ScRNAseq Pipeline overview

ScRNAseq in the Cloud

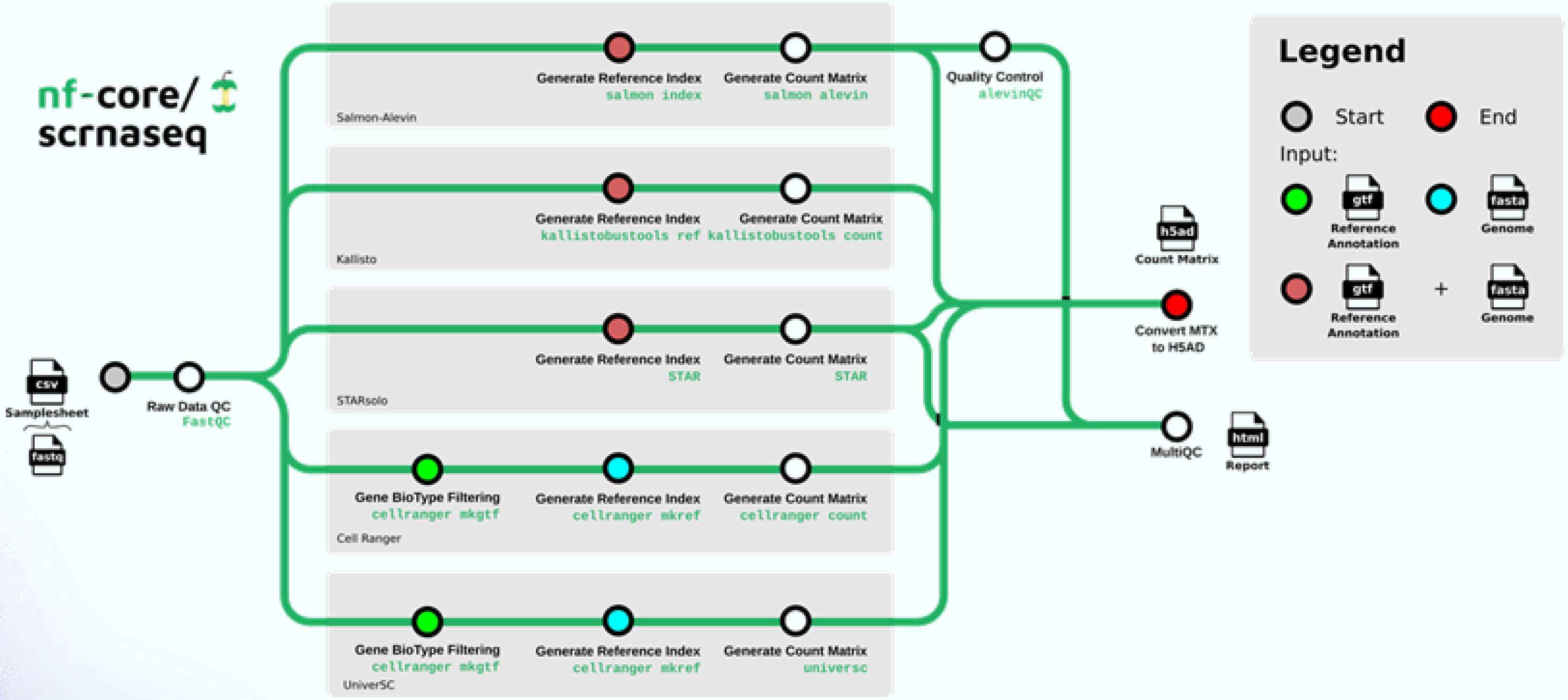
MDIBL Comparative Genomics and Data Science Core

