MobiVision v4.1 User Manual

(I) Software Introduction

MobiVision is a bioinformatics analysis software that can analyze single-cell sequencing data from the MobiNova platform, including 3'/5' transcriptome, full-length VDJ, CITE proteomics, and ChIP and ATAC sequencing data. Currently, MobiVision v4.1 has twelve subcommands and one auxiliary function, as follows:

1. Transcriptome Analysis Module, including the following seven subcommands:

- **quantify** is mainly used to process single-cell transcriptome data (.fastq format), with outputs including library quality control reports, cell-gene expression matrices (.mtx), and related files.
- **fmt_gtf** can be used to process gtf annotation files, customizing personalized gtf files by specifying the required annotation types.
- **mkindex** is used to construct reference genomes required for quantify analysis.
- **rename** can rename all files containing sampleID in the quantify analysis results, including the sample ID displayed in the quality control report.
- **mtx2csv** can convert the cell-gene expression matrix (.mtx) obtained from quantify analysis to .csv files.
- **re_call_cell** can re-screen cells from quantify output results and generate new quality control reports and cell-gene expression matrices (.mtx).
- **integrate** can integrate cell-gene expression matrices output from quantify.

2. V(D)J Analysis Module, including the following one subcommand:

- **vdj** is mainly used to process single-cell V(D)J sequencing data (.fastq format), with outputs including V(D)J library quality control reports, annotations, and clonotype-related result files.

3. CITE Analysis Module, including the following one subcommand:

- **cite** is mainly used to process CITE sequencing data (.fastq format), with outputs including library quality control reports, gene and protein expression matrix files.

4. ChIP Analysis Module, including the following two subcommands:

- **chip** is mainly used to process single-cell ChIP data (.fastq format), with outputs including library quality control reports, peak matrices, and related result files.
- **mk_chip_ref** is used to construct reference genome files required for chip analysis.

5. ATAC Analysis Module, including the following one subcommand:

- **atac** is mainly used to process single-cell ATAC data (.fastq format), with outputs including library quality control reports, peak matrices, and related result files.

6. Bam_to_fastq Auxiliary Function

- **Bam_to_fastq** is mainly used to convert bam files obtained from quantify analysis to fastq format.

(II) System Requirements

- 8-core Intel or AMD processor, x86 architecture (16 cores or more recommended)
- 64GB RAM or more
- 1TB available storage space
- Linux operating system, 64-bit CentOS, Ubuntu 20.04 or higher versions

(III) Installation Method

After extracting mobivision-v4.1.tar.gz, execute the following source command in the shell command line to use MobiVision v4.1 software. The "source" command needs to be executed each time you open a new shell window or terminal interface.

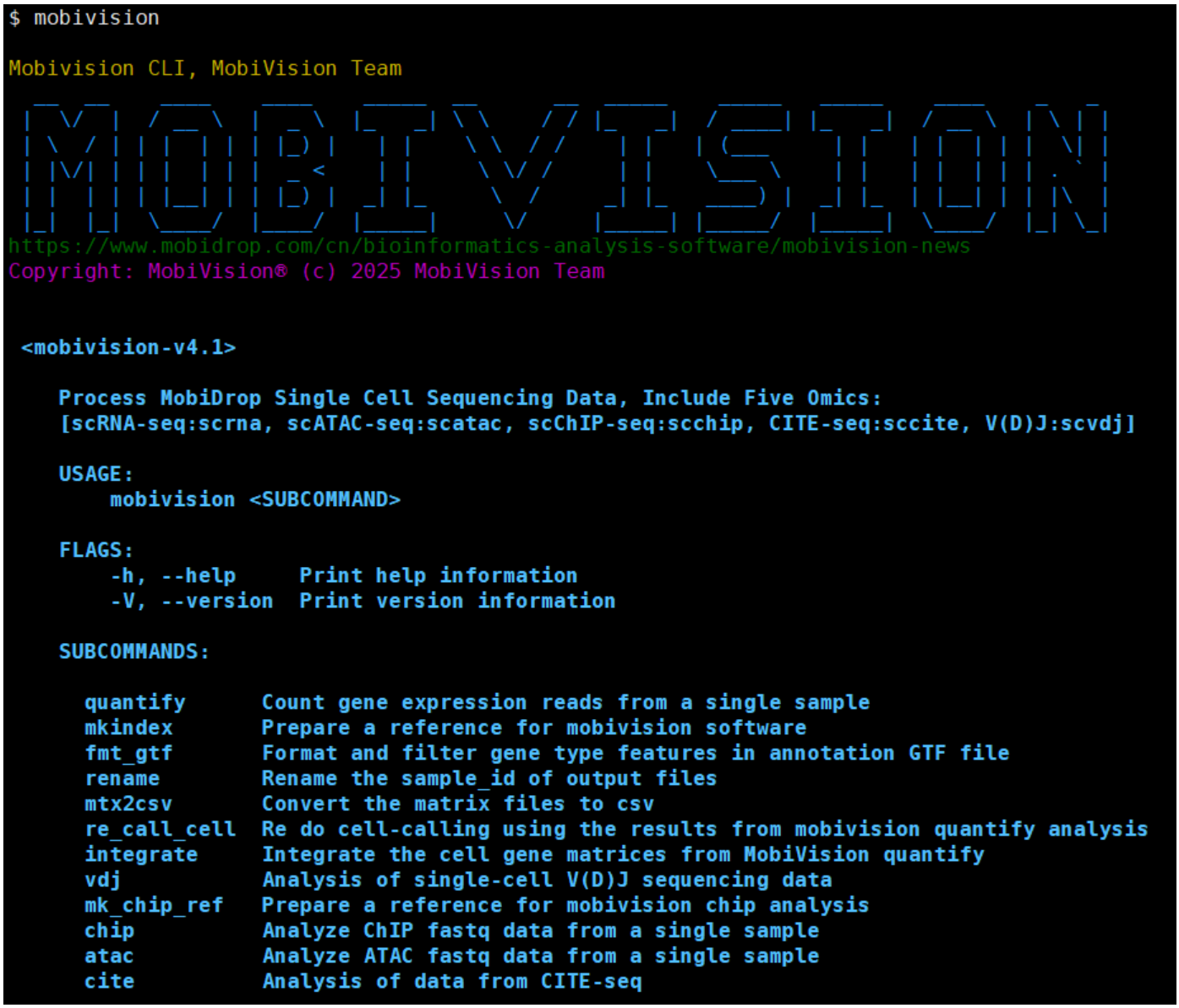
$ tar -zxvf mobivision-v4.1.tar.gz

$ source mobivision-v4.1/source.sh

Test if mobivision is installed successfully

$ mobivision --help

#When the following information is displayed, it indicates that MobiVision software has been successfully installed



(IV) Usage Instructions

1 Transcriptome Analysis Module (3'RNA, 5'RNA)

1.1 quantify

mobivision quantify is adapted for libraries and sequencing data prepared by MobiCube® high-throughput single-cell 3' transcriptome kit (3'RNA) and MobiCube® high-throughput single-cell 5' transcriptome kit (5'RNA); it can calculate the gene expression level of each cell in single-cell transcriptome libraries. Using paired-end sequencing fastq files as input, it can ultimately output html quality control reports and gene expression matrices for downstream in-depth analysis.

1.1.1 Preparing Fastq Data

Input fastq files should be named in the following format:

```
[SampleName]_[ReadType].[Suffix]
```

Where,

[SampleName] consists of letters, numbers, or underscores;

[ReadType] suggested format: R1, R2, R1_001, R2_001;

[Suffix] currently supports 4 cases: fastq.gz, fq.gz, fastq, fq.

