For lecture 7, I chose the database below:

AmtDB: a database of ancient human mitochondrial genomes

This is where you can discover an up-to-date list of the mitochondrial sequences that have been published from ancient DNA samples (aDNA). The database's major focus is on anatomically modern Homo sapiens samples from the late Paleolithic to roughly the Iron Age, with a geographic concentration on the Euroasian region. They make both the mtDNA sequences (in FASTA format) and the sample metadata (ID, date, geolocation, site, culture, mtDNA haplogroup, etc.) accessible in a comma-separated (CSV) text file for download.

Advantage:They have two major updates annually.

Disadvantage:

Their samples are not balanced originate-wise. Therefore, the results of analysis of the samples from this website might be biased towards European.

Lecture 8:

Chromosome 5 of human

wget https://hgdownload.soe.ucsc.edu/goldenPath/hg38/chromosomes/chr5.fa.gz

seqkit stats chr5.fa.gz

file format type num\_seqs sum\_len min\_len avg\_len max\_len

chr5.fa.gz FASTA DNA 1 181,538,259 181,538,259 181,538,259 181,538,259

the minimum, maximum, average and sum are equal.

# Download a GTF file.

wget https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/genes/hg38.knownGene.gtf.gz

# Uncompress the file

gunzip hg38.knownGene.gtf.gz

# How many lines in the file

cat hg38.knownGene.gtf | wc -l

3091269

#How many features per chromosome

cat hg38.knownGene.gtf | cut -f 1 | sort| uniq -c| sort -rn

260686 chr1

211594 chr2

185122 chr3

172067 chr17

170613 chr19

168949 chr11

164447 chr12

...

#How many feature types

cat hg38.knownGene.gtf | cut -f 3 | sort| uniq -c| sort -rn

1486754 exon

827935 CDS

247541 transcript

165544 3UTR

164643 5UTR

110310 start\_codon

88542 stop\_codon

...