Assignment 1

Reading

Noble 2009 A Quick Guide to Organizing Computational Biology Projects

Shade and Teal 2015 Computing Workflows for Biologists: A Roadmap

Homework

1. **Write a BASH script that will count the number of reads that start with "GGTCA" in a fastq file**

   Most of the hands-on part of the class this quarter will involve using the unix command line. If you have a linux or mac computer you're good to go. If you have a window computer, your options are:

   1. use a virtual machine (e.g. VirtualBox -- see the example here)
   2. install Cygwin
   3. rely on access to the computer cluster and install Putty.

   If you're not already familiar with unix, I strongly suggest you work though this great tutorial (runs in a browser, no unix/linux needed).

   For a brief tutorial on what bash script is and how to write one, see here or one with some more detail here. Note there are lots of these tutorials, so if neither of these helps you, google!

   For this assignment, use this fastq file from some maize data. It's four orders of magnitude smaller than a real maize sequence file, but good enough for our current purposes. You will need to uncompress it using `gzip`.

   Turn in your finished BASH script and the count of reads.

   Hint: Use `grep` and `wc`

2. **Turn your bash script into a SLURM script and run it on the cluster**

   You will need to setup an account on the Farm computer cluster. Instructions for setting up and account, how to use Farm, and how to write simple SLURM scripts can be found on my lab wiki.

   Turn in your finished SLURM script and the commandline used to run it.

Challenge

Write a script to count how many times all possible thre-nucleotide motifs (i.e. AAA, AAT, AAC ... TTT) are found at the beggining of each read.