nf-core

MDI Biological Laboratory

Metro Map

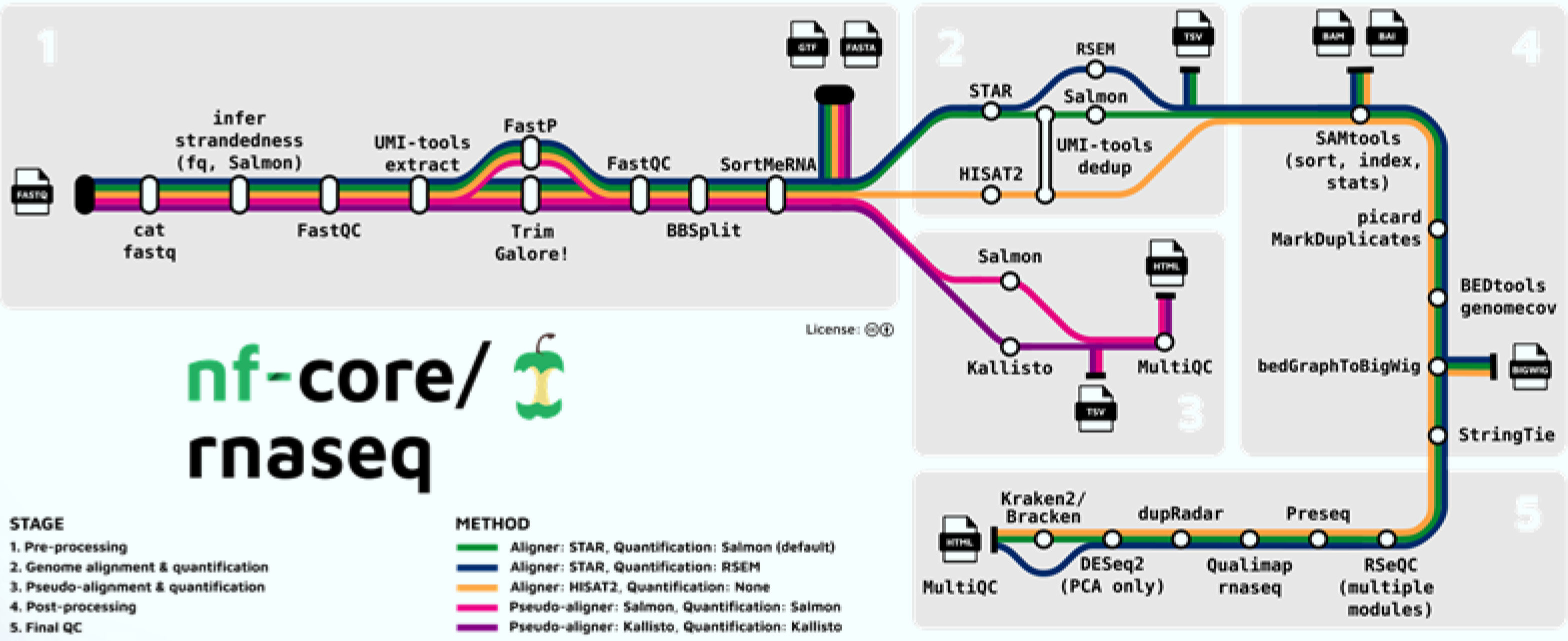
nf-core/
scrnaseq



nf-core

MDI Biological Laboratory

Metro Map (rnaseq)



STAGE

1. Pre-processing
2. Genome alignment & quantification
3. Pseudo-alignment & quantification
4. Post-processing
5. Final QC

