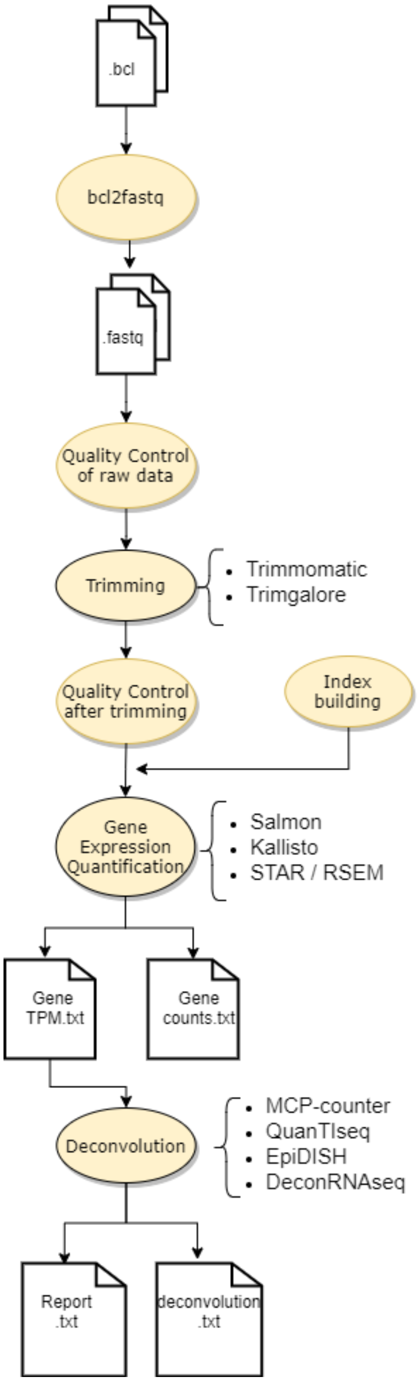
RNASeq/Deconvolution pipeline workflow and use-cases

This is a tutorial for the different possible usages of the RNASeq/Deconvolution pipeline.



After installing according to the README.md, you'll find here how to customize and run the pipeline in detail.

Workflow customization

The configuration file for the pipeline is config.yaml, where you will find all the parameter settings to customize the workflow for your purpose, e.g. BCL to FASTQ conversion, and/or RNA-seq quantification, and/or deconvolution. Depending on your usage, the corresponding parameters should be filled with values (preceded by 2 empty spaces) as described below.

First, you need to give the paths to your input data and the output folder.

- Output_Directory :

Path to the desired output folder. It will be created if not already existing.

- Input :

(1) if you start from BCL files to convert to FASTQ format, give the path to the folder containing your raw Illumina RNASeq data.

(2) If you start from FASTQ files, give the path to the folder containing the FASTQ files.

(3) If you start directly from already quantified RNAseq data to perform only deconvolution, give the path to the TPM counts file.

- THREADS :

The number of threads to use

RNA-seq quantification

If you just want to run deconvolution, you can leave all these following parameters empty and proceed to the next section directly.

BCL to FASTQ conversion

- Convert_bcl2fastq :

If you want to run conversion from raw Illumina BCL to FASTQ files, write "yes" or "no", with the quote marks. This step uses the bcl2fastq from Illumina. See https://support.illumina.com/sequencing/sequencing_software/bcl2fastq-conversion-software.html for more details.

- Sample_Sheet :

This parameter is only required if you chose Convert_bcl2fastq = "yes", in which case, provide the path to the .csv samplesheet file provided by Illumina. You can leave this field empty otherwise.

Quantification

- Do_rnaseq :

If you want to perform the RNA-seq quantification : "yes", otherwise "no".

- Samples :

If you want to perform the RNA-seq quantification (Do_rnaseq = “yes”) starting from FASTQ files, provide here the path to the file listing your samples’ names corresponding to the FASTQ files you want to analyze, one per line. E.g. :

```
Sample_1

Sample_2

Sample_3
```

Note that if you used the Convert_bcl2fastq option “yes”, the pipeline will create this file from the Illumina .csv samplesheet. It will be created under the main directory, where the Snakefile is located.

