NGS Course 2022 - Galaxy server exercises

Introduction

White clover (*Trifolium repens*) is an allotetraploid. This means that it contains genomes originating from two different species within the same nucleus. Normally, white clover is an outbreeding species, but a self-compatible line was used for sequencing the white clover genome (https://academic.oup.com/plcell/article/31/7/1466/5985684). This line is designated S10 in your data, indicating that this is the 10th self-fertilized generation. In addition, you have data from a wild clover accession (ecotype) called Tienshan (Ti), which is collected from Chinese mountains and is adapted to alpine conditions.

From https://link.springer.com/article/10.1007/s00122-021-03955-3

Install IGV from https://software.broadinstitute.org/software/igv/download.

Now, please set up your Galaxy account here:

https://usegalaxy.org/

Next, find the files for the course by accessing the following link https://usegalaxy.org/u/lavinia_fechete/h/ngs2022-1 and import the History to your own account by clicking the "+" sign in the right corner. There should be 29 files in the imported history.

You will be working with two types of sequencing data.

The first is PacBio Hifi reads, which are long and accurate. You can find them under Hifi_reads_white_clover.fastq.

The second type is Illumina RNA-seq reads, which are short and accurate and should be aligned using a spliced aligner. There are 24 of these files, 12 for each of the two genotypes mentioned above. The files are named [genotype]_[treatment]_[replicate].fastq. Treatment 1 is before and treatment 2 is after exposure to frost, respectively.

In addition to the sequencing data, there are also the 4 reference files. Here, you will find three fasta files containing the homoeologous contigs, i.e. two contigs that represent similar regions in the two subgenomes. Contig1 is from the T. occidentale-derived subgenome and Contig2 is from the T. pallescens-derived subgenome. The reference genome was generated using the PacBio Hifi reads mentioned above.

The file white_clover_genes.gtf contains the gene annotations for the two contigs.

Writing the report

In order to get a certificate for the NGS course completion, it is mandatory that you hand in

a Final report describing the analyses carried out during the course.

The submission is done through Brightspace, under Course Tools, Assignments.

When you are writing the report, please remember to read the whole report and edit it

thoroughly before you submit it. It has to be clear, concise and easy to follow - just like any

other scientific publication.

The report should be no longer than five typewritten pages in PDF format and must be

handed in on the 1st of July. It should be structured in the following way:

1. Introduction (short)

2. Results and Discussion

The purpose of the report is to convince us that you carried the data analysis out properly,