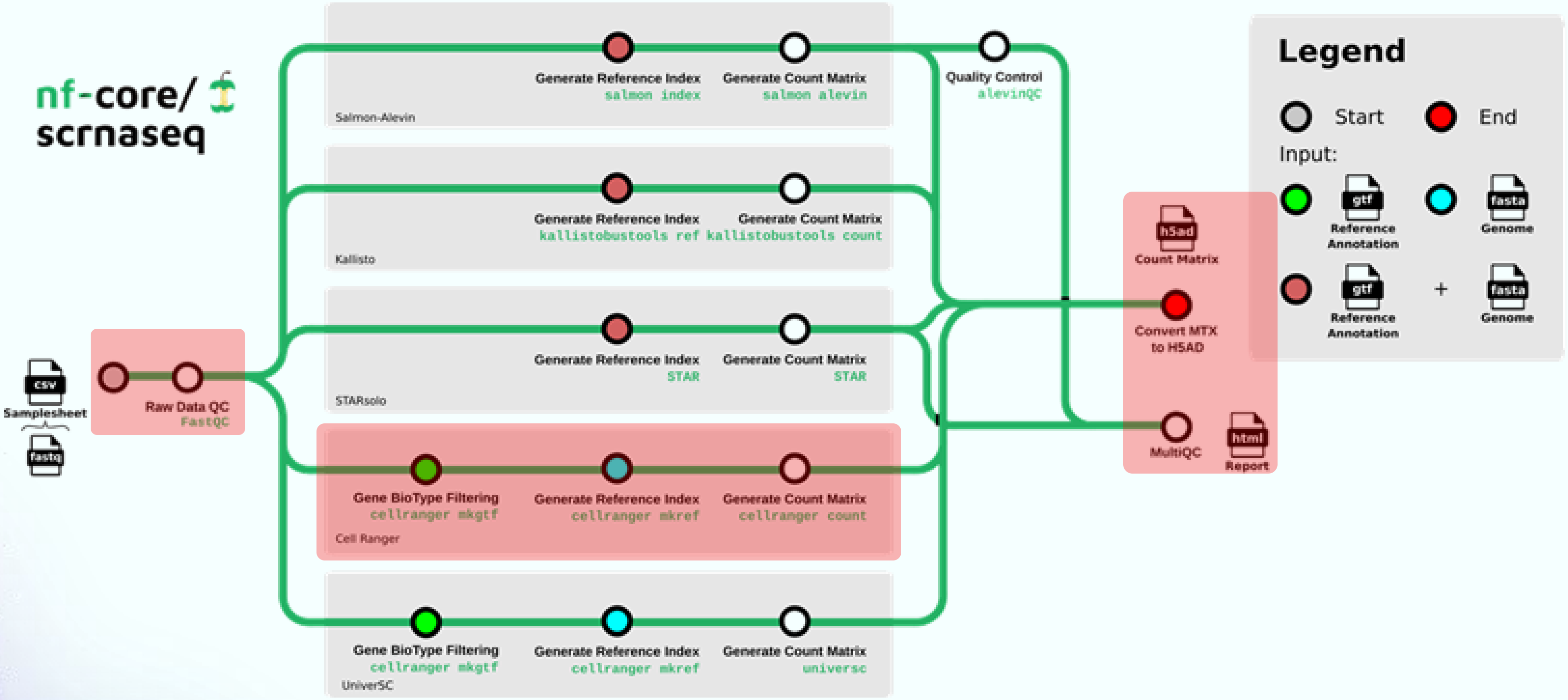
METHOD

- 🟩 Aligner: STAR, Quantification: Salmon (default)
- 🟦 Aligner: STAR, Quantification: RSEM
- 🟧 Aligner: HISAT2, Quantification: None
- 🟥 Pseudo-aligner: Salmon, Quantification: Salmon
- 🟪 Pseudo-aligner: Kallisto, Quantification: Kallisto

nf-core

MDI Biological Laboratory

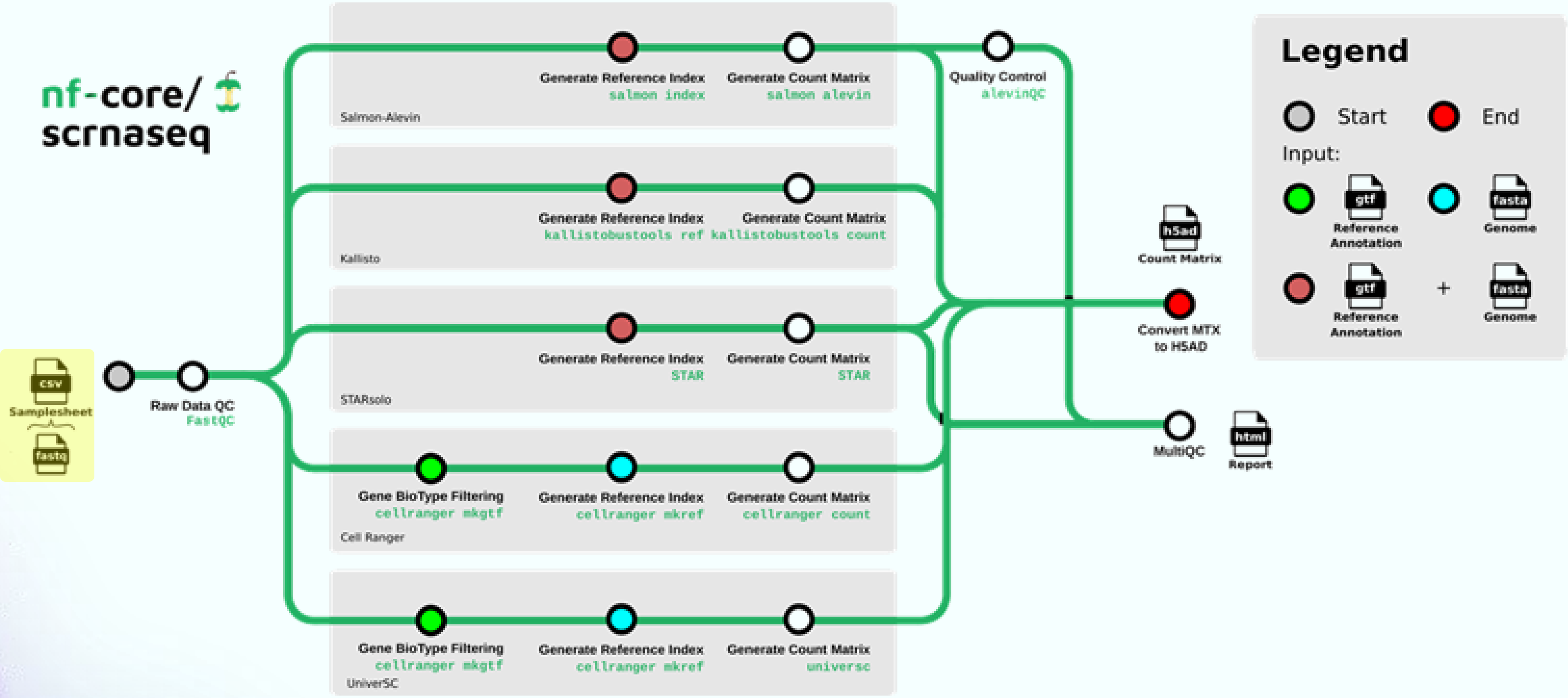
Metro Map



nf-core

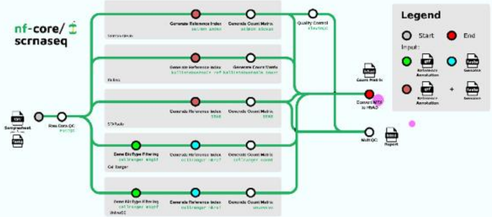
MDI Biological Laboratory

Inputs



nf-core/
scrnaseq
Samplesheet
csv
fastq
Raw Data QC
FastQC
Generate Reference Index
salmon index
Generate Count Matrix
salmon alevin
Salmon-Alevin
Quality Control
alevinQC
Generate Reference Index
kallistobustools ref
Generate Count Matrix
kallistobustools count
Kallisto
Generate Reference Index
STAR
Generate Count Matrix
STAR
STARsolo
Gene BioType Filtering
cellranger mkgtf
Generate Reference Index
cellranger mkref
Generate Count Matrix
cellranger count
Cell Ranger
Gene BioType Filtering
cellranger mkgtf
Generate Reference Index
cellranger mkref
Generate Count Matrix
universc
UniverSC
h5ad
Count Matrix
Convert MTX to H5AD
MultiQC
html
Report
Legend
Start
End
Input:
gtf
Reference Annotation
fasta
Genome
gtf
Reference Annotation
+
fasta
Genome

