BINF 8211 - Homework 1

K. Bodie Weedop

1. Search the UCSC genome browser at [https://genome.ucsc.edu/] and answer the following questions for the human gene TP53, using the hg38 human genome.

1. How many transcripts are reported by each annotation database of the NCBI RefSeq genes (curated subset) and the GENCODE V32?

    - NCBI RefSeq genes (curated subset): 15
    - GENCODE V32: 19

    (Answers from visual inspection of genome browser)

2. How many transcripts agree between the two databases?

    - 4

```
# Get columns I need to check for agreement
cat tp53Hg38Gencodev32 | cut -f2-10 > gencodeColumnsNeeded
cat tp53Hg38RefSeqCurated | grep -w "TP53" | cut -f3-11 > refSeqColumnsNeeded

# Sort the two files that we just created
cat gencodeColumnsNeeded | sed 's/\t/**/g' | sort > gencodeColumnsNeeded.sorted
cat refSeqColumnsNeeded | sed 's/\t/**/g' | sort > refSeqColumnsNeeded.sorted

comm -12 gencodeColumnsNeeded.sorted refSeqColumnsNeeded.sorted | wc -l
```

3. Choose a transcript from the NCBI RefSeq gene annotation: indicate the transcript ID and answer the following question about this transcript: How many exons are UTR and how many exons are coding?

    - Transcript ID: NM_001276698.2
    - UTR Exons: 3 (7668401, 7670608, 7673206)
    - Coding Exons: 5 (7673534, 7673700, 7674180, 7674858, 7675052)

```
# Get line of the transcript
cat tp53Hg38RefSeqCurated | grep "NM_001276698.2"  > singleTranscript

# Get the start and end coordinates for the coding region
cdsStart=$(cat singleTranscript | cut -f7)
cdsEnd=$(cat singleTranscript | cut -f8)
```

```
# Get the UTR exons
cat singleTranscript | cut -f10 | awk -F',' -v cdsStart=$cdsStart -v
cdsEnd=$cdsEnd 'x=1 {while( x<NF ){ if ( $x<cdsStart || $x>cdsEnd ){ print
$x } x++}} | wc -l'

# Get the CDS exons
cat singleTranscript | cut -f10 | awk -F',' -v cdsStart=$cdsStart -v
cdsEnd=$cdsEnd 'x=1 {while( x<NF ){ if ( $x>cdsStart ){ print $x } x++}}' |
wc -l
```

2. Download the human gene annotation GTF file “Homo_sapiens.GRCh38.102.chr.gtf.gz” from Ensembl at [http://useast.ensembl.org/info/data/ftp/index.html], analyze the file, and answer the questions below:

```
# Download the file from Ensembl
humanGeneAnnotation=Homo_sapiens.GRCh38.102.chr.gtf

# Download file if I don't have it
if [ ! -f "$humanGeneAnnotation" ]; then
  wget ftp://ftp.ensembl.org/pub/release-
102/gtf/homo_sapiens/Homo_sapiens.GRCh38.102.chr.gtf.gz
  gunzip Homo_sapiens.GRCh38.102.chr.gtf.gz
fi
```

1. Which version of the human genome is the GTF file from (hg38, hg19, etc.)?

   - hg38

2. Which annotation version is this file?

   - 102

3. How many genes and transcripts in total are in this GTF file?

   - 292573

```
cat Homo_sapiens.GRCh38.102.chr.gtf | awk -F'\t' '{if($3=="gene" ||
$3=="transcript" ){print $0}}' | wc -l
```

4. How many genes and transcripts are protein-coding?

   - 177062

```
cat Homo_sapiens.GRCh38.102.chr.gtf | awk -F'\t' '{if($3=="gene" ||
$3=="transcript" ){print $0}}' | grep -c "protein_coding"
```