User’s Manual for scatacseqflow

**Overview**

The goal of scatacseqflow is to provide an end-to-end pipeline to ease the analysis of single-cell ATAC-seq data analysis. This pipeline takes minimal input from user and applies various processing steps including raw fastq sequence preprocessing, clustering, motif enrichment, and more. Backed by Nextflow engine, scatacseqflow integrates numerous commonly practiced analytical software and is designed to work compatibly across different sequencing platforms including 10xgenomics, biorad, *etc.*.

Integrated steps:

1. Preprocessing:

The goal includes:

* 1. Prepare alignment bam file and fragment file from raw fastq files.
  2. Perform FastqQC and bam QC.

Available methods:

1. default: this will perform the following steps: barcode correction; debarcoding, trimming, mapping, QC.
2. 10xgenomics: leverages its official cellranger-atac count software
3. biorad: leverages its official ATAC-seq Analysis Toolkit, which includes: FASTQ Quality Control, FASTQ Debarcoding, Read Trimming, Alignment, Alignment QC, (under development: Bead Filtration, Bead Deconvolution, Cell Filtration, Peak Calling, Atac-Seq QC, and Report).
4. Downstream analysis:

Clustering, motif enrichment, …

**Usage**

**Prerequisite**

This pipeline requires Nextflow pipeline engine and one container engine (either Docker or Singularity) to function. Follow the instructions below to install all dependencies before continuing.

1. Download and install Nextflow engine: <https://www.nextflow.io/docs/latest/getstarted.html>

After installation, upgrade Nextflow to 20.06.0-edge by typing the following command:

NXF\_VER=20.06.0-edge nextflow self-update

1. Download and install Singularty or Docker engine:

Singularity: <https://sylabs.io/guides/3.5/user-guide/quick_start.html>

Docker: <https://docs.docker.com/get-docker/>

Note that for Docker to function properly on MacOS with Nextflow, be sure to follow this tutorial to correctly mount /Users first: <https://www.viget.com/articles/how-to-use-docker-on-os-x-the-missing-guide/>

**Basics**

This section covers the most fundamentals to get the pipeline running after installation of all dependencies.

1. First, clone the pipeline by:

git clone https://github.com/hukai916/scatacseqflow.git

1. Ensure Nextflow version > 20.06.0:

nextflow -v

1. Show help:

cd scatacseqflow

nextflow run main.nf –help

1. Specifying configurations:

There are two types of options in Nextflow: pipeline option and parameter option. While the pipeline option (supplied with single dash) regulates the behavior of the workflow (e.g. by specifying -resume, the previously analyzed results will be cached.), the parameter option (supplied with double dash) is specific to certain processes (e.g. by --input path\_to\_sample, the path\_to\_sample will be used as input). You can also set the default parameter options in the nextflow.config file, under the params scope.

1. Specifying profiles:

Nextflow profiles are sets of predefined pipeline parameters, which can be set by -profile profile\_name. For example, to run pipeline using singularity containers: nextflow run main.nf -profile singularity; to run pipeline with lsf executor (default is local, which will run on your local machine): nextflow run main.nf -profile lsf. Note that if multiple profiles, separate them by comma.

For complete Nextflow tutorial, see documentation: <https://www.nextflow.io/docs/latest/index.html>

**Preprocessing**

**Method1**: default (under development)

Invoked by: “-preprocess default”

Input:

1. A samplesheet.csv: supplied by “--input full\_path\_to\_samplesheet.csv”

A comma separated file containing data meta info, the header must be “sample\_name,path\_fastq\_1,path\_fastq\_2,path\_barcode”, please use assets/ samplesheet\_test.csv as template. Note that you may specify multiple samples, each with a new row. For each sample, multiple fastq files (e.g. multiple sequencing lanes) can be listed under each column, and they must be separated by semicolon (;), those files will be merged into one single fastq file during the preprocessing step.

1. 3’ adapter sequence supplied by “--trim3”
2. 5’ adapter sequence supplied by “--trim5”

Output:

**Method2**: 10xgenomics

Invoked by: “-preprocess 10xgenomics”

Input:

1. A samplesheet.csv: supplied by “--input full\_path\_to\_samplesheet.csv”

A comma separated file containing data meta info, the header must be “sample\_name,path\_fastq\_1,path\_fastq\_2,path\_barcode”, please use assets/ samplesheet\_test.csv as template. Note that you may specify multiple samples, each with a new row. For each sample, multiple fastq files (e.g. multiple sequencing lanes) can be listed under each column, and they must be separated by semicolon (;), those files will be merged into one single fastq file during the preprocessing step.

1. A folder containing 10xgenomics provided reference: supplied by “--ref\_cellranger full\_path\_to\_reference\_folder”

A 10xgenomics provided reference that is designed to work with cellranger-atac, follow the followings to get one: <https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/advanced/references>

Output: bam, fragment file, clustering, *etc.*

**Method3**: biorad (under development)

Note that this option is wrapped around the ***Bio-Rad ATAC-Seq Analysis*** ***Toolkit*** (<https://www.bio-rad.com/webroot/web/pdf/lsr/literature/Bulletin_7191.pdf>) and must be used with compatible sequencing data.

Invoked by: “-preprocess biorad”

Input:

1. A samplesheet.csv: supplied by “--input full\_path\_to\_samplesheet.csv”

A comma separated file containing data meta info, the header must be “sample\_name,path\_fastq\_1,path\_fastq\_2,path\_barcode”, please use assets/ samplesheet\_test.csv as template. Note that you may specify multiple samples, each with a new row. For each sample, multiple fastq files (e.g. multiple sequencing lanes) can be listed under each column, and they must be separated by semicolon (;), those files will be merged into one single fastq file during the preprocessing step.

1. A folder containing BWA index file: supplied by “--ref\_bwa\_index full\_path\_to\_bwa\_index\_folder”
2. A fasta file containing reference genome: supplied by “--ref\_bwa\_fasta”
3. A pipeline parameter supplied by “--biorad\_genome”, which takes the value of either “hg19”, “hg38”, “mm10”, or “hg19-mm10”.

Output: bam, fragment file, clustering, *etc.*

For complete Nextflow tutorial, see documentation:

1. Clone hukai916/Samplesheet.csv:

Raw sequencing

**Preprocessing**

Samplesheet.csv:

Raw sequencing fastq files including barcode sequences.