March 25, 1013

Reference-Based RNA-Seq Data Analysis Workshop, Session 2

Exercise: Using Tophat/Cufflinks/edgeR to analyze RNAseq data

Step 1. One of CBSU BioHPC Lab workstations has been allocated for your workshop exercise. The allocations are listed on the workshop exercise web page:
http://cbsu.tc.cornell.edu/lab/doc/workstations_rnaseq_2013_03.htm

Please consult the PDF file with instructions on how to access and use the Lab workstations for the exercises posted on this page.

If you would like to carry computations outside the workshop you will need to reserve a Linux

workstation of the BioHPC Lab as described in BioHPC user guide
http://cbsu.tc.cornell.edu/lab/use.aspx.

Step 2. Copy the exercise files to the directory /workdir. As /workdir is shared by many users, you will need to create a sub directory under /workdir. You will need the following files:

A_s_2_sequence.txt.gz, A_s_3_sequence.txt.gz, B_s_2_sequence.txt.gz, and B_s_3_sequence.txt.gz: Illumina data files in FASTQ format (samples are from two conditions, A & B; each condition contains two biological replicates, 2&3)

rice7.fa: Rice genome sequence in FASTA format

rice7.gff3: Gene annotation file in gff3 format

```
cd /workdir

mkdir MyUserName

cd MyUserName

cp /shared_data/RNAseq/* .
```

Note: a) Replace MyUserName with your own login user name; b) File names are sometimes very long. To make typing easier, you can type the first few letters then use the “Tab” key to auto-finish the file name; c) * can be used a wildcard for file names in the commands.

Step 3. Build a reference genome index using **bowtie-build2**, and then run **tophat**. You only need to run bowtie-build2 once for the same genome file.

```
nohup bowtie2-build rice7.fa rice7 &
```

In this exercise, we will run tophat for each data file. Since these data files are expected to come from the same transcriptome, we will only need to build the transcritptome sequence files once and reuse them for tophat runs for the rest of the data files.

To do so, we need to run tophat for the first data file with the –G option. A transcriptome sequence file will be built and a Bowtie index will be created. The resulting transcriptome index and the associated data files will be stored under a directory, named "known" (--transcriptome-index <directory>).

```
nohup tophat -p 3 -o A_s_2_guided -G rice7.gff3 --no-novel-juncs --transcriptome-index=known rice7
A_s_2_sequence.txt.gz >& tophat_1.log &
```

For the rest three data files, we can just run tophat by reusing data under the "known" directory (--transcriptome-index). You do not need to run tophat with the –G option again for these files.

```
nohup tophat -p 3 -o A_s_3 --no-novel-juncs --transcriptome-index=./known/rice7 rice7
A_s_3_sequence.txt.gz >& tophat_2.log &

nohup tophat -p 3 -o B_s_2 --no-novel-juncs --transcriptome-index=./known/rice7 rice7
B_s_2_sequence.txt.gz >& tophat_3.log &

nohup tophat -p 3 -o B_s_3 --no-novel-juncs --transcriptome-index=./known/rice7 rice7
A_s_3_sequence.txt.gz >& tophat_4.log &
```

Tophat will produce the same file name for each run. We will use the "mv" command to move each "accepted_hit.bam" file to the current directory and change the file name to its associated data file name.

```
mv A_s_2_guided/accepted_hits.bam A_s_2.bam

mv A_s_3/accepted_hits.bam A_s_3.bam

mv B_s_2/accepted_hits.bam B_s_2.bam

mv B_s_3/accepted_hits.bam B_s_3.bam
```

Step 4. Using **Samtools** to index the bam file, then visualize the alignment using **IGV**.

This step is optional. It is for visualizing the read alignment to the genome . After running samtools index, you will see a new file called A_s_2.bam.bai.

```
samtools index A_s_2.bam
```

You have three options to visualize the results with IGV: 1) run IGV on Lab workstation with VNC; 2) run IGV on Lab workstation with ssh and xming; 3) download output files to your own computer and run IGV there.

1. To run IGV over VNC on the Lab workstation, you need to connect to the workstation using VNC (please consult BioHPC Lab user guide for more info on VNC connections). IGV is accessible as an icon on the left on Linux desktop. Double click the icon to start IGV.

2. To run IGV on Lab workstation with ssh and xming, you need to connect to the workstation using ssh with X11 forwarding enabled. Start Xming on your local computer (or allow external X11 display if your local computer is Linux). Start IGV from the command line in the ssh window.

