Investigating the transcriptional effect of DNA virus infection in resistant and susceptible lines of green algae (*Ostreococcus tauri*).

Project 3 Proposal - GitGud: Mason Tatro, Ben Yunker, Naomi Williamson, Ellie Spahr

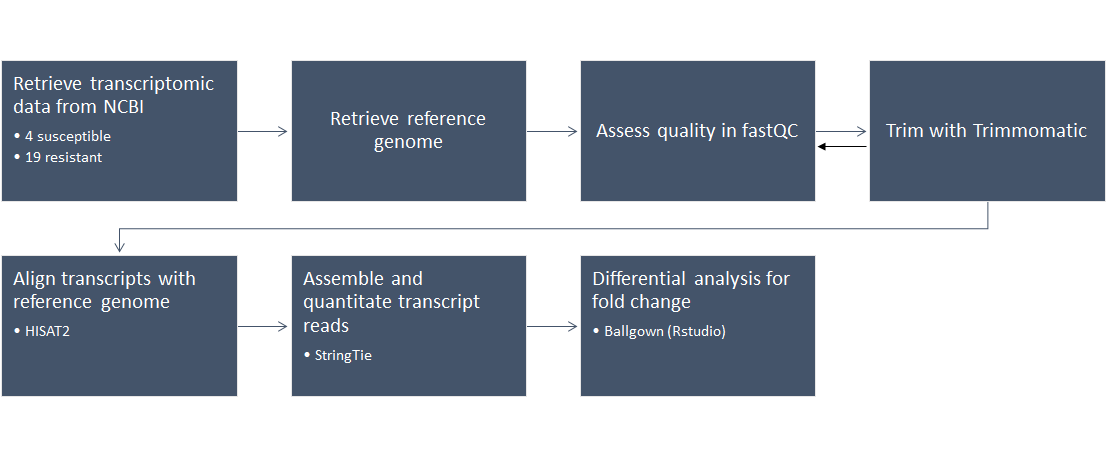
**Background**

*Ostreococcus tauri* is a unicellular marine algal species. It is one of the smallest free-living eukaryotes, having only a singular chloroplast and mitochondrion, a genome with an unusually large proportion of coding regions, and few genetic redundancies.1 *O. tauri* is found around the globe in low-nutrient, oligotrophic waters and is easily reared in a laboratory setting, making it a popular model for studying green algae.

In aquatic environments, algal blooms are often regulated in part by DNA viruses; this has become a proposed mechanism for control of large environmentally-detrimental blooms and for releasing algal-bound nutrients back into the ecosystem.2 As atmospheric CO2 concentration and water temperatures increase as a byproduct of climate change, harmful algal blooms are expected to become more frequent and severe.3 Although DNA viruses may be a promising biological control mechanism to mitigate the increasing ecological instability imposed by harmful algal blooms, concentrated treatment of blooms could increase the rate of viral resistance development within the target populations. To examine potential mechanisms for viral resistance and how it may potentially be overcome, we wish to first examine resistance-associated gene expression profiles during infection in a model algal organism that commonly exhibits DNA-virus-resistant strains. Our algal model, *O. tauri*, is closely associated to one such large DNA virus, OtV5. To examine differences in gene expression profiles between algae susceptible and resistant to large DNA viruses, we are examining the transcriptomes of clonal lines inoculated with OtV5 and evaluating fold expression changes between susceptible and resistant lines.

By first investigating differential gene expression between susceptible and resistant algae lines during viral infection, we hope to then use this information to examine regulatory differences and, finally, determine common mechanisms behind the viral resistance in future studies.

**Methods**

**

*Figure 1: Analytical Pipeline for Differential Gene Expression.* A summary of the workflow for pairwise analysis of RNA-Seq data.

*Data Acquisition*

Transcriptomic data for *O. tauri* will be retrieved from the NCBI SRA Database (BioProject: PRJNA344946). Clonal lines were inoculated with OtV5, sorted into resistant and susceptible lineages, RNA was isolated from biological replicates to create cDNA libraries, and samples were sent for generation of Illumina HiSeq 2000 paired RNA-Seq data. For this study, RNA-Seq data was generated for four replicates within susceptible lines and 19 replicates for resistant lines. The size of each replicate data set varies between isolate from anywhere between ~900M bases to 2.3G bases.

The *O. tauri* reference genome was also retrieved from NCBI’s Genome database (GCF\_000214015.3). Several genomes exist for this organism; we have selected the most recent, thoroughly annotated version (updated in 2015).