Inputs



Samples

samplesheet.csv
sampleID,fastq_1,fastq_2

- This defines the samples that will be processed
- Each entry needs a Read 1 and Read 2
- Samples need to be g-zipped.

References

<organism>.fasta.gz

- Text-based file representing nucleotide (or protein) sequences. In our case organized by chromosome.

<organism>.gtf.gz

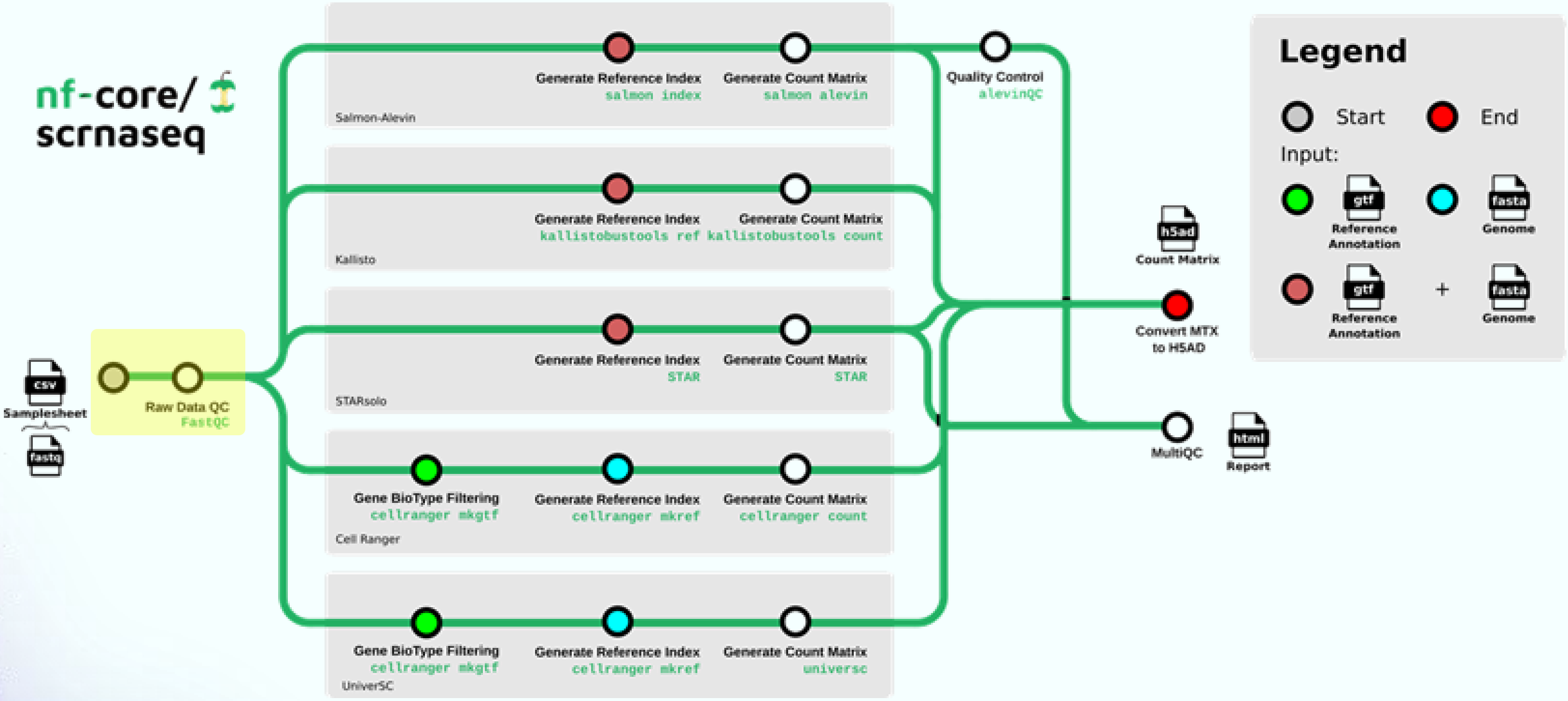
- Gene Transfer Format file describes gene structure information specifically location of genes.

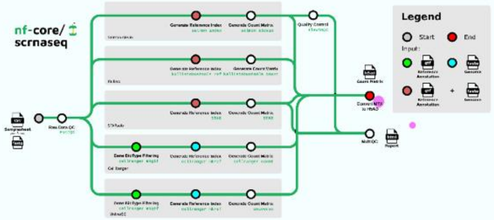
Options

- ALIGNER:
  - We will be choosing **cellranger** – specifically designed for 10x generated data.
  - **Cellranger** uses **starSOLO** as the alignment algorithm but make setup simple and easy.
  - Produces QC reports per sample.

