# Assignment #5

## Objective

Introduction to Nextflow

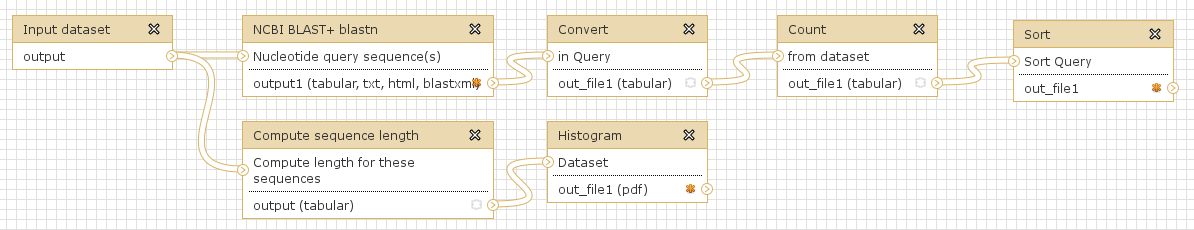
## Due

Before 9am on Tuesday April 21st.

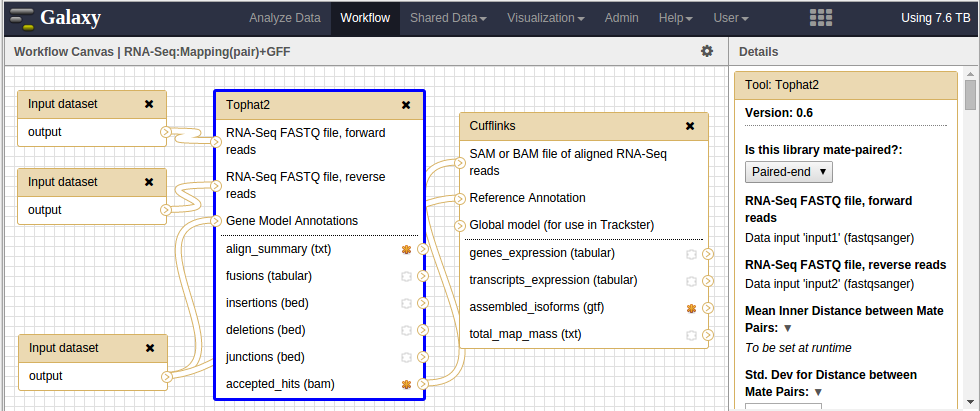
## Introduction

Nextflow (<https://www.nextflow.io/>) is a tool for building computational workflows. A workflow is a set of programs that are executed one after the other with the output of one tool being used for the next tool. A very simple example of a workflow is the case that you’ve already learned where the pipe ‘|’ operator is used on the command-line to feed the output of one tool into another. However, workflows can get a lot more complicated.

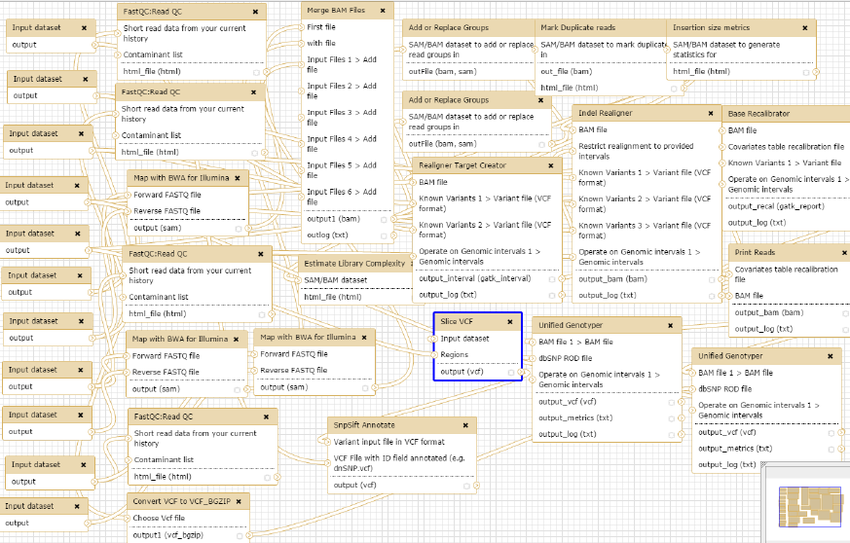
A very popular workflow system is called the Galaxy Project (<https://galaxyproject.org/>). The Galaxy project provides a web-based tool to allow users to stich together programs to execute one after the other using ouput from one tool as input to the other. The following is an example of a simple workflow that executes the NCBI BLAST+ tool for DNA/protein sequence similarity alignment.



