**Demultiplex-by-Index-Header Script**

Author: Wei Chun (John) Chen

|  |
| --- |
| **Command-line**  nohup ./main.sh -w /folderpath -h dmx -p <parallel runs> -j <projectID> -a Undetermined-R1.fastq.gz -b Undetermined-R2.fastq.gz -s SampleSheet.csv -l <barcode length> -m <mismatch> -k <index type> > /folderpath/<ProjectID>.log.out 2>&1 &  **Arguments**  -w Working directory path  -h Script ID for matching job ID (please use “dmx” as default)  -p Max number of parallel runs (10 as default, may use more)  -j Project folder name (same name as log.out file)  -a Undetermined reads R1: fastq.gz file  -b Undetermined reads R2: fastq.gz file  -s Samplesheet: csv file  -l Barcode length: 6; 8; 10  -m Mismatch: 0 = no mismatch; 1 = one mismatch  -k Index type: 1 = single index; 2 = dual indexes  **Example**  nohup ./main.sh -w /fast-data/BI/RUO\_wchen/test/demultiplex-v1.1 -h dmx -p 20 -j 22167-01 -a /fast-data/BI/RUO\_wchen/FastQ/undetermined-reads/Undetermined-121922-x-fc1\_S0\_L004\_R1\_001.fastq.gz -b /fast-data/BI/RUO\_wchen/FastQ/undetermined-reads/Undetermined-121922-x-fc1\_S0\_L004\_R2\_001.fastq.gz -s /fast-data/BI/RUO\_wchen/samplesheet/SampleSheet-121922-x-fc1-L04-test2.csv -l 6 -m 1 -k 2 > /fast-data/BI/RUO\_wchen/test/demultiplex-v1.1/22167-01.log.out 2>&1 &  **Note**  The name of log.out file should to the same as project folder name. For example, if the project folder name is 22167-01, then the log.out file will be 22167-01.log.out.  FastQ files will be generated in the “fastq” folder under the project folder.  Other files that will also be generated under the project folder:  mm1.txt – Variation of Index 1  mm2.txt – Variation of Index 2  samplelist.txt – List containing samples and barcodes from the samplesheet  runtime.txt – Total run time of the script. |

**Introduction**

Demultiplex-by-Index-Header script splits undetermined reads by their index header, and provides an alternative way of performing demultiplex locally instead of re-demultiplexing samples from run using the more computational-demanding bcl2fastq program. Variable barcode length, mismatch, and type can be assessed using this script.

**Advantage**

• Consume less resources using built-in Bash command zcat and grep.

• Allow less time to complete by parallel processing. \*

**Limitation**

• Less customization in order to reduce computational complexity.

• Data with huge size may still take longer time to process when compared to using demultiplex application on BaseSpace. \*

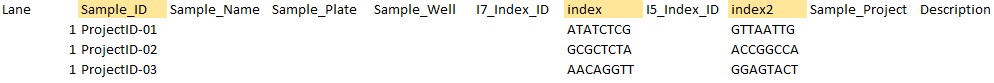
**Required Input Files**

• Undetermined reads R1: fastq.gz file (compressed)

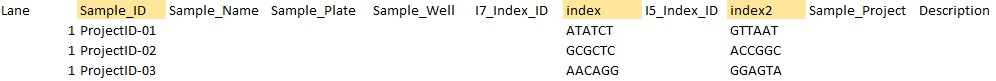
• Undetermined reads R2: fastq.gz file (compressed)

• Samplesheet (see examples below): csv file

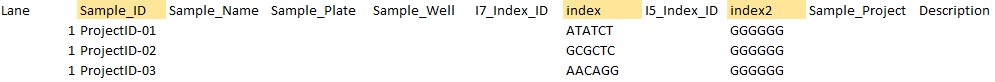
Create a samplesheet from the template. Note that columns Sample ID, Index (i7), and Index2 (i5) are needed. This samplesheet shows dual indexes with nucleotide length of 8 each.



This samplesheet shows duals indexes with length of 6.



This samplesheet shows single index with length of 6. Note that column Index2 should be filled with a place-holder index, and will not affect the results.



**Required Scripts** (placed in same folder)

• Main.sh – Master script

• convert2txt.py

• dmx0-R1.sh

• dmx0-R2.sh

• dmx6-R1.sh

• dmx6-R1.sh

• dmx8-R1.sh

• dmx8-R1.sh

• dmx10-R1.sh

• dmx10-R1.sh

**Programming Language**

• Python

• Bash

**Additional Note**

To count reads: echo $(zcat file.fastq.gz|wc -l)/4|bc

To check current jobs: ps

To cancel individual job: kill <job-ID>

To cancel all jobs in user account: pkill -u <user-ID>

Cancel all jobs only when no other important jobs are running.