For example:

```
/Data/Sample_fastq/

├── R22001221-20220105-Emix-3-20220105-S-F5-2-2_R1.fq.gz

└── R22001221-20220105-Emix-3-20220105-S-F5-2-2_R2.fq.gz
```

1.1.2 MobiVision quantify Usage

The mobivision quantify command input parameters include the path to the folder containing the index (-i), the maximum number of threads to call (-t), the path to the fastq files (-f), and the output folder path (-o). After modifying the red sample code with actual parameters, run:

$ mobivision quantify -i /media/mz-3/db/gencode_hsa/GRCh38 \
-t 12 \
-f /media/mz-3/mzRaid10/Mobi_data/Sample_R22014039-C07 \
-o /media/mz-3/analysis/R22014039-C07

All parameter descriptions are as follows:

	Parameter	Description
Required	-f, --fastqDir	Path to fastq files. For example: /media/mz-3/mzRaid10/Mobi_data/Sample_R22014039-C07.
	-i, --indexDir	Path containing MobiCube transcriptome reference sequence files. If downloading the tar.gz format index file provided by MobiCube, you need to extract the tar.gz first, then use the extracted file path as the path for the -i parameter. For example: /media/mz-3/db/gencode_hsa/GRCh38.
	-o, --outDir	Output file path. For example: /media/mz-3/analysis/R22014039-C07.
	-t, --threads	Maximum number of threads used by the software. For example: 12.
Optional	--intron	Include gene introns for gene expression analysis. Options include "excluded" and "included". Default setting is "included".
	--cellnumber	Force the use of this cell number for cell screening.
	--cr2.2	Use CR2.2 algorithm for cell screening. If not specified, EmptyDrops algorithm is used by default.
	-s, --sampleID	User-specified sample ID for result files.
	--nobam	Parameter to set whether to retain BAM files, default is to retain STAR-generated BAM files; if this parameter is enabled, STAR output BAM files will not be retained.
	--kit	Set the RNA kit version used to construct the library and specify the library type for analysis (5'RNA or 3'RNA). If not specified, default is "Unknown", and the system will automatically determine whether to analyze 5'RNA or 3'RNA.
Help	-h, --help	Display help information.

mobivision quantify specifies the path to fastq files through -f. If there is only one pair of fastq files in that path, that pair of fastq files will be analyzed by default; if there are multiple pairs of fastq files in that path, these files will be merged into one pair of fastq files and saved in a folder named "merge/". Subsequent analysis will use the merged fastq files as input. Therefore, it is recommended to place multiple sequencing data from the same sample source in the same folder.

In the mobivision quantify command, there are three algorithms for cell selection:

1. Using "--cellnumber value", mobivision quantify will use this value as the forced extracted cell number,
2. If using "--cr2.2", then the CellRanger2.2 algorithm is used for cell filtering,
3. In other cases, the default EmptyDrops algorithm is used.

By default, genes in the analysis results include introns. If only exons are considered as genes, you can add the "--intron excluded" parameter. The default setting is "--intron included".

Output result files are as follows:

```
├── _flagdone                                                    Flag
file for successful run
```

```
├── _log                                                                    Run
log file

└── SAMPLEID_outs
Output result file root directory

  ├── filtered_cell_gene_matrix
Filtered matrix file root directory

  │   ├── barcodes.tsv.gz
Filtered barcode file

  │   ├── features.tsv.gz
Filtered gene list file

  │   └── matrix.mtx.gz
Filtered expression matrix file

  ├── SAMPLEID_Aligned.sort.bam
Alignment result file

  ├── SAMPLEID_Aligned.sort.bam.bai
Alignment result file index

  ├── SAMPLEID_filtered.h5ad
Filtered data matrix in h5ad format

  ├── SAMPLEID_Report.html
Quality control report in html format

  ├── SAMPLEID_Report.json
Quality control report in json format

  ├── SAMPLEID_summary.csv
Library information in csv format

  ├── raw_cell_gene_matrix
Unfiltered matrix file root directory

  │   ├── barcodes.tsv.gz
Unfiltered barcode file

  │   ├── features.tsv.gz
Unfiltered gene list file

  │   └── matrix.mtx.gz
Unfiltered expression matrix file

  ├── result_mito_percentage.csv
Mitochondrial percentage information file
```

```
└── run_analysis_cmds.txt                                                    Complete
command line record
```

1.2 fmt_gtf

mobivision fmt_gtf is used to process genome annotation files (GTF files). The processed annotation files can be used to construct reference transcriptomes. The mobivision fmt_gtf command input parameters include input GTF file (-i), output GTF file (-o), and gene types to be retained in the output GTF file (-t). After modifying the red sample code with actual parameters, run:

$ mobivision fmt_gtf -i /media/db/gencode.v38.primary_assembly.annotation.gtf.gz \
-o /media/db_out/gencode.v38.primary_assembly.annotation.filtered.gtf \
-t protein_coding \
-t lncRNA \
-t antisense \
-t IG_LV_gene \
-t IG_V_gene \
-t IG_V_pseudogene \
-t IG_D_gene \
-t IG_J_gene \
-t IG_J_pseudogene \
-t IG_C_gene \
-t IG_C_pseudogene \
-t TR_V_gene \
-t TR_V_pseudogene \
-t TR_D_gene \
-t TR_J_gene \
-t TR_J_pseudogene \
-t TR_C_gene

All parameter descriptions are as follows:

	Parameter	Description
Required	-i	Input GTF file name, supports .gtf and .gtf.gz file formats.
	-o	Output GTF file name, default is "out_filtered.gtf".
Optional	-t	Gene types to be retained in the output GTF file, supports multiple specifications.
Help	-h, --help	Display help information.

1.3 mkindex

mobivision mkindex uses genome fasta files and GTF annotation files as input to construct MobiVision-compatible reference transcriptomes. The mobivision fmt_gtf command input parameters include genome name (-n), genome fasta file (-f), and genome GTF annotation file (-g). The constructed reference transcriptome will be saved in a folder named after the genome name (-n), which is located in the path

where the command is run. After modifying the red sample code with actual parameters, run:

$ mobivision mkindex -n GRCh38_gencode \
-f /media/db_out/GRCh38.primary_assembly.genome.fa.gz \
-g /media/db_out/gencode.v38.primary_assembly.annotation.filtered.gtf

1.3.1 Multi-species Construction

If you need to construct samples combining two or more species, the -n-f-g parameters need to be specified twice or multiple times. The genome names, genome files, and genome annotation files specified by -n-f-g need to correspond one by one according to the order. Sample code reference is as follows:

$ mobivision mkindex -n GRCh38_gencode \
-f /media/db_out/GRCh38.primary_assembly.genome.fa.gz \
-g /media/db_out/gencode.v38.primary_assembly.annotation.filtered.gtf \
-n GRCm39_gencode \
-f /media/db_out/GRCm39.primary_assembly.genome.fa.gz \
-g /media/db_out/gencode.vM39.primary_assembly.annotation.filtered.gtf

1.3.2 Parameter Description

All parameter descriptions are as follows:

	Parameter	Description
Required	-n	Genome name consists of [a-zA-Z0-9*]. If this reference genome consists of multiple different genomes, please specify the -n parameter multiple times, for example -n -n . The output reference genome folder will be named* and_.
	-f	Genome fasta file. If multiple fasta genome files need to be input, please specify the -f parameter multiple times. Genome fasta file supported formats include .fasta, .fa.gz, and .fna.gz.
	-g	GTF file (.gtf or .gtf.gz). If multiple GTF genome annotation files need to be input, please specify the -g parameter multiple times. If -n and -f are specified multiple times, then -g must be specified the same number of times in the corresponding order. Note: When constructing reference genomes with mkindex, the mobivision software has requirements for input gtf files, for example: gtf files must have more than 6 columns, the third column needs to contain gene, exon, transcript information, and the last column needs to contain gene_id, gene_name, gene_type and other information
Optional	-r	Output reference genome version number, such as "ref_v1.0", "ref-2022-06".
	-m	Maximum memory (GB) used by STAR to construct reference genomes.
Help	-h, --help	Display help information.

