Exercise 1. Alignment with TOPHAT

Part 1. Prepare the working directory.

1. Find out the name of the computer that has been reserved for you (https://cbsu.tc.cornell.edu/ww/machines.aspx?i=57). Everyone should have a BioHPC account to access the computer. The user ID is normally your Cornell NetID. If you do not know the password, using this site to reset the password. http://cbsu.tc.cornell.edu/lab/labpassreset.aspx .
2. Connect to your computer. There is a detailed instruction at http://cbsu.tc.cornell.edu/lab/doc/Remote_access.pdf, (Read the section under "Connection by ssh". There are separate instructions for Windows and Mac users) The host name of the computer is "xxxxxxx.tc.cornell.edu" (replace "xxxxxxx" with the computer name assigned to you).
   To access BioHPC computers from outside Cornell campus:
   a. If you have a Cornell NetID, you can set up VPN before running PUTTY/TERMINAL. (http://www.it.cornell.edu/services/vpn/howto/index.cfm).
   b. If you do not have a Cornell NetID, first use PUTTY/TERMINAL to connect to cbsulogin.tc.cornell.edu. Then from PUTTY/TERMINAL window, type "ssh xxxxxxx" (replace "xxxxxxx" with the computer name) and press return.
3. From the command line, create a working directory by typing the following commands. The "cp" command copies all files for this exercise to your working directory (**replace "my_user_ID" in the commands with your actual user ID**).

```
mkdir /workdir/my_user_ID/

cd /workdir/my_user_ID/

cp /shared_data/RNAseq/exercise1/* ./

ls
```

Part 2. Do quality control on the fastq file

1. Run fastqc on the fastq file

```
fastqc a.fastq.gz
```

2. The fastqc software would create a new directory called “a_fastqc”. You can download that directory to your laptop. To do this, you need the software called FileZilla (Here is the link to install FileZilla on Windows: ftp://cbsuftp.tc.cornell.edu/FileZilla_3.9.0.6_win32-setup.exe , Here is the link to other platforms: https://filezilla-project.org/download.php?show_all=1).
   Instruction to use FileZilla within Cornell campus:
   Host name: xxxxxxx.tc.cornell.edu
   UserName and Password: your user ID and password
   Port: 22
   After click “Quickconnect”, the left panel show files in your laptop, the right panel show files in the remote BioHPC computer. Next to “Remote site” on top of the right panel, enter “/workdir/my_user_ID/” and press “return”. You will see the “a_fastqc” directory and drag it into the left panel.

   Instruction to use FileZilla outside Cornell campus:
   First copy the a_fastqc directory to your home directory. From PUTTY/TERMINAL window, type : “cp -r /workdir/my_user_ID/a_fasta /home/my_user_ID”. Then, use FileZilla:
   Host name: cbsulogin.tc.cornell.edu
   UserName and Password: your user ID and password
   Port: 22
   After click “Quickconnect”, the left panel show files in your laptop, the right panel show files in the remote BioHPC computer. Next to “Remote site” on top of the right panel, enter “/home/my_user_ID/” and press “return”. You will see the “a_fastqc” directory and drag it into the left panel.

3. Open the html file ”fastqc_report.html” on your laptop by double clicking the file.

Part 3. Run alignment software TOPHAT

1. Inspect the files in the working directory (/workdir/my_user_ID. If you are not in working directory already, type “cd /workdir/usre_user_ID” first)

```
ls -l
```

Here are some of the files in the directory.

a.fastq.gz: RNA-seq data from sample a
b.fastq.gz: RNA-seq data from sample b
testgenome.fa: The reference genome file in fasta format.
testgenome.gff3: the gff3 genome annotation files.

testgenome.*.bt2: bowtie2 indexed reference genome, which will be used by the TOPHAT software.

If you are interested in finding out what are the contents in the files, use the following command to check the files. (When inspecting files with "more" command, press "space" to move on to next page, press "q" to exit).

```
gunzip -c a.fastq.gz | more

more testgenome.fa

more testgenome.gff3
```

- The a.fastq.gz is a GZIP compressed file and cannot be inspected directly with the "more" command. A combination of "gunzip -c" and "more" commands are used to inspect the file. Note the pipe character "|" in the middle to connect the two commands.

