SCOPE-tools

发布 *0.2.3*

2020 年 11 月 07 日

目录

SCOPE-tools 是一套用于处理 SCOPE(Single Cell Omics Preparation Entity) 海量单细胞测序技术产出数据的工具。原始的 fastq 数据，经过 cellbarcode 提取与校正，序列比对和基因定量后，得到单细胞的表达矩阵，用于后续进一步的分析。

CHAPTER 1

安装

1. 要求
   - 32G 内存或者更高 (根据参考基因组大小确定)
   - Linux
   - conda 环境管理器
2. 使用 conda 安装
   - 创建环境

     ```
     conda create -n scope
     ```

   - 激活环境

     ```
     conda activate scope
     ```

   - 安装 scopetools

     ```
     conda install -c singleronbio scope-tools
     ```

   - 验证安装完成

     ```
     scope -h

     Usage: scope [OPTIONS] COMMAND1 [ARGS]... [COMMAND2 [ARGS]...]...
     ```

(下页继续)

(续上页)

```
  Single Cell Omics Preparation Entity Tools

Options:
  --version   Show the version and exit.
  -h, --help  Show this message and exit.

Commands:
  STAR           STAR short help
  barcode        extract barcode and umi short help
  cluster        cluster short help
  count          count short help
  cutadapt       cutadapt short help
  featureCounts  featureCounts short help
  run            run short help
```

3. 准备参照基因组

- Homo sapiens

```
wget ftp://ftp.ensembl.org/pub/release-99/fasta/homo_sapiens/dna/Homo_
↪sapiens.GRCh38.dna.primary_assembly.fa.gz
wget ftp://ftp.ensembl.org/pub/release-99/gtf/homo_sapiens/Homo_sapiens.
↪GRCh38.99.gtf.gz

mkdir -p references/Homo_sapiens/Ensembl/GRCh38
gzip -c -d Homo_sapiens.GRCh38.dna.primary_assembly.fa.gz > references/
↪Homo_sapiens/Ensembl/GRCh38/Homo_sapiens.GRCh38.fa
gzip -c -d Homo_sapiens.GRCh38.99.gtf.gz > references/Homo_sapiens/
↪Ensembl/GRCh38/Homo_sapiens.GRCh38.99.gtf

conda activate scope

gtfToGenePred -genePredExt -geneNameAsName2 references/Homo_sapiens/
↪Ensembl/GRCh38/Homo_sapiens.GRCh38.99.gtf /dev/stdout | \
    awk '{print $12"\t"$1"\t"$2"\t"$3"\t"$4"\t"$5"\t"$6"\t"$7"\t"$8"\t"$9
↪"\t"$10}' > references/Homo_sapiens/Ensembl/GRCh38/Homo_sapiens.GRCh38.
↪99.refFlat

STAR \
    --runMode genomeGenerate \
```

(下页继续)

(续上页)

```
    --runThreadN 6 \
    --genomeDir references/Homo_sapiens/Ensembl/GRCh38 \
    --genomeFastaFiles references/Homo_sapiens/Ensembl/GRCh38/Homo_
↪sapiens.GRCh38.fa \
    --sjdbGTFfile references/Homo_sapiens/Ensembl/GRCh38/Homo_sapiens.
↪GRCh38.99.gtf \
    --sjdbOverhang 100
```

- Mus musculus

```
wget ftp://ftp.ensembl.org/pub/release-99/fasta/mus_musculus/dna/Mus_
↪musculus.GRCm38.dna.primary_assembly.fa.gz
wget ftp://ftp.ensembl.org/pub/release-99/gtf/mus_musculus/Mus_musculus.
↪GRCm38.99.gtf.gz

mkdir -p references/Mus_musculus/Ensembl/GRCm38
gzip -c -d Mus_musculus.GRCm38.dna.primary_assembly.fa.gz > references/
↪Mus_musculus/Ensembl/GRCm38/Mus_musculus.GRCm38.fa
gzip -c -d Mus_musculus.GRCm38.99.gtf.gz > references/Mus_musculus/
↪Ensembl/GRCm38/Mus_musculus.GRCm38.99.gtf

conda activate scope

gtfToGenePred -genePredExt -geneNameAsName2 references/Mus_musculus/
↪Ensembl/GRCm38/Mus_musculus.GRCm38.99.gtf /dev/stdout | \
    awk '{print $12"\t"$1"\t"$2"\t"$3"\t"$4"\t"$5"\t"$6"\t"$7"\t"$8"\t"$9
↪"\t"$10}' > references/Mus_musculus/Ensembl/GRCm38/Mus_musculus.GRCm38.
↪99.refFlat

STAR \
    --runMode genomeGenerate \
    --runThreadN 6 \
    --genomeDir references/Mus_musculus/Ensembl/GRCm38 \
    --genomeFastaFiles references/Mus_musculus/Ensembl/GRCm38/Mus_
↪musculus.GRCm38.fa \
    --sjdbGTFfile references/Mus_musculus/Ensembl/GRCm38/Mus_musculus.
↪GRCm38.99.gtf \
    --sjdbOverhang 100
```

CHAPTER 2

使用

1. 示例数据

   示例的样本数据存储于 Open Science Framework.

