Exercise 1. RNA-seq Read Mapping with TOPHAT and STAR

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=88). 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. Examine qualities of the RNA-seq data files

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 the following command: “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 read mapping software

We are going to use two different read mapping software. TOPHAT is widely used in the early days of RNA-seq data analysis. It is slow but consumes less memory. STAR is much faster, but need a machine with large memory (30GB for human genome).

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

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

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

gunzip -c a.fastq.gz | wc -l

more testgenome.fa

more testgenome.gff3
```

- The a.fastq.gz is a GZIP compressed binary 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.
- "wc" is the command to count the number of lines in a file. "gunzip –c WTa.fastq.gz | wc -l" command would give you the number of lines in the file. As every sequence record takes up 4 lines in the fastq file, the line number divided by 4 gives you the number of sequencing reads in the file.

2. First we will map the reads to a reference genome using TOPHAT.

These FASTQ files are RNA-seq data from two samples. The real RNA-seq data would normally take hours to process. They are special files 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).

First index reference genome with bowtie2-build:

```
bowtie2-build testgenome.fa testgenome
```

Run the following two TOPHAT commands:

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

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

After this step, you will find two new directories created by TOPHAT: "a_dir" and "b_dir". In each directory, there are two files are needed for next step.

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

align_summary.txt: Read 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). We will inspect the .bam files later.

```
mv a_dir/accepted_hits.bam ./TOPHAT_a.bam

mv b_dir/accepted_hits.bam ./TOPHAT_b.bam
```

3. Map the reads to reference genome using STAR.

   On the BioHPC computers, STAR is installed in the directory /programs/STAR. By putting STAR in the path, you can run the software by typing the command without typing the full path of the software. Every time you open a new SSH session, you will need to run this command.

```
export PATH=/programs/STAR:$PATH
```

Then index reference genome with STAR.

```
mkdir STARgenome

STAR --runMode genomeGenerate --runThreadN 2 --genomeDir STARgenome \

--genomeFastaFiles testgenome.fa --sjdbGTFfile testgenome.gff3  \

--sjdbGTFtagExonParentTranscript Parent --sjdbOverhang 49
```

The parameters:

--runMode genomegenerate or read align mode, and default is read alignment

--runThreadN number of threads

--genomeDir output directory of indexed genome file

--genomeFastaFiles reference genome file

--sjdbGTFfile genome annotation file and it should be GTF format.

--sjdbOverhang Normally you can use the value (reads_length -1). It is the length of the genomic sequence around the annotated junction to be used for the splice junctions database

In the next step, we will align sequencing reads to the indexed genome.

```
STAR  --genomeDir STARgenome --runThreadN 2 --readFilesIn a.fastq.gz \
--readFilesCommand zcat --outFileNamePrefix a_ --outFilterMultimapNmax 1 \
--outReadsUnmapped unmapped_a  --outSAMtype BAM SortedByCoordinate


STAR  --genomeDir STARgenome --runThreadN 2 --readFilesIn b.fastq.gz \
--readFilesCommand zcat --outFileNamePrefix b_ --outFilterMultimapNmax 1 \
--outReadsUnmapped unmapped_b  --outSAMtype BAM SortedByCoordinate
```

--genomeDir: reference genome index directory

--runThreadN: number of threads

--readFilesIn: input file

--readFilesCommand zcat: input file is a decompressed .gz file

After running STAR software, many new files have been produced. The two files with the alignment results are a_Aligned.sortedByCoord.out.bam and b_Aligned.sortedByCoord.out.bam

Part 4. Visualize the BAM file.

We will use bam files produced by TOPHAT as an example.

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 TOPHAT_a.bam

samtools index TOPHAT_b.bam
```

After this step, you will see a ".bai" file created for each ".bam" file.

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 web site (http://www.broadinstitute.org/igv/), and download the Windows package or Mac app. You will need to register with your email address for the first. Double click the IGV file to start IGV. (After you double click the file, it might take a minute for IGV to start.) If it complains that you do not have JAVA on your computer, go to https://www.java.com/en/ to get JAVA.

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 exercise. Real data files are much bigger, and normally take a few hours to finish. It is not practical to run one command after another. You can to create a batch command ("a shell script") that includes 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”. If the file is made on a Windows computer, you need to make sure to save the file as a LINUX style text file. From NotePad++, used the "Edit -> EOL Conversion -> UNIX" option.

You can use FileZilla(win & mac) to upload the file to your home directory. To make things easier, both software include a function to directly save edited file to the remote LINUX machine. Here are the lines in your shell script (tophat):

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

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

mv a_dir/accepted_hits.bam ./TOPHAT_a.bam

mv b_dir/accepted_hits.bam ./TOPHAT_b.bam

samtools index ./TOPHAT_a.bam

samtools index ./TOPHAT_b.bam
```

If you are not comfortable with LINUX tools to create shell script. You can use TextWrangler (http://www.barebones.com/products/textwrangler/download.html) on Mac or NotePad++ (https://notepad-plus-plus.org/) to make the shell script, then save/upload them to the remove Linux server. Both software include function for you to directly save edited file to the remote LINUX machine. If the file is made on a Windows computer, you need to make sure to save the file as a LINUX style text file. From NotePad++, used "Edit -> EOL Conversion -> UNIX" option.

To run the shell script:

```
nohup sh runtophat.sh >& log &
```

- 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”).