1.3.3 Adding HIV Genes to Human Reference Genome

In transgenic or viral infection studies, it may be necessary to add genes of interest to existing genomes for alignment. Taking adding HIV genes to human reference genome as an example, the operation method is as follows:

First, use the wget command to download the fasta and gtf files required to construct this reference genome from the gencode official website and extract them. After modifying the red sample code with actual parameters, run:

$wget
http://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_38/GRCh38.primary_assembly.genome.fa.gz
$wget http://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_38/gencode.v38.primary_assembly.annotation.gtf.gz

$gunzip -c GRCh38.primary_assembly.genome.fa.gz > GRCh38.primary_assembly.genome.fa
$gunzip -c gencode.v38.primary_assembly.annotation.gtf.gz > gencode.v38.primary_assembly.annotation.gtf

Second, download the gene sequence of interest, use the cat command to create a new fa file and copy the gene sequence to the fa file, press Ctrl+D to save. Replace the first line's long name with a shorter name for easy identification, and count the base number of the sequence for making the gtf file. After modifying the red sample code with actual parameters, run:

$cat > HIV.fa
$sed -i '1d' HIV.fa
$sed -i '1i >HIV' HIV.fa
$cat HIV.fa | grep -v "^>" | tr -d "\n" | wc -c

Then, refer to the basic format of gtf files on the ensembl official website to construct the .gtf file for the gene of interest. After modifying the red sample code with actual parameters, run:

$echo -e 'HIV\tunknown\texon\t1\t999\t.\t+\t.\tgene_id "HIV-1"; transcript_id "HIV-1"; gene_name "HIV-1"; gene_biotype "protein_coding";' > HIV.gtf

Then, merge the original reference genome and target gene fa files and gtf files. After modifying the red sample code with actual parameters, run:

$cp -rf GRCh38.primary_assembly.genome.fa GRCh38.hiv.genome.fa
$cp -rf gencode.v38.primary_assembly.annotation.gtf GRCh38.hiv.genome.gtf
$cat HIV.fa >> GRCh38.hiv.genome.fa
$cat HIV.gtf >> GRCh38.hiv.genome.gtf

Finally, use the mobivision mkindex command to construct new genome reference information. After modifying the red sample code with actual parameters, run:

$ mobivision mkindex \
-n GRCh39-HIV \
-f GRCh38.hiv.genome.fa \
-g GRCh38.hiv.genome.gtf \
-r GRCh39-HIV0923

1.4 re_call_cell

The result files from mobivision quantify can be used as input files for mobivision re_call_cell to re-screen cells (cell-calling) and obtain cell-gene expression matrices (reanalysis_matrix), html quality control reports, and other related result files. The mobivision re_call_cell command input parameters include the path to mobivision quantify result files (-i) and the output path for re_call_cell results (-o). After modifying the red sample code with actual parameters, run:

```
$ mobivision re_call_cell -i 230106-WJ-E12-W04_combined_outs \
-o 230106-WJ-E12-W04_reCallCell_outs \
--cellnumber 1000
```

All parameter descriptions are as follows:

	Parameter	Description
Required	-i	Path to mobivision quantify result files.
	-o	Output path for mobivision re_call_cell result files.
Optional	--cellnumber	Force the use of this cell number for cell screening.
	--ed	Use EmptyDrops algorithm for cell screening. If not specified, CellRanger2.2 algorithm is used by default.
Help	-h, --help	Display help information.

In the mobivision re_call_cell command, there are three algorithms for cell selection:

1. Using "--cellnumber value", mobivision re_call_cell will use this value as the forced extracted cell number,
2. If using "--ed", then the EmptyDrops algorithm is used for cell filtering,
3. In other cases, the default CellRanger2.2 algorithm is used.

1.5 rename

After completing mobivision quantify, mobivision rename can rename all files containing sampleID in the file names, and can also change the "Sample ID" information displayed in the html report. The mobivision rename command input parameters include the new sampleID (-i) and the output result directory from mobivision quantify (-d). After modifying the red sample code with actual parameters, run:

```
$ mobivision rename -i human_PBMC_Control \
-d /media/mz/analysis/R1-S0-1_outs
```

All parameter descriptions are as follows:

	Parameter	Description
Required	-i	New SampleID for result files.
	-d	Output result folder from mobivision quantify.
Help	-h, --help	Display help information.

1.6 mtx2csv

After completing mobivision quantify, the result files include cell-gene expression matrices filtered cell gene matrix and raw cell gene matrix. mobivision mtx2csv can convert cell-gene expression matrices to csv files and output them. The mobivision mtx2csv command input parameters include the folder containing the cell-gene expression matrix (-i), which contains matrix files, barcodes files, and features files, and the output csv file name (-o). After modifying the red sample code with actual parameters, run:

```
$ mobivision mtx2csv -i /media/mz/analysis/R1-S0-1_outs/filtered_cell_gene_matrix \
-o /output/R1-S0-1.csv
```

All parameter descriptions are as follows:

	Parameter	Description
Required	-i	Folder containing the cell-gene expression matrix, including matrix files, barcodes files, and features files.
	-o	Csv file name.
Help	-h, --help	Display help information.

1.7 integrate

After completing mobivision quantify for multiple samples, the result files include cell-gene expression matrices filtered cell gene matrix. mobivision integrate can integrate and analyze cell-gene expression matrices from multiple samples and output qlentille files and h5ad files. The mobivision integrate command input parameters include a list recording the paths of multiple sample expression matrices (-l), the path containing input matrix files (-d), the selected integration algorithm (-a), and the output file path (-o). After modifying the red sample code with actual parameters, run:

```
$ mobivision integrate -l sample.list -o /output/
```

All parameter descriptions are as follows:

	Parameter	Description
Required	-l	List recording the paths of multiple sample expression matrices, with the first column as sample name and the second column as the filtered_cell_gene_matrix path corresponding to the sample
	-d	Filtered_cell_gene_matrix path corresponding to the sample
	-o	Output file path
Optional	-a	Selected integration algorithm, default is: harmony
Help	-h, --help	Display help information.

(IV) Usage Instructions

1 Transcriptome Analysis Module (3'RNA, 5'RNA)

1.1 quantify

mobivision quantify is adapted for libraries and sequencing data prepared by MobiCube® high-throughput single-cell 3' transcriptome kit (3'RNA) and MobiCube® high-throughput single-cell 5' transcriptome kit (5'RNA); it can calculate the gene expression level of each cell in single-cell transcriptome libraries. Using paired-end sequencing fastq files as input, it can ultimately output html quality control reports and gene expression matrices for downstream in-depth analysis.

1.1.1 Preparing Fastq Data

Input fastq files should be named in the following format:

```
[SampleName]_[ReadType].[Suffix]
```

Where,

[SampleName] consists of letters, numbers, or underscores;

[ReadType] suggested format: R1, R2, R1_001, R2_001;

[Suffix] currently supports 4 cases: fastq.gz, fq.gz, fastq, fq.