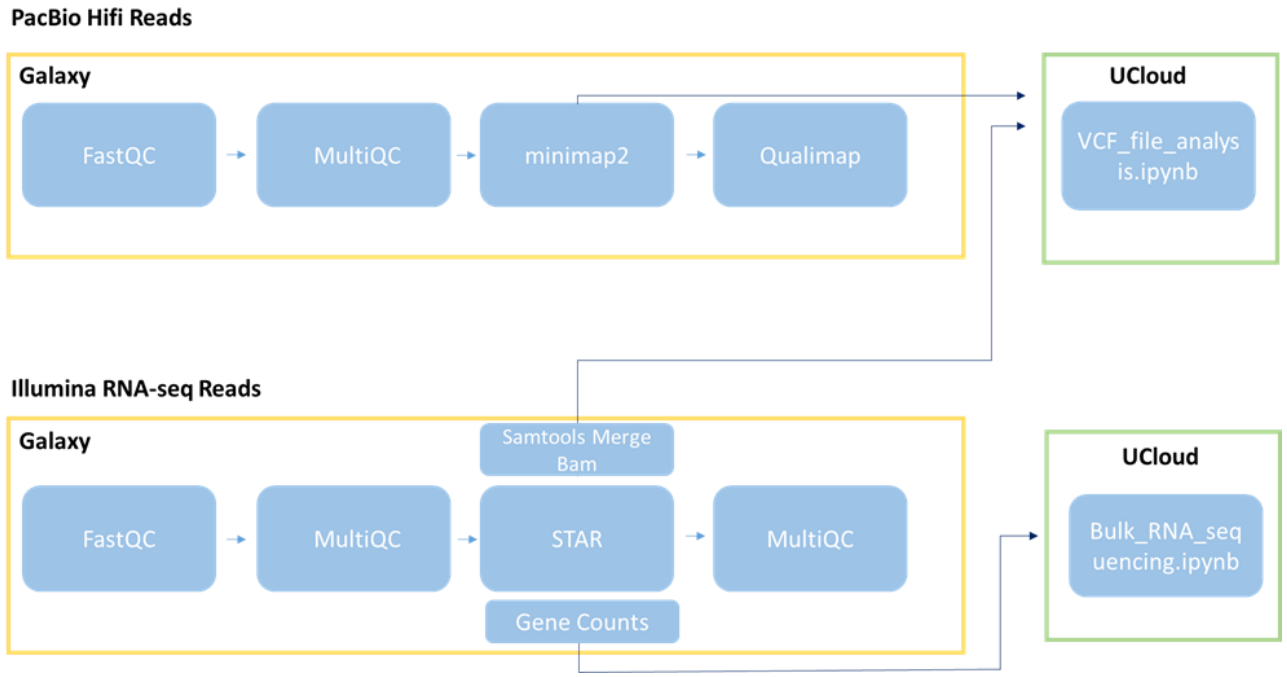
understood what you were doing and that you were able to critically evaluate the methods

used and the results obtained. We are very much interested in your interpretation of the analysis results, so don't be afraid of presenting your thoughts in the report.

Questions in italics should be answered in the report.

EXERCISES

Overview of the exercises



Quality control and mapping

Quality control

Run ***FastQC*** on the PacBio Hifi reads and on two of the Illumina RNA-seq libraries. FastQC does quality control of the raw sequence data, providing an overview of the data which can help identify if there are any problems that should be addressed before further analysis.

Search for "**FastQC** Read Quality reports" under the Tools menu. Under "Raw read data from your current history" select Multiple Datasets and choose the three libraries you want to run. You do not need to change any of the other parameters.

FastQC provides a report for each sample, however, in order to have a better comparison between the Hifi and Illumina data, we would combine the three FastQC reports into one using **MultiQC**, which allows to "Aggregate results from bioinformatics analyses across many samples into a single report".

Once you open the **MultiQC** tool, select FastQC as the tool used to generate the output, and then select the three "RawData" outputs generated by **FastQC**. Visualize the Webpage generated by MultiQC.

Hint: You can find a "Help" button that offers additional information about the plots for each panel. Focus on the following panels: "Per base sequence quality", "Per sequence quality scores"…. (**"**Per base sequence content**"** always gives a FAIL for RNA-seq data).

- *What do you notice with respect to the sequence quality scores? And are there any other quality issues worth noting?*

Hifi data mapping

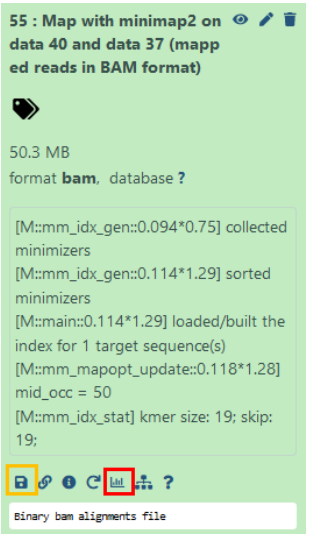
Map the PacBio Hifi reads (Hifi_reads_white_clover.fastq) to the white clover reference sequence (Contig1&2) using ***minimap2 (Map with minimap2)***.

Under reference genome select "Use a genome from history and build index", afterwards you can select the reference. Run two mapping rounds, using two different preset options, "PacBio/Oxford Nanopore read to reference mapping" (map-pb) and the "Long assembly to reference mapping. Divergence is far below 5%" settings (asm5).