Please consult BioHPC Lab user guide for more info on ssh/Xming/X11.

3. **To run IGV on local computer,** you will need to copy all the ".bam" files as well as "rice7.fa" and "rice7.gff3" to your local computer.

Run IGV

a) Start the IGV software (depends on the way you run IGV, see above).

b) Import the genome and gene annotation.

IGV provides a number of genomes that are hosted on a server at the Broad Institute. The drop-down menu listed all available genomes hosted on the IGV server. In our case, the correct version of genome (rice v.7) is not available, we will have to load it by the steps listed below:

1) Copy "rice7.fa" and "rice7.gff3" to the directory "/home/ MyUserName/igv/genomes"
2) Click *Genomes>Create .genome File*.
3) Enter an ID and a descriptive name for the genome.
4) Enter the path to the FASTA file for the genome. A genome index file,"rice7.fa.fai" will be created during the import process.
5) Optionally, specify the cytoband file and the annotation (gene) file.
6) If the sequence (chromosome) names differ between your FASTA and annotation files, you might need to create an alias file to provide a mapping between the different names. Certain well-known aliases are built into IGV and do not require an alias file.

7) Select the directory in which to save the genome archive (*.genome) file and click *Save*.

c) Load data

Load the bam and bai files. If you have multiple samples, you can load each one as an individual track.

d) Navigate around IGV

- More information on IGV is available at: http://www.broadinstitute.org/software/igv/

Step 5. Run cuffdiff on the BAM files.

We have 2 samples, A & B, each with 2 replicates, s_2 & s_3. To run cuffdiff, we need to supply corresponding SAM files as a single **comma-separated** list. We will apply upper-quartile normalization (-N option) to the data.

```
nohup cuffdiff  -p 3 -N  rice7.gff3 A_s_2.bam,A_s_3.bam  B_s_2.bam,B_s_3.bam  -o cuffdiff_out
>& cuffdiff.log  &
```

Step 6. Transfer the output files to our local computer.

We will use free software FileZilla to transfer all the cuffdiff output files to our local computer. (To download FileZilla, go to http://filezilla-project.org/download.php?type=client). We will open the files using Excel.

Check output files and find DE genes/isoforms in “gene_exp.diff” and “isoform_exp.diff”.

Step 7. Use cummeRbund to visualize the cuffdiff results. (optional)

More information can be found in http://compbio.mit.edu/cummeRbund/manual_2_0.html.

Step 8. Use edgeR to analyze count data and find DE genes.

As output from cuffdiff cannot be read by EdgeR directly. You will need to convert the output files from cuffdiff into a format that EdgeR can read.

```
cd cuffdiff_out

parse_cuffdiff_readgroup.pl
```

After this step, you will see two new files created: edgeR_count.xls and edgeR_FPKM.xls

You can run EdgeR either on your own computer or the BioHPC computers.

To run EdgeR on your own computer, you will need to have R and Bioconductor R package installed in your local computer (see http://www.r-project.org/).

Here is the instruction to run EdgeR on BioHPC computers. From the terminal, type “R” and press return. Now, you are in R console.

```
library("edgeR")

x <- read.delim("edgeR_count.xls",  row.names='Gene')

group <- factor(c(1,1,2,2))

y <- DGEList(counts=x,group=group)

y <- calcNormFactors(y)

y <- predFC(y,  prior.count.total=2*ncol(y))

pdf("myplot.pdf")

plotMDS(y,  col=c(rep("black",2), rep("red",2)))

dev.off()

quit()
```

You can transfer the pdf file myplot.pdf to your own computer and open it.

A very comprehensive tutorial of edgeR can be found in: http://cgrlucb.wikispaces.com/file/view/edgeR_Tutorial.pdf.

Note that there are a few errors in the edgeR tutorial document.

1. Parameter”dispersion” for function”exactTest” need to specified differently, i.e. dispersion = “auto”,“common”,or "tagwise".

After you feel more comfortable with running the software introduced in our session, it is highly recommended that you read their manual pages:

http://tophat.cbcb.umd.edu/manual.html

http://cufflinks.cbcb.umd.edu/manual.html

http://www.bioconductor.org/packages/release/bioc/html/edgeR.html

Please remember that /workdir is shared by many users, and each workstation has different /workdir.

After you finish a session, you will need to copy the files that you want to keep to your home directory **/home/MyUserName**. Once the files are copied to your home directory, you will be able to access them from any workstations.