For example:

```
/Data/Sample_fastq/

├── R22001221-20220105-Emix-3-20220105-S-F5-2-2_R1.fq.gz

└── R22001221-20220105-Emix-3-20220105-S-F5-2-2_R2.fq.gz
```

1.1.2 MobiVision quantify Usage

The mobivision quantify command input parameters include the path to the folder containing the index (-i), the maximum number of threads to call (-t), the path to the fastq files (-f), and the output folder path (-o). After modifying the red sample code with actual parameters, run:

```
$ mobivision quantify -i /media/mz-3/db/gencode_hsa/GRCh38 \
-t 12 \
-f /media/mz-3/mzRaid10/Mobi_data/Sample_R22014039-C07 \
-o /media/mz-3/analysis/R22014039-C07
```

All parameter descriptions are as follows:

	Parameter	Description
Required	-f, --fastqDir	Path to fastq files. For example: /media/mz-3/mzRaid10/Mobi_data/Sample_R22014039-C07.
	-i, --indexDir	Path containing MobiCube transcriptome reference sequence files. If downloading the tar.gz format index file provided by MobiCube, you need to extract the tar.gz first, then use the extracted file path as the path for the -i parameter. For example: /media/mz-3/db/gencode_hsa/GRCh38.
	-o, --outDir	Output file path. For example: /media/mz-3/analysis/R22014039-C07.
	-t, --threads	Maximum number of threads used by the software. For example: 12.
Optional	--intron	Include gene introns for gene expression analysis. Options include "excluded" and "included". Default setting is "included".
	--cellnumber	Force the use of this cell number for cell screening.
	--cr2.2	Use CR2.2 algorithm for cell screening. If not specified, EmptyDrops algorithm is used by default.
	-s, --sampleID	User-specified sample ID for result files.
	--nobam	Parameter to set whether to retain BAM files, default is to retain STAR-generated BAM files; if this parameter is enabled, STAR output BAM files will not be retained.
	--kit	Set the RNA kit version used to construct the library and specify the library type for analysis (5'RNA or 3'RNA). If not specified, default is "Unknown", and the system will automatically determine whether to analyze 5'RNA or 3'RNA.
Help	-h, --help	Display help information.

mobivision quantify specifies the path to fastq files through -f. If there is only one pair of fastq files in that path, that pair of fastq files will be analyzed by default; if there are multiple pairs of fastq files in that path, these files will be merged into one pair of fastq files and saved in a folder named "merge/". Subsequent analysis will use the merged fastq files as input. Therefore, it is recommended to place multiple sequencing data from the same sample source in the same folder.

In the mobivision quantify command, there are three algorithms for cell selection:

1. Using "--cellnumber value", mobivision quantify will use this value as the forced extracted cell number,

2. If using "--cr2.2", then the CellRanger2.2 algorithm is used for cell filtering,
3. In other cases, the default EmptyDrops algorithm is used.

By default, genes in the analysis results include introns. If only exons are considered as genes, you can add the "--intron excluded" parameter. The default setting is "--intron included".

Output result files are as follows:

```
├── _flagdone                                                                    Flag
file for successful run

├── _log                                                                          Run
log file

└── SAMPLEID_outs
Output result file root directory

  ├── filtered_cell_gene_matrix
Filtered matrix file root directory

  │   ├── barcodes.tsv.gz
Filtered barcode file

  │   ├── features.tsv.gz
Filtered gene list file

  │   └── matrix.mtx.gz
Filtered expression matrix file

  ├── SAMPLEID_Aligned.sort.bam
Alignment result file

  ├── SAMPLEID_Aligned.sort.bam.bai
Alignment result file index

  ├── SAMPLEID_filtered.h5ad
Filtered data matrix in h5ad format

  ├── SAMPLEID_Report.html
Quality control report in html format

  ├── SAMPLEID_Report.json
Quality control report in json format

  ├── SAMPLEID_summary.csv
Library information in csv format

  ├── raw_cell_gene_matrix
Unfiltered matrix file root directory

  │   ├── barcodes.tsv.gz
Unfiltered barcode file
```

```
  │   ├── features.tsv.gz
Unfiltered gene list file

  │   └── matrix.mtx.gz
Unfiltered expression matrix file

  ├── result_mito_percentage.csv
Mitochondrial percentage information file

└── run_analysis_cmds.txt                                                Complete
command line record
```

1.2 fmt_gtf

mobivision fmt_gtf is used to process genome annotation files (GTF files). The processed annotation files can be used to construct reference transcriptomes. The mobivision fmt_gtf command input parameters include input GTF file (-i), output GTF file (-o), and gene types to be retained in the output GTF file (-t). After modifying the red sample code with actual parameters, run:

$ mobivision fmt_gtf -i /media/db/gencode.v38.primary_assembly.annotation.gtf.gz \
-o /media/db_out/gencode.v38.primary_assembly.annotation.filtered.gtf \
-t protein_coding \
-t lncRNA \
-t antisense \
-t IG_LV_gene \
-t IG_V_gene \
-t IG_V_pseudogene \
-t IG_D_gene \
-t IG_J_gene \
-t IG_J_pseudogene \
-t IG_C_gene \
-t IG_C_pseudogene \
-t TR_V_gene \
-t TR_V_pseudogene \
-t TR_D_gene \
-t TR_J_gene \
-t TR_J_pseudogene \
-t TR_C_gene

All parameter descriptions are as follows:

	Parameter	Description
Required	-i	Input GTF file name, supports .gtf and .gtf.gz file formats.
	-o	Output GTF file name, default is "out_filtered.gtf".
Optional	-t	Gene types to be retained in the output GTF file, supports multiple specifications.
Help	-h, --help	Display help information.

1.3 mkindex

mobivision mkindex uses genome fasta files and GTF annotation files as input to construct MobiVision-compatible reference transcriptomes. The mobivision fmt_gtf command input parameters include genome name (-n), genome fasta file (-f), and genome GTF annotation file (-g). The constructed reference transcriptome will be saved in a folder named after the genome name (-n), which is located in the path where the command is run. After modifying the red sample code with actual parameters, run:

```
$ mobivision mkindex -n GRCh38_gencode \
-f /media/db_out/GRCh38.primary_assembly.genome.fa.gz \
-g /media/db_out/gencode.v38.primary_assembly.annotation.filtered.gtf
```

1.3.1 Multi-species Construction

If you need to construct samples combining two or more species, the -n-f-g parameters need to be specified twice or multiple times. The genome names, genome files, and genome annotation files specified by -n-f-g need to correspond one by one according to the order. Sample code reference is as follows:

```
$ mobivision mkindex -n GRCh38_gencode \
-f /media/db_out/GRCh38.primary_assembly.genome.fa.gz \
-g /media/db_out/gencode.v38.primary_assembly.annotation.filtered.gtf \
-n GRCm39_gencode \
-f /media/db_out/GRCm39.primary_assembly.genome.fa.gz \
-g /media/db_out/gencode.vM39.primary_assembly.annotation.filtered.gtf
```

1.3.2 Parameter Description

All parameter descriptions are as follows:

	Parameter	Description
Required	-n	Genome name consists of [a-zA-Z0-9*]. If this reference genome consists of multiple different genomes, please specify the -n parameter multiple times, for example -n -n . The output reference genome folder will be named* and_.
	-f	Genome fasta file. If multiple fasta genome files need to be input, please specify the -f parameter multiple times. Genome fasta file supported formats include .fasta, .fa.gz, and .fna.gz.
	-g	GTF file (.gtf or .gtf.gz). If multiple GTF genome annotation files need to be input, please specify the -g parameter multiple times. If -n and -f are specified multiple times, then -g must be specified the same number of times in the corresponding order. Note: When constructing reference genomes with mkindex, the mobivision software has requirements for input gtf files, for example: gtf files must have more than 6 columns, the third column needs to contain gene, exon, transcript information, and the last column needs to contain gene_id, gene_name, gene_type and other information
Optional	-r	Output reference genome version number, such as "ref_v1.0", "ref-2022-06".
	-m	Maximum memory (GB) used by STAR to construct reference genomes.
Help	-h, --help	Display help information.

