

cellranger-单细胞RNA测序数据

背景介绍

cellranger是一款整合了多种分析流程的单细胞分析软件，可以用来处理scRNA-seq、scATAC-seq以及单细胞免疫组库等多种类型的单细胞数据。在这里，仅对cellranger用scRNA-seq数据作简要介绍。

cellranger处理scRNA-seq数据的分析流程可以划为四步：拆分数据（mkfastq）、细胞定量（count）、定量整合（aggr）、数据下游分析（reanalysis）。但是目前，单细胞数据下游分析一般交给Seurat或Scanpy，因此cellranger的主要用途是输入测序数据生成feature-barcode表达谱。

软件下载和安装

安装cellranger

```
 wget -O cellranger-5.0.1.tar.gz "https://cf.10xgenomics.com/releases/cell-exp/cellranger-5.0.1.tar.gz?Expires=1610895624&Policy=eyJTdGF0ZW1lbnQiOlt7IlJlc291cmNljoiaHR0cHM6Ly9jZi4xMWhnZW5vbWljcy5jb20vcmVsZWFzZXMuY2VsbC1leHAvY2VsbHJhbmdlci01LjAuMS50YXluZ3oiLCJDb25kaXRpb24iOnsiRGF0ZUxlC3NUaGFuljp7IkFXUzpFcG9jaFRpbWUiOjE2MTA4OTU2MjR9fX1dfQ__&Signature=KPtDuB-k3KtxmHvDHT816TBtXRi~N0lyz2Yq9yKDtlcHJrSXMZ9nb9WSfyXuZbLRrXBJaPDJIN5veag9OFGMTHIFqTqj1JKawYRy2Js9dTc1znXFrxpj7JSDarChejC4IEm1be1kEQAn5egRHUTGSSxCQe1BNyEPDU30S972Fny3gUMSbtCpRycs7bVq0LLFpW6YduhhF4-6xhpC76Wnv7sPmkB4m812~siN5juAPW4brkbbEErUy5A5rpSxwJ7yLA8vhoUshMtRfBZ6gdTGsgHrYAKk4RrLGggNdc2dKuwOljtCtpgk2IPiWIxKao3zDhUJEvNUl2OJAMYHxb8Q__&Key-Pair-Id=APKAI7S6A5RYOXBWRPDA"
```

```
 tar -zxvf cellranger-5.0.1.tar.gz
```

```
 vi ~/.bashrc
```

```
 export PATH=$PATH:/home/wanglinxiao/cellranger-5.0.1/bin
```

下载参考序列信息

人： wget <https://cf.10xgenomics.com/supp/cell-exp/refdata-gex-GRCh38-2020-A.tar.gz>

小鼠： wget <https://cf.10xgenomics.com/supp/cell-exp/refdata-gex-mm10-2020-A.tar.gz>

下载的参考序列信息包括基因组序列信息（.fa）、基因组注释信息（.gff3）以及对应的索引信息。

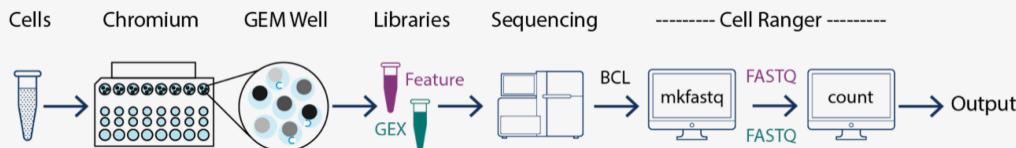
使用说明及参数介绍

一、测序流程

在介绍cellranger之前，我们首先应该对单细胞RNA测序流程有一定掌握。只有掌握了测序流程，当我们拿到数据时，才能更好的进行细胞定量以及下游的分析。

(1) 单样本, 单个测序文库, 单个测序管道 (lane)

