

Barcode-Converter Tool User Guide

Overview

Barcode-Converter is a specialized tool designed for single-cell sequencing data that converts CellBarcodes from different platforms to standard formats, ensuring data compatibility and smooth analysis workflows.

How It Works

[!NOTE]

Conversion Principle: The tool first identifies CellBarcodes in the R1 reads of input FASTQ files, allowing for one base mismatch error, then completes CellBarcode conversion through whitelist correspondence relationships.

Quick Start

Basic Conversion Example

Example 1: Converting DD Data to 10X 3' Library

```
/path/to/conv.0.1.2 \
--fq1 ./demo_dd_S39_L001_R1_001.fastq.gz \
--fq2 ./demo_dd_S39_L001_R2_001.fastq.gz \
--w11 ./P3CB.barcode.txt.gz \
--w12 3M-february-2018.txt.gz \
--rs 17C+T \
-t 12 \
-o output/
```

Parameter Description:

- `--w11`: Specifies the whitelist file for input data
- `--w12`: Specifies the whitelist file for output data
- `--rs`: Specifies the starting position of converted CellBarcode
- `-t`: Specifies the number of threads
- `-o`: Specifies the output directory

Batch Conversion for Multiple Files

Example 2: Converting Multiple Files from the Same Sample

```
/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]

When converting multiple files, the order of R1 and R2 files must be consistent, with multiple file paths separated by spaces.

Multi-omics Data Conversion

Example 3: Using Existing CellBarcode Correspondence

```
# 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]

For multi-omics data conversion, it is recommended to first convert transcriptome RNA library data and use its output `map.txt` file as input for other data types to ensure consistency of barcode correspondence.

Parameter Reference

Parameter	Type	Description	Default
<code>--fq1 <FQ1>...</code>	Required	Input R1 FASTQ files, supports multiple files (space-separated)	-
<code>--fq2 <FQ2>...</code>	Required	Input R2 FASTQ files, supports multiple files (space-separated)	-
<code>--rs <RS></code>	Optional	R1 read structure format: numbers/+ and letters - Numbers: base count - +: remaining bases - C: CellBarcode bases - T: other bases	<code>17C+T</code>
<code>--wl1 <WL1></code>	Conditionally Required	Whitelist file for input FASTQ reagent type For DD series products: <code>barcode/P3CB.barcode.txt</code>	-
<code>--wl2 <WL2></code>	Conditionally Required	Whitelist file for output FASTQ reagent type - 3' library: <code>3M-february-2018.txt.gz</code> - 5' library: <code>737K-august-2016.txt.gz</code>	-
<code>--map <MAP></code>	Conditionally Required	Barcode mapping file (TSV format) Contains two columns: input whitelist	output whitelist
<code>--no-multi</code>	Optional	Reallocate multiple matching barcodes	Default enabled
<code>-t, --threads <THREADS></code>	Optional	Number of threads	<code>10</code>
<code>-o, --out <OUT></code>	Optional	Output directory	<code>./</code>
<code>-h, --help</code>	Optional	Print help information	-
<code>-V, --version</code>	Optional	Print version information	-

[!WARNING]

At least one of the parameter combinations `--wl1` and `--wl2` or `--map` must be specified.

Output Files

After conversion, the output directory will contain the following files:

Main Output Files

- `<OUT>/*.fastq.gz`
 - Converted FASTQ files
 - Ready for downstream analysis
- `<OUT>/multi_*.fastq.gz`
 - Intermediate files containing sequences with multiple matching barcodes
 - Possible barcodes are connected with "_"
- `<OUT>/map.txt`
 - Barcode mapping file (TSV format)

- Column 1: Input whitelist
- Column 2: Output whitelist

Important Notes

Whitelist Selection

[!IMPORTANT]

Different products use different whitelist files. The `--w11` and `--w12` parameters must be set correctly.

10X Genomics Whitelist Locations:

- Definition file: `cellranger-*/lib/python/cellranger/chemistry_defs.json`
- Or: `cellranger-5.0.1/lib/python/cellranger/chemistry.py`
- Whitelist directory: `cellranger-*/lib/python/cellranger/barcodes/`

[!WARNING]

For Cell Ranger V9.0 and above, 3' library whitelists must use the latest `3M-february-2018_TRU.txt.gz`, otherwise recognition errors will occur.

Barcode Allocation Strategy

Automatic Allocation Logic:

1. When CellBarcode count in `--w11` greater than count in `--w12`:
 - Use 10M Reads to count CellBarcodes
 - If input data CellBarcode count > whitelist count: use high-frequency CellBarcodes for correspondence
 - If input data CellBarcode count \leq whitelist count: use existing CellBarcodes for correspondence, randomly assign the rest
2. When `--no-multi` is enabled:
 - Reallocate after counting CellBarcode reads
 - Sort by read count from high to low
 - Assign to CellBarcode with highest read count
 - If first and second CellBarcode read counts are equal, skip allocation

Version Update Notes

[!NOTE]

conv.0.1.2 Version Improvements:

- Fixed high memory usage issue when using fewer worker threads
- Adjusted `read_ahead` `chunk_size` and `chunk_queue_size` from default 100 to square of worker thread count

Support Scope

[!CAUTION]

Important Limitations:

- This tool only supports DD CellBarcode data conversion
- **Does not support MM data** - MM data still requires the original Python version

Best Practices

Multi-omics Data Conversion Workflow

1. Step 1: Convert Transcriptome Data

```
# Convert RNA library
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 \
  --rs 17C+T -t 12 -o rna_output/
```

2. Step 2: Use Mapping File for Other Data Types

```
# Convert TCR data
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/
```