1.3.3 Adding HIV Genes to Human Reference Genome

In transgenic or viral infection studies, it may be necessary to add genes of interest to existing genomes for alignment. Taking adding HIV genes to human reference genome as an example, the operation method is as follows:

First, use the wget command to download the fasta and gtf files required to construct this reference genome from the gencode official website and extract them. After modifying the red sample code with actual parameters, run:

$wget http://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_38/GRCh38.primary_assembly.genome.fa.gz
$wget http://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_38/gencode.v38.primary_assembly.annotation.gtf.gz

$gunzip -c GRCh38.primary_assembly.genome.fa.gz > GRCh38.primary_assembly.genome.fa
$gunzip -c gencode.v38.primary_assembly.annotation.gtf.gz > gencode.v38.primary_assembly.annotation.gtf

Second, download the gene sequence of interest, use the cat command to create a new fa file and copy the gene sequence to the fa file, press Ctrl+D to save. Replace the first line's long name with a shorter name for easy identification, and count the base number of the sequence for making the gtf file. After modifying the red sample code with actual parameters, run:

```
$cat > HIV.fa
$sed -i '1d' HIV.fa
$sed -i '1i >HIV' HIV.fa
$cat HIV.fa | grep -v "^>" | tr -d "\n" | wc -c
```

Then, refer to the basic format of gtf files on the ensembl official website to construct the .gtf file for the gene of interest. After modifying the red sample code with actual parameters, run:

```
$echo -e 'HIV\tunknown\texon\t1\t999\t.\t+\t.\tgene_id "HIV-1"; transcript_id "HIV-1"; gene_name "HIV-1"; gene_biotype "protein_coding";' > HIV.gtf
```

Then, merge the original reference genome and target gene fa files and gtf files. After modifying the red sample code with actual parameters, run:

```
$cp -rf  GRCh38.primary_assembly.genome.fa GRCh38.hiv.genome.fa
$cp -rf  gencode.v38.primary_assembly.annotation.gtf GRCh38.hiv.genome.gtf
$cat HIV.fa >> GRCh38.hiv.genome.fa
$cat HIV.gtf >> GRCh38.hiv.genome.gtf
```

Finally, use the mobivision mkindex command to construct new genome reference information. After modifying the red sample code with actual parameters, run:

```
$ mobivision mkindex \
-n GRCh39-HIV \
-f GRCh38.hiv.genome.fa \
-g GRCh38.hiv.genome.gtf \
-r GRCh39-HIV0923
```

1.4 re_call_cell

The result files from mobivision quantify can be used as input files for mobivision re_call_cell to re-screen cells (cell-calling) and obtain cell-gene expression matrices (reanalysis_matrix), html quality control reports, and other related result files. The mobivision re_call_cell command input parameters include the path to mobivision quantify result files (-i) and the output path for re_call_cell results (-o). After modifying the red sample code with actual parameters, run:

```
$ mobivision re_call_cell -i 230106-WJ-E12-W04_combined_outs \
-o 230106-WJ-E12-W04_reCallCell_outs \
--cellnumber 1000
```

All parameter descriptions are as follows:

	Parameter	Description
Required	-i	Path to mobivision quantify result files.
	-o	Output path for mobivision re_call_cell result files.
Optional	--cellnumber	Force the use of this cell number for cell screening.
	--ed	Use EmptyDrops algorithm for cell screening. If not specified, CellRanger2.2 algorithm is used by default.
Help	-h, --help	Display help information.

In the mobivision re_call_cell command, there are three algorithms for cell selection:

1. Using "--cellnumber value", mobivision re_call_cell will use this value as the forced extracted cell number,
2. If using "--ed", then the EmptyDrops algorithm is used for cell filtering,
3. In other cases, the default CellRanger2.2 algorithm is used.

1.5 rename

After completing mobivision quantify, mobivision rename can rename all files containing sampleID in the file names, and can also change the "Sample ID" information displayed in the html report. The mobivision rename command input parameters include the new sampleID (-i) and the output result directory from mobivision quantify (-d). After modifying the red sample code with actual parameters, run:

```
$ mobivision rename -i human_PBMC_Control \
-d /media/mz/analysis/R1-S0-1_outs
```

All parameter descriptions are as follows:

	Parameter	Description
Required	-i	New SampleID for result files.
	-d	Output result folder from mobivision quantify.
Help	-h, --help	Display help information.

1.6 mtx2csv

After completing mobivision quantify, the result files include cell-gene expression matrices filtered cell gene matrix and raw cell gene matrix. mobivision mtx2csv can convert cell-gene expression matrices to csv files and output them. The mobivision mtx2csv command input parameters include the folder containing the cell-gene expression matrix (-i), which contains matrix files, barcodes files, and features files, and the output csv file name (-o). After modifying the red sample code with actual parameters, run:

$ mobivision mtx2csv -i /media/mz/analysis/R1-S0-1_outs/filtered_cell_gene_matrix \
-o /output/R1-S0-1.csv

All parameter descriptions are as follows:

	Parameter	Description
Required	-i	Folder containing the cell-gene expression matrix, including matrix files, barcodes files, and features files.
	-o	Csv file name.
Help	-h, --help	Display help information.

1.7 integrate

After completing mobivision quantify for multiple samples, the result files include cell-gene expression matrices filtered cell gene matrix. mobivision integrate can integrate and analyze cell-gene expression matrices from multiple samples and output qlentille files and h5ad files. The mobivision integrate command input parameters include a list recording the paths of multiple sample expression matrices (-l), the path containing input matrix files (-d), the selected integration algorithm (-a), and the output file path (-o). After modifying the red sample code with actual parameters, run:

$ mobivision integrate -l sample.list -o /output/

All parameter descriptions are as follows:

	Parameter	Description
Required	-l	List recording the paths of multiple sample expression matrices, with the first column as sample name and the second column as the filtered_cell_gene_matrix path corresponding to the sample
	-d	Filtered_cell_gene_matrix path corresponding to the sample
	-o	Output file path
Optional	-a	Selected integration algorithm, default is: harmony
Help	-h, --help	Display help information.

2 V(D)J Analysis Module

2.1 vdj

mobivision vdj is adapted for libraries and sequencing data prepared by MobiCube® high-throughput single-cell V(D)J kit; it can analyze single-cell V(D)J sequencing data from the MobiNova platform. Using paired-end sequencing fastq files as input, it can ultimately output html quality control reports, annotations, and clonotype-related result files.

2.1.1 Preparing Fastq Data

Input fastq files should be named in the following format:

```
[SampleName]_[ReadType].[Suffix]
```

Where,

[SampleName] consists of letters, numbers, or underscores;

[ReadType] suggested format: R1, R2, R1_001, R2_001;

[Suffix] currently supports 4 cases: fastq.gz, fq.gz, fastq, fq.