One Sample, One GEM Well, One Flowcell

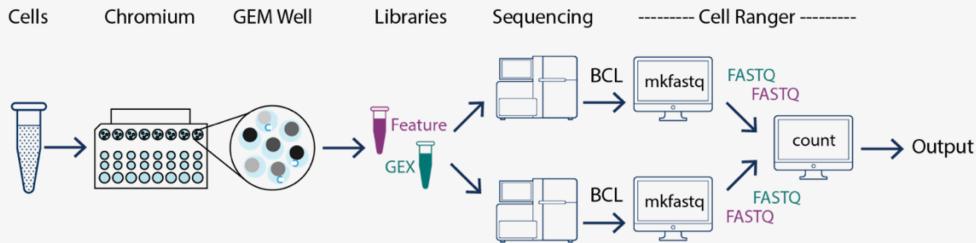


In this example you have one **sample** that is processed through one **GEM well** (a set of partitioned cells from a single 10x Chromium™ Chip channel) and sequenced on one **flowcell**. In this case you would generate FASTQs using `cellranger mkfastq`, and run `cellranger count` as described in [Single-Sample Analysis](#).

一个样本建立了一个测序文库，在将该文库放到一条lane上进行测序。细胞定量时指定样本名称即可

(2) 单样本, 单个测序文库, 多个测序管道 (lane)

One Sample, One GEM well, Multiple Flowcells

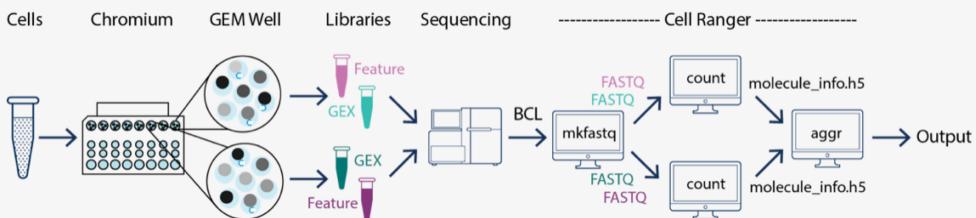


In this example you have one sample that is processed through one GEM well then you generate one library which is sequenced across multiple flowcells. This may be done to increase sequencing depth, for example. In this case all of the reads can be combined in a single instance of the `cellranger count` pipeline. This process is described in [Specifying Input Fastqs](#).

有时由于一个lane不能放得下一个样本，因此经常会把一个样本放到不同lane上进行测序。细胞定量时指定样本名即可，不需要加lane参数。

(3) 单样本, 多个测序文库, 单个测序管道 (lane)

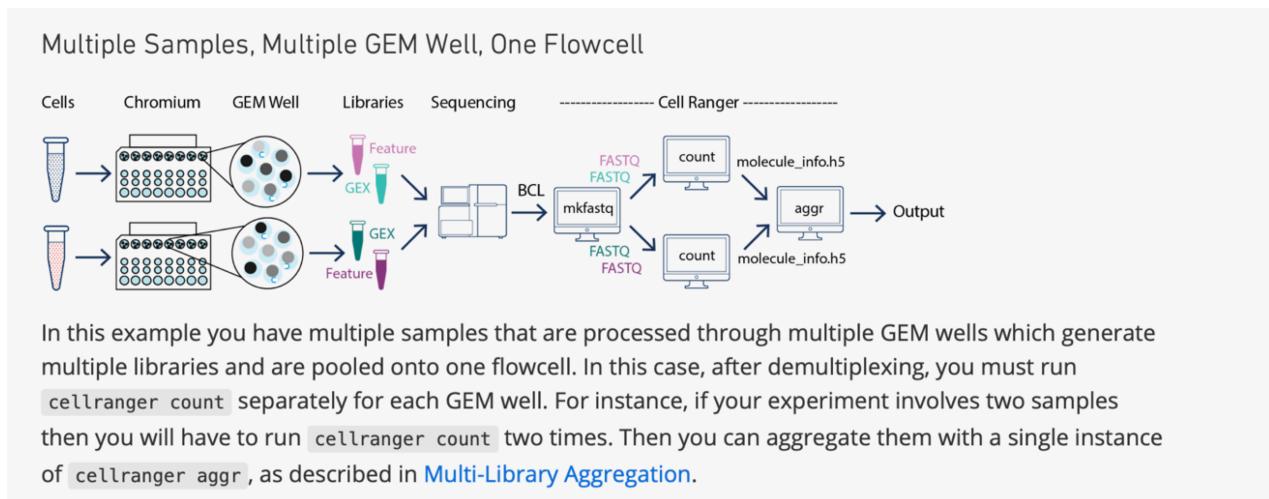
One Sample, Multiple GEM Wells, One Flowcell



