CBB520_Fall_2018 assignment 2 Finding SNPS and INDELS

In this course one area we focus on is methods of DNA/RNA sequencing. This is your chance to work with some of that data.

In this assignment, you are asked to produce a program/script/pipeline that will download a certain set of genomic sequence data from the NCBI short read archive and then compare it to a reference genome.

Please, please note the following, as in past years there has been some confusion about this type of assignment. The assignment is to write a program to carry out the various steps, not just to get to some sort of answer like: There are 128,000 SNPs.

Thus a program that works – full credit. Answer to particular questions below, but no functioning program – 0 credit.

Where should you write/host this software? I strongly suggest using one of the Duke OIT VM's that you can sign up for at: vcm.duke.edu. However if you want to host it on a different Linux based machine, where you can likewise run the script from a web page, and the machine is up 24/7 so I can see the assignment, then that would be fine.

The program should be in one of: perl/C/C++/python/PHP, and this whole thing should be running on Linux or some similar operating system like MAC OSX, but NOT ON WINDOWS.

What if you are comfortable working with windows, but not so much with Linux? Then this is a wonderful educational opportunity. Linux is the standard operating system in bioinformatics/genomics, so that is why the assignment needs to be done in that environment.

For this assignment, lets use Illumina data for strain YJF153 available from NCBI as short read archive SRR4841864.
Lets compare this to the S288c reference genome, available from the yeast genome database (www.yeastgenome.org) and from NCBI.

Your program should:

1) Run from your web site (i.e. as a CGI application, or PHP)
2) Download the fastq files from the NCBI SRA archive using the SRA toolkit..

Hint: –split-spots (do not ignore this hint)
3) Identify and report how many high quality pairs are in this dataset, if we

define "high quality" as being each member of the pair has at least 50 consecutive bases of quality score 25+., and what sequence coverage this represents, assuming the genome size is 13 million bases.

4. Carry out a reference genome alignment (versus S288c) Hint: BWA

5. Identify number of SNPs between these genomes based on the BWA analysis. Hint samtools
6. Count up how many of each of the 12 types of SNPs are found between these strains. i.e. A->T; A->C; A->G; C->T . . .
7. Identify the number of single base indels (inserstions/deletions) and count up how many of each type there are. i.e. +A, +C, +G, -C, etc
8. Are the SNPs/single base INDELS uniformly distributed across the genome? Using 10kb windows across the genome count up how many SNPS and INDELs in each window. Calculate the mean and standard deviation. Given the number of windows, for a Poisson distribution how many windows would you expect to have more SNPS/INDELS than 4 standard deviations from the mean? How many are there?
9. There should be a link on you web site where I can see the script of your

CGI/PHP code, if it is not visible from a browser looking at the page source.