

Barcode-Converter 工具使用指南

概述

Barcode-Converter 是一个专为单细胞测序数据设计的条形码转换工具，主要用于将不同平台的 CellBarcode 转换为标准格式，确保数据兼容性和分析流程的顺利进行。

工作原理

[!NOTE]

转换原理：工具首先识别输入 FASTQ 文件中 R1 的 CellBarcode，允许一个碱基的错误匹配，然后通过白名单对应关系完成 CellBarcode 的转换。

快速开始

基础转换示例

示例 1：DD 数据转换为 10X 3' 文库

```
/path/to/conv.0.1.2 \  
--fq1 ./demo_dd_S39_L001_R1_001.fastq.gz \  
--fq2 ./demo_dd_S39_L001_R2_001.fastq.gz \  
--wl1 ./P3CB.barcode.txt.gz \  
--wl2 3M-february-2018.txt.gz \  
--rs 17C+T \  
-t 12 \  
-o output/
```

参数说明：

- `--wl1`：指定输入数据的白名单文件
- `--wl2`：指定输出数据的白名单文件
- `--rs`：指定转换后的 CellBarcode 起始位点
- `-t`：指定线程数
- `-o`：指定输出目录

多文件批量转换

示例 2：同一样本多组文件转换

```
/path/to/conv.0.1.2 \  
--fq1 ./demo_dd_S39_L001_R1_001.fastq.gz ./demo_dd_S39_L001_R1_002.fastq.gz \  
--fq2 ./demo_dd_S39_L001_R2_001.fastq.gz ./demo_dd_S39_L001_R2_002.fastq.gz \  
--wl1 ./P3CB.barcode.txt.gz \  
--wl2 3M-february-2018.txt.gz \  
--rs 17C+T \  
-t 12 \  
-o output/
```

[!TIP]

多文件转换时，R1 和 R2 文件顺序必须保持一致，以空格分隔多个文件路径。

多组学数据转换

示例 3：使用已有 CellBarcode 对应关系

免疫组数据

Step 1: Convert 5' RNA library

```
conv.0.1.2 --fq1 rna_R1.fastq.gz --fq2 rna_R2.fastq.gz --wl1 P3CB.barcode.txt.gz --wl2  
737K-august-2016.txt.gz --rs 17C+T -t 12 -o rna_output/
```

Step 2: Convert TCR library

```
conv.0.1.2 --fq1 tcr_R1.fastq.gz --fq2 tcr_R2.fastq.gz --map rna_output/map.txt --rs  
17C+T -t 12 -o tcr_output/
```

Step 3: Convert BCR library

```
conv.0.1.2 --fq1 bcr_R1.fastq.gz --fq2 bcr_R2.fastq.gz --map rna_output/map.txt --rs  
17C+T -t 12 -o bcr_output/
```

[!IMPORTANT]

多组学数据转换时，建议先转换转录组 RNA 文库数据，将其输出的 `map.txt` 文件作为其他类型数据的输入，确保 barcode 对应关系的一致性。

参数详解

参数	类型	描述	默认值
<code>--fq1 <FQ1>...</code>	必需	输入的 R1 FASTQ 文件，支持多个文件（空格分隔）	-
<code>--fq2 <FQ2>...</code>	必需	输入的 R2 FASTQ 文件，支持多个文件（空格分隔）	-
<code>--rs <RS></code>	可选	读取 R1 的结构，格式：数字/+和字母组成 - 数字：碱基数 - +：剩余碱基 - C：CellBarcode 碱基 - T：其他碱基	<code>17C+T</code>
<code>--w11 <WL1></code>	条件必需	输入 FASTQ 的试剂类型对应的白名单文件 DD 系列产品选 择： <code>barcode/P3CB.barcode.txt</code>	-
<code>--w12 <WL2></code>	条件必需	输出 FASTQ 的试剂类型对应的白名单文件 - 3' 文库： <code>3M-february-2018.txt.gz</code> - 5' 文库： <code>737K-august-2016.txt.gz</code>	-
<code>--map <MAP></code>	条件必需	条形码映射文件（TSV 格式） 包含两列：输入白名单	输出白名单
<code>--no-multi</code>	可选	对多个匹配的 barcode 进行重新分配	默认执行
<code>-t, --threads <THREADS></code>	可选	线程数	<code>10</code>
<code>-o, --out <OUT></code>	可选	输出目录	<code>./</code>
<code>-h, --help</code>	可选	打印帮助信息	-
<code>-V, --version</code>	可选	打印版本信息	-

[!WARNING]
必须至少指定 `--w11` 和 `--w12` 或 `--map` 中的一个参数组合。

输出文件说明

转换完成后，输出目录将包含以下文件：

主要输出文件

- `<OUT>/*.fastq.gz`
 - 转换后的 FASTQ 文件
 - 可直接用于下游分析
- `<OUT>/multi_*.fastq.gz`

- 包含多个匹配条形码的序列中间文件
- 可能的条形码以 "_" 连接

3. `<OUT>/map.txt`

- 条形码映射文件 (TSV 格式)
- 第一列: 输入白名单
- 第二列: 输出白名单

重要注意事项

白名单选择

[!IMPORTANT]

不同产品使用不同的白名单文件, `--w11` 和 `--w12` 参数必须设置正确。

10X Genomics 白名单位置:

- 定义文件: `cellranger-*/lib/python/cellranger/chemistry_defs.json`
- 或: `cellranger-5.0.1/lib/python/cellranger/chemistry.py`
- 白名单目录: `cellranger-*/lib/python/cellranger/barcodes/`

[!WARNING]

Cell Ranger V9.0 以上版本的 3' 文库白名单需要使用最新的 `3M-february-2018_TRU.txt.gz`, 否则会出现识别错误。

条形码分配策略

自动分配逻辑:

1. 当 `--w11` 中 CellBarcode 数目 大于 `--w12` 中数目时:
 - 取 10M Reads 统计 CellBarcode
 - 如果输入数据 CellBarcode 数目 > 白名单数目: 取高频 CellBarcode 对应
 - 如果输入数据 CellBarcode 数目 ≤ 白名单数目: 取出现过的 CellBarcode 对应, 余下随机分配
2. 启用 `--no-multi` 时:
 - 统计完 CellBarcode 的 reads 数目后重新分配
 - 按 reads 数目从高到低排序
 - 分配给 reads 数最多的 CellBarcode
 - 如果第一和第二 CellBarcode reads 数目相同, 则放弃分配

版本更新说明

[!NOTE]

conv.0.1.2 版本改进:

- 修复工作线程较小时内存占用过高问题

- 将 `read_ahead` 的 `chunk_size` 和 `chunk_queue_size` 从默认 100 调整为工作线程数的平方

支持范围

[!CAUTION]

重要限制：

- 本工具仅支持 DD CellBarcode 白名单数据的转换
- 不支持 MM 数据，MM 数据仍需使用原始 Python 版本

最佳实践建议

多组学数据转换流程

1. 第一步：转换转录组数据

```
# 转换 RNA 文库
conv.0.1.2 --fq1 rna_R1.fastq.gz --fq2 rna_R2.fastq.gz \
  --w11 P3CB.barcode.txt.gz --w12 737K-august-2016.txt.gz 【5文库】 \
  --rs 17C+T -t 12 -o rna_output/
```

2. 第二步：使用映射文件转换其他类型数据

```
# 转换 TCR 数据
conv.0.1.2 --fq1 tcr_R1.fastq.gz --fq2 tcr_R2.fastq.gz \
  --map rna_output/map.txt 【上步转化的barcode 对应关系】 --rs 17C+T -t 12 -o
tcr_output/
```