Together, they will be used to make our index.

FASTQC


nf-core/
scrnaseq
Samplesheet
csv
fastq
Raw Data QC
FastQC
Generate Reference Index
salmon index
Generate Count Matrix
salmon alevin
Salmon-Alevin
Quality Control
alevinQC
Generate Reference Index
kallistobustools ref
Generate Count Matrix
kallistobustools count
Kallisto
Generate Reference Index
STAR
Generate Count Matrix
STAR
STARsolo
Gene BioType Filtering
cellranger mkgtf
Generate Reference Index
cellranger mkref
Generate Count Matrix
cellranger count
Cell Ranger
Gene BioType Filtering
cellranger mkgtf
Generate Reference Index
cellranger mkref
Generate Count Matrix
universc
UniverSC
h5ad
Count Matrix
Convert MTX
to H5AD
MultiQC
html
Report
Legend
Start
End
Input:
gtf
Reference
Annotation
fasta
Genome
gtf
Reference
Annotation
+
fasta
Genome



FASTQC

What

- Tool for assessing the quality of raw sequencing.
- Commonly used for high-throughput sequencing such as ScRNAseq.

Output

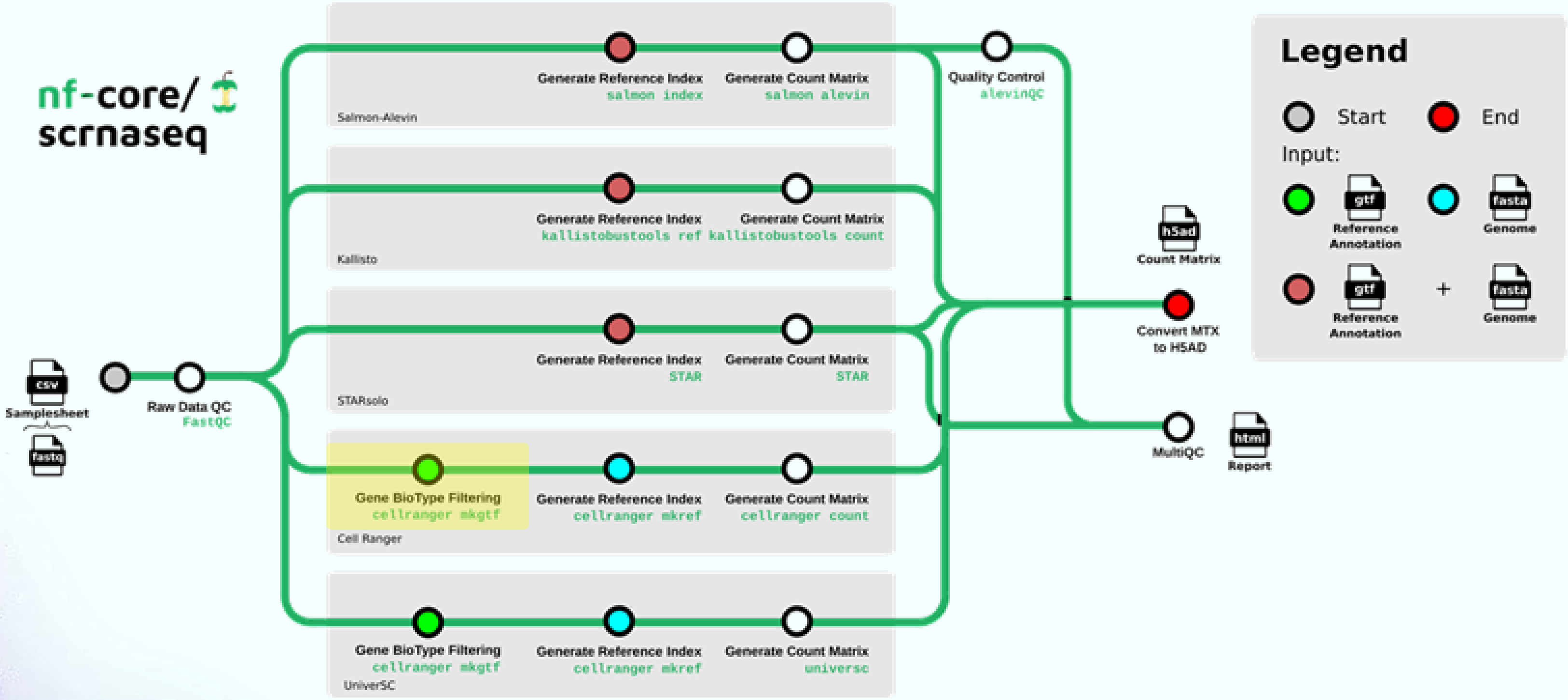
- HTML file
- Each sample has 2 files:
  - Read 1
  - Read 2

Readouts

- Basic Statistics
- Per base sequence quality
- Per sequence quality score
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

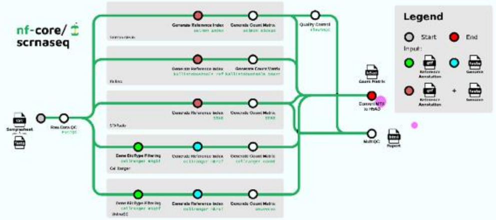
⚠ The .fastq files do not get modified in any way during this step.
FASTQC only interprets the .fastq files.

cellranger mkgtf



nf-core

MDI Biological Laboratory



cellranger mkgtf

What

- Pre-processing step run by **cellranger** to prepare the .gtf file.