All files will be uploaded to WVU’s Spruce HPC Network for further analysis in the bioconda environment.

*Quality Control and Trimming*

Quality will be assessed using FastQC (v0.11.7). Trimmomatic (v0.38) will be used to trim any adapter sequences and regions of poor quality from the RNA-seq reads. Afterwards, FastQC will be conducted once again to determine quality and if more trimming is needed.

*Transcript Mapping, Assembly, and Differential Analysis*

Transcripts will be aligned to the reference genome using HISAT2 (v2.1.0), which will provide splice-aware mapping. As HISAT2 will produce a .sam output, Samtools (v0.1.19) will be used to convert .sam files to .bam. StringTie (v1.3.4) will take the .bam mapped transcript information to assemble and quantitate transcript reads based on the positional information informed by HISAT2. Differential analysis for fold change in gene expression between our samples will be generated in RStudio using Ballgown.

*Division of Labor*

The proposal document was prepared by Naomi and Ellie. Ben will conduct initial FastQC, as well as data trimming and post-trim quality control runs. Transcript alignment through HISAT2 will be coded and run by Mason. Ellie will assemble transcripts and estimate read coverage using StringTie. The final differential analysis will be written and run in Ballgown by Naomi. All parties will be responsible for interpreting and summarizing their results, with the final result interpretation shared among the group. Group contribution will be expected in the final report, with emphasis on each member’s portion of the pipeline and final result analysis.

*Conflict Resolution*

Disagreements will be resolved within the group. All group members will be given an opportunity to present ideas and reasoning and discuss the matter internally. After discussion, if no decisions have been made, disagreements will be resolved through majority vote.

In extreme cases where an internal resolution cannot be reached, issues will be mediated through the instructor.

**Expected Outcomes**

We expect to generate an output of differential transcript expression levels from a pairwise analysis of RNA-Seq data for resistant and susceptible OtV5-inoculated *O. tauri*. By examining fold change in gene expression across our two conditions, we will use the associated functional annotations from our analysis to inform conclusions about how biological function is altered in our organism as a result of DNA virus resistance.

As our transcriptome samples have all been inoculated with the DNA virus OtV5, we expect to capture viral RNA within our RNA-Seq samples. We expect for these reads to be filtered out during the mapping process (using HISAT2) as they should not bear homology to the *O. tauri* reference. Furthermore, we are most interested in the gene expression profile of virus-resistant *O. tauri*; these transcripts should not be highly represented in these sample sets. In transcriptomic analyses, repeat elements can cause issues during the mapping of repeat-rich transcripts. The *O. tauri* genome has a markedly low proportion of repeat regions and transposable elements (TEs); Chromosome 19 and a portion of chromosome 2 make up 77% of all TEs within the genome and LINEs and SINEs were not detected at all.1 The low content and chromosomal segregation of repeat-rich elements should reduce complications faced by these integral genomic features.

We are utilizing the HISAT2/StringTie/Ballgown pipeline instead of the TopHat suite due to the increased efficiency and speed it provides. Although some packages, such as HISAT2, may use more memory over their older counterpart, we anticipate the overall size of our files in this pipeline to require long wall times. Certain packages, such as StringTie are both faster and more memory efficient than other available software, such as Cufflinks. As StringTie has been estimated to perform most processes on an average desktop computer, it should have no problem handling the complete dataset on Spruce.4 In our last project, our group ran into issues troubleshooting efficiently during long walltime runs; by choosing efficient packages this time, we hope to progress through the analytical portion of our pipeline quickly so more time can be devoted to interpreting results. Our chosen organism should also help us accomplish this goal. Though we are working with a eukaryote, its small, efficient genome should reduce the memory and time required to run each process compared to other eukaryotic data sets.

**References**

1. Derelle, E., Ferraz, C., Rombauts, S., Rouzé, P., Worden, A. Z., Robbens, S., ... & Saeys, Y. (2006). Genome analysis of the smallest free-living eukaryote Ostreococcus tauri unveils many unique features. *Proceedings of the National Academy of Sciences*, *103*(31), 11647-11652.
2. Suttle, C. A. (2005). Viruses in the sea. *Nature*, *437*(7057), 356.
3. Griffith, A. W., & Gobler, C. J. (2019). Harmful algal blooms: A climate change co-stressor in marine and freshwater ecosystems. *Harmful Algae*.
4. Pertea, M., Kim, D., Pertea, G. M., Leek, J. T., & Salzberg, S. L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature protocols*, *11*(9), 1650.