# splitBarcode Manual

Version: v0.1.6

Security Level: External Use

Author: Wansi Mao, Qin Sheng

Edit: Alexander Butarbutar

Date: 29/10/2019

# 1 Software Download

https://github.com/MGI-tech-bioinformatics/splitBarcode

# 2 Software Introduction

Software Name : splitBarcode

Release : v0.1.6 (Oct 2019)

Function : splitBarcode performs demultiplexing of barcoded fastq on MGI data. This script can handle single or double barcode in Single End (SE) or Paired End (PE) format.

Supported OS: Windows 10 or Linux (centos 7.x). There is no need to install or compile the software, simply copy the corresponding version of the software.

# 3 Parameter Introduction

***Required Parameters:***

- `index.txt`： index list
- `sample_1.fq.gz`： input fastq file

By default, the script will run in SE mode if only these two parameters are provided (using the last 10 bp as the barcode and a barcode mismatch tolerance is 1). Please be sure to use the correct barcode sequence as defined in the `index.txt` file.

- `-b <start_postion length allowed_mismatch>` Use space to separate the 3 parameters. For single barcode, provide a single `-b` parameter. For double barcode, provide two `-b` parameters. By default it will split the barcode based on the last 10 bp, with a maximum allow mismatch of 1 bp.
- `-n <Maximum threads number>` The maximum number of threads to process compression at the same time. The value must be an integer greater than 0, such as 20.
- `-m <Maximum memory>` The maximum memory limitation. Unit is GB. The value must be an integer or floating point number greater than 1, such as 100

Note: Software will allocate memory based on the `-n` and `-m` parameter setting. Recommend memory and threads number is: **`-n 30 -m 200.`**

***Optional Parameters:***

- `-2` Fastq file for reverse read (for PE sequencing).
- `-o <output directory>` The output directory for the split result. Default is the directory where the input fastq file is located.
- `-r` Enable splitting of barcode based on the reverse complement of the barcode sequence. This function is disabled by default.

## 4 *Index.txt* File Format

The `index.txt` file supports the following format

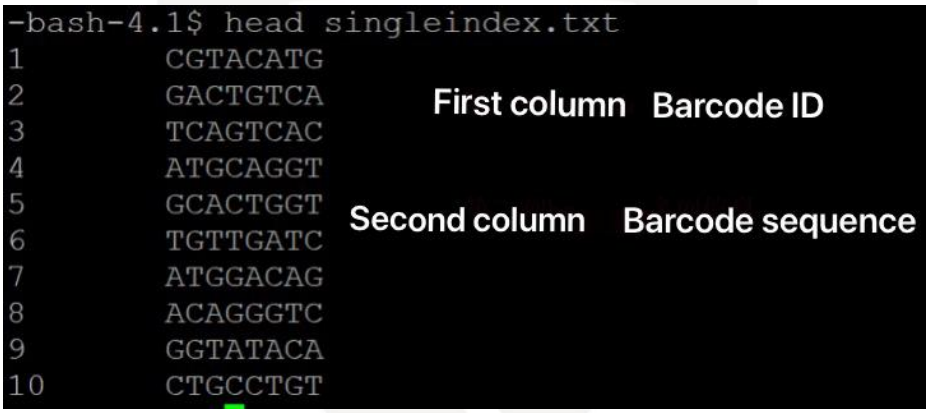
- First column: barcode ID.
- Second column: barcode sequence.

***Note:***

1) Please use numbers in the first column, do not use strings or special symbols, which software cannot recognize.
2) The second column should only contain barcode used in this specific lane, do not mix barcodes from other lanes, the barcodes should be unique, otherwise the data cannot split successfully.

### 4.1 Single Barcode Split

① When using MGI barcode, the barcode sequence is provided with the information on library preparation. Please use `-r` option, to enable the reverse complement.