In the image above, each box represents some program or script. The “noodles” that connect each program represent the redirection of output from one tool to another. This is similar to piping that we learned about for the command-line tools. Note that the output from the first “Input dataset” tool on the left is used as input into two different tools: one for running BLAST+ and the other for computing the length of all sequences in the input. Unlike using the bar ‘|’ on the command-line a workflow can split and have multiple “parallel” branches such as in the example above. The next image shows another workflow that uses the Tophat2 and cufflinks packages to create a simple RNA-seq alignment workflow. Notice here that the Tophat2 program takes multiple inputs.

Also notice that the Cufflinks tool takes one input from Tophat2 and one input from a different tool (the noodle goes behind the Tophat2 box). Thus, workflows not only branch but branches may also converge, and a tool may receive inputs from anywhere in the workflow.

Workflows are advantageous because they prevent human error. Imagine if you had to repeat the same steps for 100 different input files by hand. The odds of a mistake are high, but also the time you spend manually executing each tool one after the other becomes unmanageable. Consider the case where workflows can get really complicated as in the example below. This type of analysis becomes too much to deal with manually.



Programs such as Galaxy that allow you to build workflows and re-execute those workflows help reduce human-error, reduce the time it takes to execute an analysis, and support reproducibility, which is extremely important in science! Once the workflow is constructed, you can re-use it as many times as needed but you can also share it with others. Provided the input data and tool settings are the same the workflow should generate the same results each time.

## Tasks

For this assignment we will be exploring Nextflow (<https://www.nextflow.io/>) to build a workflow. Unlike Galaxy, NextFlow is purely text-based which makes it appropriate to build workflows to execute on any high-performance computer such as Kamiak (for reference, Kamiak uses the SLURM scheduler). One way that Nextflow improves on reproducibility is using “containers”. (Although Nextflow can work without containers).

Task 1: Read the paper titled “Nextflow enables reproducible computational workflows”.

Task 2: Nextflow was built with the bioinformaticist in mind and has built-in support for common biological data types. As a result, you can find a variety of example workflow for biology here: <https://github.com/nextflow-io/awesome-nextflow>. Take a moment to browse through the workflows to see if anything may be of interest to your own research projects.

Task 3: Lets learn a little about Nextflow by reading the online documentation.

1. Read about the basic concepts of Nextflow here: <https://www.nextflow.io/docs/latest/basic.html>
2. Nextflow uses the Groovy programming language. Just like in Python there are a lot of useful constructs you can make. Take some time to browse through the features that you can use by reading the “Pipeline script” section of the documentation: <https://www.nextflow.io/docs/latest/script.html>. NOTE: Most programming languages have similar concepts and learning a new language is mostly learning a new syntax. You should see a lot here that looks familiar.
3. Like the boxes in the Galaxy images above, a “tool” or “step” is encapsulated in a “process” in nextflow. A process is very similar to a function when coded. Read about processes here: <https://www.nextflow.io/docs/latest/process.html>. NOTE: you do not need to read the entire page. There’s lots to learn about Nextflow and we’ll just focus on the basics. **Only read through to the section titled “Conditional scripts”.**
4. Like the “noodles” in the Galaxy images above, Nextflow, processes send data using what they term as “channels”.
   1. Read the section about “Inputs” here: <https://www.nextflow.io/docs/latest/process.html#inputs>. NOTE: again, let’s just read a portion of this. Only read through to the section titled “Multiple input files”
   2. Read the section about “Outputs here: <https://www.nextflow.io/docs/latest/process.html#outputs>. Only read through to the section titled “Multiple output files”.

There is a lot more to Nextflow workflows than we have time to cover in class! But we will learn enough to help you understand how to get started with them. Next week in class we will explore execution of workflows together.

Task 4: Let’s practice creating a workflow. Follow the step-by-step instructions on the Nextflow Getting Started page: <https://www.nextflow.io/docs/latest/getstarted.html>. After you have completed the tutorial, rerun the Nextflow tutorial script with the --with-report argument. It should generate a file new HTML file. Download this file to your local machine and open it with a web browser. It should show resource usage during execution of the workflow.

You are allowed to work together via Slack: ask questions to try to solve problems you may encounter and help others*.*

Full credit will be awarded for either of the following cases:

1. A finished nextflow script that runs and report HTML file.
2. A partial solution but documented interactions in a Word file listing any efforts you made to work with others that include:

* Questions you asked on Slack
* Responses you gave to others on Slack
* Solutions you found that were already posted on Slack and tagged by you with a thumbs up (or other emoji)

Partial credit will be awarded If you have attempted the assignment and turned in a solution but it does not work and you *do not* have documented interactions on Slack. Remember there is a 5-point per day penalty for each day late.

**What to Turn in**

Please send your completed nextflow script and HTML report file to Dr. Ficklin via Slack.