Apr 16, 2020

Quality control for metagenomics data

GigaByte

✓ Peer-reviewed method

In 4 collections

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

Qi Wang[1]

[1]BGI

GigaScience Press

BGI

Qi Wang

OPEN ACCESS

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

External link: https://doi.org/10.46471/gigabyte.12

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

Manuscript citation:
Qi Wang, Qiang Sun, Xiaoping Li, Zhefeng Wang, Haotian Zheng, Yanmei Ju, Ruijin Guo, Songlin Peng, Huijue Jia, Linking gut microbiome to bone mineral density: a shotgun metagenomic dataset from 361 elderly women, Gigabyte, 2021 https://doi.org/10.46471/gigabyte.12

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

Created: April 16, 2020

Last Modified: April 16, 2020

Protocol Integer ID: 35776

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

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 start

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, $\text{OA}_{\text{seed}}$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 $\text{OA}_{\text{fragment}}$, is above a defined accuracy threshold. In this study, weset $\text{OA}_{\text{fragment}}$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'
```