② If the barcode sequences were provided by third-party, and the barcode sequence is normally in reverse complement format. As such, the `-r` parameter is typically not needed. If the barcode length is 6 bp, 0 is recommended for the `allowed_mismatch` parameter.

```
-bash-4.1$ head singleindex.txt
1       CGTACATG
2       GACTGTCA
3       TCAGTCAC
4       ATGCAGGT
5       GCACTGGT
6       TGTTGATC
```

If the script is unsuccessful in demultiplexing the fastq file, open the `SequenceStat.txt` file and compare the barcode sequence with high undecoded ratio (Percentage (%) column) to with the barcode information from the Library Preparation. If the barcode sequence does not match, you may need to orient the barcode in its reverse complement order and / or use the -r parameter

Barcode information from `SequenceStat.txt`

| #Sequence | Barcode | Count | Percentage(%) |
|---|---|---|---|
| TACTGT | undecoded | 42103587 | 5.489020 |
| CACTGT | undecoded | 42036475 | 5.466622 |
| GACTGT | undecoded | 41625939 | 5.381600 |
| AACTGT | undecoded | 41563383 | 5.369847 |

Barcode information from Library Preparation

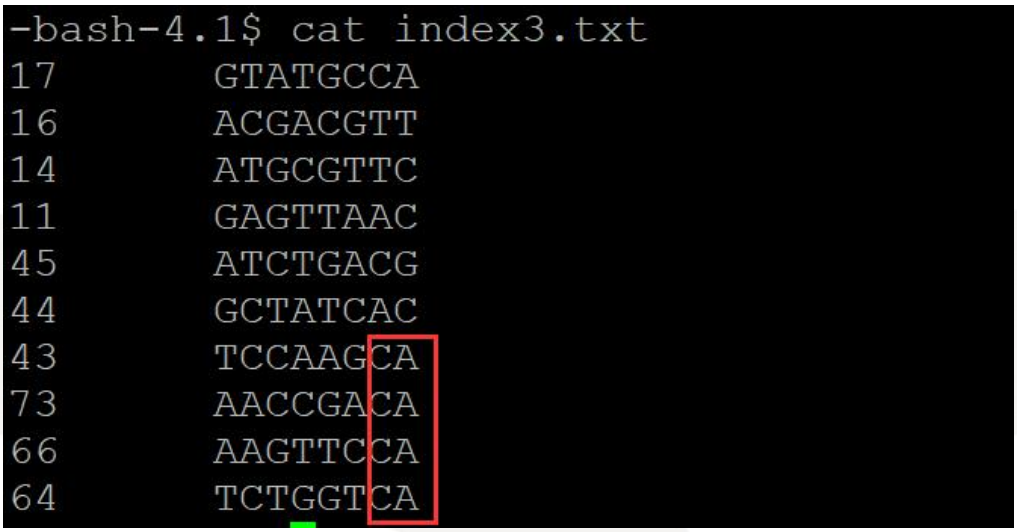
| SampleID | Barcode_seq | Data Size(M) | Note |
|---|---|---|---|
| 1 | ACAGTC | 6666.667 | |
| 2 | ATTGGC | 6666.667 | |
| 3 | CAAGGA | 6666.667 | |

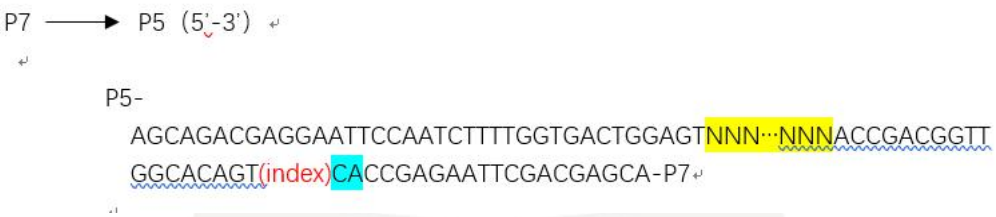
③ The splitBarcode software currently does not handle an `index.txt` file with 2 different barcode length. If two different barcode lengths are used for mixed library preparation. 1) Create two different `index.txt` file, each file should contain the same barcode length. The `splitBarcode` will need to be executed twice, once per each `index.txt` file

```
-bash-4.1$ cat index1.txt
29      AGCGTTGA
28      CCATTACT
27      AACGGCTG
26      TGGACTGT
23      AGATGCAC
22      GTACGACT
21      CGATAGAT
20      ACCGGTTC
19      AAGTGCTC
18      GCAGCTAT
```

```
-bash-4.1$ cat index2.txt
11      CCGAAG
10      AATTCG
9       GTGACG
8       TAGACG
7       CTAAGG
6       TGGCGG
5       GCATGG
3       TCGTAG
2       GCTAAT
1       AACGAT
```

2) If two different barcode lengths are used, and some of the short-length barcode was elongated during library preparation. It is necessary to check elongate the position in the short-length barcode, as shown in the figure below. The library is prepared as *index+CA*, so the default index is filled as *index+CA*. Please use the `-r` parameter and verify the results.