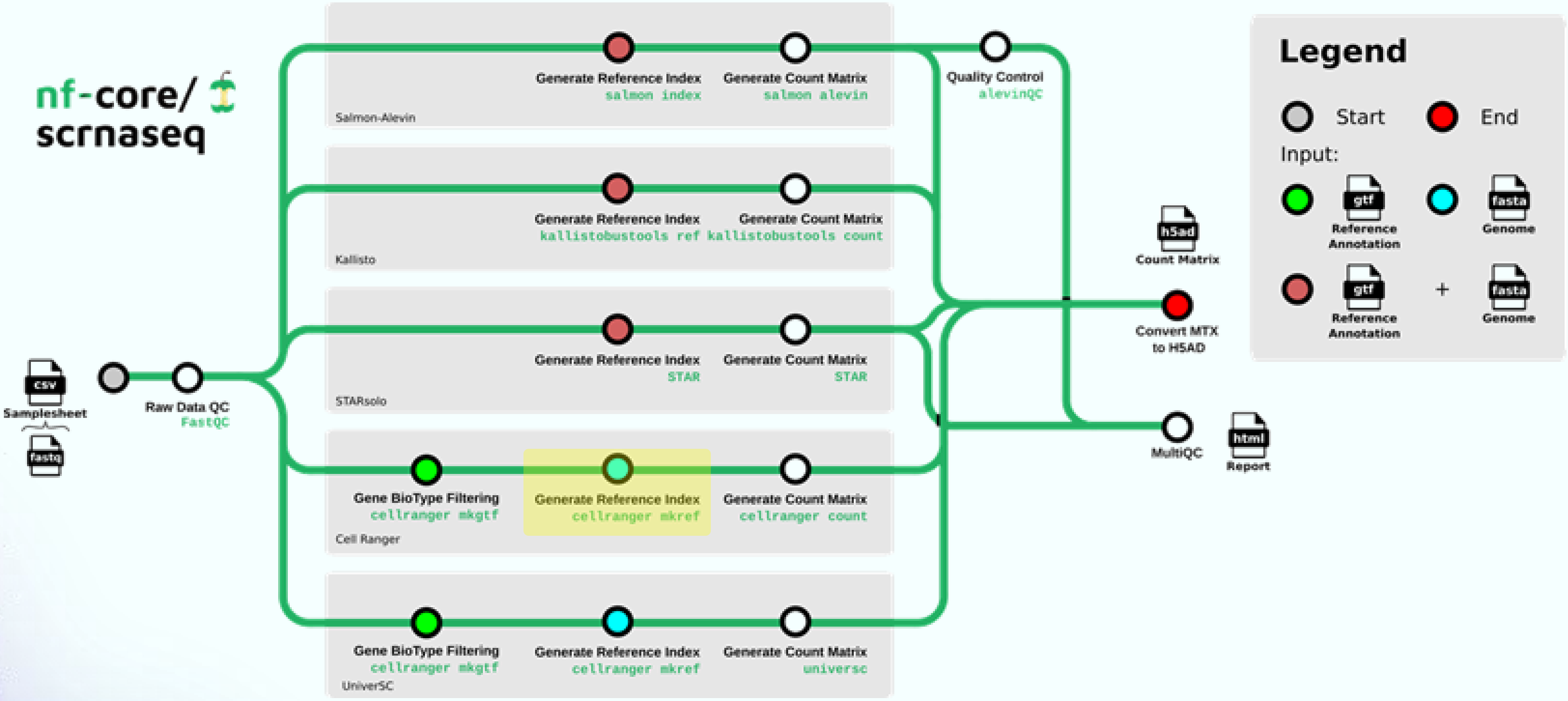
Why

- .gtf files do not always conform to a strict organization.
- Often, there is additional information in the .gtf file that is not needed for **cellranger**.

