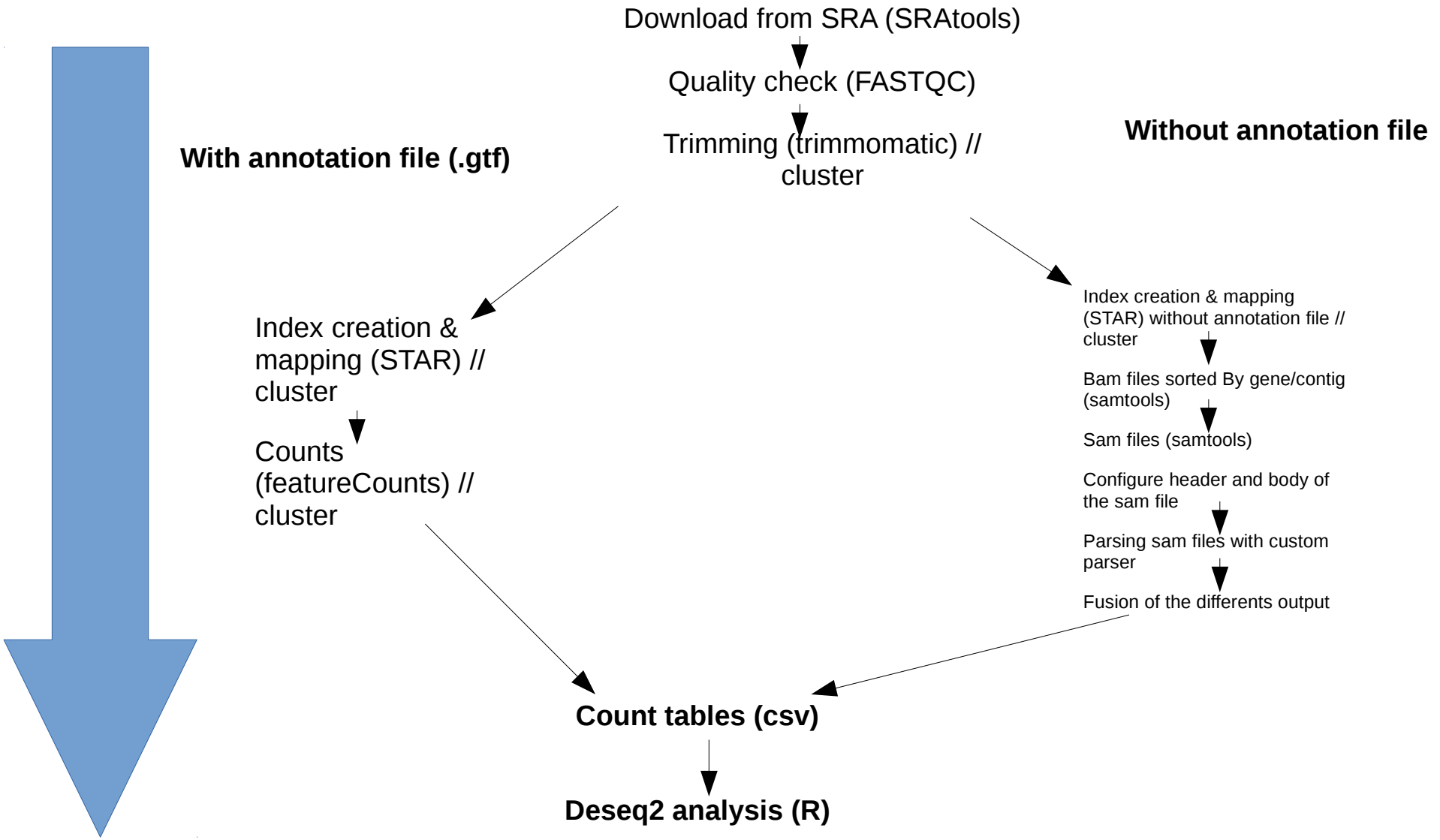
Workflow


Download from SRA (SRAtools)
Quality check (FASTQC)
Trimming (trimmomatic) // cluster
With annotation file (.gtf)
Without annotation file
Index creation & mapping (STAR) // cluster
Counts (featureCounts) // cluster
Index creation & mapping (STAR) without annotation file // cluster
Bam files sorted By gene/contig (samtools)
Sam files (samtools)
Configure header and body of the sam file
Parsing sam files with custom parser
Fusion of the differents output
Count tables (csv)
Deseq2 analysis (R)

