DIFFERENTIAL GENE EXPRESSION ANALYSIS

Module 4:

PRE-PROCESSING FORMAT CHECK OF .GFF ANNOTATION AND .FASTA GENOME

```
# LOOK AT THE FORMAT OF A GFF (ANNOTATION) FILE

https://en.wikipedia.org/wiki/General_feature_format

# OPEN GFF FILE AND VIEW

cd /home/$USER/DGE_Virtual/human_reference

less gencode.v29.chr_patch_hapl_scaff.annotation.gff3

# ENSURE HEADERS IN GENOME AND ANNOTATION FILES ARE IDENTICAL

# Once you have downloaded the FASTA & GFF3 files for your genome of interest it is very
# important to make sure the sequence region names are the same in both files.

# For example, if Chromosome 1 is named "CHR1" in the FASTA file and "Chr1" or just "1" in the GFF3
# file or vice versa then there will be issues during the downstream analysis.

# For a genome .fasta file:

grep "^>" GRCh38.p12.genome.fa | sort | head
>chr10 10
>chr1 1
>chr11 11
>chr12 12
>chr13 13
>chr14 14
>chr15 15
>chr16 16
>chr17 17
>chr18 18

# For an annotation .gff file,

grep -v "^#" gencode.v29.chr_patch_hapl_scaff.annotation.gff3| awk '{print $1}' | sort | uniq | head
chr1
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18

# Once you check the headers, can you use a simple command to count the different features in
# a gff file?
# Hint: Use grep!
```

EXERCISE: Find which genome/annotation pair show header mismatch

```
cd /home/elavelle/DGE_Virtual/exercise_genomes/

# There are four files in the above folder

# GCF_000006745.1_ASM674v1_genomic.fna
# GCF_000006745.1_ASM674v1_genomic.gff

# GCF_000013425.1_ASM1342v1_genomic.fna
# GCF_000013425.1_ASM1342v1_genomic.gff

# Identify the genome/gff pair that have mismatched sequence headers.
```

GENERATING READ COUNTS USING FEATURECOUNTS

```
# GENERATE READ COUNTS FOR HISAT2 ALIGNMENTS

/home/$USER/DGE_Virtual/

mkdir hisat2_featureCounts/

source activate featurecounts

cd /home/$USER/DGE_Virtual/hisat2_alignments/

featureCounts --help

# USE THE FOLLOWING FLAGS WHEN RUNNING "featureCounts"

# -a input .gtf/.gff file
# -o create an output .txt file
# -T number of threads
# -t feature (exon, gene, etc)
# -g attribute (choose appropriate information on gene id so you can use that in pathway analysis)

featureCounts \
-a /home/$USER/DGE_Virtual/human_reference/gencode.v29.chr_patch_hapl_scaff.annotation.gff3 \
-o /home/$USER/DGE_Virtual/hisat2_featureCounts/count_matrix.tsv \
-T 4 \
-t gene \
-g gene_name \
*.sam
```

PRE-PROCESSING FEATURECOUNTS OUTPUT FOR DESEQ2

```
# featureCounts put in an extra row and some columns we want to get rid of before doing
# differential expression analysis.

less count_matrix.tsv

# Use tail to take every row starting with the second, then extract only the columns of interest.

tail -n +2 count_matrix.tsv | awk '{print $1 "\t" $7 "\t" $8 "\t" $9 "\t" $10 "\t" $11 "\t" $12\
"\t" $13 "\t" $14 "\t" $15 "\t" $16 "\t" $17 "\t" $18}' > DESEQ2_matrix.tsv

mkdir /home/$USER/DGE_Virtual/DESEQ2

mv DESEQ2_matrix.tsv /home/$USER/DGE_Virtual/DESEQ2/
```