**# Retrieve from a newly generated metadata file ATAC-seq peaks (bigBed narrow, pseudoreplicated peaks,**

**# assembly GRCh38) for stomach and sigmoid\_colon for the same donor used in the previous sections.**

**# Hint: have a look at what we did here. Make sure your md5sum values coincide with the ones provided by ENCODE.**

**# Create directories for data and bigBed.files**

mkdir data

mkdir data/bigBed.files

**#Download metadata.tsv**

../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=sigmoid+colon&biosample\_ontology.term\_name=stomach&type=Experiment"

**#Check metadata.tsv**

head -1 metadata.tsv

**# Get bigBed narrow, pseudo-replicated peaks,**

**# assembly GRCh38**

**# save the files in the following directory analyses/bigBed.peaks.ids.txt**

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

grep -F ATAC-seq metadata.tsv |\

grep -F "bigBed\_narrowPeak" |\

grep -F "pseudoreplicated\_peaks" |\

grep -F "GRCh38" |\

awk 'BEGIN{FS=OFS="\t"}{print $1, $11}' |\

sort -k2,2 -k1,1 |\

sort -k2,2 -u > analyses/bigBed.peaks.ids.txt

**# just checking…**

**root@4a0841ed0fe3:/home/stdimitr/epigenomics\_uvic/ATAC-seq/analyses# cat bigBed.peaks.ids.txt**

**ENCFF287UHP sigmoid\_colon**

**ENCFF762IFP stomach**

**# Download them from the encode-project**

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

**#Check the integrity of the files with md5sum.txt**

../bin/selectRows.sh <(cut -f1 analyses/bigBed.peaks.ids.txt) metadata.tsv | cut -f1,46 > data/bigBed.files/md5sum.txt

cat data/bigBed.files/md5sum.txt |\

while read filename original\_md5sum; do

md5sum data/bigBed.files/"$filename".bigBed |\

awk -v filename="$filename" -v original\_md5sum="$original\_md5sum" 'BEGIN{FS=" ";OFS="\t"}{print filename, original\_md5sum, $1}'

done > tmp

mv tmp data/bigBed.files/md5sum.txt

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

**# Create a directory tailored to bed.files – then, read the files from the bigBed.peaks.ids.txt and save them to a specific directory for bigbed files**

**# data/bigBed.files**

mkdir data/bed.files

cut -f1 analyses/bigBed.peaks.ids.txt |\

while read filename; do

bigBedToBed data/bigBed.files/"$filename".bigBed data/bed.files/"$filename".bed

done

**# Create a directory for storing peaks.analysis**

mkdir analyses/peaks.analysis

**# For every tissue, run an intersection analysis using BEDTools**

**# 1) the number of peaks that intersect promoter regions**

cut -f-2 analyses/bigBed.peaks.ids.txt |\

while read filename tissue; do

echo "$tissue"

bedtools intersect -a data/bed.files/"$filename".bed -b ../ChIP-seq/annotation/gencode.v24.protein.coding.non.redundant.TSS.bed -u |\

sort -u -k1,1 -k2,2 |\

wc -l

done

------------------------------------------

**Answer:**

**# sigmoid\_colon**

**# 21500**

**# stomach**

**# 22211**

**# 2) the number of peaks that fall outside gene coordinates**

cut -f-2 analyses/bigBed.peaks.ids.txt |\

while read filename tissue; do

echo "$tissue"

bedtools intersect -a data/bed.files/"$filename".bed -b ../ChIP-seq/annotation/gencode.v24.protein.coding.gene.body.bed -v |\

sort -u -k1,1 -k2,2 |\

wc -l

done

-------------------------------------

**Answer:**

**# sigmoid\_colon**

**# 25635**

**# stomach**

**# 25665**