Version 3

Aug 25, 2021

MultiQuas (*Multi*ple reference *quas*ispecies reconstruction protocol) V.3

Marco Cacciabue[1]

[1]Instituto de Agrobiotecnología y Biología Molecular (IABIMO, INTA-CONICET)

In Development Share

This protocol is published without a DOI.

FMDV_ARG_Lab

Marco Cacciabue
Instituto de Agrobiotecnología y Biología Molecular (IABIMO,...

DISCLAIMER

DISCLAIMER – FOR INFORMATIONAL PURPOSES ONLY; USE AT YOUR OWN RISK

The protocol content here is for informational purposes only and does not constitute legal, medical, clinical, or safety advice, or otherwise; content added to protocols.io is not peer reviewed and may not have undergone a formal approval of any kind. Information presented in this protocol should not substitute for independent professional judgment, advice, diagnosis, or treatment. Any action you take or refrain from taking using or relying upon the information presented here is strictly at your own risk. You agree that neither the Company nor any of the authors, contributors, administrators, or anyone else associated with protocols.io, can be held responsible for your use of the information contained in or linked to this protocol or any of our Sites/Apps and Services.

ABSTRACT

The following protocol summarizes the major steps to run the MultiQuas pipeline to evaluate viral variability and reconstruct the viral quasispecies from NGS data (particularly Miseq reads). It is based on the assumption that 1 o more known references are available. These references could be obtained using other haplotype reconstruction softwares. Nonetheless, it is recommended that only a few trusted references are used.

PROTOCOL CITATION

Marco Cacciabue 2021. MultiQuas (Multiple reference quasispecies reconstruction protocol). **protocols.io**
https://protocols.io/view/multiquas-multiple-reference-quasispecies-reconstr-bxpvpmn6
Version created by Marco Cacciabue

LICENSE

CREATED

Aug 25, 2021

LAST MODIFIED

Aug 25, 2021

PROTOCOL INTEGER ID

52693

MATERIALS TEXT

QuRe

source by Mattia C. F. Prosperi

FastQC 0.11.9

by Simon Andrews

Align_to_references.sh

```
#!/bin/bash


start=`date +%s`


bbduk.sh in1=$2 out1=reads_1.fq in2=$3 out2=reads_2.fq ref=
[path/to/bbmap/instalation]/bbmap/resources/adapters.fa ktrim=r k=23 mink=11
hdist=1 tpe tbo qtrim=rl trimq=20 minlen=50 maq=20


bowtie2-build $1 VFAref

bowtie2 --no-discordant --no-mixed -p $4 -x VFAref -1 reads_1.fq -2 reads_2.fq |
samtools view -@ 4 -bT $1 - > SAMPLE.bam


samtools sort -@ 4 -m 2G SAMPLE.bam > SAMPLE_sorted.bam

samtools view -@ 4 -h -F 4 -b SAMPLE_sorted.bam > SAMPLE_map.bam

samtools index -@ 4 SAMPLE_map.bam SAMPLE_map.bai

samtools depth -d10000000 SAMPLE_map.bam > coverage.txt




lofreq viterbi -f $1 -o SAMPLE_map_viterbi.bam SAMPLE_map.bam



samtools sort  -@ 4 -m 2G SAMPLE_map_viterbi.bam >
SAMPLE_map_viterbi_sorted.bam

samtools index -@ 4 SAMPLE_map_viterbi_sorted.bam
SAMPLE_map_viterbi_sorted.bai


lofreq indelqual --dindel -f $1 -o SAMPLE_map_viterbi_sorted_indels.bam
SAMPLE_map_viterbi_sorted.bam

samtools index -@ 4 SAMPLE_map_viterbi_sorted_indels.bam
SAMPLE_map_viterbi_sorted_indels.bai

lofreq call-parallel --pp-threads $4 --call-indels --use-orphan  -f $1
SAMPLE_map_viterbi_sorted_indels.bam -o variants.vcf


end=`date +%s`
echo Execution time was `expr $end - $start` seconds.
```

bbduk

source by Brian Bushnell

samtools 1.12

source

bcftools 1.12

source

Bowtie2 2.4.4

source

DISCLAIMER:

