Quality control for metagenomics data

Qi Wang[1]

[1]BGI

Apr 16, 2020

1 Works for me dx.doi.org/10.17504/protocols.io.be68jhhw

Qi Wang

ABSTRACT

Quality control for metagenomics data,including: remove low quality reads and host contamination reads.

GUIDELINES

Quality control for metagenomics data,including: remove low quality reads and host contamination reads.

SAFETY WARNINGS

No

BEFORE STARTING

The user should provided the single or paired metagenomics data.

1 Step1: remove low quality reads

We firstly calculate the accuracy probabilities of each base using the following equations:

1)$Q = -10 \log_{10} E$

2)$P = 1 - E$

Where Q is thePhred quality score of each base, E is the error probability of each base.

Then we calculate the overall accuracy probability(OA)of each readusing the following equation:

For each read, an initial 30-mer seed sequence is selected at the 5' end of the read and its overall accuracy, defined as OA_{seed}, is calculated. To ensure high data quality, $OA_{seed}$is defined as 0.9 using m equal to 0with zero low quality bases allowed. Once the seed position of the read has been defined, the seed would extend to keep the longest contiguous read fragment in which the OA, defined as $OA_{fragment}$, is above a defined accuracy threshold. In this study, weset $OA_{fragment}$equal to or greater than 0.8 using $m$equal to 1.

And the Perl scripts for overall accuracy based QC pipeline are freely available for download and reuse from Github (https://github.com/Scelta/OAFilter).

2 Step2: remove host contamination reads by one command:

'bowtie2 --very-sensitive -p $thread -x $host_bowtie2_index -1 $sample_r2 -2 $sample_r1 2> 02.rmhost/bowtie2.log | samtools view -h | samtools sort -n |samtools fastq -N -c 5 -f 12 -1 02.rmhost/$name.rmhost.1.fq.gz -2 02.rmhost/$name.rmhost.2.fq.gz'

Citation: Qi Wang (04/16/2020). Quality control for metagenomics data. https://dx.doi.org/10.17504/protocols.io.be68jhhw