

MobiVision v4.1 使用说明书

(一) 软件介绍

MobiVision生信分析软件，可分析MobiNova平台来源的单细胞测序数据-3'/5'转录组，全长VDJ，CITE蛋白组和ChIP、ATAC测序数据。目前，MobiVision v4.1共拥有十二个子命令以及一个辅助功能，分别为：

1、转录组分析模块，包含以下七个子命令：

- **quantify**主要用于处理单细胞转录组数据（.fastq格式），输出的结果为文库质控报告、细胞基因表达矩阵（.mtx）等相关文件。
- **fmt_gtf**可用于处理gtf注释文件，通过指定所需的注释类型，定制个性化gtf文件。
- **mkindex**用于构建quantify分析所需的参考基因组。
- **rename**可将quantify的分析结果进行重命名，包括质控报告中所显示的样本ID。
- **mtx2csv**可将quantify分析得到的细胞基因表达矩阵（.mtx）转换为.csv文件。
- **re_call_cell**可对quantify输出的结果重新进行细胞筛选，并生成新的质控报告及细胞基因表达矩阵（.mtx）。
- **integrate**可对quantify 输出的的细胞基因表达矩阵进行整合。

2、V(D)J分析模块，包含以下一个子命令：

- **vdj**主要用于处理单细胞V(D)J测序数据（.fastq格式），输出的结果为V(D)J文库质控报告、注释以及克隆型等相关结果文件。

3、CITE分析模块，包含以下一个子命令：

- **cite**主要用于处理CITE测序数据（.fastq格式），输出的结果为文库质控报告、基因以及蛋白的表达矩阵文件。

4、ChIP分析模块，包含以下两个子命令：

- **chip**主要用于处理单细胞ChIP数据（.fastq格式），输出为文库质控报告、peak矩阵等相关结果文件。
- **mk_chip_ref**用于构建chip分析所需的参考基因组文件。

5、ATAC分析模块，包含以下一个子命令：

- **atac**主要用于处理单细胞ATAC数据（.fastq格式），输出为文库质控报告、peak矩阵等相关结果文件。

6、Bam_to_fastq辅助功能

- **Bam_to_fastq**主要用于将quantify分析得到的bam文件转化成fastq格式。

（二）系统要求

- 8核心的Intel或AMD处理器，x86架构 (推荐16核或以上)
- 64GB或以上
- 1TB的可用空间
- Linux操作系统，64-bit CentOS、Ubuntu 20.04或更高版本

（三）安装方法

解压mobivision-v4.1.tar.gz后，在shell命令行中执行一下source命令，即可使用MobiVision v4.1软件。每次打开一个新的shell窗口或终端界面时都需要执行一下“source”命令。

```
$ tar -zxvf mobivision-v4.1.tar.gz
$ source mobivision-v4.1/source.sh
```

```
#测试mobivision是否安装成功
$ mobivision --help
```

当以下信息显示时，说明MobiVision软件已经成功安装

```

$ mobivision

Mobivision CLI, MobiVision Team

MOBIVISION
https://www.mobidrop.com/cn/bioinformatics-analysis-software/mobivision-news
Copyright: MobiVision® (c) 2025 MobiVision Team

<mobivision-v4.1>

Process MobiDrop Single Cell Sequencing Data, Include Five Omics:
[scRNA-seq:scrna, scATAC-seq:scatac, scChIP-seq:scchip, CITE-seq:scsite, V(D)J:scvdj]

USAGE:
  mobivision <SUBCOMMAND>

FLAGS:
  -h, --help      Print help information
  -V, --version    Print version information

SUBCOMMANDS:

  quantify      Count gene expression reads from a single sample
  mkindex        Prepare a reference for mobivision software
  fmt_gtf        Format and filter gene type features in annotation GTF file
  rename        Rename the sample_id of output files
  mtx2csv        Convert the matrix files to csv
  re_call_cell   Re do cell-calling using the results from mobivision quantify analysis
  integrate      Integrate the cell gene matrices from MobiVision quantify
  vdj            Analysis of single-cell V(D)J sequencing data
  mk_chip_ref    Prepare a reference for mobivision chip analysis
  chip           Analyze ChIP fastq data from a single sample
  atac           Analyze ATAC fastq data from a single sample
  cite           Analysis of data from CITE-seq

```

(四) 使用说明

1 转录组分析模块 (3'RNA、5'RNA)