Output

- A slimmed down .gtf that is properly formatted to ensure that the subsequent **cellranger** processes are run correctly and efficiently.

cellranger mkref



nf-core/ scrnaseq

Samplesheet
csv
fastq

Raw Data QC
FastQC

Salmon-Alevin

Generate Reference Index
salmon index

Generate Count Matrix
salmon alevin

Quality Control
alevinQC

Kallisto

Generate Reference Index
kallistobustools ref

Generate Count Matrix
kallistobustools count

STARsolo

Generate Reference Index
STAR

Generate Count Matrix
STAR

Cell Ranger

Gene BioType Filtering
cellranger mkgtf

Generate Reference Index
cellranger mkref

Generate Count Matrix
cellranger count

UniverSC

Gene BioType Filtering
cellranger mkgtf

Generate Reference Index
cellranger mkref

Generate Count Matrix
universc

h5ad
Count Matrix

Convert MTX to H5AD

MultiQC

html
Report

Legend

Start End

Input:

gtf Reference Annotation

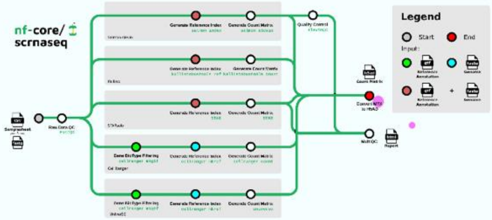
fasta Genome

gtf Reference Annotation + fasta Genome

nf-core

MDI Biological Laboratory

cellranger mkref



What

- Builds an index or *map* to be used for aligning your reads.
- Takes in both the .fasta (sequence data) and .gtf (gene annotations).

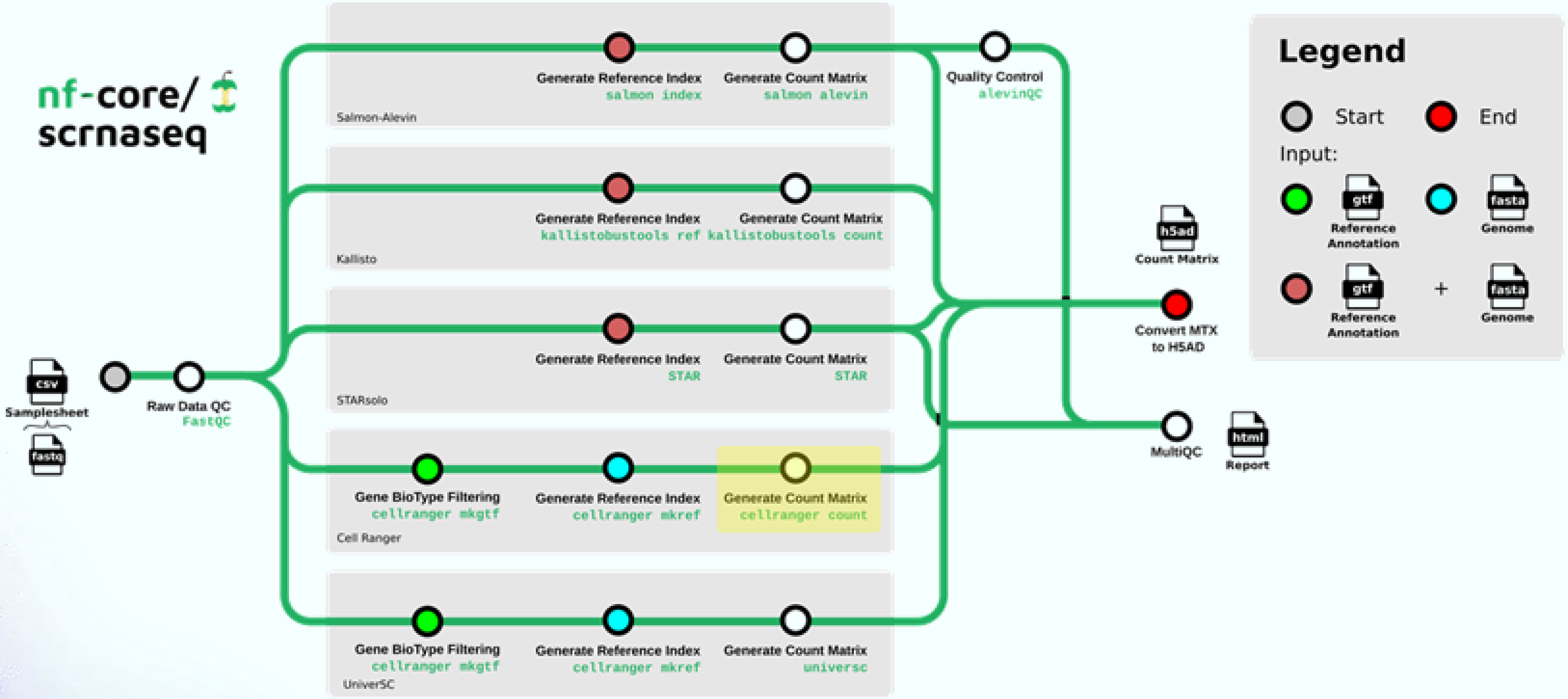
Why

- Having an index ensures the alignment step runs efficiently and accurately.

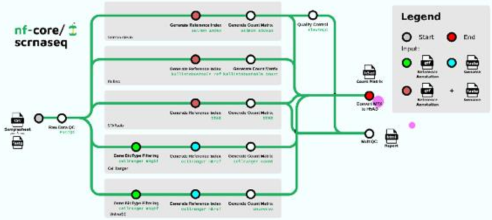
Output

- A structured directory containing the index and auxiliary files to be used in the **cellranger count** step.

cellranger count



cellranger count



What

- This is the most computationally intensive task in the pipeline.
- Takes the reads (from the .fasta files) and uses the index created in **cellranger mkref** to map what gene each read matches.