For example:

```
/Data/Sample_fastq/

├── R22001221-20220105-Emix-3-20220105-S-F5-2-2_R1.fq.gz

└── R22001221-20220105-Emix-3-20220105-S-F5-2-2_R2.fq.gz
```

2.1.2 MobiVision vdj Usage

The mobivision vdj command input parameters include the path to fastq files (-f), the path to the folder containing the index (-i), the maximum number of threads to call (-t), and the output folder path (-o). After modifying the red sample code with actual parameters, run:

$ mobivision vdj -f /PATH_TO_FASTQ/230113-LYY-N7T-O04/convert/mobi \
-i /PATH_TO_REFERENCE/GRCh38_vdj/ \
-o /PATH_TO_OUTPUT

All parameter descriptions are as follows:

	Parameter	Description
Required	-f, --fastqDir	Path to fastq files.
	-i, --referencePath	Path containing MobiCube transcriptome reference sequence files. For example: /home/refer/GRCh38_vdj/.
	-o, --outDir	Output file path.
	-t, --threads	Number of threads used by the software, default value: 8.
Optional	-s, --sampleID	User-specified sample ID for result files.
	-c, --chainType	Specify the chain type of the library: 'TR' for T cell receptor, 'IG' for B cell receptor, 'ALL' for simultaneous analysis of T cell receptor and B cell receptor, 'auto' for automatic chain type detection.
	-p, --innerPrimerPath	Specify the primer sequences on the C gene of the species, format is txt, one primer sequence per line. This needs to be specified when the species is other than human, mouse, or rabbit already included in MobiCube kits, such as monkey, alpaca, etc.
Help	-h, --help	Display help information.

Individual TCR/BCR output result files are as follows:

```
230113-LYY-N4T-O01
├── _flagdone
                      Flag file for successful run
├── _log
                        Run log file
├── run_analysis_cmds.txt
                       Run command line record
└── 230113-LYY-N4T-O01_outs
                       Output result file root directory
    ├── 230113-LYY-N4T-O01_airr_rearrangement.tsv                  AIRR format annotation
results
    ├── 230113-LYY-N4T-O01_all_contig_annotations.csv             CSV format annotation
results for all contigs
    ├── 230113-LYY-N4T-O01_all_contig.fasta                              FASTA format
sequence file for all contigs
    ├── 230113-LYY-N4T-O01_clonotypes.csv                                Clonotype result
file
    ├── 230113-LYY-N4T-O01_filtered_contig_annotations.csv     Filtered contig annotation
results
    ├── 230113-LYY-N4T-O01_filtered_contig.fasta                        FASTA format
sequence file for filtered contigs
    ├── 230113-LYY-N4T-O01_metrics_summary.csv                    CSV format analysis
summary file
```

```
    ├── 230113-LYY-N4T-O01_Report.html                                    HTML format
quality control report
    └── 230113-LYY-N4T-O01_Report.json                                     JSON format
quality control report
```

Combined library TCR/BCR analysis results are as follows (IG_result is BCR result file, TR_result is TCR result file):

```
230113-LYY-N4T-O01
├── _flagdone
                      Flag file for successful run
├── _log
                       Run log file
├── run_analysis_cmds.txt
                       Run command line record
└── 230113-LYY-N4T-O01_outs
                       Output result file root directory
├── 230113-LYY-N4T-O01_metrics_summary.csv                      CSV format analysis
summary file
├── 230113-LYY-N4T-O01_Report.html                                HTML format quality
control report

├── 230113-LYY-N4T-O01_Report.json                                 JSON format quality
control report

└── IG_result

│    ├── 230113-LYY-N4T-O01_airr_rearrangement.tsv              AIRR format annotation
results

│    ├── 230113-LYY-N4T-O01_all_contig_annotations.csv         CSV format annotation
results for all contigs

│    ├── 230113-LYY-N4T-O01_all_contig.fasta                          FASTA format
sequence file for all contigs

│    ├── 230113-LYY-N4T-O01_clonotypes.csv                            Clonotype result
file

│    ├── 230113-LYY-N4T-O01_filtered_contig_annotations.csv Filtered contig annotation
results

│    ├── 230113-LYY-N4T-O01_filtered_contig.fasta                   FASTA format sequence
file for filtered contigs

└── TR_result

├── 230113-LYY-N4T-O01_airr_rearrangement.tsv            AIRR format annotation results

├── 230113-LYY-N4T-O01_all_contig_annotations.csv       CSV format annotation results for
all contigs
```

```
├── 230113-LYY-N4T-O01_all_contig.fasta                          FASTA format sequence file
for all contigs

├── 230113-LYY-N4T-O01_clonotypes.csv                            Clonotype result file

├── 230113-LYY-N4T-O01_filtered_contig_annotations.csv   Filtered contig annotation
results

└── 230113-LYY-N4T-O01_filtered_contig.fasta                 FASTA format sequence file
for filtered contigs
```

3 CITE Analysis Module

3.1 cite

mobivision cite is adapted for libraries and sequencing data prepared by MobiCube® high-throughput single-cell CITE-seq (A-Human) kit; it can calculate transcriptome and cell surface protein expression information at the single-cell level. Using transcriptome fastq files and protein Tag fastq files as input, it can ultimately output html quality control reports and gene expression matrices for downstream in-depth analysis.

3.1.1 Preparing Fastq Data

Input fastq files contain two groups of library data: transcriptome and proteome. These two types of data need to be placed in different folders and named in the following format:

```
    [SampleName]_[ReadType].[Suffix]
```

Where,

[SampleName] consists of letters, numbers, or underscores;

[ReadType] suggested format: R1, R2, R1_001, R2_001;

[Suffix] currently supports 4 cases: fastq.gz, fq.gz, fastq, fq.

For example,

Transcriptome data is as follows:

:::
/data/Mobi/test_data/rna_data

├── R22001221-20220105-Emix-3-20220105-S-F5-2-2_R1.fq.gz

└── R22001221-20220105-Emix-3-20220105-S-F5-2-2_R2.fq.gz
:::

Proteome data is as follows:

:::
/data/Mobi/test_data/adt

├── R22001221-20220105-ADT-3-20220105-S01_R1.fq.gz

└── R22001221-20220105-ADT-3-20220105-S01_R2.fq.gz

:::

3.1.2 MobiVision cite Usage

The mobivision cite command input parameters include the path to transcriptome fastq files (-f), the path to proteome fastq files (-fb), the path to the folder containing the index (-i), the maximum number of threads to call (-t), the path to the taglist file (-b), and the output folder path (-o). After modifying the red sample code with actual parameters, run:

$ mobivision cite -f /data/Mobi/test_data/rna_data \
-fb /data/Mobi/test_data/adt \
-i /data/References/GRCh38-2023 \
-t 8 \
-b tag_ref50.csv \
-o output06

All parameter descriptions are as follows:

	Parameter	Description
Required	-f, --fastqDir	Path to transcriptome fastq files. For example: /media/mz-3/mzRaid10/Mobi_data/Sample_R22014039-C07.
	-fb,	Path to proteome Tag fastq files. For example: /media/mz-3/mzRaid10/Mobi_data/Sample_T22014039-T01
	-i, --indexDir	Path containing MobiCube transcriptome reference sequence files, requires index constructed with MobiVision v2.2 or later versions. If downloading the tar.gz format index file provided by MobiCube, you need to extract the tar.gz first, then use the extracted file path as the path for the -i parameter. For example: /media/mz-3/db/gencode_hsa/GRCh38.
	-b, --tagBarcodeInfoFile	Path to antibody Barcode sequence file, file is in csv format with "id,sequence" as header, each row corresponds to an antibody name and the corresponding barcode sequence, for example "CD56,TTTTAAATCGAT".
	-o, --outDir	Output file path. For example: /Data/Results/T22014039-T01.
Optional	--intron	Include gene introns for gene expression analysis. Options include "excluded" and "included". Default setting is "included".
	-m, --mismatchNum	Number of tolerated mismatched bases when aligning antibody tag sequences, default value is 1.
	-p, --captureType	Antibody tag capture configuration, default value is "Antibody", optional values are "Antibody" and "Multiplexing".
	-t, --threads	Maximum number of threads used by the software. For example: 12.
Help	-h, --help	Display help information.

mobivision cite default analysis method for transcriptome data includes gene introns. If only exons are considered as genes, you can add the "--intron excluded" parameter. The default setting is "--intron included".

