Graduate School - Sequence analysis
Galaxy tutorial

Barbera DC van Schaik
Bioinformatics and Biomedical Computing
Epidemiology and Data Science
Amsterdam UMC
b.d.vanschaik@amsterdamumc.nl

March 09, 2021

Galaxy (`http://galaxyproject.org/`) is a web portal where several bioinformatics tools, that are usually only executed from the command-line, can be used in a relative easy way. A large part of the available tools focus on the analysis of next generation sequencing data and many tutorials have been shared by the community. For this session we have selected one tutorial to get you started with Galaxy. In the other exercise you will analyse a NGS dataset. It is possible that the Galaxy server is very busy. In that case a demonstration will be given and you can finish the rest of these exercises at home.

1 Getting started

In this tutorial you will see the possibilities of Galaxy without having to write your own analysis programs. Work through the exercises on this page `https://galaxyproject.org/tutorials/g101/` up to "Creating and editing a workflow".

2 Sequence alignment and visualization

The following tutorial walks you through a basic pipeline for analysing a NGS dataset. First you will prepare a sequence dataset for alignment and inspect quality metrics. Next you will align sequences to a reference genome, followed visualization of the results. While you go through the exercises **examine the description of the tools** and have a look at the **description of the datasets** in your history. With the "eye" sign you can see what the data looks like. By clicking on the name of the dataset you can inspect the data format and the amount of sequences/entries that are in your dataset. In this view there is sometimes the possibility to view the data in a genome browser. Furthermore you can download the results. The "i" button gives information about file creation date and the data size.

2.1 NGS: QC and manipulation

In this section you will import a public dataset, convert the dataset into a format that the other tools can use, and create a QC report for the sequences.

1. Go to the public Galaxy web server `https://usegalaxy.org/`
2. Choose "Data libraries" from "Shared data" in the top menu
3. Select "Sample NGS datasets" and import the "Human Illumina dataset" to your current history
4. Go back to "Analyze data" and run the "FASTQ Groomer" tool on the dataset. This will make the dataset suitable for the tools that we will use next.
5. Run the "FastQC: read QC" tool on the groomed fastq dataset and inspect the results.

2.2 NGS: Mapping

There are several aligners available for mapping sequences to a reference genome or other database. In this exercise you will align the dataset against the human genome and visualize the results in a genome browser.

1. Select "BWA-MEM" from the "NGS: mapping" menu
2. You will need to select the right reference genome to compare the sequences with. In this case select the "Human (Homo sapiens) (b37) hg19" reference genome.
3. The dataset is a single-end read experiment, this can be defined in the tool window.
4. ... Time for coffee ...
5. Run "flagstat" (NGS: SAM tools) to get a basic summary of the alignment. What percentage could be aligned to the genome?
6. You can visualize the alignments by choosing "display at UCSC main" when you click on the alignment (BWA-MEM) dataset. Go to the mitochondrial chromosome (type chrM in the search box) and click on the "BWA-MEM" track to inspect the alignments.
7. Have a look at the other tools to see what else you can do with the alignment dataset.

3 Want more?

There are many more tutorials (`https://wiki.galaxyproject.org/Learn`) and screencasts (`https://wiki.galaxyproject.org/Learn/Screencasts`) available for several usage scenarios.