Perfomance Test

sample: PRJEB35434ENA from ENA (European Nucleotide Archive), EBI

accession: ERR3668534

1. DADA2 result (fasta)

1) figaro

$ figaro.py -i data/input -o data/output -a 350 -f 20 -r 20

: Run time: 0:00:18.742778

Forward read length: 270
Reverse read length: 270
{"trimPosition": [**219, 191**], "maxExpectedError": [2, 2], "readRetentionPercent": 85.52, "score": 83.51630894034139}
{"trimPosition": [218, 192], "maxExpectedError": [2, 2], "readRetentionPercent": 85.51, "score": 83.50544436128347}
{"trimPosition": [217, 193], "maxExpectedError": [2, 2], "readRetentionPercent": 85.5, "score": 83.49699413312732}
{"trimPosition": [220, 190], "maxExpectedError": [2, 2], "readRetentionPercent": 85.49, "score": 83.49457978222554}
{"trimPosition": [216, 194], "maxExpectedError": [2, 2], "readRetentionPercent": 85.47, "score": 83.47405779956058}
{"trimPosition": [221, 189], "maxExpectedError": [2, 2], "readRetentionPercent": 85.47, "score": 83.4668147468553}
{"trimPosition": [215, 195], "maxExpectedError": [2, 2], "readRetentionPercent": 85.43, "score": 83.4422100968155}

2) DADA2 (https://benjjneb.github.io/dada2/tutorial.html)

① R install: sudo apt install r-base-core (R version 4.3.3)

② DADA2 install:

```
> if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")
  BiocManager::install("dada2")
```

③ Run DADA2

```
> library("dada2")
> setwd("~/PixelCut/performance_test/01_figaro")
> path <- "./data/input"
> list.files(path)
```

```
> fnFs <- sort(list.files(path, pattern="_R1.fastq", full.names = TRUE))
> fnRs <- sort(list.files(path, pattern="_R2.fastq", full.names = TRUE))
> sample.names <- sapply(strsplit(basename(fnFs), "_"), `[`, 1)
> filtFs <- file.path(path, "filtered", paste0(sample.names, "_F_filt.fastq.gz"))
> filtRs <- file.path(path, "filtered", paste0(sample.names, "_R_filt.fastq.gz"))
> names(filtFs) <- sample.names
> names(filtRs) <- sample.names
> out <- filterAndTrim(fnFs, filtFs, fnRs, filtRs, truncLen=c(219, 191),
              maxN=0, maxEE=c(2,2), truncQ=2, rm.phix=TRUE,
              compress=TRUE, multithread=TRUE) # On Windows set multithread=FALSE
Creating output directory: ./data/input/filtered
> out
                                                reads.in  reads.out
ERR3668534_S1_L001_R1_001.trimmed.fastq    160910     135571
ERR3668535_S1_L001_R1_001.trimmed.fastq    170440     146906

> errF  <- learnErrors(filtFs, multithread=TRUE)
> errR  <- learnErrors(filtRs, multithread=TRUE)
> dadaFs  <- dada(filtFs, err=errF, multithread=TRUE)
> dadaRs  <- dada(filtRs, err=errR, multithread=TRUE)
> mergers <- mergePairs(dadaFs, filtFs, dadaRs, filtRs, verbose=TRUE)
> seqtab  <- makeSequenceTable(mergers)
> seqtab.nochim <- removeBimeraDenovo(seqtab, method="consensus", multithread=TRUE,
                          verbose=TRUE)
> sum (seqtab.nochim)/sum(seqtab)
> seqs <- colnames(seqtab.nochim)
> seq_names  <- paste0 ("seq", seq(1, length(seqs)))
> fasta_output <- paste0(">", seq_names, "₩n", seqs, "₩n", collapse = "")
> writeLines(fasta_output, "figaro.fasta")
> quit()
$ cd ~/PixelCut/performance_test/01_figaro
$ grep -c '>' figaro.fasta
11 <- consensus sequence count
```

2) PixelCut

$ python PixelCut.py ERR3668534_1_fastqc.html ERR3668534_2_fastqc.html

Read1 Length: 283
Read2 Length: 279
[Read1] truncation position: **274**
[Read2] truncation position: **257**
Total time: 0:00:07.497741

$ python PixelCut.py ERR3668535_1_fastqc.html ERR3668535_2_fastqc.html

Read1 Length: 283
Read2 Length: 279
[Read1] truncation position: 274
[Read2] truncation position: 266
Total time: 0:00:07.140748

2) DADA2 (https://benjjneb.github.io/dada2/tutorial.html)

① R install: sudo apt install r-base-core (R version 4.3.3)

② DADA2 install:

```
> if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")
  BiocManager::install("dada2")
```