   目前已经上传 SCOPEv2 样本数据，SCOPEv1 样本数据，其余类型样本 TENX(10X), dropseq, indrop, BD Rhapsody 待测试上传.

2. 快速使用

```
scope run \
    --fq1 ./rawdata/R2005073_L1_1.fq.gz \
    --fq2 ./rawdata/R2005073_L1_2.fq.gz \
    --outdir ./ \
    --bctype SCOPEv2 \
    --annot ./references/Homo_sapiens/Ensembl/GRCh38/Homo_sapiens.GRCh38.99.
↪gtf \
    --refFlat ./references/Homo_sapiens/Ensembl/GRCh38/Homo_sapiens.GRCh38.
↪99.refFlat \
    --genomeDir ./references/Homo_sapiens/Ensembl/GRCh38
    --sample samplename
```

   - **参数说明**
     - –fq1：read1 fastq 文件路径
     - –fq2：read2 fastq 文件路径

  - –outdir: 输出路径
  - –bctype: 预置的接头类型
  - –annot: 基因组注释文件, gtf 格式
  - –refFlat: refFlat 文件路径
  - –genomeDir: 参考基因组路径, 包含 STAR 所建索引
  - –sample: 样本名称
- 示例报告
- 输出文件

3. 使用说明

SCOPE-tools 包含 7 个子命令, 分别是 sample, barcode, cutadapt, STAR, featureCounts, count, cluster 和 run.

```
scope -h

Usage: scope [OPTIONS] COMMAND1 [ARGS]... [COMMAND2 [ARGS]...]...

  Single Cell Omics Preparation Entity Tools

Options:
  --version   Show the version and exit.
  -h, --help  Show this message and exit.

Commands:
  STAR           STAR short help
  barcode        extract barcode and umi short help
  cluster        cluster short help
  count          count short help
  cutadapt       cutadapt short help
  featureCounts  featureCounts short help
  run            run short help
  sample         sample short help
```

- sample

  设置报告中的样本信息.

  - **示例**

```
scope sample \
    --transcriptome Homo_sapiens \
    --sample samplename \
    --outdir ./
```

- **参数说明**
  - –outdir: 输出路径
  - –version: 软件版本
  - –description: 功能描述信息
  - –sample: 样本名称
  - –transcriptome: 转录组名称

- barcode

  基于 read1 序列信息 (barcode 序列, linker 序列, 质量值和 polyT 长度) 过滤, 提取并矫正 barcode, 将矫正后的 barcode 和原始的 UMI 序列添加到 read2 的 ID 中.

  过滤规则:

  1. 过滤 polyT 碱基数目 <10 的 reads
  2. 过滤 barcode 和 UMI 碱基质量值低于 14 的个数 >2
  3. 过滤两段接头中任何一段中错配碱基数 >1
  4. 过滤三段 barcode 中错配碱基数和 >1

  barcode 矫正规则:

  将未出现在 whitelist 中的 barcode, 矫正为与 whitelist 中 汉明距离 为 1 的 barcode.

  - **示例**

```
scope barcode \
    --fq1 ./rawdata/R2005073_L1_1.fq.gz \
    --fq2 ./rawdata/R2005073_L1_2.fq.gz \
    --sample samplename \
    --outdir ./ \
    --bctype SCOPEv2
```

  - **参数说明**
    - –fq1: read1 fastq 文件路径
    - –fq2: read2 fastq 文件路径

* –sample: 样本名称
* –outdir: 输出路径
* –bctype: 预置的接头类型
* –pattern: 自定义的接头结构, 字母 C, L, U, T 分别表示 cell barcode、linker、UMI、T 碱基, 数字表示碱基长度. C8L16C8L16C8L1U8T18 即表示以下结构:

  CCCCCCCCLLLLLLLLLLLLLLLLCCCCCCCCLLLLLLLLLLLLLLLLCCCCCCCCLUUUU

  从第一个碱基开始，为 C 位置碱基构成 cell barcode, 为 U 位置的碱基构成 UMI, 第一段为 L 的碱基为 linker1, 第二段为 L 的碱基为 linker2, 末尾 T 为 polyT.
* –whitelist 自定义的 cell barcode 白名单文件
* –linker: 自定义的 linker 白名单文件
* –lowQual: 定义为低质量碱基的质量值, 默认: 14
* –lowNum: 允许的低质量碱基的个数, 默认: 2

- cutadapt

  调用 cutadapt 对 read2 进行质控.

  – trim 接头序列

  – trim 两端低质量碱基

  – **示例**

```
scope cutadapt \
    --fq ./samplename/01.barcode/samplename_2.fq.gz \
    --sample samplename \
    --outdir ./ \
    --adapter p5=AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC␣
↪polyT=A{18} \
    --overlap 5
```

  – **参数说明**

    * –fq: barcode 处理后的 read2 fastq 文件
    * –sample: 样本名称
    * –outdir: 输出路径
    * –adapter: 接头序列, 可多次使用, 默认: p5=AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC polyT=A{18}

    * –overlap: 认为检测到接头时重叠碱基数, 默认: 5
    * –minimum-length: 允许的最短序列长度, 默认: 20
    * –nextseq-trim: trim 使用的质量值 (忽略 G, 针对双色试剂, 如 NextSeq), 默认: 20
    * –thread: 线程数, 默认: 2

- STAR

  调用 STAR 将 read2 序列定位到基因组上.

