APR 16, 2020

OPEN ACCESS

DOI:
dx.doi.org/10.17504/protocols.io.be68jhhw

Protocol Citation: Qi Wang 2020. Quality control for metagenomics data.
protocols.io
<https://dx.doi.org/10.17504/protocols.io.be68jhhw>

License: This is an open access protocol distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Protocol status: Working
 We use this protocol and it's working

Created: Apr 16, 2020

Last Modified: Apr 16, 2020

PROTOCOL integer ID:
 35776

Keywords: Quality control, Metagenomics data, remove low quality reads, remove host contamination reads

Quality control for metagenomics data

In 4 collections

Qi Wang¹

¹BGI

GigaScience Press

BGI

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

BEFORE START INSTRUCTIONS

The user should provide the single or paired metagenomics data.

Step1: remove low quality reads :

1 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 the Phred 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 0 with 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, we set

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'
```