③ Run DADA2

```
> library("dada2")
> setwd("~/PixelCut/src/cli_version/01_PixelCut/ver_0.01")
> path <- "./data/input"
> list.files(path)
> fnFs <- sort(list.files(path, pattern="_1.fastq.gz", full.names = TRUE))
> fnRs <- sort(list.files(path, pattern="_2.fastq.gz", full.names = TRUE))
> sample.names <- sapply(strsplit(basename(fnFs), "_"), `[`, 1)
> filtFs <- file.path(path, "filtered", paste0(sample.names, "_F_filt.fastq.gz"))
> filtRs <- file.path(path, "filtered", paste0(sample.names, "_R_filt.fastq.gz"))
> names(filtFs) <- sample.names
> names(filtRs) <- sample.names
> out <- filterAndTrim(fnFs, filtFs, fnRs, filtRs, truncLen=c(274, 257),
                       maxN=0, maxEE=c(2,2), truncQ=2, rm.phix=TRUE,
```

```
            compress=TRUE, multithread=TRUE) # On Windows set multithread=FALSE
Creating output directory: ./data/input/filtered
> out
                                    reads.in reads.out
ERR3668534_1.fastq.gz     160980      112004
ERR3668535_1.fastq.gz     171088      127297

> errF   <- learnErrors(filtFs, multithread=TRUE)
> errR   <- learnErrors(filtRs, multithread=TRUE)
> dadaFs   <- dada(filtFs, err=errF, multithread=TRUE)
> dadaRs   <- dada(filtRs, err=errR, multithread=TRUE)
> mergers <- mergePairs(dadaFs, filtFs, dadaRs, filtRs, verbose=TRUE)
> seqtab   <- makeSequenceTable(mergers)
> seqtab.nochim <- removeBimeraDenovo(seqtab, method="consensus", multithread=TRUE,
                        verbose=TRUE)
> sum (seqtab.nochim)/sum(seqtab)
> seqs <- colnames(seqtab.nochim)
> seq_names   <- paste0 ("seq", seq(1, length(seqs)))
> fasta_output <- paste0(">", seq_names, "₩n", seqs, "₩n", collapse = "")
> writeLines(fasta_output, "PixelCut.fasta")
> quit()
$ cd ~/ PixelCut/src/cli_version/01_PixelCut/ver_0.01
$ grep -c '>' PixelCut.fasta
718 <- consensus sequence count
```

2) PixelCut – (use figaro test data)

$ python PixelCut.py ERR3668534_R1_fastqc.html ERR3668534_R2_fastqc.html

```
Read1 Length: 270
Read2 Length: 270
[Read1] truncation position:   266
[Read2] truncation position:   266
Total time: 0:00:07.140748
```

$ python PixelCut.py ERR3668535_R1_fastqc.html ERR3668535_R2_fastqc.html

```
Read1 Length: 270
Read2 Length: 270
[Read1] truncation position:   266
[Read2] truncation position:   266
Total time: 0:00:06.237060
```

2) DADA2 (https://benjjneb.github.io/dada2/tutorial.html)

① R install: sudo apt install r-base-core (R version 4.3.3)

② DADA2 install:

```
> if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")
  BiocManager::install("dada2")
```

③ Run DADA2

```
> library("dada2")
>setwd("~/PixelCut/src/cli_version/01_PixelCut/ver_0.01/data/input_trimmed_fastqc/PixelCut")
> path <- "./data/input"
> list.files(path)
> fnFs <- sort(list.files(path, pattern="_S1_L001_R1_001.trimmed.fastq", full.names = TRUE))
> fnRs <- sort(list.files(path, pattern="_S1_L001_R2_001.trimmed.fastq", full.names = TRUE))
> sample.names <- sapply(strsplit(basename(fnFs), "_"), `[`, 1)
> filtFs <- file.path(path, "filtered", paste0(sample.names, "_F_filt.fastq"))
> filtRs <- file.path(path, "filtered", paste0(sample.names, "_R_filt.fastq"))
> names(filtFs) <- sample.names
> names(filtRs) <- sample.names
> out <- filterAndTrim(fnFs, filtFs, fnRs, filtRs, truncLen=c(266, 266),
                maxN=0, maxEE=c(2,2), truncQ=2, rm.phix=TRUE,
```

```
            compress=TRUE, multithread=TRUE) # On Windows set multithread=FALSE
Creating output directory: ./data/input/filtered
> out
                                                reads.in reads.out
  ERR3668534_S1_L001_R1_001.trimmed.fastq     160910       105650
  ERR3668535_S1_L001_R1_001.trimmed.fastq     170440       122848

> errF    <- learnErrors(filtFs, multithread=TRUE)
> errR    <- learnErrors(filtRs, multithread=TRUE)
> dadaFs    <- dada(filtFs, err=errF, multithread=TRUE)
> dadaRs    <- dada(filtRs, err=errR, multithread=TRUE)
> mergers <- mergePairs(dadaFs, filtFs, dadaRs, filtRs, verbose=TRUE)
> seqtab    <- makeSequenceTable(mergers)
> seqtab.nochim <- removeBimeraDenovo(seqtab, method="consensus", multithread=TRUE,
                            verbose=TRUE)
> sum (seqtab.nochim)/sum(seqtab)
> seqs <- colnames(seqtab.nochim)
> seq_names <- paste0 ("seq", seq(1, length(seqs)))
> fasta_output <- paste0(">", seq_names, "₩n", seqs, "₩n", collapse = "")
> writeLines(fasta_output, "PixelCut_trimmed.fasta")
> quit()
$ cd ~/PixelCut/src/cli_version/01_PixelCut/ver_0.01/data/input_trimmed_fastqc/PixelCut
$ grep -c '>' PixelCut_trimmed.fasta
```

702 <- consensus sequence count