Exercise 1. Using Tophat/Cufflinks to analyze the RNA-seq data.

1. A CBSU workstation has been served for all workshop attendants. You should have received an email with the name of the computer that has been reserved for you. The workshop reservation is active for 3 days.

2. Using Putty (Windows) or Mac terminal to start ssh connection to the workstation. The host name is: computer_name.tc.cornell.edu. If you do not know how to connect, follow the instruction at http://cbsu.tc.cornell.edu/lab/doc/Remote_access.pdf.

3. Create a work directory on the computer that you reserved. **The work directory must be created under the directory called "/workdir".**

```
mkdir /workdir/ MyUserName
```

Note: Replace MyUserName with your own login user name;

4. Go to your work directory with the "cd" command, and copy the raw data files for this project to your work directory.

We are going to use three data files for this workshop.

Genomic sequence: s1_sequence.txt.gz

RNA-seq sample 1: 001_s2_sequence.txt.gz

RNA-seq sample 2: 002_s2_sequence.txt.gz

The reference genome and gene annotation used for this project is located at /local_data/tair10/. As the /local_data directory is mounted on a local drive, you do not need to copy it over to your workdir.

```
cd /workdir/ MyUserName

cp /shared_data/cbsuworkshop/biohpcapps/*  ./

ls
```

Note:

- To make typing easier, you can type the first few letters then use the "Tab" key to auto-finish the file name;
- "*" can be used as a wildcard for file names in the commands.
- "./" represent the current directory. "../" represent the parental directory.

5. Run RNAseq pipeline.

```
./rnaseq.sh
```

Note:

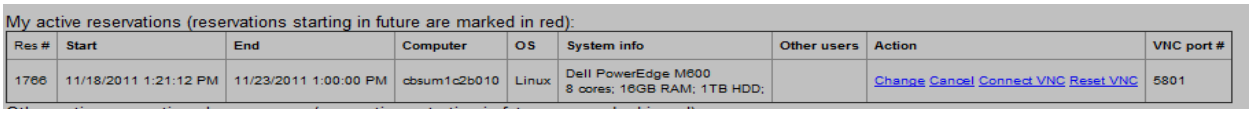
1) The rnaseq.sh is a script. It instructs the computer to run through a list of commands.

2) This script uses the TopHat software to align reads to the reference genome, then use cufflinks to quantify the expression level. You can use the “more rnaseq1.sh” command to check the detailed steps in the script.

6. Check the results.

a. Visualize the alignment with IGV

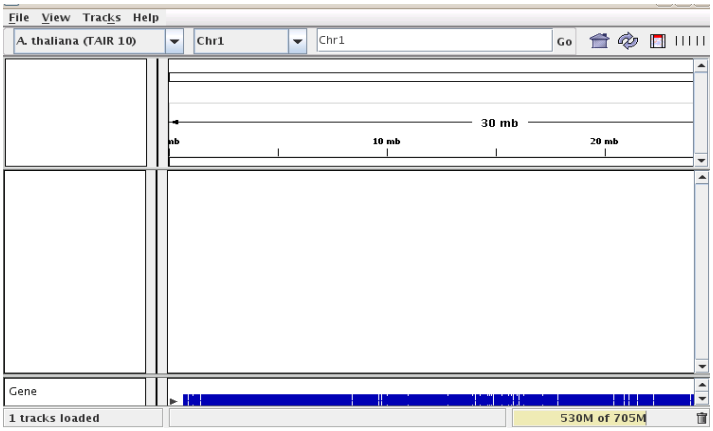
- Download the VNC viewer if you do not have it on your computer.
  Windows: RealVNC http://www.realvnc.com/products/free/4.1/index.html)
  Mac: Chicken of the VNC http://sourceforge.net/projects/cotvnc/
- From a web browser, go to CBSU homepage: cbsu.tc.cornell.edu. Click “CBSU/3CPG BioHPC Lab” -> “My reservations”. Click “Connect VNC”. You will need to write down two things from the table: Computer and VNC port#



My active reservations (reservations starting in future are marked in red):

Res #	Start	End	Computer	OS	System info	Other users	Action	VNC port #
1766	11/18/2011 1:21:12 PM	11/23/2011 1:00:00 PM	cbsum1c2b010	Linux	Dell PowerEdge M600 8 cores; 16GB RAM; 1TB HDD;		Change Cancel Connect VNC Reset VNC	5801

- Open RealVNC, when prompt for server, enter <computer_name>.tc.cornell.edu:<port>. (Replace <computer_name> with the computer name of the my reservation table. Replace <port> with the VNC port#, removing the “580” part from the port#. For example: cbsum1c2b010.tc.cornell.edu:1)
- When prompt for password, enter your bioHPC password.
- You will see the Linux desktop. Double click the “IGV”. (Make sure you only double click once. The response is slow. You might have to wait 10 to 20 seconds before the IGV window open).