In this example you have one sample that is processed through multiple GEM wells. This is often done when conducting technical replicate experiments. The libraries from the GEM wells are then pooled onto one flowcell and sequenced. In this case you demultiplex the data from the sequencing run and then run the libraries from each GEM well through a separate instance of `cellranger count`. Once those are completed, you can perform a combined analysis using `cellranger aggr`, as described in [Multi-Library Aggregation](#). (See figure above.)

有时，为了技术重复，会对一个样本构建多个测序文库。细胞定量时样本名指定各自文库的样本名，定量后可在进行合并。

(4) 多样本，多文库，单个测序管道 (lane)



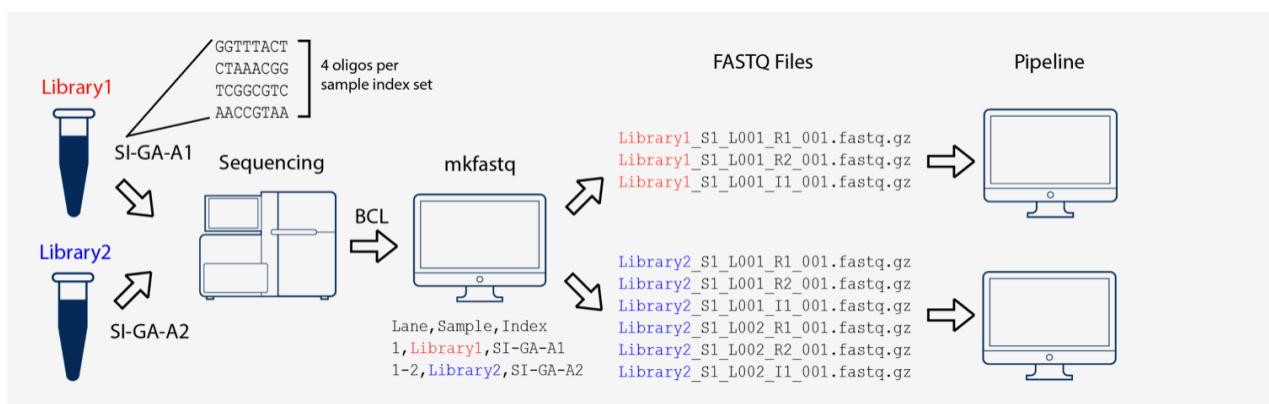
多个样本构建的测序文库放置到一条lane上进行测序。细胞定量时样本名指定各自文库的样本名，定量后可在进行合并。