1.1 quantify

mobivision quantify 适配于MobiCube® 高通量单细胞 3' 转录组试剂盒 (3'RNA) 以及 MobiCube® 高通量单细胞 5' 转录组试剂盒 (5'RNA) 制备的文库和测序数据；可以计算单细胞转录组文库中每个细胞的基因表达水平。以双端测序的fastq文件作为输入，最终可输出html质控报告以及基因表达矩阵，可用于下游作深入分析。

1.1.1 准备Fastq数据

输入的fastq需按以下方式命名:

[SampleName]_[ReadType].[Suffix]

其中,

[SampleName]由字母、数字或下划线组成;

[ReadType]建议格式: R1、R2、R1_001、R2_001;

[Suffix]目前支持4种情况: fastq.gz、fq.gz、fastq、fq。

例如:



/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使用方法

mobivision quantify命令输入的参数包括index所在的文件夹路径(-i), 调用的最多线程数(-t), fastq文件所在的路径(-f)以及输出的文件夹路径(-o)。将红色样例代码修改为实际参数后, 运行:

```
$ 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
```

所有参数说明如下:

	参数	描述
必需	-f, --fastqDir	fastq文件所在路径。例如: /media/mz-3/mzRaid10/Mobi_data/Sample_R22014039-C07。
	-i, --indexDir	包含墨卓转录组参考序列文件所在的路径。如果下载的是墨卓官方格式的index文件, 需要先将tar.gz解压, 然后以解压后的文件路径作为路径。例如: /media/mz-3/db/gencode_hsa/GRCh38。
	-o, --outDir	输出文件路径。例如: /media/mz-3/analysis/R22014039-C07。
	-t, --threads	软件使用的最大线程数目。例如: 12。

可选	--intron	<包含intron的选项> 将基因的内含子考虑进去，用于基因表达量的分析. 选项包括"excluded"和"included". 默认设置是"included".
	--cellnumber	<最大细胞数> 强制使用这个细胞数用于细胞的筛选。
	--cr2.2	使用CR2.2的算法进行细胞筛选。如果不指定，默认使用EmptyDrops算法。
	-s, --sampleID	<指定样本名> 用户指定的结果文件的样本id。
	--nobam	设置是否保留BAM文件的参数，默认保留STAR生成的BAM文件；若不指定，则不会保留STAR输出的BAM文件。
	--kit	设置用于构建文库的RNA试剂盒版本，以及指定分析的文库类型（5'和3'RNA）。如果不指定，默认为“Unknown”，系统将自动判定分析5'和3'RNA。
帮助	-h, --help	显示帮助信息。

mobivision quantify 通过-f 指定fastq文件所在的路径。若该路径下仅有一对fastq文件，将会默认分析该对fastq文件；若该路径下存在多对fastq文件，那么这些文件将被合并为一对fastq文件，并保存在名为“merge/”的文件夹中。后续的分析将以合并后的fastq文件作为输入。因此，建议将同一样本来源的多次测序数据置于同一文件夹。

在mobivision quantify命令中有三种挑取细胞的算法，

1. 使用“--cellnumber 数值”，mobivision quantify就会用这个数值作为强制提取的细胞数，
2. 如果使用“--cr2.2”，那么就采用CellRanger2.2算法用于细胞过滤，
3. 其它情况下采用默认的EmptyDrops算法。

默认分析结果中的基因包含内含子，若仅考虑将外显子为基因，可通过添加“--intron excluded”参数，默认设置为“--intron included”。

输出结果文件如下：

...	—— _flagdone	运行成功的flag文件
	—— _log	运行日志文件
	—— SAMPLEID_outs	输出结果文件根目录

└── filtered_cell_gene_matrix	过滤后的matrix文件根目录
└── barcodes.tsv.gz	过滤后的barcode文件
└── features.tsv.gz	过滤后的基因列表文件
└── matrix.mtx.gz	过滤后的表达矩阵文件
└── SAMPLEID_Aligned.sort.bam	比对结果文件
└── SAMPLEID_Aligned.sort.bam.bai	比对结果文件的index
└── SAMPLEID_filtered.h5ad	h5ad格式的过滤后数据矩阵
└── SAMPLEID_Report.html	html格式的质控报告
└── SAMPLEID_Report.json	json格式的质控报告
└── SAMPLEID_summary.csv	csv格式的文库信息
└── raw_cell_gene_matrix	未过滤的matrix文件根目录
└── barcodes.tsv.gz	未过滤的barcode文件
└── features.tsv.gz	未过滤的基因列表文件
└── matrix.mtx.gz	未过滤的表达矩阵文件
└── result_mito_percentage.csv	线粒体百分比信息文件
└── run_analysis_cmds.txt	完整命令行记录

