ECOGEO 'Omics Training: 2.0 Quality Control Version 2

Ben Tully

Abstract

Quality impact results is the universal first step for all sequencing methods. This protocol is for sequencing results in FASTQ format. FastQC allows for visualization of quality scores. Sequence trimming allows for removal of residual primer sequences and increases overall quality scores. Trimmomatic does both. CutAdapt cuts adapters from sequences. Sickle creates a sliding window for quality and length trimming of FASTQ files.

Open this protocol inside the virtual machine (details in 'Start Instructions') for easy copy, paste of commands into the command line terminal window.

Citation: Ben Tully ECOGEO 'Omics Training: 2.0 Quality Control. **protocols.io**
dx.doi.org/10.17504/protocols.io.fixbkfn
Published: 18 Aug 2016

Before start

Before starting, please visit the ECOGEO website for more information on this 'Introduction to Environmental 'Omics' training series. The site contains a pre-packaged virtual machine that can be downloaded and used to run all of the protocols in this protocols.io collection. In addition to the VM, the website contains video and presentations from our initial 'Intro to Env 'Omics' workshop held at the Univ. of Hawai'i at Manoa on 25-26 Jul 2016.

Please email ‘ecogeo-join@earthcube.org’ to join the ECOGEO listserv for future updates.

Protocol

FastQC

Step 1.

Move to quality control directory:

cmd COMMAND

```
$ cd /home/c-debi/ecogeo/qualitycontrol
```

FastQC

Step 2.

View sample FASTQ file:

cmd COMMAND

```
$ less igm1000002065_R1.fastq
```

EXPECTED RESULTS

```
@SRR492065.1453522 HWI-EAS385_0095_FC:2:33:6836:12013
length=100
ATTTTTGCTCAATCAATTTTGACATAGAAATGCCATTCGGACACAAAATCACCGCCTTAAT
TATTTTCTCTTCCAAATCATTGTGATCATTACTAATCAA
+
IIIIIIIIIIIIIIIIGIIHIIIIHIIIIIIIIIGIIIGIHIIIIIIIIIIIIBIIFHHIG
IIIHFHIHIHBIIIIIIIIGIFGDHIIIIIIIIHIGIEI
```

Step 3.

FastQC - allows for visualization of quality scores:

cmd COMMAND

```
$ fastqc
```

NOTES

Elisha Wood-Charlson 08 Aug 2016

The rest of the steps are explained in the presentation slides, video.

Trimmomatic

Step 4.

Trimmomatic

Input parameters

Java -jar : command used to run Java based programs

Location of program : /home/c-debi/ecogeo/BioinfPrograms/Trimmomatic-0.35/

PE vs SE : PE can maintain order of PE sequences

-phred33 : explicitly telling Trimmomatic which quality score system we are using

Inputs : igm1000002065_R1.fastq & igm1000002065_R2.fastq

Outputs : simplified (allows for quickly testing different parameters) generates FASTQ of surviving PE and SE results

Outputs : R1_pe & R2_pe contain paired reads after quality trimming

R1_se & R2_se contain reads who lost their mate during trimming

ILLUMINACLIP:

Location of adapter file : /home/c-debi/ecogeo/BioinfPrograms/Trimmomatic-0.35/adapters/

TruSeq3-PE.fa : FASTA file containing adapters used by MiSeq & HiSeq

2 : maximum mismatch value

30 : palindrome clip threshold – Trimmomatic will use F & R primers and check for presence at beginning and end of reads

10 : simple clip threshold – specifies accuracy required of match when not palindromic

SLIDINGWINDOW : checks for decrease in quality

10 = Slide over 10 bp windows of the sequence

28 = if the average quality score <28 trim the sequence

MINLEN : remove sequences <50bp in length

cmd COMMAND

```
$ java -jar /home/c-debi/BioinfPrograms/Trimmomatic-0.35/trimmomatic-0.35.jar PE -
phred33 igm1000002065_R1.fastq igm1000002065_R2.fastq R1_pe R1_se R2_pe R2_se ILLUMINACLIP:
/home/c-debi/BioinfPrograms/Trimmomatic-0.35/adapters/TruSeq3-
PE.fa:2:30:10 SLIDINGWINDOW:10:28 MINLEN:50
```

Trimmomatic

Step 5.

Trimmed output, open FastQC.

cmd COMMAND

```
$ fastqc
```

Trimmomatic

Step 6.

File > Open ... R1_pe & R2_pe (Filter: All Files)

Check the Quality Score Assessment

Rename Trimmomatic output:

cmd COMMAND

```
$ mv R1_pe igm1000002065_R1_trim.fastq
$ mv R2_pe igm1000002065_R2_trim.fastq
```

CutAdapt + Sickle

Step 7.