  – **示例**

```
scope STAR \
    --fq ./samplename/02.cutadapt/samplename_2.fq.gz \
    --sample samplename \
    --outdir ./ \
    --refFlat ./references/Homo_sapiens/Ensembl/GRCh38/
↪Homo_sapiens.GRCh38.99.refFlat \
    --genomeDir ./references/Homo_sapiens/Ensembl/GRCh38
```

  – **参数说明**

    * –fq: cutadapt 处理后的 read2 fastq 文件
    * –sample: 样本名称
    * –readFilesCommand: STAR 读取输入文件的命令, 默认: zcat
    * –genomeDir: 参考基因组路径, 包含 STAR 所建索引
    * –runThreadN: 线程数, 默认: 2
    * –outdir: 输出路径
    * –refFlat: refFlat 文件路径

- featureCounts

  调用 featureCounts 将定位到基因组上的 reads, 进一步定位到基因上.

  – **示例**

```
scope featureCounts \
    --bam ./samplename/03.STAR/samplename_Aligned.
↪sortedByCoord.out.bam \
    --annot ./references/Homo_sapiens/Ensembl/GRCh38/
↪Homo_sapiens.GRCh38.99.gtf \
    --sample samplename \
    --outdir ./
```

- **参数说明**
  - –bam: STAR 比对并排序后的 bam 文件
  - –sample: 样本名称
  - –annot: 基因组注释文件, gtf 格式
  - –nthreads: 线程数, 默认: 2
  - –format: 输入文件的格式, 默认: BAM
  - –outdir: 输出路径

- count

  对同一 barcode 内比对到同一基因的 UMI 序列进行校正, 之后进行 UMI 计数; 细胞数目评估 (cell-calling); 输出单细胞基因表达矩阵.

  UMI 矫正规则:

  1. 对同一 barcode 中，比对到同一 gene_id 下的 umi 间进行校正
  2. 若不存在 mismatch 为 1 的 UMI，则取 **原始 UMI** 为最终 UMI;
  3. 若存在 mismatch 为 1 的 UMI，取 **readcount 数最大**的为最终 UMI;
  4. 若 readcount 数相同，则取 **序列字符排序最大**的为最终 UMI。

  细胞数目评估规则:

  1. 以 UMI count 降序对 barcode 排序
  2. 取第预设细胞数目 *0.01 为基准细胞
  3. 取基准细胞的 UMI count*0.1 为阈值, 大于阈值则判定为细胞

  - **示例**

```
scope count \
    --bam ./samplename/04.featureCounts/samplename_name_
↪sorted.bam \
    --cells 3000 \
    --sample samplename \
    --outdir ./
```

  - **参数说明**
    - –bam: featureCounts 输出的 bam 文件
    - –sample: 样本名称
    - –cells: 预估细胞数, 默认: 3000
    - –outdir: 输出路径

- cluster

  调用 scanpy 对表达矩阵进行分析, 得到细胞 QC 和初步的聚类图

  - **示例**

```
scope cluster \
    --matrix ./samplename/05.count/samplename_matrix.mtx␣
↪\
    --barcodes ./samplename/05.count/samplename_barcodes.
↪tsv
    --genes ./samplename/05.count/samplename_genes.tsv \
    --outdir ./
    --sample samplename
```

  - **参数说明**

    * –outdir: 输出路径
    * –sample: 样本名称
    * –matrix: 单细胞基因表达稀疏矩阵路径
    * –barcodes: 表达稀疏矩阵路径
    * –genes:

- run

  快速使用 SCOPE-tools 运行流程, 对数据进行分析.

  - **示例**

```
scope run \
    --fq1 ./rawdata/R2005073_L1_1.fq.gz \
    --fq2 ./rawdata/R2005073_L1_2.fq.gz \
    --outdir ./ \
    --bctype SCOPEv2 \
    --annot ./references/Homo_sapiens/Ensembl/GRCh38/
↪Homo_sapiens.GRCh38.99.gtf \
    --refFlat ./references/Homo_sapiens/Ensembl/GRCh38/
↪Homo_sapiens.GRCh38.99.refFlat \
    --genomeDir ./references/Homo_sapiens/Ensembl/GRCh38
    --sample samplename
```

  - **参数说明**

    * –fq1: read1 fastq 文件路径

* –fq2：read2 fastq 文件路径
* –outdir: 输出路径
* –bctype: 预置的接头类型
* –annot: 基因组注释文件, gtf 格式
* –refFlat: refFlat 文件路径
* –genomeDir: 参考基因组路径, 包含 STAR 所建索引
* –sample: 样本名称