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-BIO160028 and TG-MCB170160

Code Performance and Resource Usage

This document is based on 70-80% usage (PSC Bridges LM: 113 jobs, RM: 1042 jobs) of startup allocation TG-BIO160028 on PSC Bridges since we began our allocation on 11/26/2018 and 100% usage of 150,000 SUs on Jetstream from 2016-2018 with TG-BIO160028 and 72% usage of 100,000 SUs on Jetstream with TG-MCB170160.

Analyses run

The following table includes processes run previously and their resources used that are directly related to the processes we are requesting additional resources for in this proposal.

Transcriptomic analysis

Jetstream resources were estimated, as long-running tools are often run overnight, and instances used for training and workshops are kept running during instruction and troubleshooting, so active program runtimes are not continuous. Bridges jobs are listed for the species with the longest time and highest resource usage.

Process	Species	Data size	Resource	SU/run used
Workshop tutorial (single commands run by hand), *de novo* transcriptome assembly pipeline	*Nematostella vectensis*	30,000 reads subset	Jetstream, m1.medium	12 = 6 CPU x 8hrs workshop time
dib-MMETSP, automated *de novo* transcriptome assembly pipeline (Johnson et al. 2018)	*Chaetoceros neogracile* (Phylum: Bacillariophyta), MMETSP0751	33,150,207 reads	Jetstream, s1.xxlarge	264 = 44 CPU x 6hrs
Digital normalization with khmer	*Fundulus grandis*	467,432,867 trimmed reads	Bridges RM; jobID 4524354	48.10
De novo transcriptome assembly with Trinity	*Fundulus heteroclitus* (MDPL population)	79,341,401 reads (after digital normalization)	Bridges LM; jobID 4555613	232.07

Transcriptome annotation, dammit	*Fundulus heteroclitus* (MDPP population)	668,487 contigs	Bridges LM; jobID 4710488	57.48
Transcriptome and genome evaluation, BUSCO	*Fundulus grandis*	809,060 contigs	Bridges RM; jobID 4612712	292.97
Transcriptome evaluation, Transrate	*Fundulus grandis*	809,060 contigs	Bridges RM; jobID 4609550	33.33
Expression quantification, salmon	Loop with all 9 individuals from the *Fundulus grandis* species	467,432,867 trimmed reads	Bridges RM; jobID 4707781	14.37
Genome alignment of RNAseq reads, STAR	*Fundulus grandis*, BW treatment, sample 1	81,736,962 trimmed reads	Bridges RM; jobID 4702491	6.76
Species trees, agalma	Subset of data from 6 invertebrate siphonophore and cnidarian species	RNAseq reads	Jetstream, m1.medium	12 = 6 CPU x 2hrs

Genomic analysis:

Process	**Data type**	**Data size**	**Resource**	**SUs used**
De novo genome assembly of a bacteria, ONT reads with canu and Pilon correction	*Tenacibaculum* bacteria (3Mb)		Jetstream, s1.large	80 = 10 CPU x 8 hrs
MaSuRCA	Killifish genome (1.1Gb)	4x ONT coverage subset with 50x coverage Illumina data	MSU hpcc	168 hrs = 1 TB for 1 week; MSU deleted my log file :(
Racon and Pilon	Killifish genome (1.1Gb)	4x ONT coverage subset with 50x coverage Illumina data	MSU hpcc	Job failed “Caused by: java.lang.OutOfMemoryError: GC overhead limit

				exceeded " with 500 GB memory

Extrapolation for Future work

We have not had a chance to benchmark everything that we want to yet for the *de novo* genome assemblies, which is why we're writing this proposal requesting more computing resources on these platforms. As mentioned in the main document, because long reads from Oxford Nanopore Technologies (ONT) are "third-generation" sequencing technology, software tools that are being developed at a fast rate to keep up with the developing technology. The raw signal files used to call bases to generate fastq sequence files must be saved because the base-calling software algorithms are improving. We have not tested the new base-calling software, and thus have not included it in our pipeline here. But, in the next year, if the software improves, we will need to return to the beginning of our pipeline to acquire more accurate bases for our reads.