It may be a good idea to edit the tasks' names in Galaxy to be more reasonable, such as "map_pb_contig1_2" and "asm5_contig1_2". You can do this by clicking "Edit Attributes".

Next, create reports of the mapping results by running "**QualiMap BAM QC**" on the two obtained BAM files. Visualize the reports generated by QualiMap.

You can also inspect the alignment files in IGV. First, you will need to download the reference fasta sequence **Contigs1&2** and import it into IGV. Click on the Contigs1&2 fasta file and use the Download button marked with yellow on the figure on the right. Go to IGV and use the "Genomes → Load Genome from file" menu and select the relevant fasta file. To view the alignments directly in IGV click on the Visualize button on the selected dataset (marked in red) and choose the "1. display with IGV (local)" option. You need to have IGV opened locally on your computer. Have a look at the polymorphic regions and think about if they are true polymorphisms.



- *What do you observe when comparing the two BAM files? Which setting do you think is the more appropriate to use, and why?*

Next, map the white clover PacBio Hifi reads to **contig1** and **contig2 separately,** using the setting you selected at the previous step. As the two contigs represent the two white clover subgenomes, this mapping will allow you to see the two subgenome haplotypes and call subgenome SNPs. Run QualiMap also on these two bam files and inspect the results.

Visualize all three alignments (HiFi reads mapped to Contigs1&2, contig1 and contig2) in IGV.

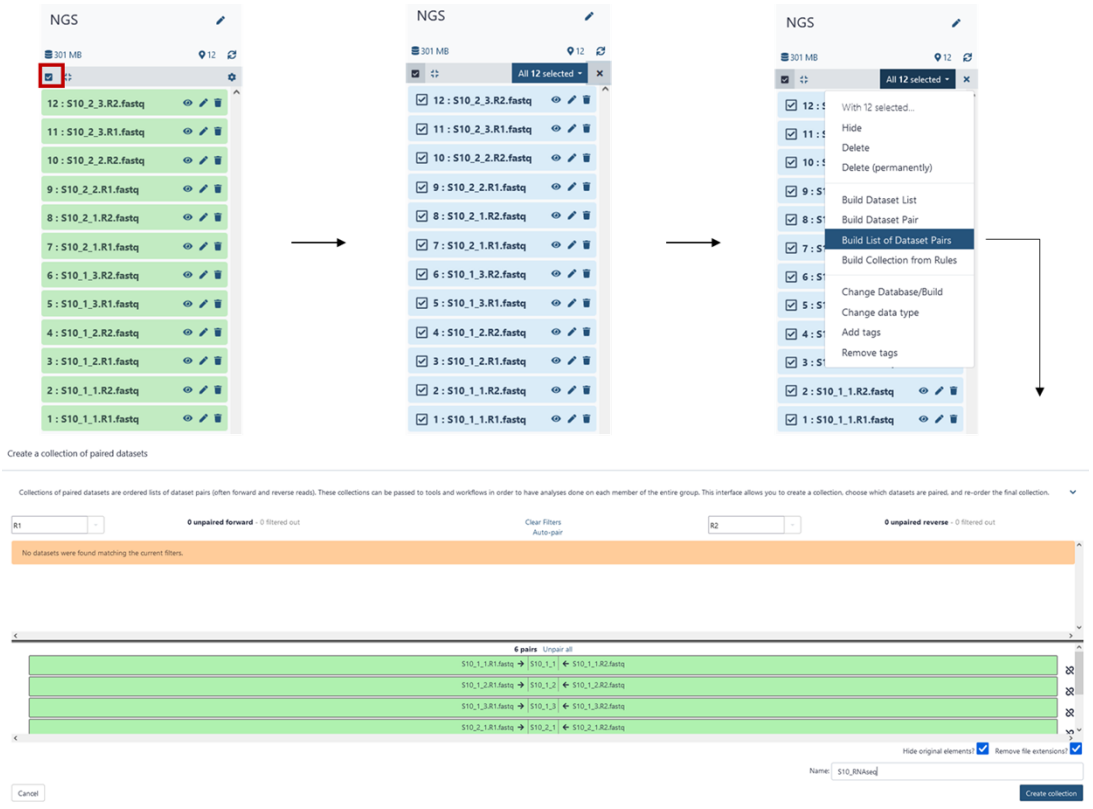
- *Why do you see fluctuations in coverage and large regions without any apparent subgenome SNPs?*
- *What are the major differences between the stats for the reads mapped to Contigs1&2 versus contig1 and contig2? What is your interpretation of the differences?*