1.2 fmt_gtf

mobivision fmt_gtf用于处理基因组注释文件（GTF文件），处理后的注释文件可用于构建参考转录组。mobivision fmt_gtf命令输入的参数包括输入GTF文件(-i)，输出GTF文件（-o）及需要在输出GTF文件中保留的基因类型（-t）。将红色样例代码修改为实际参数后，运行：

```
$ 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
```

所有参数说明如下：

	参数	描述
必需	-i	输入的GTF文件名称，支持.gtf和.gtf.gz两种文件格式。
	-o	输出的GTF文件名称，默认为“out_filtered.gtf”。
可选	-t	需要在输出的GTF文件中保留的基因类型，支持多次指定。
帮助	-h, --help	显示帮助信息。

1.3 mkindex

mobivision mkindex以基因组fasta文件和GTF注释文件作为输入，构建MobiVision兼容的参考转录组。mobivision fmt_gtf命令输入的参数包括基因组名称（-n），基因组fasta文件（-f）和基因组GTF注释文件（-g）。构建得到的参考转录组将被保存于以基因组名称（-n）命名的文件夹中，该文件夹位于运行命令所在的路径下。将红色样例代码修改为实际参数后，运行：

```
$ 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
```

若需构建两个或者多个物种合并的样本，-n\ -f\ -g参数需两次或多次指定，-n\ -f\ -g所指定的基因组名称、基因组文件、基因组注释文件需根据先后顺序一一对应，样例代码参考如下：

```
$ 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
```

所有参数说明如下：

	参数	描述
必需	-n	基因组名称由【a-zA-Z0-9_】所组成。如果该参考基因组由多个不同的基因组所组成，请多次指定-n参数，例如 -n <genome1> -n <genome2>，输出的参考基因组文件夹将被命名为<genome1>_and_<genome2>。
	-f	基因组fasta文件。若需输入多个fasta基因组文件，请多次指定-f参数。基因组fasta文件支持的格式包括.fasta、.fa.gz和.fna.gz。
	-g	GTF文件 (.gtf 或 .gtf.gz)。若需输入多个GTF基因组注释文件，请多次指定-g参数。如果-n和-f指定了多次，那么-g必须按照对应的顺序，指定相同的次数。
可选	-r	输出参考基因组版本号，如"ref_v1.0", "ref-2022-06"。
	-m	使用STAR构建参考基因组所使用的最大内存（GB）。
帮助	-h, --help	显示帮助信息。

在转基因或者病毒感染等研究中，可能需要把感兴趣的基因添加到现有基因组中进行比对。以将某分型HIV基因添加至人的参考基因组为例，操作方法如下：

首先，使用wget命令从gencode官方网站下载构建该参考基因组所需的fasta和gtf文件，并进行解压。将红色样例代码修改为实际参数后，运行：

```
$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
```

其次，下载感兴趣的基因序列，使用cat命令新建fa文件并将基因序列复制到fa文件中，输入Ctrl+D保存。将第一行较长的名称替换为方便识别的较短的名称，并统计序列的碱基数目用于gtf文件的制作。将红色样例代码修改为实际参数后，运行：

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

接着，参考ensembl 官网的gtf文件的基本格式，构建感兴趣基因的.gtf文件。将红色样例代码修改为实际参数后，运行：

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

然后，合并原有参考基因组和目标基因的fa文件和gtf文件。将红色样例代码修改为实际参数后，运行：

```
$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
```

最后，利用mobivision mkindex命令构建新的基因组reference信息。将红色样例代码修改为实际参数后，运行：

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

1.4 re_call_cell

mobivision quantify的结果文件，用户可作为mobivision re_call_cell的输入文件，重新进行细胞筛选（cell-calling），并获得细胞-基因表达矩阵（reanalysis_matrix）、html质控报告及其他相关结果文件。mobivision re_call_cell命令输入的参数包括mobivision quantify的结果文件路径（-i）、re_call_cell结果输出路径（-o）。将红色样例代码修改为实际参数后，运行：

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

所有参数说明如下：

	参数	描述
必需	-i	mobivision quantify的结果文件路径。
	-o	mobivision re_call_cell输出的结果文件路径。
可选	--cellnumber	<最大细胞数> 强制使用这个细胞数用于细胞筛选。
	--ed	使用EmptyDrops的算法进行细胞筛选。如果不指定，默认使用CellRanger2.2算法。
帮助	-h, --help	显示帮助信息。