## 4.2 Double Barcode Split

① When using MGI Barcode, the barcode sequence is provided with the information on library perpetration. Barcode 2 is provided in the reverse complement sequence and barcode 1 is in the original sequence. As such, please use `-r` parameter. If the barcode length is in the range of 6-8 bp, 0 is recommended for `allowed_mismatch` parameter.

```
-bash-4.1$ cat doubleIndex.txt
1       ACTGCATACAGTTGAC
2       ACTGCATAGTCAGAGT
3       ACTGCATATGAGTGTG
4       ACTGCATATTGCTCTC
5       ACTGCATACTGTAGTA
6       ACTGCATACGGATAAG
7       ACTGCATAGCAGAGCC
8       ACTGCATAGACTGCTA
9       ACTGCATACATAAATG
10      ACTGCATATACTCTCG
```

② When MGI's APP-A library preparation is used, the barcode sequence will be constructed with the reverse complement of i5 sequence and the original i7 sequence from sequence information table below (***i5 sequence+ i7 sequence***). Note: there is no need to use the `-r` parameter.

```
-bash-4.1$ cat doubleIndex.txt
1       ACTGCATACAGTTGAC
2       ACTGCATAGTCAGAGT
3       ACTGCATATGAGTGTG
4       ACTGCATATTGCTCTC
5       ACTGCATACTGTAGTA
6       ACTGCATACGGATAAG
7       ACTGCATAGCAGAGCC
8       ACTGCATAGACTGCTA
9       ACTGCATACATAAATG
10      ACTGCATATACTCTCG
```

If the script is unsuccessful in demultiplexing the fastq file, open the `SequenceStat.txt` file and compare the barcode sequence with high undecoded ratio (Percentage (%) column) to with the barcode information from the Library Preparation. If the barcode sequence does not match, you may need to orient the barcode in its reverse complement order and / or use the -r parameter

Barcode information from `SequenceStat.txt`

```
-bash-4.1$ less result_pe_double_index11/L01_132/SequenceStat.txt
TATCCTCTTCTGTCGA          undecoded          106904   0.032364
```

Barcode information from Library Preparation

| ID | SampleID | I5 index seqnence | I7 index seqnence | Sample type |
|---|---|---|---|---|
| 1 | YSD1006 | AGAGGATA | TCTGTCGA | Amplification |
| 2 | YSD1007 | ACTGCATA | ACTCTGAC | Amplification |

③ The splitBarcode software currently does not handle an `index.txt` file with 2 different barcode length. If two different barcode lengths are used for mixed library preparation. 1) Create two different `index.txt` file, each file should contain the same barcode length. The `splitBarcode` will need to be executed twice, once per each `index.txt` file

# 5 Software Usage

## 5.1 Paired End (PE) Double Barcode

For PE150 sequencing with double barcode, the read length of PE_1 is 150 bp and PE_2 is 166 bp. The double barcode sequences are located in the last 16 bp of PE_2.

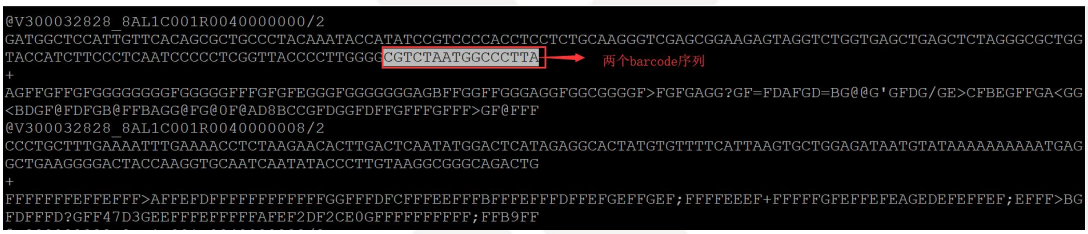
Use the following command to split the barcode:

```
./splitBarcode doubleIndex.txt test_1.fq.gz \
   -2 test_2.fq.gz \
   -o result_pe_double_index \
   -b 300 8 1 \
   -b 308 8 1 \
   -n 30 \
   -m 200
```