Download the three bam files generated in the mapping process (*Contigs1&2, Contig1 and Contig2*) to your computer using the download button.

If you want to load BAM files in IGV directly from your computer, you will also need to index the files. You can index the bam files by running "**bam-to-bai** converter". Note that the bam file and the bai file need to have the same prefix name in the same folder, such as "XXXX.bam" and "XXXX.bam.bai". After naming the downloaded bam and bai properly, go to IGV and use "File → Open" menu and select the bam file that you need to inspect. You can load multiple bam files into the IGV at the same time.

RNA-seq mapping

First, group the 24 RNA-seq libraries into two dataset lists, one list of pairs for S10 libraries and another for Tienshan libraries. This will allow us to work with multiple samples simultaneously. You can do this by selecting the libraries for each genotype and choosing "Build Lists of Dataset Pairs". Make sure that the R1 and R2 libraries for each sample are grouped together.



Map the RNA-seq data as a list of pairs to the reference Contigs1&2 using ***"RNA STAR"***.

First, choose the option Pair-end (as collection). Use "DNA_Contrig1_2" as the reference genome and the white clover annotations (.gtf file) as the "Gene Model".

Here, we need to change a few parameters before running the mapping:

1. As we are working with a very short reference, we need to change the length of the SA pre-indexing string, to 9.

2. Under "Per gene/transcript output" select "Per gene read counts".

3. Note that plant introns are very rarely more than 5000 bp and that you are mapping to two homoeologous contigs that show high similarity, especially in genic regions. You can find the setting for changing the maximum intron size under Algorithmic settings (Extended parameter list).

Running the mapping can take some time, and it should produce four types of files. Note that multiple files are stored in the main output, as we are now working with collections.

- *What are the relevant parameters to change from the default settings, and what did you change them to?*

*Run **MultiQC**,* select run on the 2 log collections generated by STAR and review the output.

Merge the 6 BAM files generated by STAR for each genotype (TI and S10) using ***MergeSamFiles*** and download the two output files BAM to your computer together with their index files (.bai).

You will also need to download the "Per gene read counts" output, we will use this for differential gene expression further on. You can download the whole collection at once using the Download button.

Galaxy can also be used to create an automatic workflow that will map the RNA-seq data, and merge the BAM files. This workflow can be useful when running multiple samples. You can generate a workflow from the analysis already completed in a history, by going to Settings → Extract workflow. You can also create a workflow from scratch using the Workflow editor.

You will now transition from Galaxy to Jupyter Notebooks to continue the data analysis. This gives greater flexibility with respect to the analyses that can be carried out, but also requires a bit more familiarity with writing scripts. Please check the main page in Brightspace for instructions about accessing UCloud (https://cloud.sdu.dk). You should continue the exercises with the VCF_file_analysis.ipynb notebook, followed by the bulk_RNA_sequencing.ipynb notebook.

Your own question

Identify one research question that is not mentioned in the exercise manual or notebooks, but which you find interesting. Write some code in the notebook to answer it and describe both the question, code and answer in the report.

Tables for the report

Based on your analyses using the Jupyter Notebooks, please fill in the Excel tables you can find on Brightspace under the 'Data analysis exercise' and include them in the report.