# 4. EN-TEx ATAC-seq data: downstream analyses

Inside epigenomics_uvic, run

```
sudo docker run -v $PWD:$PWD -w $PWD --rm -it dgarrimar/epigenomics_course
```

## ▼ Tasks:

▼ Establish directories for housing bigBed data files and peaks analysis files (ensure files are arranged consistently as per the ChIP seq method).

```
cd ATAC-seq
```

```
mkdir data
mkdir data/bigBed.files
mkdir data/bigWig.files
```

```
mkdir analyses
```

```
# generate metadata
../bin/download.metadata.sh "https://www.encodeproject.org/metadata/?replicates.library.biosample.donor.uuid=d370683e-81e7-473f-8475-7716d027849b&status=released&status=submitted&status=in+progress&assay_slims=DNA+accessibility&assay_title=ATAC-seq&biosample_ontology.term_name=stomach&biosample_ontology.term_name=sigmoid+colon&type=Experiment"

head -1 metadata.tsv | awk 'BEGIN{FS=OFS="\t"}{for (i=1;i<=NF;i++){print $i, i}}'
```

▼ Extract from the metadata, ATAC-seq peaks (bigBed narrow, pseudoreplicated_peaks, assembly GRCh38 (execute within ATAC-seq folder):

▼ bigBed files:

1. IDs: Files for bigBed peak calling (bigBed narrow, pseudoreplicated peaks, assembly GRCh38, the latest file for each tissue).

```
grep -F "bigBed_narrowPeak" metadata.tsv |\
grep -F "pseudoreplicated_peaks" |\
grep -F "GRCh38" |\
awk 'BEGIN{FS=OFS="\t"}{print $1, $11, $23}' |\
sort -k2,2 -k1,1r |\
sort -k2,2 -u > analyses/bigBed.peaks.ids.txt
```

2. Download files bigBed files:

```
cut -f1 analyses/bigBed.peaks.ids.txt |\
while read filename; do
wget -P data/bigBed.files "https://www.encodeproject.org/files/$filename/@@download/$filename.bigBed"
done
```

▼ bigWig files:

1. IDs: bigWig FC files (bigWig files, FC, assembly GRCh38)

```
grep -F "bigWig" metadata.tsv |\
grep -F "fold_change_over_control" |\
grep -F "GRCh38" |\
awk 'BEGIN{FS=OFS="\t"}{print $1, $11, $23}' |\
sort -k2,2 -k1,1r |\
sort -k2,2 -u > analyses/bigWig.FC.ids.txt
```

2. Download files bigWig files:

```
cut -f1 analyses/bigWig.FC.ids.txt |\
while read filename; do
wget -P data/bigWig.files "https://www.encodeproject.org/files/$filename/@@download/$filename.bigWig"
done
```

▼ Confirm that my md5sum values match the ones provided by ENCODE.

```
for file_type in bigBed bigWig; do
# retrieve original MD5 hash from the metadata
../bin/selectRows.sh <(cut -f1 analyses/"$file_type".*.ids.txt) metadata.tsv | cut -f1,46 > data/"$file_type".files/md5sum.txt

# compute MD5 hash on the downloaded files
```

```
cat data/"$file_type".files/md5sum.txt |\
while read filename original_md5sum; do
md5sum data/"$file_type".files/"$filename"."$file_type" |\
awk -v filename="$filename" -v original_md5sum="$original_md5sum" 'BEGIN{FS=" "; OFS="\t"}{print filename, original_
md5sum, $1}'
done > tmp
mv tmp data/"$file_type".files/md5sum.txt

# make sure there are no files for which original and computed MD5 hashes differ

awk '$2!=$3' data/"$file_type".files/md5sum.txt

done
```

> bigBed files:

| | | |
|---|---|---|
| ENCFF287UHP | 46f2ae76779da5be7de09b63d5c2ceb9 | 46f2ae76779da5be7de09b63d5c2ceb9 |
| ENCFF762IFP | f6a97407b6ba4697108e74451fb3eaf4 | f6a97407b6ba4697108e74451fb3eaf4 |

