**Setting up the Azure Cluster for Transcriptomics**

1. Create an instance/virtual machine on Azure using 2 cores (our maximum) and 14 GB of RAM
2. Log into your instance using ssh:
   1. ssh [azureusers@myinstance.cloudapp.net](mailto:azureusers@myinstance.cloudapp.net)
   2. Proved your password
3. Create two directories
   1. bin
   2. data
4. Using wget, to download the *Hosta* data from GenBank
   1. wget ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/SRR/SRR402/SRR402443/SRR402443.sra
   2. Place the data in the data directory
5. Move to the bin directory
6. Use wget to download the SRAToolkit from NCBI’s github
   1. get https://github.com/ncbi/sratoolkit/archive/v2.3.5-2.tar.gz
7. Unpack and unarchive the gzipped tarball
   1. gunzip
   2. tar –xvf
8. sudo apt-get update
9. sudo apt-get install make
10. sudo apt-get install build-essential
11. sudo apt-get install libxml2-dev
12. Move into the SRAToolkit directory
    1. cd sratoolkit-2.3.5-2
13. Run make
14. Since we are having problems compiling SRAToolkit properly, we will just use a compiled executable that I have provided on DropBox (fastq-dump).
15. Use sftp to login into your Azure instance
    1. sftp [azureuser@myinstance.cloudapp.net](mailto:azureuser@myinstance.cloudapp.net)
    2. If you are using PuTTY-SFTP, then you would use the open command instead of sftp.
16. Use the put command to move (or “put”) the file fastq-dump from your personal computer to your virtual machine.
    1. put fastq-dump
    2. This will place fastq-dump in the home directory of your virtual machine.
    3. If you are having problems transferring the program, use right click to identify the file properties and use the full path (address) of the file to move it to your virtual machine.
17. Make sure fastq-dump is executable.
    1. If not, use chmod +x fastq-dump to change it to an executable file.
18. Run fastq-dump on the SRA file
    1. ./fastq-dump srafile-path --split-3
       1. This command telss fastq-dump that these are paired reads.
19. You can put this process (or job) into the background so you can do other work in your virtual machine.
    1. Ctl+Z will stop the job
    2. bg will put it in the background
20. Use wget to download FASTQC while the reads are being unarchived
    1. wget <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/fastqc_v0.11.2.zip>
    2. We are using the already compiled version for Linux to save some time.
21. The Azure instance will not have “unzip” installed. Install it using apt-get
    1. sudo apt-get install unzip
22. Unzip FASTQC
    1. unzip fastqc\_v0.11.2.zip
23. Install Java to run FASTQC
    1. sudo apt-get install default-jdk
24. Run FASTQC on the *Hosta* data set (read 1)
    1. cd into the FastQC directory
    2. run: perl fastqc path-to-read-file-1 &
    3. The symbol, &, will put the job into the background
25. While FastQC is running, download the Trinity software suite.
    1. wget <http://sourceforge.net/projects/trinityrnaseq/files/trinityrnaseq_r20140717.tar.gz>
26. Install the following libraries to allow Trinity to compile
    1. sudo apt-get install zlib1g-dev
    2. sudo apt-get install libncurses5-dev
27. Move into the Trinity program directory and type: make
28. Running trimmomatic:
    1. nohup java -classpath /home/azureuser/bin/trinityrnaseq\_r20140717/trinity-plugins/Trimmomatic-0.32/trimmomatic-0.32.jar org.usadellab.trimmomatic.TrimmomaticPE -threads 2 SRR402443\_1.fastq SRR402443\_2.fastq Hosta.trimmed\_P1.fq Hosta.trimmed\_U1.fq Hosta.trimmed\_P2.fq Hosta.trimmed\_U2.fq ILLUMINACLIP:/home/azureuser/bin/trinityrnaseq\_r20140717/trinity-plugins/Trimmomatic-0.32/adapters/TruSeq2-PE.fa:3:30:10 SLIDINGWINDOW:10:20 MINLEN:40 &