CONVERTING SAM TO BAM FORMAT AND SORTING SEQUENCES ACCORDING TO THE GENOMIC COORDINATES

After the alignment is produced, reads are in random order with respect to their position in the reference genome. To examine the alignment with IGV or Tablet bam should be sorted and indexes of bam and fasta files should be created.

Practical work:

 # conda create -n alignment
conda install -c bioconda samtools
conda install -c bioconda tablet

```
# convert sam file to bam
samtools view -o aligned.out.bam aligned.out.sam

# sort bam file according to the location of reads on the
reference genome
samtools sort -o aligned.out.sorted.bam aligned.out.bam

# create index of bam file with samtools (it is required by
Tablet to visualize the alignment)
samtools index aligned.out.sorted.bam

# create index of reference genome fasta file
samtools faidx reference_genome.fasta


# get information on the SAM file
samtools flagstat output.sam
# description for flag value 163 can be obtained with:
samtools flags 163
```

##SAM format
https://www.metagenomics.wiki/tools/samtools/bam-sam-file-format
FLAG explanation
https://broadinstitute.github.io/picard/explain-flags.html

Include a screenshot of the sorted.bam file to confirm that the reads were sorted based on the reference genome position.
Which column in the sam or bam file contains the leftmost mapping position?
How to retrieve the unmapped reads from a bam file?

CHECK THE ALIGNMENT WITH TABLET OR IGV

Open sorted.bam file with Tablet, import GTF file and examine some regions of the reference genome, how the reads were aligned. Provide some screenshots.

Practical work:

Additional sources of information:
https://www.biostarhandbook.com/

COUNT THE NUMBER OF READS PER GENE USING FEATURECOUNTS

Bioinformatic tool featureCounts will be used for counting the aligned reads per gene (htseq-count (https://htseq.readthedocs.io/en/release_0.11.1/count.html) is also frequently used).

FeatureCounts is a part of subread package.
http://subread.sourceforge.net/

Before running the featureCounts it should be checked if your dataset is strand-specific or not. You can do this with infer_experiment.py (part of RSeQC package, http://rseqc.sourceforge.net/#).

RSeqQC requires a bam file and a bed file. Bed file can be produced with BEDOPS (gtf2bed function, https://bedops.readthedocs.io/en/latest/content/reference/file-management/conversion/gtf2bed.html).

If you get an error, which suggests that the first line is missing the transcript_id field try with the following solution:

```
$ awk '{ if ($0 ~ "transcript_id") print $0; else print $0"
transcript_id \"\";"; }' input.gtf | gtf2bed - > output.bed
```

(solution obtained from https://www.biostars.org/p/206342/)

(The above mentioned method for converting bam to bed doesn't work for geneBody_coverage.py which requires a bed file with 12 columns. A solution for this is provided here: https://gist.github.com/gireeshkbogu/f478ad8495dca56545746cd391615b93 using the ucsc-genepredtobed tool)

bedparse
https://bedparse.readthedocs.io/en/stable/Usage.html
$
All mentioned tools can be installed using conda package manager.

Practical work:

```
conda install -c bioconda bedparse
conda install -c bioconda rseqc
conda install -c bioconda subread


$ bedparse gtf2bed genomic.gtf > genomic.bed

$ awk '{ if ($0 ~ "transcript_id") print $0; else print $0" transcript_id
\"\";"; }' genomic.gtf | gtf2bed - > output.bed

# infer_experiment.py (a part of RseQC package)
infer_experiment.py -r output.bed -i aligned.sorted.bam

Results of infer_experiment.py
This is PairEnd Data
Fraction of reads failed to determine: 0.0398
Fraction of reads explained by "1++,1--,2+-,2-+": 0.0238
Fraction of reads explained by "1+-,1-+,2++,2--": 0.9365

Compare your results with the examples of paired-end non strand specific
and paired-end strand specific dataset
http://rseqc.sourceforge.net/


# count number of fragments per gene using a featureCounts tool
featureCounts -a genomic.gtf -o readCounts.tsv -s 2 -p aligned.sorted.bam

Summary statistics of featureCounts
//================================= Running
=================================\\
||
||
|| Load annotation file genomic.gtf...                 ||
||    Features : 313952
||
||    Meta-features : 32833
||
||    Chromosomes/contigs : ?
||
||
||
|| Process BAM file SRR58310_Aligned.sorted.bam...                 ||
||    Strand specific : reversely stranded
```

```
||
||    Paired-end reads are included.
||
||    Total alignments : 318069
||
||    Successfully assigned alignments : 285182 (89.7%)
||
||    Running time : 0.01 minutes
||
||
||
|| Write the final count table.
||
|| Write the read assignment summary.
||
||
||
|| Summary of counting results can be found in file
"readCounts.tsv.summary"  ||
||
||
\
\=========================================================================
====//
```

Check which are the 10 most expressed genes. Is there a correspondence with the results of Salmon?

Count matrix can be obtained with “quantifiers” as well. Examples of quantifiers are Salmon (https://combine-lab.github.io/salmon/) and Kallisto (https://pachterlab.github.io/kallisto/).

References:

Liao et al. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. https://academic.oup.com/bioinformatics/article/30/7/923/232889

Conesa et al. 2016. A survey of best practices for RNA-seq data analysis. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8