Exercise 1. Using Tophat/Cufflinks to analyze the RNAseq data.

1. Reserve a Linux workstation of the BioHPC lab, and log on to the workstation. If you have problems of doing this, read the user guide (http://cbsu.tc.cornell.edu/lab/use.aspx), or come to the office hour at 625 Rhodes Hall every Monday afternoon from 2 to 4PM.

Go to the CBSU web site: http://cbsu.tc.cornell.edu/. Mouse over the "CBSU/3CPG BioHPC Lab" label on the menu bar and click "Reservations". Using Putty (Windows) or Mac terminal to start ssh connection to the workstation.

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:

s_1_sequence.txt and s_5_sequence.txt : Illumina data files in FASTQ format

maize.fa: Maize genome sequence in FASTA format

ZmB73_5a.59_WGS.gtf : Gene annotation file in gtf format

```
cd /workdir

mkdir MyUserName

cd MyUserName

cp /shared_data/ cbsuworkshop_2/s*.txt ./

cp /shared_data/ cbsuworkshop_2/maize.fa ./

cp /shared_data/ cbsuworkshop_2/ZmB73_5a_WGS.gtf ./
```

Note: a)Replace MyUserName with your own login user name; b)The file names are 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.

3. Build a reference genome index using bowtie-build, then run tophat. You only need to run bowtie-build once for the same genome file. In this exercise, we will run tophat twice, first time guided by the gtf gene annotation file, second time without the gene annotation.

```
bowtie-build maize.fa maize &

tophat –p 3 -o s1_guided -G ZmB73_5a_WGS.gtf --no-novel-juncs
maize s_1_sequence.txt &

tophat –p 3 -o s1_unguided maize s_1_sequence.txt &

mv s1_guided/accepted_hits.bam s1_tophat.bam
```

Note: Run tophat on s_5_sequence.txt data file. As samtools would produce the same file name for each run. You might want to change to avoid confusion. In this example, I used the "mv" command to change the file name to s1_tophat_sam, and at the same time move the file one directory up.

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

```
samtools index s1_tophat.bam
```

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 s1_tophat.bam.bai. You will need to copy both the s1_tophat.bam and s1_tophat.bam.bai files to your local computer. You will also need the maize.fa and ZmB73_5a_WGS.gtf file to use IGV software.

1) start the IGV software

http://www.broadinstitute.org/software/igv/StartIGV

2) Import the genome and gene annotation.

http://www.broadinstitute.org/software/igv/LoadGenome

(Using the fasta file as the sequence file, the gtf file as the gene file).

3) load data

http://www.broadinstitute.org/software/igv/LoadData

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

4)navigate around IGV

http://www.broadinstitute.org/software/igv/Navigate

5. Run cuffdiff on the BAM files.

```
cufflinks  -G ZmB73_5a_WGS.gtf s1_tophat.bam

cuffdiff ZmB73_5a_WGS.gtf  s1_tophat.bam s5_tophap.bam
```

Note: Use cufflinks if you have only one sample. Use cuffdiff if you have multiple samples.

6. Transfer the output file transcripts.expr to your local computer. You can open this file using Excel. This file gave you the FPKM values for each gene to represent the expression level.

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 /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/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. CBSU is in transition of moving our standard SNP calling pipeline to the GATK package used by the 1000 Human Genome project. GATK has several advantages over Samtools based methods. As we are still testing GATK, we do not have a documentation ready. Talk to us if you have a project now that need to use GATK.

If you did not do the exercise 1, please do the first two steps of exercise 1 before starting exercise 2.

1. Build a reference genome index using bwa index , then run bwa alignment. In this run, we will allow 2 mismatches for reach alignment.

```
bwa index -a bwtsw maize.fa &

bwa aln -n 2 -t 3 maize.fa s_1_sequence.txt > s_1.sai &

bwa samse maize.fa s_1.sai s_1_sequence.txt > s_1.sam &
```

2. Convert the SAM file to BAM file using SAMtools, then sort and index the BAM file.

```
samtools view -bS -o s_1.bam s_1.sam

samtools sort s_1.bam s_1.sorted

samtools index s_1.sorted.bam
```

You can visualize the sorted BAM by following the step 4 in exercise 1.

3. Run mpileup and bcftools to call SNP and INDELs.

```
samtools mpileup -uf bwa/maize.fa s1.sorted.bam |bcftools view -bvcg - > s1.raw.bcf

bcftools view s1.raw.bcf | vcfutils.pl varFilter -D100 > s1.vcf
```

Note: There are actually three sub steps in this block. Samtools mpileup is used to projecting the whole depth of sequencing reads to the each nucleotides of the reference genome, and output in a bcf format. Then using bcftools 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. Use the annovar software if you want to annotate the SNP and indels. Following instructions at http://www.openbioinformatics.org/annovar/ Annovar software requires genomes available through UCSC genome browser. Maize and Arabidopsis genomes are not on the UCSC site. Contact us if you need custom made UCSC database for running annovar.

```
cp -R /shared_data/cbsuworkshop_2/zm2 ./

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

annotate_variation.pl --buildver zm2 s1.annovar zm2/
```

Note: We have created maize gene annotation files in UCSC format. In the first step in this block, you copy the maize annotation files into your working directory. Then use the convert2annovar.pl tool to convert the vcf format to annovar format. In the last 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 Tom Brutnell lab of BTI.