在mobivision re_call_cell命令中有三种挑取细胞的算法，

1. 使用“--cellnumber 数值”，mobivision re_call_cell就会用这个数值作为强制提取的细胞数，
2. 如果使用“--ed”，那么就采用EmptyDrops算法用于细胞过滤，
3. 其它情况下采用默认的CellRanger2.2算法。

1.5 rename

完成mobivision quantify后，mobivision rename可以重命名所有文件名中包含sampleID的文件，同时可更改html报告中所展示的“Sample ID”信息。mobivision rename命令输入的参数包括新的sampleID（-i）和mobivision quantify输出的结果目录（-d）。将红色样例代码修改为实际参数后，运行：

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

所有参数说明如下：

	参数	描述
必需	-i	结果文件新SampleID。
	-d	mobivision quantify输出的结果文件夹。
帮助	-h, --help	显示帮助信息。

1.6 mtx2csv

完成mobivision quantify后，结果文件包含细胞基因表达矩阵filtered cell gene matrix 和 raw cell gene matrix。mobivision mtx2csv可以将细胞基因表达矩阵转换为csv文件并输出。mobivision mtx2csv命令输入的参数包括细胞基因表达矩阵所在的文件夹（-i），包含matrix文件、barcodes文件和features文件，以及输出csv文件名（-o）。将红色样例代码修改为实际参数后，运行：

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

所有参数说明如下：

	参数	描述
必需	-i	细胞基因表达矩阵所在的文件夹，包含matrix文件、barcodes文件和features文件。
	-o	csv文件名。
帮助	-h, --help	显示帮助信息。

1.7 integrate

完成多个样本的mobivision quantify后，结果文件包含细胞基因表达矩阵filtered cell gene matrix。mobivision integrate可以将多个样本的细胞基因表达矩阵进行整合分析并输出qlentille文件和h5ad文件。mobivision integrate命令输入的参数包括记录了多个样本表达矩阵路径的list

(-l) , 包含输入matrix文件的路径 (-d) 、选择的整合算法 (-a) , 以及输出文件路径 (-o) 。将红色样例代码修改为实际参数后, 运行:

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

所有参数说明如下:

	参数	描述
必需	-l	记录了多个样本表达矩阵路径的list, 第一列为样本名, 第二列为样本对应的filtered_cell_gene_matrix路径
	-d	样本对应的filtered_cell_gene_matrix路径
	-o	输出文件路径
可选	-a	选择的整合算法, 默认为: harmony
帮助	-h, --help	显示帮助信息。

2 免疫组分析模块 (V(D)J)

2.1 vdj

mobivision vdj适配于MobiCube® 高通量单细胞 V(D)J 试剂盒制备的文库和测序数据; 可以分析MobiNova平台来源的单细胞V(D)J测序数据。以双端测序的fastq文件作为输入, 最终可输出html质控报告、注释以及克隆型等相关结果文件。

2.1.1 准备Fastq数据

输入的fastq需按以下方式命名:

[SampleName]_[ReadType].[Suffix]

其中,

[SampleName]由字母、数字或下划线组成;

[ReadType]建议格式: R1、R2、R1_001、R2_001;

[Suffix]目前支持4种情况：fastq.gz、fq.gz、fastq、fq。

例如：



/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使用方法

mobivision vdj命令输入的参数包括fastq文件所在的路径(-f)，index所在的文件夹路径(-i)，调用的最多线程数(-t)以及输出的文件夹路径(-o)。将红色样例代码修改为实际参数后，运行：

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

所有参数说明如下：

	参数	描述
必需	-f, --fastqDir	fastq文件所在路径。
	-i, --referencePath	包含墨卓转录组参考序列文件所在的路径。例如：/home/refer/GRCh38_vdj/。
	-o, --outDir	输出文件路径。
	-t, --threads	软件使用的线程数目，默认值：8。
可选	-s, --sampleID	<指定样本名> 用户指定的结果文件的样本id。
	-c, --chainType	指定文库的链类型：'TR'表示T细胞受体，'IG'表示B细胞受体，'ALL'表示同时分析T细胞受体和B细胞受体，'auto'表示自动检测链类型。

	-p, -- innerPrimerPath	指定物种的C基因上的Primer序列，格式为txt，每行一条引物序列。当物种为墨卓试剂盒已包含的人，小鼠或兔子以外的情况时需指定，比如猴子，羊驼等等。
帮助	-h, --help	显示帮助信息。