Output result files are as follows:

```
├── _flagdone                                                    Flag file
for successful run

├── _log                                                          Run log
file

├── run_analysis_cmds.txt                                      Complete command
line record

└── SAMPLEID_outs                                               Output result
file root directory

  ├── SAMPLEID _R2_fractions.csv              Tag library detected all tag sequence types
and statistical results

  ├── SAMPLEID _Report.html                     HTML format quality control report

  ├── SAMPLEID _Report.json                      JSON format quality control report

  ├── SAMPLEID _valid_tags_fractions.csv   Tag library detected statistical results for
each antibody tag sequence

  └── sample_feature_bc_matrix                    Transcriptome and proteome expression
matrix files

      ├── barcodes.tsv.gz                                      Filtered barcode file

      ├── features.tsv.gz                                        Filtered gene and
antibody list file

      └── matrix.mtx.gz                                       Filtered expression
matrix file
```

4 ChIP Analysis Module

4.1 chip

mobivision chip is adapted for libraries and sequencing data prepared by MobiCube® high-throughput single-cell ChIP-seq kit; it can achieve single-cell level gene regulatory site analysis. Using paired-end sequencing fastq files as input, it can ultimately output html quality control reports, peak matrices, and related result files for downstream in-depth analysis.

4.1.1 Preparing Fastq Data

Input fastq files should be named in the following format:

```
[SampleName]_[ReadType].[Suffix]
```

Where,

[SampleName] consists of letters, numbers, or underscores, it is not recommended to include R1, R2;

[ReadType] suggested format: R1, R2, R1_001, R2_001;

[Suffix] currently supports 4 cases: fastq.gz, fq.gz, fastq, fq.

For example:

:::
/Data/Sample_fastq/

├── R22001221-20220105-Emix-3-20220105-S-F5-2-2_R1.fq.gz

└── R22001221-20220105-Emix-3-20220105-S-F5-2-2_R2.fq.gz
:::

4.1.2 MobiVision chip Usage

The reference genome construction method for mobivision atac is the same as that for mobivision chip, and you can directly use the mk_chip_ref subcommand. For detailed usage instructions, please refer to the mk_chip_ref section above.

4.1.3 MobiVision atac Usage

The mobivision chip command input parameters include the path to the folder containing the index (-i), the maximum number of threads to call (-t), the path to fastq files (-f), and the output folder path (-o). After modifying the red sample code with actual parameters, run:

```
$ mobivision chip -i /share/Reference/mobi_chip_ref/GRCh38 \
-t 12 \
-f /share/Data/Sample_SQ23009375-230428C-S-YXH-L06 \
-o /share/Outs/230428C-S-YXH-L06
```

All parameter descriptions are as follows:

	Parameter	Description
Required	-f, --fastqDir	Path to fastq files. For example: /share/Data/Sample_SQ23009375-230428C-S-YXH-L06
	-i, --referencePath	Path containing MobiCube ChIP reference sequence files. If downloading the tar.gz format index file provided by MobiCube, you need to extract the tar.gz first, then use the extracted file path as the path for the -i parameter. For example: /share/Reference/mobi_chip_ref/GRCh38
	-o, --outDir	Output file path. For example: /share/Outs/230428C-S-YXH-L06
	-t, --threads	Maximum number of threads used by the software. For example: 12
Optional	--peaktype	Expected peak type, includes 'broad' and 'narrow' options. If not specified, 'narrow' is used by default.
	--cellnumber	Force the use of this cell number for cell screening. If not specified, CR2.2 algorithm is used by default for cell screening.
	--control	Path to IgG sample fastq files. If this parameter is used, IgG mode is activated by default for peak signal background correction.
Help	-h, --help	Display help information.

Output result files are as follows:

```
├── _flagdone                                                    Flag file
for successful run

├── _log                                                          Run log
file

├── run_analysis_cmds.txt                                      Complete command
line record

└── SAMPLEID_out                                                Output result
file root directory

  ├── filtered_cell_fragments_matrix              Filtered cell-fragment matrix file
root directory

  │    ├── barcodes.tsv.gz                              Filtered cell file

  │    ├── fragments.tsv.gz                             Filtered fragment file

  │    └── matrix.mtx.gz                                Filtered cell-fragment matrix
file

  ├── filtered_cell_peaks_matrix                     Filtered cell-peak matrix file root
directory

  │    ├── barcodes.tsv.gz                             Filtered cell file

  │    ├── peaks.tsv.gz                                Filtered peak file
```

```
   │    ├── filtered.h5ad                                H5AD format filtered data matrix
   │    └── matrix.mtx.gz                                Filtered cell-peak matrix file
   ├── SAMPLEID.bam                                      Alignment result file
   ├── SAMPLEID.bw                                       Alignment result visualization
file
   ├── SAMPLEID.filtered.bed.gz                          Deduplicated and filtered fragment
file
   ├── SAMPLEID_Report.html                              HTML format quality control report
   ├── SAMPLEID_Report.json                              JSON format quality control report
   ├── summary.csv                                       CSV format library information
   ├── raw_cell_peaks_matrix                             Unfiltered cell-peak matrix file
root directory
   │    ├── barcodes.tsv.gz                              Unfiltered cell file
   │    ├── peaks.tsv.gz                                 Unfiltered peak file
   │    └── matrix.mtx.gz                                Unfiltered cell-peak matrix file
   ├── fragmentsInCells.tsv.gz                           Fragment file obtained after
filtering cells
   └── SAMPLEID.narrowPeak/broadPeak                Full library peak file
```

4.2 mk_chip_ref

mobivision mk_chip_ref uses genome fasta files as input to construct MobiVision-chip compatible reference genome files. The mobivision mk_chip_ref command input parameters include genome name (-n), genome fasta file (-f), and the path to output reference genome files (-r). The constructed reference genome files will be saved in a folder named after the genome name (-n), which is located in the path specified by -r. After modifying the red sample code with actual parameters, run:

$ mobivision mk_chip_ref -n GRCh38 \
-f /media/db_out/GRCh38.primary_assembly.genome.fa.gz \
-r /share/Reference/chip

4.2.1 Multi-species Construction

If you need to construct samples combining two or more species, the -n-f parameters need to be specified twice or multiple times. The genome names and genome files specified by -n-f need to correspond one by one according to the order. Sample code reference is as follows:

```
$ mobivision mk_chip_ref -n GRCh38 \
-n GRCm39 \
-f /media/db_out/GRCh38.primary_assembly.genome.fa.gz \
-f /media/db_out/GRCm39.primary_assembly.genome.fa.gz \
-r /share/Reference/chip
```

4.2.2 Parameter Description

All parameter descriptions are as follows:

	Parameter	Description
Required	-n	Genome name consists of [a-zA-Z0-9]+. *If this reference genome consists of multiple different genomes, please specify the -n parameter multiple times, for example -n <species 1> -n . The output reference genome folder will be named* and_.
	-g	GTF file (.gtf or .gtf.gz). If multiple GTF genome annotation files need to be input, please specify the -g parameter multiple times. If -n and -f are specified multiple times, then -g must be specified the same number of times in the corresponding order. Note: When constructing reference genomes with mkindex, the mobivision software has requirements for input gtf files, for example: gtf files must have more than 6 columns, the third column needs to contain gene, exon, transcript information, and the last column needs to contain gene_id, gene_name, gene_type and other information
	-f	Genome fasta file. If multiple fasta genome files need to be input, please specify the -f parameter multiple times. Genome fasta file supported formats include .fasta, .fa.gz, and .fna.gz.
	-r	Output path for reference genome.
Help	-h, --help	Display help information.

