Worksheet_9.2-Processing MACS called peak bed files for DAStk analysis
Author: Joe Cardiello (joseph.cardiello@colorado.edu)

DAStk manual: https://github.com/Dowell-Lab/DAStk
DAStk paper: https://www.mdpi.com/1420-3049/23/5/1136/htm

Before you start:

1. Make a directory to store your DAStk related files under /scratch/:

```
mkdir '/scratch/Users/<USERNAME>/ChIPorATAC-seq/dastk/'
```

2. Go to the directory with your ChIP or ATAC-seq scripts you created previously:

```
cd /scratch/Users/USERNAME/ChIPorATACseq/scripts
```

Note: you will need to have peak calls from two data sets in order to use DAStk for comparison of data peak enrichment changes around transcription factor motifs. This means you will need to then run the following script you make to process both of these peak called files.

Processing a Macs2 called peak file for use with DAStk analysis:

3. Rename the template file to a new file for DAStk processing:

```
rsync /scratch/Workshop/SR2019/scripts/template.sbatch
/scratch/Users/USERNAME/<ChIPorATAC-seq>/scripts/DAStk_process_file.sbatch
```

4. Use vim to view the new file::

```
vim /scratch/Users/<USERNAME>/ChIPorATAC-seq/scripts/MACS2_BroadPeaks.sbatch
```

5. Edit the sbatch parameters:
    a. Write a useful job name

```
#SBATCH --job-name=DAStk_processFile  # Job name
```

    b. Enter your email

```
#SBATCH --mail-type=ALL # Mail events
 #SBATCH --mail-user=<YOUR EMAIL>
```

    c. Edit the error and output directory to match your appropriate directory

```
#SBATCH --
output=/scratch/Users/<USERNAME>/e_and_o/DAStkProcess.%j.out

#SBATCH --
error=/scratch/Users/<USERNAME>/e_and_o/DAStkprocess.%j.err
```

d. For processing these files, these settings should work for the files we will analyze:

```
#SBATCH --nodes=
#SBATCH --ntasks=1
#SBATCH --mem=8gb
#SBATCH --partition compute
#SBATCH --time=00:40:00
```

6. Under "LOAD MODULES" enter the following command to load python3 where dastk is installed for this instance:

```
module load python/3.6.3
```

7. Under 'Job Specific Variables' section, enter the following variables:
   a. Provide the directory where the MACS2 derived bed files were saved:

```
BEDS='/scratch/Users/<USERNAME>/ChIP/macs2/'
```

   b. Provide a path to the new output DAStk directory you made:

```
OUTDIR='/scratch/Users/<USERNAME>/ChIP/dastk/'
```

   c. Provide a directory with the curated transcription factor motifs of interest, or all well established motifs for the genome you're studying

```
MOTIF='/scratch/Workshop/hg38/best_curated_Human_TFs_p1e-
6_grch38'
```

   d. Provide a rootname to indicate that the file to be processed will be written after the sbatch script name

```
ROOTNAME=$1
```

   e. Provide the genome of the sample

```
GENOME='hg38'
```

8. For running the process ATAC/ChIP seq DAStk process:
   a. Begin with the process_atac command, even if using ChIP-seq data:

```
process_atac \
```

   b. For this example we are analyzing just one chromosome of data, so indicate that we will just use one thread (if using larger files, more threads likely helpful)

```
--threads 1 \
```

   c. Provide the species genome

```
--genome $GENOME \
```

d. Provide the peak file along with the ending of your peak called file
*** this ending may vary depending on whether you ran broad or narrow MACS2 peaks calling

```
--atac-peaks ${BEDS}/${ROOTNAME}_chr1_peaks_clean.broadPeak \
```

e. Provide the path to the motif directory

```
--motif-path $MOTIF \
```

f. Provide your output directory to the function:

```
--output $OUTDIR
```

9. Save your script!

ESC then: :wq!

10. Run the script on the DMSO ChIP-seq sample!

```
sbatch DAStk_process_file.sbatch SRR5855055
```

11. Also run it for the largazole treatment sample, so we can use DAStk later!

```
sbatch DAStk_process_file.sbatch SRR5855056
```