DISCLAIMER – FOR INFORMATIONAL PURPOSES ONLY; USE AT YOUR OWN RISK

The protocol content here is for informational purposes only and does not constitute legal, medical, clinical, or safety advice, or otherwise; content added to protocols.io is not peer reviewed and may not have undergone a formal approval of any kind. Information presented in this protocol should not substitute for independent professional judgment, advice, diagnosis, or treatment. Any action you take or refrain from taking using or relying upon the information presented here is strictly at your own risk. You agree that neither the Company nor any of the authors, contributors, administrators, or anyone else associated with protocols.io, can be held responsible for your use of the information contained in or linked to this protocol or any of our Sites/Apps and Services.

Brief pipeline description

1 Reads are trimmed and filtered using

bbduk

source by Brian Bushnell

2 Filtered and trimmed reads are aligned to a set of user-defined references (multifasta format) with

Bowtie2 2.4.4
source

3 Reads are then split into different classes (one for each reference and one for the unmapped reads) using

SAMtools 1.8
Linux
source by Wellcome Trust Sanger Institute

4 For each class, reads are merged using

PEAR - Paired-End reAd mergeR
source by Alexandros Stamatakis

and then haplotypes are reconstructed with

QuRe
source by Mattia C. F. Prosperi

Important note: If the time limit is reached this step is repeated with a subset of the reads in order to reduce computation time and resources required by QuRe. Steps include 0.99, 0.9, 0.8, 0.7, 0.6 and 0.5 proportions of the toal reads for each class in that order.

5 Next, the proportions of each haplotype class (predicted by QuRe) are adjusted to reflect the number of reads of the corresponding class.

6 All reconstructed haplotypes are aligned to the first reference in the multifasta file using

mafft 7.487
source by Kazutaka Katoh

7 Additionally, reads are then aligned to the first reference in the multifasta file. Single Nucleotide Variants (SNVs) are called using

Lofreq 2

source by Andreas Wilm

8 Finally, concordance between the SNVs (expected minor allele frequency) from the predicted quasispecies and the Lofreq variants using an in-house R script. Higher value of R-squared indicates a better quasispecies reconstruction. Important: this step does not meant as a validation of the obtained results, but it allows the user to choose between different haplotypes reconstructions.

Installing docker

9 In order to run the pipeline, a wrapper file is available (bash) which automatically perfoms all the above numbered steps. A docker image is available that includes all the necessary dependencies. If you do not yet have docker installed, do so at this time, and ensure that is in your PATH. For more information please visit https://www.docker.com/get-started

10 The docker image ("multiquas") is available at Docker hub
To pull the image, use the command below:

docker pull

```
docker pull cacciabue/multiquas:latest
```

This will download and install the corresponding docker image. Only has to be run the first time (it may take several minutes depending on your internet connection)

10.1 Alternatively, If you don't want to use Docker, you can install all dependencies by yourself (only for linux users). The dependencies are:

- BCFtools v1.8 (or later version) http://www.htslib.org/download/
- Samtools v1.8 (or later version) http://www.htslib.org/download/
- Bowtie2 v2.2.4 http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
- PEAR https://cme.h-its.org/exelixis/web/software/pear/doc.html
- seqtk https://github.com/lh3/seqtk
- bbmap https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbduk-guide/
- Lofreq v2 https://csb5.github.io/lofreq/
- mafft https://mafft.cbrc.jp/alignment/software/
- R v4.1 (or later version) https://www.r-project.org/
- R package seqinr https://cran.r-project.org/web/packages/seqinr/index.html
- R package ape https://cran.r-project.org/web/packages/ape/
- R package VariantAnnotation https://bioconductor.org/packages/release/bioc/html/VariantAnnotation.html
- R package Biostrings https://bioconductor.org/packages/release/bioc/html/Biostrings.html
- R package ggplot2 https://ggplot2.tidyverse.org/
- R package ggrepel https://cran.r-project.org/web/packages/ggrepel/vignettes/ggrepel.html

and contact the author for the most up-to-date wrapper scripts.