二、数据拆分

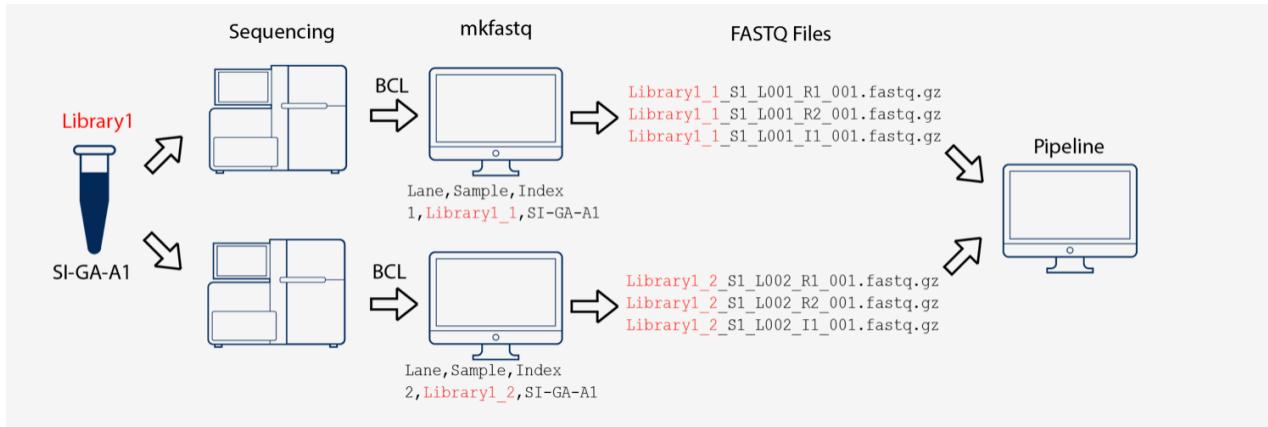
目的

在测序时，flowcell中的每个lane可能会同时加入多个样本进行测序。因此在测序完成后，应根据每个样本的index对数据进行拆分。

示意流程



将两个样本混合在一个flowcell中进行测序，根据样本的各自index将样本分开



将一个样本放置于两个flowcell中进行测序，得到样本在不同lane（L001,L002）的测序数据

重要参数

| 参数命令 | 说明 |
|---------------|----------------------------|
| --run | 原始BCL文件所在目录 |
| --id | 输出文件目录的名称 |
| --samplesheet | 包含测序lane, 样本名称, 样本index等信息 |

下图是一个samplesheet的简单示例

```
Lane,Sample,Index
1,test_sample,SI-TT-D9
```

三、细胞定量

目的

细胞定量，cellranger将质控、比对、定量整合到一个命令中，使操作更加方便。

重要参数

| 参数命令 | 说明 |
|-----------------|-------------------------------|
| --id | 输出文件目录的名称 |
| --fastqs | 输入文件fastq所在的目录名 |
| --sample | 样本名，第一个'_'之前的字符 |
| --transcriptome | 参考文件所在目录名 |
| --lanes | 如果只需要分析特定lane的测序数据，可用lane进行指定 |

示例

```
cellranger count --id=sample345 \
    --transcriptome=/opt/refdata-cellranger-GRCh38-1.2.0 \
    --fastqs=/home/scRNA/runs/HAWT7ADXX/outs/fastq_path \
    --sample=mysample \
```

重要的输出文件

| 输出文件名 | 说明 |
|---------------------------|--|
| web_summary.html | 以网页版形式对样本信息的总结，包括测序质量，基因数量以及细胞数目等 |
| metrics_summary.csv | 同样是对样本信息的总结，csv格式 |
| filtered_gene_bc_matrices | 是一个目录，存储细胞定量的结果，有三个文件：barcodes.tsv.gz, features.tsv.gz, matrix.mtx.gz。这三个文件是下游分析软件Seurat或Scanpy的输入文件 |
| raw_feature_bc_matrix | 也是存储细胞定量结果的目录，但是这里的细胞没有经过过滤，因此一般不用。 |
| molecule_info.h5 | 下一步定量整合时需要用到的文件 |

四、定量整合

目的

当处理多个生物学样本或一个样本存在多个文库时，应分别对每个文库单独count定量，然后将定量结果aggr起来。

重要参数

| 参数命令 | 说明 |
|----------|--|
| --id | 输出文件目录的名称 |
| --csv | 需要整合的样本信息，下附有该csv文件的格式 |
| --mapped | 将不同样本的定量信息进行整合时，需进行标准化，使得基因在不同细胞间表达量可比 |

```
library_id,molecule_h5
LV123,/opt/runs/LV123/outs/molecule_info.h5
LB456,/opt/runs/LB456/outs/molecule_info.h5
LP789,/opt/runs/LP789/outs/molecule_info.h5
```

示例

```
cellranger aggr --id=AGG123 \
--csv=AGG123_libraries.csv \
--normalize=mapped
```

重要的输出文件

| 输出文件名 | 说明 |
|----------------------------|--|
| web_summary.html | 对整合数据的一个网页版形式的总结 |
| aggregation.csv | 说明数据是哪些样本整合而来 |
| filtered_feature_bc_matrix | 是一个目录，存储细胞定量的结果，有三个文件： barcodes.tsv.gz, features.tsv.gz, matrix.mtx.gz。这三个文件 是下游分析软件Seurat或Scanpy的输入文件 |
| raw_feature_bc_matrix | 也是存储细胞定量结果的目录，但是这里的细胞没有经过过滤，因 此一般不用。 |