Genome-Analysis, Project Plan

By: Hugo Swenson, husw7546

Introduction

I chose paper II, as i have a keen interest in evolutionary biology, and feel asif will have more motivation towards the project doing something which i thoroughly find intriguing. This will hopefully lead to increased productivity, as well as a basic understanding on how to perform studies aimed toward evolutionary biology and genome-assembly. The various steps involved will lead to a varied and informative approach regarding how to interpret large genomic data, and how to utilize it as to get a qualitative approach.
Overall, i feel the difficulty level of this project will be fairly high, but the results very educational.

Methodology

I will perform most work utilizing the linux subsystem for windows 10, through the “Bash on Ubuntu on Windows” tool. If necessary, i will also utilize the UU-linux computers provided in the dedicated computer-rooms on BMC/EBC.

As there is a total of 41 work days (excluding holidays and other classes) to dedicate to this project, a realistic approach of 4 hours a days would result in a total of 164 work hours, excluding unforeseen events and lab hours.
These total hours needs to be allocated not only on the analysis, but also on the interpretation of the data, the understanding and utilizing and coding required to utilize certain tools. As well as the log/wiki and presentation.
As the timeframe required for coding often is very hard to estimate, a generous amount of time should be allocated towards this. Furthermore, due to the restrictions of the UPPMAX server and our reservation, the work should be scheduled in such a manner that a request for heavy work load should be done at the end of each lab-session, whereas light work can be performed the following day using the results obtained from the heavy workload, as to optimize the workflow.
For proper documentation, at the end of each week, i will do a compilation of the work done and the results obtained, as to measure my own progress and constantly stay updated upon my own progress.

Tools

The basic requirements for this paper is as follows:
- Genome assembly of Illumina reads.
- Assembly quality assessment.
- Transcriptome assembly.
- Structural and functional annotation.
- Differential expression analyses.
- Biological interpretation of the results.

Firstly, FastQC will be used to assess the quality of the data used, something which needs to be done in an early stage to prevent a snowball effect from taking place if bad data should be present.
For the genome assembly, SOAPdenovo will be used after FastQC, coupled with Trimmomatic for read pre-processing (if i have enough time, i will try Spades as well, for a higher quality assembly).The choice of SOAPdenovo is, despite its poorer results when compared to other assemblers such as Spades, due to its low memory performance reducing time. Ideally, after assembly quality check, as an extra, the tool Exonerate should be used to clear up intron/exons. Furthermore, PASA should be used if possible to identify/confirm alternatively spliced transcripts.
For the transcriptome assembly, RNA-Spades or Trinity are good options to use coupled with trimmomatic, with respect to the original study, Trinity will most likely be used for this purpose, with the option of trying RNA-Spades if enough time is available.

To evaluate the quality of the assemblies, the tool Quast will be used, followed by the Maker2 tool for proper annotation of the eukaryotic genome studied.
This will be followed by a homology search using BLAST, BLASTN and BLASTP.
For mapping the genome, TopHat will be used. For read counting, HTSeq will be used followed by expression analysis using DESeq2 combined with h-clust to cluster genes showing similar expression.

Additionally, for the ChIP-seq, BEDTools will be used after bowtie to obtain read-counts and to merge peaks from all samples, in turn followed by BEDOPS -to partition all the peaks.
This will be followed by using the tool Picard MarkDuplicates to remove duplicate reads.

If i have the required time, as a final step for the comparative genomics, LASTZ should be run, followed by the UCSC Genome Browser with the purpose of chaining, netting converting the alignment files to a MAF format. When this has been done, the Multiz-TBA tools roast command should be used to convert the Individual MAF files into a multiple MAF file.
When this has been performed, the phyloFit tool should be used to create a model for conserved + non conserved sequences, followed by use of the phastCon tool with the purpose identification and generation of base-by-base conversion scores.
Finally, the PhyloP tool should be used for the identification of BAR's (Bat accelerated regions) followed by the use of DAVID for BAR analysis and functional annotation.

Time allocation

From the provided Student_Manual_180323.pdf paper, approximate time-allocations have been estimated.
For the proposed methodology, using the estimated times provided in the student manual, this would result in a total of roughly 24 hours of total computational time at UPPMAX (1419 minutes).
Obviously, this is subject to change, as the estimates may not be accurate, but should be taken into consideration. For the proposed methodology, the main culprits timewise, can be seen as the Maker2 assembly, as well as the DNA-assembly.

As to not waste lab-time, i will try to queue these processes at the end of the lab, or on my free time to optimize the workflow. Instead focusing on the “smaller” work when the lab assistants are available.

The “dark matter” of this estimate is of course the time which will need to be allocated towards the understanding, and debugging of the various programs used. To combat this, i will use a generous estimate when allocating the time, stealing somewhat from the time i normally would give to the presentation, as graphical design combined with oral presentation of a topic are moments i know i can handle well and efficiently.

As such, my proposed time plan will handle the read quality control and reads pre-processing during the first week, alongside properly learning how to use UPPMAX and Github (15+20+180+? minutes).
During the second week, i will opt to perform RNA assembly and annotation alongside error probing and quality assessment using the original study as a reference (180 + 15 + 210 + 60 + 60 +60 minutes). Needless to say, from the UPPMAX required time, this week will be the heaviest, and may not be realistic, but assuming i properly understand how to use UPPMAX from week one, this is not impossible.

The third week, will be mainly used for alignment and differential expression, followed by the the Chl-P seq and quality control over the fourth week (60 + 5 + ? + ? minutes) .
This will give me roughly two weeks for exploration, debugging, and presentation, which is a time-frame i feel is adequate for the mountains of error i will likely run into.