Preparing analysis

11 Depending on your operating system follow these steps

On windows: open a windows terminal (WIN + R), type "cmd" and press enter. A Windows terminal

11.1 should be up and running.

11.2 Create a folder to work with and *navigate* to the location for your new *folder*. For example:

mkdir

```
mkdir test_dir
cd test_dir
```

linux

11.3 Copy the fastq files and the reference file (1 or more sequences to use as references in multifasta format) to the test_dir folder.

12 On linux the steps are similar.

12.1 Open a terminal (ctrl + alt + T), A bash terminal should be up and running.

12.2 Create a folder to work with and *navigate* to the location for your new *folder*. For example:

mkdir

```
mkdir test_dir
cd test_dir
```

linux

12.3 Copy the fastq files and the reference file (1 or more sequences to use as references in multifasta format) to the test_dir folder.

Running MultiQuas workflow

13 The following command will create a docker container, mount the test_dir folder into the container and perform all the necessary step of the MultiQuas workflow.

⚠ Backups files of the fastq reads and reference should be made and store in a different location before running the following command in order to prevent loss of data.

13.1 For Windows:

run reconstruction

```
docker run -it --volume $(pwd):/nexus cacciabue/multiquas:latest
reconstruction.sh complete -o OUTPUT_FOLDER -1 R1.fq -2 R2.fq  -r
REFERENCE.fasta
```

The user should replace each of the following:

- OUTPUT_FOLDER: folder name inside test_dir to automatically save all files.
- R1.fq and R2.fq: corresponding names of the paired-end reads (use the correct extension in case you have .fastq)
- REFERENCE.fasta: name of the fasta file to use.

Optional paramenters can be set:

- -l: use to indicate a specific label (default: SAMPLE)
- -p/--proc: Number of threads to use (default=2)
- -m/--mem: ram memmory available to use by the java machine (default=4).
- --min_quality: Phred value cutoff for filtering the reads (default=25)
- --timeout: seconds (per reference) before shutting down QuRe in case it freezes (default=600). Each time, the reads are downsampled and the step is repeated (6 times per reference)

13.2 For linux:

run reconstruction

```
docker run -it --volume $(pwd):/nexus cacciabue/multiquas:latest
reconstruction.sh complete -o OUTPUT_FOLDER -1 R1.fq -2 R2.fq  -r
REFERENCE.fasta
```

The user should replace each of the following:

- OUTPUT_FOLDER: folder name inside test_dir to automatically save all files.
- R1.fq and R2.fq: corresponding names of the paired-end reads.
- REFERENCE.fasta: name of the fasta file to use.

Optional paramenters can be set:

- -l: use to indicate a specific label (default: SAMPLE)
- -p/--proc: Number of threads to use (default=2)
- -m/--mem: ram memmory available to use by the java machine (default=4).
- --min_quality: Phred value cutoff for filtering the reads (default=25)
- --timeout: seconds (per reference) before shutting down QuRe in case it freezes (default=600). Each time, the reads are downsampled and the step is repeated (6 times per reference)

Output files

14 Regardless of the operating system a set of folders should be created:

- Filtering: Filtered and trimmed reads are stored here.
- Multiple_Aligning: Alignment files are stored here.
- unmapped: reads extracted from the alignment are stored here.
- Additionally, for each reference a corresponding folder will be created.

- lofreq: the alignment to the first reference is store here. Also you will find the variant.vcf file here.

A set of output files will also be saved in test_dir/OUTPUT_FOLDER. Most relevant are:

- SAMPLE_haplotypes.fasta (and SAMPLE_haplotypes_aligned.fasta): the reconstructed haplotypes (first sequence is always the first sequence in the REFERENCE.fasta file and should NOT be considered an haplotype).
- SAMPLE_adjusted_proportions.txt: proportions of each reconstructed haplotype.
- SAMPLE_graphs.png: Concordance graph between lofreq and the reconstructed haplotypes. It should include a linear regression equation (the higher R-squared the better)