Module 3 IBIX-NGS: Next Generation Sequencing Informatics

Assignment

Background:

Oat production has steadily increased in Northern Europe, including the UK, because oats are considered superior to wheat and barley for dietary fiber, and the richest sources of soluble fiber beta-glucan, rich in essential amino acids, and contain a range of unique antioxidants and vitamin E compounds. Thus, they have become a cereal of social and economic importance. However, there have been concerns in the EU that oats can become contaminated with type A trichothecene mycotoxins (T-2 and HT-2) produced mainly by *Fusarium langsethiae*.

Objectives:

The objective of this study is to perform a transcriptomic analysis of the impact of climate change on mycotoxin production. To simulate climate change, we are studying interacting conditions of temperature (20 vs. 30 °C) and water activity (a_w, 0.995–0.998) on T-2 and HT-2 toxins-related genes on *F. langsethiae*.

You have been supplied with 18 files of RNA-Seq raw sequence Paired-End reads, for *F. langsethiae*. under two conditions; Normal (3 replicates at 20°C, 0.995aw) and two climate change scenarios (3 replicates: 25°C, 0.995aw, and 5 replicates: 25°C, 0.98aw).

Library Name	**Sample name**	**°c**	**aw**
B009951	20-0.995-1	**20**	**0.995**
B009952	20-0.995-2	**20**	**0.995**
B009953	20-0.995-3	**20**	**0.995**
B509951	25-0.995-1	**25**	**0.995**
B509952	25-0.995-2	**25**	**0.995**
B509953	25-0.995-3	**25**	**0.995**
B50981	25-0.98-1	**25**	**0.98**
B50982	25-0.98-2	**25**	**0.98**
B50983	25-0.98-3	**25**	**0.98**

You have also been provided with a reference genome in FASTA format `flang.fasta`, as well as two annotation files. The Augustus gene model annotation file `flang_gene_models_AUGUSTUS.gff` is required for the reads alignment, while the functional annotation file (`flang_functional_annotation.gff`) includes gene names, description, molecular functions, for the Augustus gene models. Please use this file for the downstream analysis in R for your DE genes/transcripts (**Do not use this file for alignment**!). You

have also been provided with the transcript sequences in aa format (`flang_gene_models_AUGUSTUS.aa`) in case you need it for transcriptome-based alignment.

The RNA-Seq samples, reference genome and annotation files are accessible on Crescent via the following shared folder:

`/project/Fady_Mohareb/bioinformatics/I-BIX-NGS_dataForAssignment_flang`

Details about the reference genome is accessible via the genome publication:

Zuo Y, Verheecke-Vaessen C, Molitor C, Medina A, Magan N, Mohareb F*. De novo genome assembly and functional annotation for Fusarium langsethiae. BMC Genomics. 2022 Feb 22;23(1):158. doi: 10.1186/s12864-022-08368-0

You are required to develop a bioinformatics RNA-Seq analysis pipeline using best practices, justifying your choice of parameters and tools incorporated within the pipeline.

Deliverables:

Part I : Analysis Report outlining and discussing your results: [80 marks] (up to 2000 words).

Your analysis should include four main building blocks as follows:

1. **QC [5 marks]:** your pipeline should generate quality reports for the raw reads of each individual fastq file, and perform any necessary low-quality reads removal and trimming as you see fit. You should report the no. of reads and the overall sequence quality plots before and after the pre-processing step.
2. **Read alignment [15 marks]:** you should use the reference genome and annotation provided. You should apply test at least two aligner software and compare the results obtained from each one (i.e. the % of reads uniquely mapped for each). Only use one set of aligner's output for the rest of your analysis (i.e the best of the two aligners' output).
3. **Differential Expression and downstream analysis [50 Marks]:** The first step performed in differential expression is to perform an overall differential gene expression between the two conditions using tools of your choice. You should include:
    a. A multivariate analysis (e.g. HCA/PCA) showing how good/bad the replicates within each condition are clustered **[5 marks].**
    b. the total number of over- and under-expressed genes for each contrast. Feel free to try different thresholds for logFC and pValue/FDR and justify your choice **[10 marks].**
    c. Perform an intersection of the differentially expressed genes across all contrasts, showing the no. of common and unique differentially expressed

genes at 25°C aw0.995 and aw0.98 vs. control **[5 marks].** Support your results with a Venn diagram **[5 marks].**

d. The main focus of your discussion should be around the treatment impact on the trichothecene genes expression. Your discussion should also report any DE genes within the metabolic pathways related to mycotoxin production (even if some of these don't appear in your top list). Support your results with cross-references from the literature. A good starting point is to report and discuss differential expression for the genes: TRI6, TRI10, TRI15, TRI3, TRI9, TRI13, TRI14 and TRI16 **[25 marks].**

4. **RNA-Seq variant calling [10 Marks]:** As the final step of your analysis, provide an RNA-Seq based variant calling analysis in order to identify SNPs and InDels common between your samples compared to the reference and report potential SNPs that impact your DE genes of interest in Step 3.d. In order to do so, it is recommended that you follow the GATK pipeline best practice documents for Calling variants in RNA-Seq (https://gatk.broadinstitute.org/hc/en-us/articles/360035531192-RNAseq-short-variant-discovery-SNPs-Indels-). Make sure you properly annotate the list of variants according to their corresponding genes and impact as shown during the SNP calling practical.

Tips:

- You do not necessarily need to re-run the mapping step; you can simply use RNA-Seq aligner output files (BAM files) to follow the rest of the protocol.

- You may notice that the *Fusarium* databases available via snpEff do not match the reference that you have received, which means that you must create your own snpEff database. To make things easier for you, I have also provided you with a snpEff database matching the reference genome you will be using (`flang_SNPEff_database.zip`).

- If you are going to use SNPEff for annotation, make sure you are using the fusarium annotation for this (not the tomato one from the practical!).

Part II Analysis pipeline script [20 marks]

5. Develop a script to automate the entire process of reproducing your results. This could be in the form of a job submission script (or a series of scripts, one for each part) on Crescent. The report should also include a reference of all tools/R libraries needed as well as the version used in your analysis.
   Note: Analysis in R could be a separate script and doesn't need to be part of the Crescent job submission.

Submission:

The analysis report (including any appendices and generated plot) and the analysis scripts (qsub + R scripts) are to be submitted via the normal route on Canvas (Please don't include your raw reads, BAMs or any sequencing files within your submission!). Please do not submit the reference genome, or any of your alignment output via Canvas.

General Tips:

1. **You may want to keep a separate backup copy of your analysis as some of the steps are quite time consuming.**
2. **Please consult Crescent2 documentation on the Intranet to learn about queue scripting.**
    a. The more nodes you allocate for your job, the more the waiting time will be. A good rule of thumb is to stick to 1 node and 4-8 Cores max.
    b. As some of the analysis steps would require a long waiting time, a good practice is to manage your time properly across this week, and work on your report while you're waiting for the job to finish.

3. **Additional Tasks: Integration of gene expression with KEGG metabolic pathways will be awarded extra marks (There are some R libraries to do that, and you can even build your own functions).**