Basic Illumina Sequence Quality Control

Benjamin Tully

Abstract

Method used to perform quality control on Illumina sequences - includes adapter/primer removal and quality trimming
Goal: Remove low quality sequences to increase average quality score of reads above 28

Citation: Benjamin Tully Basic Illumina Sequence Quality Control. **protocols.io**
dx.doi.org/10.17504/protocols.io.d4e8td
Published: 17 Nov 2015

Before start

There are many tools available for this process - this is one example that has been used effectively
Required software:
Cutadapt - https://pypi.python.org/pypi/cutadapt
Trimmomatic - http://www.usadellab.org/cms/?page=trimmomatic
Recommended software:
FASTQC - http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Protocol

Step 1.

Optional (Highly Recommended): Assess raw FASTQ sequences using the program FastQC

Goal: Remove low quality sequences to increase average quality score of reads above 28

SOFTWARE PACKAGE (Linux/Unix)
FastQC, 0.11.4
Babraham Bioinformatics

Step 2.

For current sequencing centers, the 5' end of has had Illumina adapters and primers removed. But there can be inclusion of the 3' end Illumina adapter in DNA fragments with length less than the number of the cycles for the sequencer.
Remove Illumina adaptor sequences from 3' end of sequences - be sure to preserve reads of 0 length

SOFTWARE PACKAGE (Linux/Unix)
Cutadapt, 1.9
Marcel Martin
cmd COMMAND (Linux/Unix)

```
cutadapt -a AGATCGGAAGAGC -e 0.08 --overlap=3 -
o OUTPUT_FILE.R1.cutadapt.fastq.gz RAW_FILE_R1_001.fastq.gz > R1_001_summary
```

-a = 3' end adapter sequence 'Illumina adapter - AGATCGGAAGAGC' -e = error tolerance in detecting adapter (0.08 = 8%) --overlap minimum overlap to adapter sequence (default = 3) -o =

output file name

ANNOTATIONS

Benjamin Tully 17 Nov 2015

NCBI provides a robust list of other possible vector/adapter contaminants: http://www.ncbi.nlm.nih.gov/tools/vecscreen/uvcurrent/#Replist

Step 3.

Repeat step for all raw read files

Step 4.

Assess Cutadapt results with FastQC

Step 5.

Remove low quality base pairs from sequences, generally from the 3' end, using a sliding window of 10 bp that will trim all trailing bases if the average quality score drops below 28

Important: To maintain the order of paired-end Illumina sequences both reads must be input simultaneously with the same number of sequences submitted from each read pair, regardless of the length of the sequence

SOFTWARE PACKAGE (Linux/Unix)

Trimmomactic, 0.35

Anthony Bolger

COMMAND (Linux/Unix)

```
java -jar /directory/location/containing/Trimmomatic-0.33/trimmomatic-0.33.jar PE -
phred33 -
threads 32 IN_FILE.R1.cutadapt.fastq.gz IN_FILE.R2.cutadapt.fastq.gz OUT_FILE.R1.cutadapt.t
rimmomatic_paired.fastq.gz OUT_FILE.R1.cutadapt.trimmomatic_unpaired.fastq.gz OUT_FILE.R2.c
utadapt.trimmomatic_paired.fastq.gz OUT_FILE.R2.cutadapt.trimmomatic_unpaired.fastq.gz SLID
INGWINDOW:10:28 MINLEN:75
```

-phred33 = uses the Phred 33 scale for quality control - current output for all Illumina sequences - threads = the number of CPUs that can be assigned to task SLIDINGWINDOW:10:28 = use a sliding window of 10 bp to trim reads if the average quality score in that window drops below 28 MINLEN:75 = minimum length for good sequence output set at 75 bp In files = read pairs following cutadapt step Out file = 2 different file types. Sequences in 'paired.fastq.gz' have the same paired sequences in identical order for reads R1 and R2. Sequences in 'unpaired.fastq.gz' are sequences for which the partner sequence did not survive the trimming thresholds.

Step 6.

Assess file quality scores using FastQC.

Utilize Trimmomatic for any other issues - such as poor quality in the first bp, etc.