单独TCR/BCR输出结果文件如下:

230113-LYY-N4T-O01

— _flagdone	运行成功的flag文件
— _log	运行日志文件
— run_analysis_cmds.txt	运行命令行记录
— 230113-LYY-N4T-O01_outs	输出结果文件根目录
— 230113-LYY-N4T-O01_airr_rearrangement.tsv	airr格式的注释结果
— 230113-LYY-N4T-O01_all_contig_annotations.csv	csv格式的所有重叠群的注释结果
— 230113-LYY-N4T-O01_all_contig.fasta	fasta格式的所有重叠群的序列文件
— 230113-LYY-N4T-O01_clonotypes.csv	克隆型结果文件
— 230113-LYY-N4T-O01_filtered_contig_annotations.csv	过滤后的重叠群的注释结果
— 230113-LYY-N4T-O01_filtered_contig.fasta	fasta格式的过滤重叠群的序列文件
— 230113-LYY-N4T-O01_metrics_summary.csv	csv格式的分析总结文件
— 230113-LYY-N4T-O01_Report.html	html格式的质控报告
— 230113-LYY-N4T-O01_Report.json	json格式的质控报告

合并建库TCR/BCR分析结果如下(IG_result为BCR结果文件，TR_result为TCR结果文件):

230113-LYY-N4T-O01

— _flagdone	运行成功的flag文件
— _log	运行日志文件
— run_analysis_cmds.txt	运行命令行记录
— 230113-LYY-N4T-O01_outs	输出结果文件根目录
— 230113-LYY-N4T-O01_metrics_summary.csv	csv格式的分析总结文件
— 230113-LYY-N4T-O01_Report.html	html格式的质控报告
— 230113-LYY-N4T-O01_Report.json	json格式的质控报告
— IG_result	
— 230113-LYY-N4T-O01_airr_rearrangement.tsv	airr格式的注释结果
— 230113-LYY-N4T-O01_all_contig_annotations.csv	csv格式的所有重叠群的注释结果
— 230113-LYY-N4T-O01_all_contig.fasta	fasta格式的所有重叠群的序列文件

—— 230113-LYY-N4T-O01_clonotypes.csv	克隆型结果文件
—— 230113-LYY-N4T-O01_filtered_contig_annotations.csv	过滤后的重叠群的注释结果
—— 230113-LYY-N4T-O01_filtered_contig.fasta	fasta格式的过滤重叠群的序列文件
—— TR_result	
—— 230113-LYY-N4T-O01_airr_rearrangement.tsv	airr格式的注释结果
—— 230113-LYY-N4T-O01_all_contig_annotations.csv	csv格式的所有重叠群的注释结果
—— 230113-LYY-N4T-O01_all_contig.fasta	fasta格式的所有重叠群的序列文件
—— 230113-LYY-N4T-O01_clonotypes.csv	克隆型结果文件
—— 230113-LYY-N4T-O01_filtered_contig_annotations.csv	过滤后的重叠群的注释结果
—— 230113-LYY-N4T-O01_filtered_contig.fasta	fasta格式的过滤重叠群的序列文件

3 蛋白组分析模块 (CITE)

3.1 cite

mobivision cite适配于MobiCube® 高通量单细胞 CITE-seq (A-Human)试剂盒制备的文库和测序数据；可以在单细胞水平计算转录组和细胞表面蛋白的表达信息。以转录组的fastq文件和蛋白Tag的fastq文件作为输入，最终可输出html质控报告以及基因表达矩阵，可用作下游深入分析。

3.1.1 准备Fastq数据

输入的fastq包含了两组文库数据，分别是转录组和蛋白组，两类数据需放在不同的文件夹中，按以下方式命名：

[SampleName]_[ReadType].[Suffix]

其中，

[SampleName]由字母、数字或下划线组成；

[ReadType]建议格式：R1、R2、R1_001、R2_001；

[Suffix]目前支持4种情况：fastq.gz、fq.gz、fastq、fq。

例如，

转录组的数据如下：



/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

蛋白组的数据如下：



/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使用方法

mobivision cite命令输入的参数包括转录组fastq文件所在的路径(-f)，蛋白组fastq文件所在的路径(-fb)，index所在的文件夹路径(-i)，调用的最多线程数(-t)，taglist文件所在的路径(-b)以及输出的文件夹路径(-o)。将红色样例代码修改为实际参数后，运行：

```
$ 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
```