Commands

- Download from SRA (SRAtools)

  Install SRAtools and execute

  « fastq-dump SRRXXXXXXX » to download SRRXXXXXXX data set (single end), add

  « --split-files » for paired-end

- Quality check (FASTQC)

  Install FASTQC and execute :

  « fastqc file_name » with filename fasta or fastq file

- Trimming (trimmomatic) // cluster

  install trimmomatic and execute following command to cut the file at X pb :

  « fastx_trimmer -l X -i inputfile.fastq -o outputfile.fastq »

COMMANDS WITH GTF

- Index creation (STAR)

  Execute on cluster :

  > *« STAR --runMode genomeGenerate --runThreadN 40 --genomeDir INDEX_DIR --genomeFastaFiles REFERENCE_GENOME_FASTA_FILE --sjdbGTFfile ANNOTATION_GENOME_GTF_FILE --sjdbOverhang X »*

  with X the (length-1) of your reads

- mapping (STAR) // cluster

  Execute on cluster :

  > *« --runThreadN 40 --twopassMode Basic --outSAMstrandField intronMotif --outFilterMultimapNmax -1 --genomeDir INDEX_DIR --outSAMattributes All --outSAMtype BAM SortedByCoordinate --outFileNamePrefix OUTPUT_FILE_PREFIX --readFilesIn INPUT_FILES_FASTQ »*

  outFilterMultimapNmax : determine the number of multihit (-1 for complete multiple alignement)

- Counts (featureCounts) // cluster

  Install subread package and execute :

  > « featureCounts -T 20 -g gene_id -a ANNOTATION_FILE.gtf -o OUTPUT.txt INPUT.sam(.bam) »
  >
  > To count reads per gene with 20 threads from a BAM or SAM file

COMMANDS WITHOUT GTF (1/2)

- Index creation (STAR) without annotation file // cluster

  Execute on cluster :

  « STAR --runMode genomeGenerate --runThreadN 40 --genomeDir INDEX_DIR --genomeFastaFiles REFERENCE_GENOME_FASTA_FILE --sjdbOverhang X »

  with X the (length-1) of your reads

- Mapping (STAR) without annotation file // cluster

  Execute on cluster :

  « --runThreadN 40 --twopassMode Basic --outFilterMultimapNmax -1 --genomeDir INDEX_DIR --outSAMattributes All --outSAMtype BAM SortedByCoordinate --outFileNamePrefix OUTPUT_FILE_PREFIX --readFilesIn INPUT_FILES_FASTQ »

  outFilterMultimapNmax : determine the number of multihit (-1 for complete multiple alignement)

- Bam files sorted By gene/contig (samtools)

  « samtools sort sample.bam »

- Sam files (samtools)

  « samtools view -h -o out.sam in.bam »

COMMANDS WITHOUT GTF (2/2)

- Configure header and body of the sam file

  Manual configuration (with python script or UNIX commands) for custom parseur

- Parsing sam files with custom parser

  « python3 parseur.py »

  The output is 2 files per file.sam, unique alignement count and multiple alignement count

- Fusion of the differents output

  With custom python script

  « python3 fusion_file.py »

  The output is a count table delimited by tab

COMMANDS

- Organise Count table (R)

- DESeq2 analyses (R, rstudio)

  « dds=DESeqDataSetFromMatrix(countData=countdata,colData = coldata,design=~condition) »

  « deseq=DESeq(dds) »

  « res=results(deseq,name="condition_root_vs_nodule") »

  « resSig=subset(res,pvalue<0.01) »

  « write.csv(resSig,file="resSig.txt") »