Content of `doubleIndex.txt`: Combine the 2 barcode sequences together, without spaces.


| ID | SampleID | I5 index seqnence | I7 index seqnence | Sample type |
|---|---|---|---|---|
| 1 | YSD1006 | AGAGGATA | TCTGTCGA | Amplification |
| 2 | YSD1007 | ACTGCATA | ACTCTGAC | Amplification |

If you would like to check the barcode orientation, open the fastq file and examine the last 16 bp of the read. Based on this barcode sequence, you may need to update the `doubleIndex.txt` file or use the `-r` parameter



## 5.2 Paired End (PE) Single Barcode

For PE150 sequencing with single barcode, the read length for PE_1 is 150bp and PE_2 is 158bp. The barcode sequences is located in the last 8 bp of the PE_2 read.

Use the following command to split the barcode:

```
./splitBarcode Index.txt test_1.fq.gz \
   -2 test_2.fq.gz \
   -o result_pe_single_index \
   -b 300 8 1 \
   -n 30 \
   -m 200
```

Content of `singleIndex.txt`

```
-bash-4.1$ head singleindex.txt
1       CGTACATG
2       GACTGTCA
3       TCAGTCAC
4       ATGCAGGT
5       GCACTGGT
6       TGTTGATC
7       ATGGACAG
8       ACAGGGTC
9       GGTATACA
10      CTGCCTGT
```

If you would like to check the barcode orientation, open the fastq file and examine the last 16 bp of the read. Based on this barcode sequence, you may need to update the `doubleIndex.txt` file or use the `-r` parameter

```
@V300032828_8AL1C001R0040000013/2
ACATAAATGGAATCATACAGTGTGCAGACTTTGAGTCGGCTTCTTAGCACGATGCATTTTGGACCATCCATTTGGCTTGTTCTTTTTCATGGCTCAGTAGTATTCCACTGA
ATGGATGTACCACAGTTTGTTCATCATTCCTCAGCTGAACGATGACTCAGACTGT
+
FFFFFFFFFFFFFGGFFF>FFFFGFFFFFFFGFFF<FFFFFGFFFEGCF?FCGG9FEGFFFBCGFFFFFAGG>FFGFEF>FGFGFEFGCFGDFFFFDFFFFEFGCFFGFF8
FFGF>FGFCFAFFGCFFFCEFFDDFGFFGFGFF@FFFCFEFFFFFFFFFFFFFFF
@V300032828_8AL1C001R0040000017/2
CTGGCAAGCAAGCCCTTCTGCAATCTAACTTTTCTAGCCTGAATTCCCACCACTCCCCTACCTATACCATCCACTTTTGCCAAACTGTAAATACTTACTGTTTACTGAAGA
AGCTTCACCCTTGAAACGTCCTTTCTCCACCTCCACCTCCAGTTGACCTATATCA
+
FGFFFFFFGFFFFGFFGGFGFFGGGFFFGFFFGGFDFGGFCFFFFFGFFGFFFFFFFGGGGFEFFBGFGFGFGFGFG>GBFGFFEGFFGFEEGFGGFEGFACFFFGB9FFB
FCGF=GFGFEGFFDFCGG@FGCGDFCGG@F@FDF?FGCFFFFFFF?FFFFGGGGG
```

## 5.3 Single End (SE) Double Barcode

For SE sequencing with double barcode, the read length is 120 bp. The barcode sequences are located in the last 20 bp of the read.

Use the following command to split the barcode:

```
./splitBarcode doubleIndex.txt \
   -o result_se_double_index \
   -b 100 10 1 \
   -b 110 10 1
```

## 5.4 Single End (SE) Single Barcode

For SE sequencing with single barcode, the read length is 110 bp. The barcode sequences are located in the last 10b bp of the read.

Use the following command to split the barcode:

```
./splitBarcode singleIndex.txt \
   -o result_se_single_index \
   -b 100 10 1
```