Development of a Bioinformatics Pipeline to Compare Putative Enhancers Usage within Different Strains of *Drosophilia melanogaster*

Student Researcher: Justin K. Yang

Advisor: Michael Palopoli

# Bowdoin College Biology Department

Within the domain of evolutionary biology, the degree to which phenotypic variation is driven by mutations of the protein coding sequence or by changes to gene regulation of the genome are still mostly unknown and remains as one of the long standing domain debates. Cis-regulatory elements known as enhancers have been known to contribute to phenotypic diversity and lie upstream of the coding region of genes. Enhancers that are actively controlling the transcription of genes are known as putative enhancers, and are known to be found within only sections of DNA in an open chromatin confirmation (as opposed to closed conformation). The conformation of DNA sections containing enhancers has been shown to be heritable. With this in mind, our study aims to compare putative enhancer usage between different strains of *Drosophilia melanogaster* (fruit flies), and potentially even different fly species. To do this, we developed a bioinformatics pipeline to analyze a genome-wide open chromatin assay of the *D. melanogaster* genome, collected last summer by the Palopoli lab. Using a professionally created ATAC-Seq bioinformatics pipeline, provided by Lifebytes, as a guide, this project used a series of genome analysis and quality control programs, implemented on the Bowdoin High Powered Computing (HPC) to develop a robust bioinformatics pipeline for ATAC-Seq generated DNA libraries.

 The main goal of this summer project was to recreate the bioinformatics pipeline provided by the Lifebytes bioinformatics consulting company. Additionally, we aimed to use the newly developed lab pipeline to analyze the ATAC-Seq open chromatin genome-assays that had been previously created by this lab. The final goal of this summer was to further our research and understanding about the domains of evolutionary biology and genetics, specifically through research of other studies pertaining to ATAC-Seq and our current work.

 Since the Palopoli lab had previously generated all necessary ATAC-Seq open chromatin genome-assay data, our project did not incorporate any in-house lab elements, and was done entirely remotely. Zoom calls and the google drive data sharing platform were used to facilitate lab communication. We sent the lab data to the Lifebytes bioinformatics consulting company for analysis. Lifebytes created their own proprietary bioinformatics pipeline using our data and provided our lab with figures and graphs from their pipeline. Additionally, alongside the figures, Lifebytes also provided a general flowchart detailing the basic flow of data within their pipeline (Fig. 1). Using this flowchart as a guide, our lab identified all possible genome analysis programs used by Lifebytes. All tools were researched, documented, and compiled into a number of lists that were sent to the Bowdoin HPC liason DJ Merrill. DJ Merrill would then install these tools onto the HPC for our use. Tool packages and individual programs used include: Deeptools, ATAC-Seq, snakePipes, Qualimap, multiQC, Bowtie2, samTools, and sambamba. Once installed, the tools were implemented into a lab bioinformatics pipeline, through a string of bash scripts. These bash script were manually created through the use of linux text editors on the Bowdoin HPC (Fig. 2). After implementing these tools, further scripts were created to smooth the flow of information for the sake of accessibility and ease of use. Alongside all coding, our lab conducted extensive research into ATAC-Seq data analysis and bioinformatics, to ensure our pipeline was effective in data analysis.

 Our lab was successful in creating a bioinformatics pipeline capable of handling any type of ATAC-Seq data. The lab pipeline converts unreadable ‘.fastq’ files into figures and statistics, through a series of file type conversions, quality control checks, and read filtering. The pipeline is fully automated and handles all forms of data file type conversion without manual input. Additionally, all original file types are saved for viewing alongside the final figures. In this way, our ATAC-Seq bioinformatics pipeline is quite robust and will not require manual changes to accommodate for ATAC-Seq input data. The lab pipeline recreated versions of all the graphs and figures provided by Lifebytes (Fig 3; Fig. 4), as well additional figures that were never provided by Lifebytes (Fig. 3). Figures generated by the lab pipeline are not exact recreations of the Lifebytes figures, however the interpretation of our figures presents the same conclusions as the Lifebytes figures, indicating our figures are reliable (Fig. 4).

 By successfully recreating a professionally created bioinformatics pipeline, our lab proves that our project does not need to rely on professional bioinformatics companies for the analysis of our data. This will save both time and costs for future ATAC-Seq projects, as all data analysis can be run automatically from our own lab. Furthermore, small differences in bioinformatics pipelines has been shown to dramatically impact the interpretation of results (Reske et al. 2020), so the creation of a uniform pipeline, in which all of our sample’s data will be run, removes the potential for differences in pipelines created by out of lab companies. In this way, the creation of a unique bioinformatics pipeline will streamline the process of analyzing future sample, as well improving accuracy and removing potential bias of the results.



**Figure 1: Lifebytes pipeline flow chart.** The implemented ATAC-Seq lab pipeline is heavily based off of this Lifebytes flowchart and the tools described within it. The lab pipeline, is not a direct copy of this flowchart, but does include all elements shown in this flowchart. Figure provided by Lifebytes.

**Figure 2: Bash script created for the implementation of the genome aligner tool Bowtie2.** The script here inputs ‘.fastq’ files found within the “workingData” directory and outputs a series of ‘.sam’ files after file conversion to a designated output folder. This script serves as one of the 20 different scripts the comprise the ATAC-Seq bioinformatics pipeline.

Our Pipeline

Lifebytes Pipeline


## Figure 3: Comparison of Pair End (PE) Alignment Scores generated by the Lifebytes pipeline and the lab pipeline. The graphs generated by the lab pipeline (top two graphs) represent the PE alignment scores of each technical replicate. The top left graph shows PE alignment scores based on total read count and the top left graph shows PE alignment scores based on percentage. The figure provided by Lifebytes (bottom graph) shows the reads by total read count. Both graphs based on total read count show similar quantities of PE Alignment scores for each sample. Lifebytes did not provide a graph showing the type of read based on percentage of total reads. Bottom graph provided by LifeBytes.

**Figure 4: Comparison of technical replicate Principal Component Analysis (PCA) graphs from both the lab pipeline and the Lifebytes pipeline. B**oth the Lifebytes and lab PCA graph both indicate a similar level of variance between each technical replicate. Clusters of biological replicates (green and neon blue, red and yellow, purple and dark blue) are visible within both graphs. Technical replicates of the same biological replicate (E, I, K) cluster closer to samples in the same biological replicate then to other samples. Rightmost graph provided by LifeBytes.

Our Pipeline

Lifebytes Pipeline

I would like to thank our project advisor Professor Mike Palopoli for his insight, guidance, and support. I would also like to acknowledge Mohamed Ashick and LifeBytes, for their work on data analysis and for providing us with the data flow chart. Thank you to DJ Merrill for setting up all the various tools and programs we used on the HPC. Funding for this project was provided by the Maine Space Grant Consortium (MSGC). Bowdoin College is an affiliate of the Maine Space Grant Consortium and any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Aeronautics and Space Administration or of the Maine Space Grant Consortium.

I would also like to thank my friends and previous lab mates: David Brower, Andy Bolender, and a special thanks to Hannah Konkel for all her help. We would never have been able to work on this project if not for all their hard lab work. And thank you to my friends and current lab mates: Samira Iqbal, Nicholas Purchase, and Callie Burkhart for powering through all the tedium of bash scripting and seemingly making the impossible possible.