所有参数说明如下：

	参数	描述
必需	-f, --fastqDir	转录组fastq文件所在路径。例如：/media/mz-3/mzRaid10/Mobi_data/Sample_R22014039-C07。
	-fb,	蛋白组Tag fastq文件所在路径。例如：/media/mz-3/mzRaid10/Mobi_data/Sample_T22014039-T01
	-i, --indexDir	包含墨卓转录组参考序列文件所在的路径，需使用MobiVision v2.2及以后版本构建的index。如果下载的是墨卓官方提供的tar.gz格式的index文件，需要先将tar.gz解压，然后以解压后的文件路径作为-i参数的路径。例如：/media/mz-3/db/gencode_hsa/GRCh38。

	-b, --tagBarcodeInfoFile	抗体Barcode序列文件的路径，文件为csv格式，以“id,sequence”为表头，每一行对应一个抗体名称和相应的barcode序列，例如“CD56,TTTTAAATCGAT”。
	-o, --outDir	输出文件路径。例如：/Data/Results/T22014039-T01。
可选	--intron	<包含intron的选项> 将基因的内含子考虑进去，用于基因表达量的分析。选项包括“excluded”和 “included”。默认设置是“included”。
	-m, --mismatchNum	<错配碱基数> 抗体tag序列比对时的容错碱基数，默认值是1。
	-p, --captureType	<Tag序列捕获类型> 抗体tag的捕获配型，默认值是“Antibody”，可选值是“Antibody”和“Multiplexing”。
	-t, --threads	软件使用的最大线程数目。例如：12。
帮助	-h, --help	显示帮助信息。

mobivision cite对转录组数据默认分析方法包含基因的内含子，若仅考虑将外显子为基因，可通过添加“--intron excluded” 参数，默认设置为“--intron included”。

输出结果文件如下：

—— _flagdone	运行成功的flag文件
—— _log	运行日志文件
—— run_analysis_cmds.txt	完整命令行记录
—— SAMPLEID_outs	输出结果文件根目录
—— SAMPLEID_R2_fractions.csv	Tag文库测出所有的tag序列类型和统计结果
—— SAMPLEID_Report.html	html格式的质控报告
—— SAMPLEID_Report.json	json格式的质控报告
—— SAMPLEID_valid_tags_fractions.csv	Tag文库测出的各个抗体tag序列的统计结果
—— sample_feature_bc_matrix	转录组和蛋白组的表达量matrix文件
—— barcodes.tsv.gz	过滤后的barcode文件
—— features.tsv.gz	过滤后的基因和抗体列表文件
—— matrix.mtx.gz	过滤后的表达矩阵文件

4 表观组分析模块 (ChIP)

4.1 chip

mobivision chip适配于MobiCube® 高通量单细胞 ChIP-seq 试剂盒制备的文库盒测序数据；可以实现单细胞水平的基因调控位点分析。以双端测序的fastq文件作为输入，最终可输出html质控报告、peak矩阵等相关结果文件，用于下游深入分析。

4.1.1 准备Fastq数据

输入的fastq需按以下方式命名：

[SampleName]_[ReadType].[Suffix]

其中，

[SampleName]由字母、数字或下划线组成，不建议包含R1、R2；

[ReadType]建议格式：R1、R2、R1_001、R2_001；

[Suffix]目前支持4种情况：fastq.gz、fq.gz、fastq、fq。

例如：



/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使用方法

mobivision chip命令输入的参数包括index所在的文件夹路径(-i)，调用的最多线程数(-t)，fastq文件所在的路径(-f)以及输出的文件夹路径(-o)。将红色样例代码修改为实际参数后，运行：

```
$ 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
```

所有参数说明如下：

	参数	描述
必需	-f, --fastqDir	fastq文件所在路径。例 如： /share/Data/Sample_SQ23009375-230428C-S-YXH-L06
	-i, --referencePath	包含墨卓ChIP参考序列文件所在的路径。如果下载的是墨卓官方提供的tar.gz格式的index文件，需要先将tar.gz解压，然后以解压后的文件路径作为 -i 参数的路径。例如： /share/Reference/mobi_chip_ref/GRCh38
	-o, --outDir	输出文件路径。例如： /share/Outs/230428C-S-YXH-L06
	-t, --threads	软件使用的最大线程数目。例如： 12
可选	--peaktype	预计得到的峰的类型，包含'broad'和'narrow'两个选项。若不指定，默认为使用'narrow'。
	--cellnumber	<最大细胞数> 强制使用这个细胞数用于细胞的筛选。若不指定，则默认使用CR2.2算法进行细胞筛选。
	--control	IgG样本fastq文件所在路径。若使用该参数，默认激活IgG模式，用于峰信号背景矫正。
帮助	-h, --help	显示帮助信息。