★ Trimming

- Trim_with :

If Do_rnaseq = “yes”, this is a required parameter for the trimming process. Choose between Trimmomatic or Trimgalore (without quote marks).

- Adapter :

If Do_rnaseq = “yes”, this is a required parameter in case you chose Trim_with = Trimmomatic for the trimming process. Please provide the path to the fasta Illumina adapter file. E.g. : TruSeqAdapt.fa

★ Quantification method

- Quantification_with :

If Do_rnaseq = “yes”, this is a required parameter for indicating which RNAseq quantification method you would like to use : STAR, kallisto or salmon (without quote marks). Note that kallisto and salmon are faster than STAR(+RSEM), as they are pseudo-aligner.

★ Index generation

- Compute_index :

If Do_rnaseq = “yes”, this is a required parameter. Do you want to run index computation? "yes" or "no". The created index file or folder will be required for all downstream RNAseq quantification methods.

- CDNA :

(1) If you chose Compute_index = “yes” and Quantification_with = kallisto, provide the path to the CDNA file. The generated index will be created under data/genome/kallisto_transcript.idx

(2) If you chose Compute_index = “yes” and Quantification_with = salmon, please provide the path to the CDNA file. The generated index folder will be created at the location data/genome/salmon_index

- GTF :

If you chose Compute_index = “yes” and Quantification_with = STAR, please provide the path to the GTF file as it is required to create the STAR index file. The generated index folder will be created at data/genome/star

If you chose Compute_index = “no” and Quantification_with = STAR, please also provide the path to the GTF file as it is required to perform the quantification with the RSEM method.

The GTF file must be provided as a .gtf formatted file (not compressed). E.g : Homo_sapiens.GRCh38_cdna.fa.gz

- Genome :

If you chose Compute_index = “yes” and Quantification_with = STAR or salmon, please provide the path to the Reference Genome file as it is required to create the STAR index file.

If you chose Compute_index = “no” and Quantification_with = STAR, please also provide the path to the Reference Genome file as it is required to perform the quantification with the RSEM method.

The Reference Genome file must be provided as a .fa formatted file (not compressed). E.g. : Homo_sapiens.GRCh38.dna.primary_assembly.fa

Note that index generation can be time-consuming. It is possible to set the option Compute_index to “no” and provide the index path directly to the workflow.

- Index_rnaseq :

If you chose Do_rnaseq = “yes” and Compute_index = “no”, this is a required parameter to provide the index to the pipeline for RNA-seq quantification.

If you are using the kallisto, salmon or STAR quantification method, please provide the path to the kallisto index file, salmon index folder, or STAR index folder, respectively.

Deconvolution

- Do_deconv

If you want to run the deconvolution, write "yes", otherwise "no" and you can leave the following Signatures parameter empty.

- Signatures :

If you chose Do_deconv = yes, please provide the path to the folder containing the signatures. You can use the ones provided with this pipeline under data/signatures.

Run the pipeline

Make sure you have activated the snakemake environment :

```
$ conda activate snakemake
```

If you have done this correctly, you should see (snakemake) at the beginning of your command line. Now simply run :

```
$ snakemake -j [THREADS] --use-conda
```

With [THREADS] corresponding to the number of threads you want to use. Note that this parameter in the command line can be different from the parameter in the config file. For example, if you give 4 threads in the config file and 8 in the snakemake command line, 2 jobs can be executed in parallel.

A. Quality control and trimming

You will find results of the quality controls before trimming in the following folders :

- output/fastqc_raw : all samples individually
- output/multiqc_raw : summary of all samples

And the results of the quality control after trimming in :

- output/fastqc_after_trimming : all samples individually
- output/multiqc_after_trimming : summary of all samples

B. Gene expression quantification

You will find the quantified genes expression in the following files :

- output/gene_counts.txt : quantified genes
- output/TPM.txt : quantified genes normalized in transcripts per million

C. Deconvolution

You will find the results of cell types proportions estimates in the following file :

- output/deconvolution.txt

You will also find a report in .html format displaying a correlation plots among your samples and the different cell types estimates :

- output/HTML_REPORT/analyses.html