the XSEDE startup allocation TG-BIO160028) that has contributed to most of her dissertation work. Tessa Pierce has an NSF Postdoctoral Fellowship and has been a PI for the XSEDE startup allocation TG-MCB170160, which has contributed to her postdoctoral research. Ms. Johnson and Dr. Pierce have both served as co-PIs for the XSEDE education proposals listed above. Results from these analyses are necessary to contribute as preliminary analyses for future proposals requesting funding from agencies such as the NSF.

References

Bushmanova, E., Antipov, D., Lapidus, A., & Przhibelskiy, A. D. (2018). rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. *bioRxiv*, 420208.

Cantarel, B. L., Korf, I., Robb, S. M., Parra, G., Ross, E., Moore, B., ... & Yandell, M. (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome research*, *18*(1), 188-196.

Chollet, F. (2018). Keras: The python deep learning library. *Astrophysics Source Code Library*.

de Lannoy, C., de Ridder, D., & Risse, J. (2017). The long reads ahead: de novo genome assembly using the MinION. *F1000Research*, *6*.

Dunn, C. W., Howison, M., & Zapata, F. (2013). Agalma: an automated phylogenomics workflow. *BMC bioinformatics*, *14*(1), 330.

Ebbert, M. T., Jensen, T. D., Jansen-West, K., Sens, J. P., Reddy, J. S., Ridge, P. G., ... & Keene, D. (2019). Systematic analysis of dark and camouflaged genes: disease-relevant genes hiding in plain sight. *bioRxiv*, 514497.

Foster, E. D., & Deardorff, A. (2017). Open science framework (OSF). *Journal of the Medical Library Association: JMLA*, *105*(2), 203.

Kannan, S., Hui, J., Mazooji, K., Pachter, L., & Tse, D. (2016). Shannon: An information-optimal de novo rna-seq assembler. *bioRxiv*, 039230.

Keeling, P. J., Burki, F., Wilcox, H. M., Allam, B., Allen, E. E., Amaral-Zettler, L. A., ... & Beszteri, B. (2014). The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS biology*, *12*(6), e1001889.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., ... & Chen, Z. (2011). Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature biotechnology*, *29*(7), 644.

Johnson, Lisa K., Alexander, Harriet, & Brown, C. Titus. (2018). Re-assembly, quality evaluation, and annotation of 678 microbial eukaryotic reference transcriptomes, *GigaScience*, giy158, https://doi.org/10.1093/gigascience/giy158

Johnson, Lisa K., Alexander, Harriet, & Brown, C. Titus. (2018). MMETSP re-assemblies [Data set]. Zenodo. https://doi.org/10.5281/zenodo.740440

Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, *28*(1), 27-30.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *GenomeBiology*. 15:550.

MacManes, M. D. (2018). The Oyster River Protocol: a multi-assembler and kmer approach for de novo transcriptome assembly. *PeerJ*, *6*, e5428.

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference. *Nature Methods.* 14(4): 417-419.

Rodgers, R., Roach, J. L., Reid, N. M., Whitehead, A., & Duvernell, D. D. (2018). Phylogenomic analysis of Fundulidae (Teleostei: Cyprinodotiformes) using RNA-sequencing data. *Molecular phylogenetics and evolution*, *121*, 150-157.

Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., ... & Griffith, M. (2010). De novo assembly and analysis of RNA-seq data. *Nature methods*, *7*(11), 909.

Scott C, dammit: an open and accessible de novo transcriptome annotator; 2016. *In prep*. Available at: https://dib-lab.github.io/dammit/. Accessed 15 Jan 2019.

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.

Tan, J., Hammond, J. H., Hogan, D. A., & Greene, C. S. (2016). ADAGE-based integration of publicly available Pseudomonas aeruginosa gene expression data with denoising autoencoders illuminates microbe-host interactions. *MSystems*, *1*(1), e00025-15.

Tan MH, Austin CM, Hammer MP, lee YP, Croft LJ, Gan HM. 2018. Finding nemo: hybrid assembly with Oxford Nanopore and Illumina reads greatly improves the clownfish (*Amphiprion ocellaris*) genome assembly. *GigaScience*. 7(3): gix137. https://doi.org/10.1093/gigascience/gix137

Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P. A. (2008, July). Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning* (pp. 1096-1103). ACM.

Whitehead, A. (2010). The evolutionary radiation of diverse osmotolerant physiologies in killifish (Fundulus sp.). *Evolution: International Journal of Organic Evolution*, *64*(7), 2070-2085.

Zimin, A. V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S. L., & Yorke, J. A. (2013). The MaSuRCA genome assembler. *Bioinformatics*, *29*(21), 2669-2677.