输出结果文件如下：

—— _flagdone	运行成功的flag文件
—— _log	运行日志文件
—— run_analysis_cmds.txt	完整命令行记录
—— SAMPLEID_out	输出结果文件根目录
—— filtered_cell_fragments_matrix	过滤后的细胞-片段矩阵文件根目录
—— barcodes.tsv.gz	过滤后的细胞文件
—— fragments.tsv.gz	过滤后的片段文件
—— matrix.mtx.gz	过滤后的细胞-片段矩阵文件
—— filtered_cell_peaks_matrix	过滤后的细胞-peak矩阵文件根目录

—— barcodes.tsv.gz	过滤后的细胞文件
—— peaks.tsv.gz	过滤后的peak文件
—— filtered.h5ad	h5ad格式的过滤后数据矩阵
—— matrix.mtx.gz	过滤后的细胞-peak矩阵文件
—— SAMPLEID.bam	比对结果文件
—— SAMPLEID.bw	比对结果可视化文件
—— SAMPLEID.filtered.bed.gz	去重并过滤后的片段文件
—— SAMPLEID_Report.html	html格式的质控报告
—— SAMPLEID_Report.json	json格式的质控报告
—— summary.csv	csv格式的文库信息
—— raw_cell_peaks_matrix	未过滤的细胞-peak矩阵文件根目录
—— barcodes.tsv.gz	未过滤的细胞文件
—— peaks.tsv.gz	未过滤的peak文件
—— matrix.mtx.gz	未过滤的细胞-peak矩阵文件
—— fragmentsInCells.tsv.gz	过滤细胞后所得的片段文件
—— SAMPLEID.narrowPeak/broadPeak	全文库peak文件

4.2 mk_chip_ref

mobivision mk_chip_ref以基因组fasta文件作为输入，构建MobiVision-chip兼容的参考基因组文件。mobivision mk_chip_ref命令输入的参数包括基因组名称（-n），基因组fasta文件（-f）和输出参考基因组文件的路径（-r）。构建得到的参考基因组文件将被保存于以基因组名称（-n）命名的文件夹中，该文件夹位于-r指定的路径下。将红色样例代码修改为实际参数后，运行：

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

若需构建两个或者多个物种合并的样本，-n\ -f参数需两次或多次指定，-n\ -f所指定的基因组名称、基因组文件需根据先后顺序一一对应，样例代码参考如下：

```
$ 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
```

所有参数说明如下：

	参数	描述
必需	-n	基因组名称由[a-zA-Z0-9_]+所组成。如果该参考基因组由多个不同的基因组所组成，请多次指定-n参数，例如 -n <species1> -n <species2>，输出的参考基因组文件夹将被命名为 <species1>_and_<species2>。
	-f	基因组fasta文件。若需输入多个fasta基因组文件，请多次指定-f参数。基因组fasta文件支持的格式包括.fasta、.fa.gz和.fna.gz。
	-r	参考基因组的输出路径。
帮助	-h, --help	显示帮助信息。

输出结果文件如下：

—— GENOME.1.bt2 (l)	参考基因组的第一级BWT文件
—— GENOME.2.bt2(l)	参考基因组的第二级BWT文件
—— GENOME.3.bt2(l)	参考基因组的第三级BWT文件
—— GENOME.4.bt2(l)	参考基因组的第四级BWT文件
—— GENOME.chrom.sizes	参考基因组染色体大小文件
—— GENOME.genome.fa	参考基因组fasta文件
—— GENOME.genome.fa.fai	参考基因组fasta索引文件
—— GENOME.reference.json	参考基因组构建信息文件
—— GENOME.rev.1.bt2(l)	参考基因组第一级BWT的反向互补序列文件
—— GENOME.rev.2.bt2(l)	参考基因组第二级BWT的反向互补序列文件

5 表观组分析模块（ATAC）

4.1 atac

mobivision atac适配于MobiCube® 高通量单细胞 ATAC-seq 试剂盒制备的文库盒测序数据；可以实现单细胞水平的基因调控位点分析。以双端测序的fastq文件作为输入，最终可输出html质

控报告、peak矩阵等相关结果文件，用于下游深入分析。

4.1.1 准备Fastq数据

输入的fastq需按以下方式命名：

[SampleName]_[ReadType].[Suffix]