How

- Two pass solution:
- Pass one: reads are aligned to the reference genome. Spice junction (both known and novel) and logged.
- Junctions are then filtered based on number of reads supporting the junction.
- Pass two: reads are once again aligned, but with this filtered set of junctions from the first pass.

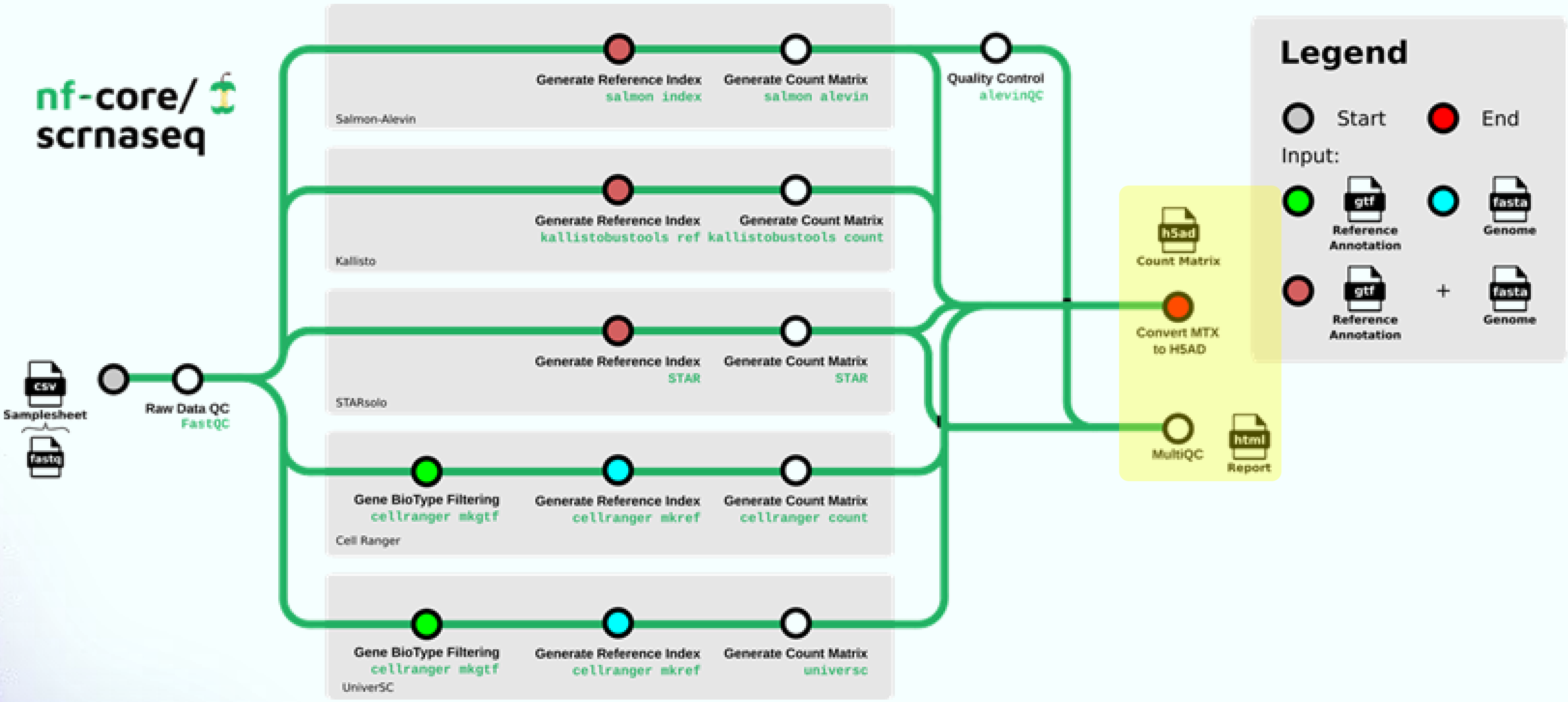
Output

- Key Output: raw and filtered counts
- Structured in a directory that contains 3 files (MEX format):
  - features.tsv.gz
  - barcodes.tsv.gz
  - matrix.mtx.gz

nf-core

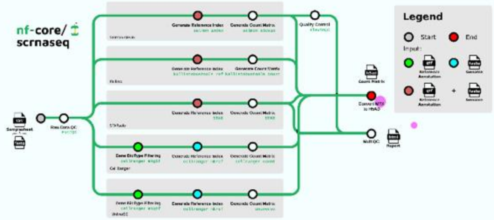
MDI Biological Laboratory

post-alignment



nf-core

MDI Biological Laboratory



post-alignment

Extra Outputs

- This pipeline produces extra outputs mostly consisting of additional file conversions.
  - MEX → H5ad
  - MEX → rds

pipeline_info

- Overview of the execution of the pipeline.
  - Report
  - Timeline
  - Dag

multiqc

- A full breakdown of the workflow reporting results and statistics on the steps run.

Questions?

nf-core

MDI Biological Laboratory