4.2.3 Output Display

Output result files are as follows:

```
├── GENOME.1.bt2 (l)          Reference genome first-level BWT file
├── GENOME.2.bt2(l)           Reference genome second-level BWT file
├── GENOME.3.bt2(l)           Reference genome third-level BWT file
├── GENOME.4.bt2(l)           Reference genome fourth-level BWT file
├── GENOME.chrom.sizes         Reference genome chromosome size file
├── GENOME.genome.fa          Reference genome fasta file
├── GENOME.genome.fa.fai       Reference genome fasta index file
```

├── GENOME.reference.json Reference genome construction information file

├── GENOME.rev.1.bt2(l) Reference genome first-level BWT reverse complement sequence file

└── GENOME.rev.2.bt2(l) Reference genome second-level BWT reverse complement sequence file

5 ATAC Analysis Module

5.1 atac

mobivision atac is adapted for libraries and sequencing data prepared by MobiCube® high-throughput single-cell ATAC-seq kit; it can achieve single-cell level gene regulatory site analysis. Using paired-end sequencing fastq files as input, it can ultimately output html quality control reports, peak matrices, and related result files for downstream in-depth analysis.

5.1.1 Preparing Fastq Data

Input fastq files should be named in the following format:

```
[SampleName]_[ReadType].[Suffix]
```

Where,

[SampleName] consists of letters, numbers, or underscores, it is not recommended to include R1, R2;

[ReadType] suggested format: R1, R2, R1_001, R2_001;

[Suffix] currently supports 4 cases: fastq.gz, fq.gz, fastq, fq.

For example:

:::
/Data/Sample_fastq/

├── R22001221-20220105-Emix-3-20220105-S-F5-2-2_R1.fq.gz

└── R22001221-20220105-Emix-3-20220105-S-F5-2-2_R2.fq.gz
:::

5.1.2 Reference Genome Construction

The reference genome construction method for mobivision atac is consistent with mobivision chip, and the mk_chip_ref subcommand can be used directly. For detailed usage methods, please refer to the mk_chip_ref content above.

5.1.3 MobiVision atac Usage

The mobivision atac command input parameters include the path to the folder containing the index (-i), the maximum number of threads to call (-t), the path to fastq files (-f), and the output folder path (-o). After modifying the red sample code with actual parameters, run:

$ mobivision atac -i /share/Reference/mobi_atac_ref/GRCh38 \
-t 12 \
-f /share/Data/Sample_SQ23009375-230428C-S-YXH-L06 \
-o /share/Outs/230428C-S-YXH-L06

All parameter descriptions are as follows:

	Parameter	Description
Required	-f, --fastqDir	Path to fastq files. For example: /share/Data/Sample_SQ23009375-230428C-S-YXH-L06
	-i, --referencePath	Path containing MobiCube ATAC reference sequence files. If downloading the tar.gz format index file provided by MobiCube, you need to extract the tar.gz first, then use the extracted file path as the path for the -i parameter. For example: /share/Reference/mobi_atac_ref/GRCh38
	-o, --outDir	Output file path. For example: /share/Outs/230428C-S-YXH-L06
	-t, --threads	Maximum number of threads used by the software. For example: 12
Optional	--peaktype	Expected peak type, includes 'broad' and 'narrow' options. If not specified, 'narrow' is used by default.
	--cellnumber	Force the use of this cell number for cell screening. If not specified, CR2.2 algorithm is used by default for cell screening.
	--control	Path to IgG sample fastq files. If this parameter is used, IgG mode is activated by default for peak signal background correction.
Help	-h, --help	Display help information.

Output result files are as follows:

```
├── _flagdone                                                    Flag file
for successful run

├── _log                                                          Run log
file

├── run_analysis_cmds.txt                                      Complete command
line record

└── SAMPLEID_out                                                Output result
file root directory

  ├── filtered_cell_fragments_matrix              Filtered cell-fragment matrix file root
directory

  │   ├── barcodes.tsv.gz                              Filtered cell file

  │   ├── fragments.tsv.gz                             Filtered fragment file

  │   └── matrix.mtx.gz                                Filtered cell-fragment matrix
file
```

```
  ├── filtered_cell_peaks_matrix                    Filtered cell-peak matrix file root
directory
  │   ├── barcodes.tsv.gz                             Filtered cell file
  │   ├── peaks.tsv.gz                                Filtered peak file
  │   ├── filtered.h5ad                               H5AD format filtered data matrix
  │   └── matrix.mtx.gz                               Filtered cell-peak matrix file
  ├── SAMPLEID.bam                                  Alignment result file
  ├── SAMPLEID.bw                                     Alignment result visualization
file
  ├── SAMPLEID.filtered.bed.gz                    Deduplicated and filtered fragment file
  ├── SAMPLEID_Report.html                        HTML format quality control report
  ├── SAMPLEID_Report.json                        JSON format quality control report
  ├── summary.csv                                     CSV format library information
  ├── raw_cell_peaks_matrix                         Unfiltered cell-peak matrix file
root directory
  │   ├── barcodes.tsv.gz                           Unfiltered cell file
  │   ├── peaks.tsv.gz                                Unfiltered peak file
  │   └── matrix.mtx.gz                               Unfiltered cell-peak matrix file
  ├── fragmentsInCells.tsv.gz                     Fragment file obtained after filtering
cells
  ├── fragmentsInPeaks.tsv.gz                     Fragment file overlapping with peaks
  └── SAMPLEID.narrowPeak/broadPeak          Full library peak file
```

(V) Reference Sequence Information

Reference Sequence Name	Species	Reference Sequence Version	Download URL
GRCh38	Human	Gencode_v38*	GENCODE - Human Release 38 (gencodegenes.org)
		Ensembl_v104	Index of /pub/release-104/fasta/homo_sapiens/dna (ensembl.org)
GRCm39	Mouse	Gencode_vM27*	GENCODE - Mouse Release M27 (gencodegenes.org)
		Ensembl_v104	Index of /pub/release-104/fasta/mus_musculus/dna (ensembl.org)
GRCh38_and _GRCm39	Human+Mouse	Gencode_v38*	GENCODE - Human Release 38 (gencodegenes.org)
		Gencode_vM27*	GENCODE - Mouse Release M27 (gencodegenes.org)
	Human+Mouse	Ensembl_v104	Index of /pub/release-104/fasta/homo_sapiens/dna (ensembl.org)
			Index of /pub/release-104/fasta/mus_musculus/dna (ensembl.org)
OryCun2	Rabbit	Gencode	-
		Ensembl_v109*	Index of /pub/release-109/fasta/oryctolagus_cuniculus/dna (ensembl.org)

Note: Reference sequence versions marked with * are provided by the MobiNova platform