There is not yet a set of best practices for these software tools for *de novo* genome assembly using ONT reads, therefore several must still be tested. As mentioned in the main document, we have tested several *de novo* genome assemblers for ONT data with a subset of data and have found that MaSuRCA to be the best. However, if we scale up our data and find that another tool is better, we would like to use the better set of tools.

SUs requested for running genome assembly tools for 4 *Fundulus* genomes, 1.1Gb each:

Program	**Description**	**# runs**	**SUs/run**	**Resources**	**Total SUs**
MaSuRCA	assembler	4	168 = 1 TB x 168 hrs	Bridges LM	672
Racon and/or Pilon	Consensus improvement	24 = 4 x 6 iterations/ species*	48 = 1 TB x 48 hrs	Bridges LM	1,152
khmer	Error correction	4	50	Bridges RM	200
sourmash	decontamination	4	50	Bridges RM	200
MAKER	annotation	4	48 = 1 TB x 48 hrs	Bridges LM	1,152
storage	ONT signal files 1TB/species + 100 GB fastq sequence data/species x 4 species + 500 GB for 2 lanes Illumina PE150 fastq data + 3 TB/species x 4 species for intermediate storage during assemblies			Bridges Pylon storage	17 TB

*Racon and Pilon are separate tools that take Illumina data and polish the final assembly to improve it. The reason why we have more runs for either Racon and Pilon in our list of SUs requested below is because these require multiple iterations for correcting the consensus genome sequence (Miller et al. 2018). We have done this and confirmed for a 3 Mb bacteria (*Tenacibaculum*) genome on Jetstream (m1.large), but not on larger assemblies because the jobs crashed without enough memory on our local MSU hpcc We have not attempted on Bridges yet because we didn't have enough SUs in our startup nor the time between 11/26/18 and the completion of this proposal (1/15/19).

SUs requested for running *de novo* transcriptome assembly tools for 700 transcriptomes:

Program	**Description**	**# runs**	**SUs/run**	**Resources**	**Total SUs**
khmer	digital normalization and error correction	700	50	Bridges RM	3,500
Spades x2 *k-mers*, Shannon	de novo transcriptome assembler	16	500	Bridges LM	8,000
Spades x2 *k*-mers, Shannon	de novo transcriptome assembler	678	50	Bridges LM	33,900
Dammit	annotation pipeline	700	50	Bridges RM	35,500
Transrate	transcriptome evaluation	1,400 (run twice for each assembly, comparing against reference)	35	Bridges RM	49,000
BUSCO	Genome and transcriptome evaluation	700	50	Bridges RM	35,000
OrthoFinder	identification of 1-1 orthologs from data sets	2 = MMETSP assemblies + killifish assemblies	250	Bridges RM	400
Salmon	gene expression quantification	700	20	Bridges RM	14,000

STAR	alignments	130 *Fundulus* individuals	15	Bridges RM	1,950
Oyster River Protocol	contig merging	700	~200 hours of s1.xxlarge at 44 SU	Jetstream	10,000
storage	2 TB raw sequence data + 6 TB for intermediate files and alignment .bam files			Bridges Pylon storage	8 TB

Bioinformatics Tools, performance and scaling

In general, the bioinformatics software tools that we use have been published and optimized by the software authors - including the PI of our lab, Dr. C. Titus Brown who developed and maintains *khmer* (digital normalization and reads error correction) and *sourmash* (decontamination) - and are made available for users, like ourselves, to generate data products such as *de novo* genome and transcriptome assemblies and annotations for our first two objectives.

The third deep learning objective is an exploratory analysis, and we do not have all of the code yet. The scripts that we have written collaboratively thus far do not require computing resources beyond a cloud instance that will accomodate the software and tables containing expression data from hundreds of species by >30,000 genes each, which is beyond what can be run on a laptop computer. Therefore, we are requesting Jetstream resources to further explore this type of analysis.

We have been successful in running these tools on systems with slurm management, such as MSU hpcc and UCD farm cluster. However, these hpc resources will not support the work that we want to do to assemble genomes with ONT and Illumina data, re-assemble more transcriptomes, and complete machine learning analyses. Therefore, we are hoping to use PSC Bridges to conduct the assemblies of four genomes.