- Select “A. thaliana (TAIR 10) from the pulldown menu. Click File->”Load from File”. Navigate to /workdir/ MyUserName /, select s1.bam and s2.bam files.
- Enter Chr1:8000-9000 in the Chromsomal region text box. You can see the reads mapped to the one of the genes. Feel free to try the zooming tool at the top right corner.

b. Open the gene expression level with Excel.

- Using FileZilla (win) or Fetch (mac) to download the files genes.fpkm_tracking and gene_exp.diff under the results directory.

  (host name: <machine_name> .tc.cornell.edu; port number 22 (sftp)).

- Open these files in Excel. The FPKM values are the normalized numbers representing the transcription level.

7. How to process your own data files?

a. Prepare the reference genome. We do have some commonly used reference genomes available on /local_data directory on each workstation. The /local_data directory is on the local drive on each computer. If you do not see your genome under /local_data, you will need to prepare the genome and annotation files before running the pipeline.

First you need to download the genome fasta file and the gff (or gtf) annotation file. If you work on animal genomes, the http://genome.ucsc.edu is a good place to find these files. Using the following command to format the genome fasta file.

```
bowtie-build maize.fa maize &
```

b. Modify the rnaseq.sh script. Download the rnaseq.sh script to your laptop. Open the file in a text editor (Windows users use Wordpad, Mac users s use Text Editor). Modify the input file name in the rnaseq.sh script. If you have more than two samples, you need to repeat the section from “tophat” to “samtools index”. Then upload the file back to the workstation.

Note: If you modify the script on a windows machine, you will need to convert the WordPad edited file into a Unix file.

```
dos2unix rnaseq.sh
```

8. After you feel more comfortable with running these software, it is highly recommended that you read the manual pages for Tophat and Cufflinks. http://cufflinks.cbcb.umd.edu/manual.html and http://tophat.cbcb.umd.edu/manual.html.

Please remember that workstation has a different /workdir. After you finish the analysis, you will need to copy the result files to your home directory /home/YourLogin. Once the files are copied to your home directory, you will be able to access them from any workstations.

Exercise 2. Using BWA/Samtools for SNP/INDEL calling

In this exercise, you will use BWA/Samtools to call SNP/INDELs from whole genome sequencing data.

Alternatively, you can use the BWA/GATK software to call SNP/INDELs, which is used by the 1000 human genome project. (For instructions using GATK , check our workshop web site: http://cbsu.tc.cornell.edu/ww/1/Default.aspx?wid=10)

1. If you skipped exercise 1, you will need to do step 2 to step 4 of the exercise 1

2. Run the script.

```
./callsnp.sh
```

Note:

1) This script uses the BWA software to align reads to the genome, then use samtools to call SNPs.

2) You can use the "more callsnp.sh" command to check the detailed steps in the script. Alignment is done with the BWA software, which output a .sam file. Samtools view is used to convert the SAM file to BAM file. Samtools sort is used to order the BAM file using the chromosome coordinates. Samtools mpileup is used to project the aligned sequence reads to each nucleotides of the reference genome, and output a bcf format. Then bcftools is used to convert bcf to vcf format. Finally, using the vcfutils.pl to filter the results. You can open the output s1.vcf file in excel. If you have multiple samples, mpileup can take in multiple bam files.

4. Check the results.

a. Following the instruction 6a in exercise 1 to visualize the alignments.

b. Download the s1.vcf file, and open in Excel.

Note: There are actually three sub steps in this block.

5. Use the annovar software if you want to annotate the SNP and indels. Following instructions at

```
convert2annovar.pl s1.vcf -format vcf4 > s1.annovar

annotate_variation.pl --buildver at10 s1.annovar /local_data/tair10/at10/
```

e through
s if you need

Note:

1) When you run the 2nd step in this block, you will get lots of error messages. It does not work! The reason is that annovar software requires all chromosome names start with "chr". In the genome sequence file we downloaded from Arabidopsis.org, the chromosome names start with "Chr". This is the one of the most common problems we encountered when running RNAseq or SNP pipelines. In this case, you will need to change the "Chr" in the vcf file to "chr". You can do it by copying the s1.vcf file to your laptop, make modifications in Wordpad or TextEditor, move it back to the workstation.

2) We have created *Arabidopsis* gene annotation files in UCSC format. In the first step in this block, you use the convert2annovar.pl tool to convert the vcf format to annovar format. In the next step, run annotation. The output files are s1.annovar.variant_function and s1.annovar.exonic_variant_function. You can open them in Excel.

After you feel more comfortable with running these software, it is highly recommended that you read the manual pages for BWA and SAMtools. http://bio-bwa.sourceforge.net/bwa.shtml and http://samtools.sourceforge.net/samtools.shtml .

Acknowledgement:

Exercise data for this workshop is provided by Jian Hua lab of the Plant Biology Department.