CutAdapt

-a : adapter sequence to searched at 3’ end of sequence (-g = front, -b = anywhere)

-e : maximum allowed error rate between a match

--overlap : minimum number of base pairs to match adapter to trim

cmd COMMAND

```
$ cutadapt -a AGATCGGAAGAGC -e 0.08 --overlap=3 -
o R1_cut igm1000002065_R1.fastq > cut_R1_summary
```

CutAdapt + Sickle

Step 8.

View CutAdapt summary file:

cmd COMMAND

```
$ less cut_R1_summary
```

EXPECTED RESULTS

```
You are running cutadapt 1.7.1 with Python 2.7.6
Command line parameters: -a AGATCGGAAGAGC -e 0.08 --overlap=3 -o
R1_cut igm1000002065_R1.fastq
Maximum error rate: 8.00%
   No. of adapters: 1
   Processed reads:       100000
   Processed bases:     10000000 bp (10.0 Mbp)
     Trimmed reads:         2451 (2.5%)
     Trimmed bases:        20795 bp (0.0 Mbp) (0.21% of total)
   Too short reads:            0 (0.0% of processed reads)
    Too long reads:            0 (0.0% of processed reads)
        Total time:      1.40 s
     Time per read:      0.014 ms
```

CutAdapt + Sickle

Step 9.

“All” Illumina sequences (thanks to cadium-gcat)

File = ncbi_univec_adaptors_primers.fasta

cmd COMMAND

```
$ cutadapt -b file:ncbi_univec_adaptors_primers.fasta -e 0.02 --overlap=5 -
o R1_cut igm1000002065_R1.fastq > cut_R1_summary
```

EXPECTED RESULTS

```
You are running cutadapt 1.7.1 with Python 2.7.6
Command line parameters: -b
file:ncbi_univec_adaptors_primers.fasta -e 0.02 --overlap=5 -o
R1_cut igm1000002065_R1.fastq
Maximum error rate: 2.00%
   No. of adapters: 40
   Processed reads:       100000
   Processed bases:     10000000 bp (10.0 Mbp)
     Trimmed reads:         5203 (5.2%)
     Trimmed bases:        35990 bp (0.0 Mbp) (0.36% of total)
   Too short reads:            0 (0.0% of processed reads)
    Too long reads:            0 (0.0% of processed reads)
        Total time:     17.20 s
    Time per read:      0.172 ms
```

CutAdapt + Sickle

Step 10.

Repeat for R2:

cmd COMMAND

```
$ cutadapt -b file:ncbi_univec_adaptors_primers.fasta -e 0.02 --overlap=5 -
o R2_cut igm1000002065_R2.fastq > cut_R2_summary
```

EXPECTED RESULTS

```
You are running cutadapt 1.7.1 with Python 2.7.6
Command line parameters: -b
file:ncbi_univec_adaptors_primers.fasta -e 0.02 --overlap=5 -o
R2_cut igm1000002065_R2.fastq
Maximum error rate: 2.00%
   No. of adapters: 40
   Processed reads:       100000
   Processed bases:     10000000 bp (10.0 Mbp)
     Trimmed reads:         5176 (5.2%)
     Trimmed bases:        33463 bp (0.0 Mbp) (0.33% of total)
   Too short reads:            0 (0.0% of processed reads)
    Too long reads:            0 (0.0% of processed reads)
        Total time:     17.58 s
    Time per read:      0.176 ms
```

CutAdapt + Sickle

Step 11.

Sickle: Sliding window quality and length trimming of FASTQ files.

-t : type of quality scores (solexa, illumina, sanger)

-q : quality score cutoff

-l : minimum length after trimming

-n : trim at first N (ambiguous base)

-o : F out

-p : R out

-s : SE out

cmd COMMAND

```
$ sickle pe -f R1_cut -r R2_cut -t sanger -q 28 -l 50 -n -o R1_cut_sickle -
p R2_cut_sickle -s se_cut_sickle
```

CutAdapt + Sickle

Step 12.

Open FastQC

File > Open ... R1_cut_sickle & R2_cut_sickle

Check the Quality Score Assessment

cmd COMMAND

```
$ fastqc
```

CutAdapt + Sickle

Step 13.

Rename Output

cmd COMMAND

```
$ mv R1_cut_sickle igm1000002065_R1_cut_sickle.fastq
$ mv R2_cut_sickle igm1000002065_R2_cut_sickle.fastq
```