Bio430\_exercise1

Cheryl

3/16/2018

## MSCI/BIO430 Bioinformatics Computing Exercise I

# Goals:

1. Practice basic unix commands
2. Log onto an ftp site from the terminal
3. Download files from an ftp site
4. Unzip the files and concatenate (combine) them together into a single file

# Background:

We will eventually be downloading our rockfish sequences from the UCB Genome Center ftp site. The format of RNAseq files is called FASTQ. These are basically really big texts files that provide us the sequence information, including ID, nucleotide sequence, and quality of each nucleotide read from the Illumina HiSeq machine in ASCII format. You can find more info here: <http://en.wikipedia.org/wiki/FASTQ_format>.

You will need to remotely log-in to the UCB ftp site using terminal, download the sequences, and place them into a designated folder on your computer. There will be multiple files for each individual fish that we will have to combine (or concatenate) them into a single file.

Today, we will go through a basic exercise to practice how to do this using human FASTQ sequences from the 1000 genomes project (<http://www.1000genomes.org/data>). Our practice data are real Illumina sequences from an anonymous woman in Great Britain.

## Exercise:

1. Open terminal in RStudio (or terminal if you have a Mac)
2. Navigate to your desktop using cd
3. Make a new directory on your desktop called ‘FASTQ’ using mkdir
4. Navigate to an ftp site where we will download some practice sequence files: You will use the command ftp. Type:

ftp ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/HG00099/sequence\_read/

1. Look at the files in the directory using ls
2. Look for these two files: SRR765993.filt.fastq.gz SRR741412.filt.fastq.gz

Download files from an ftp site using a command called get. [If you wanted to download multiple files that all ended with .gz you would use a command called mget and a wildcard expression: > mget \*.gz]. Today we will download files one by one using get:

ftp> get 'filename'

1. Close the ftp site using quit

ftp> quit

1. Doublecheck that your 2 files are in your FASTQ directory using ls
2. These files are in a compressed format ending with .gz. Use gunzip to unzip each file (or use the following wildcard command \*.gz to download all files in a folder that end with.gz):

gunzip 'filename'

1. Use ls to verify that the files are unzipped (they should now end with .fastq instead of .gz)
2. Now, combine the two files use a command called cat:

cat filename1 filename2 > newcombinedfile.fastq

1. Open the header (top portion) of the file using less
2. Hit q to quit viewing the file
3. Check to see if your concatenation worked using ls -lh to see if the new file is twice the size of the original files (l means long format; h means human readable)

~~BREAK~~~~~

1. Open the header of the file again. You’ll see groups of 4 lines that represent a single sequence. FASTQ files normally use four lines per sequence.

* Line 1 begins with a ‘@’ character and is followed by a sequence identifier and an optional description (like a FASTA title line). Line 2 is the raw sequence letters. Line 3 begins with a ‘+’ character and is optionally followed by the same sequence identifier (and any description) again. Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence.

1. How long are each of the sequences in your concatenated file?
2. Try BLASTing one sequence against the NCBI database to see what gene it is: <http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome>
3. What gene is it?