> bigWig files:

| | | |
|---|---|---|
| ENCFF997HHO | 689b9a5828c53a594c75f6534a324a6c | 689b9a5828c53a594c75f6534a324a6c |
| ENCFF841ZHA | 486b00b039875f77111f732b6a076554 | 486b00b039875f77111f732b6a076554 |

▼ For each tissue, conduct an intersection analysis using BED tools (this should be executed within the ATAC-seq folder):

▼ Initially, download the annotation file:

```
mkdir annotation
wget -P annotation "https://www.encodeproject.org/files/gencode.v24.primary_assembly.annotation/@@download/gencod
e.v24.primary_assembly.annotation.gtf.gz"

gunzip annotation/gencode.v24.primary_assembly.annotation.gtf.gz

less annotation/gencode.v24.primary_assembly.annotation.gtf
```

▼ Next, convert bigBed files of ATAC-seq peaks to BED files using the bigBedToBed command:

```
cut -f1 analyses/bigBed.peaks.ids.txt |\
while read filename; do
bigBedToBed data/bigBed.files/"$filename".bigBed data/bed.files/"$filename".bed
done
```

▼ Lastly, carry out the intersection analysis:

```
mkdir analyses/peaks.analysis
```

▼ Determine the number of peaks that intersect with promoter regions:

▼ Download the list of promoters, which are (-2kb, +2kb) from TSS, of protein-coding genes. This should be done inside the annotation folder.

```
wget -P annotation https://public-docs.crg.es/rguigo/Data/bborsari/UVIC/epigenomics_course/gencode.v24.prot
ein.coding.non.redundant.TSS.bed
```

```
cut -f-2 analyses/bigBed.peaks.ids.txt |\
while read filename tissue; do
bedtools intersect -a data/bed.files/"$filename".bed -b annotation/gencode.v24.protein.coding.non.redundant.TS
S.bed -u> analyses/peaks.analysis/peaks.promoter."$tissue".bed
done
```

▼ Results

```
wc analyses/peaks.analysis/peaks.promoter*.bed -l

# 47871 analyses/peaks.analysis/peaks.promoter.sigmoid_colon.bed
# 44749 analyses/peaks.analysis/peaks.promoter.stomach.bed
# 92620 total
```

▼ Determine the number of peaks that are located outside gene coordinates (considering the entire gene body, not just promoter regions):

▼ Generate a bed file with the coordinates of the gene body:

```
# — retrieve gene body coordinates of protein-coding genes (chr, start, end, strand)
# — remove mitochondrial genes (i.e. those located on chrM)
# — move from a 1-based to a 0-based coordinate system.

awk '$3=="gene"' annotation/gencode.v24.primary_assembly.annotation.gtf |\
grep -F "protein_coding" |\
cut -d ";" -f1 |\
awk 'BEGIN{OFS="\t"}{print $1, $4, $5, $10, 0, $7, $10}' |\
sed 's/\"//g' |\
awk 'BEGIN{FS=OFS="\t"}$1!="chrM"{$2=($2-1); print $0}' > annotation/gencode.v24.protein.coding.gene.body.b
ed
```

▼ Intersection between the bed files containing the ATAC-seq peaks and the list of gene body coordinates (it returns a bed file with those ATAC-seq peaks that are outside gene coordinates for each tissue):

```
cut -f-2 analyses/bigBed.peaks.ids.txt |\
while read filename tissue; do
bedtools intersect -a data/bed.files/"$filename".bed -b annotation/gencode.v24.protein.coding.gene.body.bed
-v > analyses/peaks.analysis/peaks.not.body."$tissue".bed
done
```

▼ Results

```
wc analyses/peaks.analysis/peaks.not.body*.bed -l

# 37035 analyses/peaks.analysis/peaks.not.body.sigmoid_colon.bed
# 34537 analyses/peaks.analysis/peaks.not.body.stomach.bed
# 71572 total
```