# Lab 2: STAR, RSeQC, RSEM and Salmon

- What is full form of STAR?
  - Spliced transcripts alignment to a reference
- How much RAM does STAR typically take to align RNA-seq data to a reference human genome?
  - Approximately 30 GB
- What is the difference in .fastq files for RNA-seq data with single-end reads versus paired-end reads?
  - The single-end reads are one fastq file per sample and paired-end reads are two fastq files per sample.
- What are the main arguments of STAR aligner and what do they do?
  - Make genome index:
    - —runThreadN: Number of threads
    - —runMode: genomeGenerate
    - —genomeDir: path to genome directory
    - —genomeFastaFiles: path to genome fasta1 path to genome fasta2 ...
    - —sjdbGTFfile: path to annotations.gtf file
    - —sjdbOverhang: Read length - 1 , but default value of 100 works for most cases
    - if sufficient RAM is not present, then the following also become relevant - - genomeSAsparseD should be 2 or 3, - - limitGenomeGenerateRAM to less than available RAM.
  - Perform alignment:
    - —genomeDir: path to genome index file

▼ —readFilesIn:  path to read1 path to read2

▼ —runThreadN: Number of threads - how many things to run in parallel (keep to half the number of CPUs, but less than total number of CPUs)

▼ —outFileNamePrefix

▼ Which python package apart from fastQC can be used to check quality of RNA-seq data and what additional metrics can it give you?

▼ RSeQC, which is a python package

▼ TIN and medTIN that measure integrity of each transcript and median of transcript integrity number across all transcripts in a sample

▼ Gene body coverage across the length of each gene: how many times each base pair across the length of a gene was covered in the data. If you don’t see much coverage across the 5′ end indicates that there is degradation in the sample.

▼ What input does RSeQC take?

▼ RSeQC takes aligned and sorted BAM files as input and a BED file containing housekeeping genes since you don’t want to look at plots of all genes in the same visualization.

▼ What are the main arguments for running Salmon for alignment?

▼ salmon quant: main command

▼ -i path_to_index which has to be created with salmon

▼ -r path to input .fastq files

▼ -p Number of threads

▼ -l A: detect if data is paired-end or single-end automatically

▼ -o Name of output directory

▼ What input files does Salmon take for alignment?

▼ Salmon takes fastq files as input and an index that we have prepared using Salmon

▼ What is the output of Salmon?

▼ a text file quant.sf that contains, transcript_id, length of the targeted transcript), effective length, counts, TPM

▼ What input does RSEM take?

▼ RSEM runs on BAM files that are output from STAR using three additional arguments `--quantMode transcriptomeSAM --outSAMtype BAM SortedByCoordinate` and a reference that needs to be built using `rsem-prepare-reference`

▼ What are the arguments for running RSEM?

▼ `rsem-calculate-expression` : main command

▼ `--no-bam-output` : to specify that we don't need a bam file as output

▼ `--time` : to specify that we need a .time file giving the runtime

▼ `-p` : specify number of threads

▼ `[input file]` : Name of input file

▼ `[reference_name]` : Reference transcriptome prepared using `rsem-prepare-reference`

▼ `[sample_name]` : Name of output file

▼ What is the output of RSEM?

▼ a text file ending with .isoforms.results that contains transcript_id, gene_id, gene length, effective length, counts, TPM, and FPKM

▼ Which tool can you use to convert between ENSEMBL IDs to HUGO gene symbols?

▼ BioMart