Mapping filtered Tag-Seq data and summing read counts

This short guide explains how to map filtered Tag-Seq fastq files using SHRiMPs gmapper command. Note that there are many other mapping softwares out there, another popular one is bowtie2, and there's no reason why you could not substitute bowtie2 here. However, we have found that gmapper is more sensitive, resulting in higher mapping percentages and greater retention of reads, for a minimal difference in speed. The subsequent step simply sums up the number of reads that mapped to each isogroup (or contig if you choose) in your reference transcriptome.

Run the mapping command on your data subset

1. Make a new copy of your job script file and open it using nano, call it map.sh
2. Then add the following command to the bottom of your job script file, note there is no return, this should all be typed in one line.

```
gmapper clean.fastq /scratch/data/TagSeq/CombinedPoritesCladeA_apr2014.fasta -N 1 -o 2 --
        fastq --strata --qv-offset 33 > mapped.sam
```

3. Then submit your job script file for the test sample

```
qsub -N map map.sh
```

   a. What do all of the options mean? How would you find out?

   b. What output files were produced?

```
tail mapped.sam

grep 'Reads Matched' map.e*
```

   c. Now, lets use our batch submission set-up to run this command on all of your cleaned read files.

```
batch.sh map.sh samples.list maps
```

Summing read counts by isogroup

4. The next step is summing up the number of reads that mapped exclusively to each isogroup (a cluster of transcript isoforms putatively belonging to the same gene). There is no return, type all of this into a single line in your map.sh file, after commenting out the gmapper command

   NOTE: if your reference does not have 'isogroup' designations see Appendix I below.

```
SAMFilterByGene.pl -i mapped.sam -m 0.8 -c p -o filter.sam -p 1 -r g -g
        /scratch/data/TagSeq/CombinedPoritesCladeA_seq2iso.tab > samfilter_out.txt
```

a. Lets again, run this command just for our data subset to look at the output.

```
qsub map.sh
```

b. How would you view the output files?

5. Now, we must run this command for all of our .sam files

```
batch.sh map.sh samples.list counts
```

6. The final step is very simple, we're just concatenating all of these counts files by column to generate a dataframe with each sample as a column and each isogroup as a row. To do this, open up your job.sh file, comment out the prior command and add the following line:

```
CombineExpression.pl */counts.tab > AllCounts.tab
```

a. Then, submit your the job and your counts table will be in your main directory.

```
qsub map.sh
```

b. Look at the beginning of your file - you'll notice the headers are rather long becuase it's just the individual file names. Let's fix that using the substitution command, 'sed'. Before executing this script, what do you think it is doing?

```
cat AllCounts.tab | sed -r 's/\/counts.tab//g' > AllCountsNH.tab
```

c. One more thing. We mapped our reads to a combined host and symbiont reference transcriptome, and we need to split these files into just the host counts. We'll use an inverse grep for this.

```
cat AllCountsNH.tab | grep -v 'kb8' > AllCounts.txt
```

7. That's it! You've just created a table of read counts, which is the standard input format for most, if not all, of the downstream analysis packages designed for differential expression analysis. Most downstream analyses use R, so you will just need to pull this counts table to your computer. Mac or linux users can actually use a terminal to do this, but a GUI client, either WinSCP or Filezilla, will also work for this purpose.
   a. To initiate a transfer, open your client and enter your hostname (which is the IP address of your HPC) and your username (the same that you use for your ssh login) and your password.
   b. Choose SFTP as your file protocol, then save and login.
   c. Open the SFTP connection to our AWS machine, and in the remote panel navigate to your scratch directory where the AllCounts.txt file is located
   d. In your local panel, move to the folder in which you would like to store your text file.
   e. Then just drag and drop.

Appendix I: Adding isogroups to a reference transcriptome

8. Use cd-hit to cluster contigs if your reference transcriptome does not have assembler derived isogroups.

    a. Cluster contigs at 99% similarity taking 30% of length of either the longer or shorter sequence

```
cd-hit-est -i transcriptome.fasta -o transcriptome_clust.fasta -c 0.99 -G 0 -aL 0.3 -aS 0.3
```

    b. Then add these cluster designations to the fasta headers

```
isoogroup_namer.pl transcriptome.fasta transcriptome_clust.fasta.clstr
        >transcriptome_seq2iso.tab
```