其中，

[SampleName]由字母、数字或下划线组成，不建议包含R1、R2；

[ReadType]建议格式：R1、R2、R1_001、R2_001；

[Suffix]目前支持4种情况：fastq.gz、fq.gz、fastq、fq。

例如：



/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 atac使用方法

mobivision atac命令输入的参数包括index所在的文件夹路径(-i)，调用的最多线程数(-t)，fastq文件所在的路径(-f)以及输出的文件夹路径(-o)。将红色样例代码修改为实际参数后，运行：

```
$ 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
```

所有参数说明如下：

	参数	描述
必需	-f, --fastqDir	fastq文件所在路径。例 如： /share/Data/Sample_SQ23009375-230428C-S-YXH-L06
	-i, --referencePat	包含墨卓ATAC参考序列文件所在的路径。如果下载的是墨卓官方提供的tar.gz格式的index文件，需要先将tar.gz

	h	解压，然后以解压后的文件路径作为 -i 参数的路径。例如：/share/Reference/mobi_atac_ref/GRCh38
	-o, --outDir	输出文件路径。例如：/share/Outs/230428C-S-YXH-L06
	-t, --threads	软件使用的最大线程数目。例如：12
可选	--peaktype	预计得到的峰的类型，包含'broad'和'narrow'两个选项。若不指定，默认为使用'narrow'。
	--cellnumber	<最大细胞数> 强制使用这个细胞数用于细胞的筛选。若不指定，则默认使用CR2.2算法进行细胞筛选。
	--control	IgG样本fastq文件所在路径。若使用该参数，默认激活IgG模式，用于峰信号背景矫正。
帮助	-h, --help	显示帮助信息。

输出结果文件如下:

—— _flagdone	运行成功的flag文件
—— _log	运行日志文件
—— run_analysis_cmds.txt	完整命令行记录
—— SAMPLEID_out	输出结果文件根目录
—— filtered_cell_fragments_matrix	过滤后的细胞-片段矩阵文件根目录
—— barcodes.tsv.gz	过滤后的细胞文件
—— fragments.tsv.gz	过滤后的片段文件
—— matrix.mtx.gz	过滤后的细胞-片段矩阵文件
—— filtered_cell_peaks_matrix	过滤后的细胞-peak矩阵文件根目录
—— barcodes.tsv.gz	过滤后的细胞文件
—— peaks.tsv.gz	过滤后的peak文件
—— filtered.h5ad	h5ad格式的过滤后数据矩阵
—— matrix.mtx.gz	过滤后的细胞-peak矩阵文件
—— SAMPLEID.bam	比对结果文件
—— SAMPLEID.bw	比对结果可视化文件
—— SAMPLEID.filtered.bed.gz	去重并过滤后的片段文件
—— SAMPLEID_Report.html	html格式的质控报告
—— SAMPLEID_Report.json	json格式的质控报告

summary.csv	csv格式的文库信息
raw_cell_peaks_matrix	未过滤的细胞-peak矩阵文件根目录
barcodes.tsv.gz	未过滤的细胞文件
peaks.tsv.gz	未过滤的peak文件
matrix.mtx.gz	未过滤的细胞-peak矩阵文件
fragmentsInCells.tsv.gz	过滤细胞后所得的片段文件
fragmentsInPeaks.tsv.gz	与peak重合的片段文件
SAMPLEID.narrowPeak/broadPeak	全文库peak文件

(五) 参考序列信息

参考序列名称	物种	参考序列版本	下载网址
GRCh38	人	Gencode_v38*	GENCODE - Human Release 38 (gencodegenes.org)
		Ensembl_v104	Index of /pub/release-104/fasta/homo_sapiens/dna (ensembl.org)
GRCm39	鼠	Gencode_vM27*	GENCODE - Mouse Release M27 (gencodegenes.org)
		Ensembl_v104	Index of /pub/release-104/fasta/mus_musculus/dna (ensembl.org)
GRCh38_and_GRCm39	人+鼠	Gencode_v38*	GENCODE - Human Release 38 (gencodegenes.org)
		Gencode_vM27*	GENCODE - Mouse Release M27 (gencodegenes.org)
	人+鼠	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	兔子	Gencode	-
		Ensembl_v109*	Index of /pub/release-109/fasta/oryctolagus_cuniculus/dna (ensembl.org)

注：以*结尾的参考序列版本为MobiNova平台提供的参考序列版本