2. Map the reads to reference genome using TOPHAT.

This FASTQ files are RNA-seq data from two samples. The real RNA-seq data would normally take several hours to process. Special files were prepared for this workshop so that the exercise can be finished in minutes (The files only include reads from first 20mb region from a genome).

Run the following two tophat commands:

```
tophat -o A -G testgenome.gff3 --no-novel-juncs testgenome a.fastq.gz

tophat -o B -G testgenome.gff3 --no-novel-juncs testgenome b.fastq.gz
```

After this step, you will find that two new directories are created: "A" and "B". In each directory, there are two files are needed for next step.

accepted_hits.bam: Alignment results. This file will be used for next steps.

align_summary.txt: Alignment statistics. Use the "more" command to inspect. The numbers are important QC measures. For this exercise, you will see 100% of the reads can be aligned, for real data, you will more likely see the percentage between 70-90%.

As TOPHAT creates alignment results for different samples with the same files name "accepted_hits.bam", to make things easier for next step, it would be a good idea to change the

file names and move to the same directory. Here is the Linux command to change file name and move to a different directory. ("./" refers to current directory)

```
mv A/accepted_hits.bam ./a.bam

mv B/accepted_hits.bam ./b.bam
```

Part 4. Visualize the BAM file

1. Index the bam files
   We are going to use the IGV software to visualize the BAM files. For IGV to read the BAM files, the ".bam" files need to be indexed. We will use the samtools software:

```
samtools index a.bam

samtools index b.bam
```

2. Using FILEZILLA to download the "*.bam" ,"*.bai", "testgenome.fa", "testgenome.gff3" files to your laptop computer.

3. IGV is a JAVA software that can be run on Windows, MAC or a Linux computer. To launch IGV on your laptop, go to IGV, and click "Downloads". You will need to register with your email address for the first. Then click "Launch with 750MB" to download the igv.jnlp file. Double click the igv.jnlp file to start IGV. (After you double click igv.jnlp, it might take a minute or two for IGV to start.)

4. Most commonly used genomes are already in IGV. For this testgenome, we will need to create our own genome database. Click "Genomes"->"Create .genome" file. Fill out the following fields:

   Unique identifier: testgenome

Descript name: testgenome
Fasta: use the “Browse” button to find the testgenome.fa file
Gene file: use the “Browse” button to find the testgenome.gff3 file

Then save the genome database on your computer.

5. From menu “File” -> “Load file”, open the “a.bam” and “b.bam”.
   Inspect the following regions by enter the text in the box next to “Go” and click “Go”.
   chr1:5017000-5026000

Part 5. Run the pipeline as a shell script

We are using a very small data file for this homework. Real data files are much bigger, and normally take a few hours to do alignment. It is not practical to run one command after another. You can need to create a batch command to include all the steps.

In order to do this, you can use a text editor to make a text file with the following lines. We recommend Mac users to use “TextWrangler” (http://www.barebones.com/products/textwrangler/), Windows users can use “Notepad+” (a free software http://notepad-plus-plus.org/) or EditPlus (not free). You can give the script a name, normally with the extension “sh”, e.g.“runtophat.sh”. Then use FileZilla(win & mac) to transfer the file to your home directory.

Here are the lines in your shell script:

```
tophat -o A -G testgenome.gff3 --no-novel-juncs testgenome a.fastq.gz

tophat -o B -G testgenome.gff3 --no-novel-juncs testgenome b.fastq.gz

mv A/accepted_hits.bam ./a.bam

mv B/accepted_hits.bam ./b.bam

samtools index a.bam

samtools index b.bam
```

If the file was created on a windows computer, you will need to convert it to a standard Unix text file.

```
cd /home/my_user_ID

dos2unix runtophat.sh
```

Now, run the shell script by this command in the workdir

```
cd /workdir/my_user_ID

nohup sh /home/my_user_ID/runtophat.sh >& mylog &
```

- Please note that the runtophat.sh script is stored in your home directory, so that you can have a permanent record of the file. We run the script from workdir.
- By using the wrappers “nohup ” and “>& mylog &” before and after the actual command, you can safely disconnect your laptop, and check the results after the run is finished. (use the command “top” to check whether the run is finished, press “q” to exit “top”).