LoReTTA ver. 0.1

Description

LoReTTA (Long Read Template Targeted Assembler) is a *de novo* assembler specifically designed to deal with PacBio reads generated from viral genomes. A user-friendly GUI (graphic user interface) makes it very easy to use although a command line version is also available.

LoReTTA uses a reference genome to guide the read assembly which results in superior specificity and sensitivity even in the presence of reads generated from additional organisms (e.g. host reads in clinical sample datasets, other microorganisms in metagenomics datasets, etc).

Installation

LoReTTA takes advantage of the Anaconda framework to create an isolated environment where all the dependencies and executable files are installed.

After downloading and unzipping the latest software release, simply entry the LoReTTA folder and launch the installation script as follow (replace X.X with the release number you downloaded) :

```
cd LoReTTA-X.X

bash install_linux.sh
```

The installation will last several minutes (depending also on your internet connection speed) and will produce two executable files: LoReTTA (graphic version) and LoReTTA_cl (command line version). To launch the former just type (within the installation directory):

```
./LoReTTA
```

LoReTTA can be launched from different locations in your disk as long as you add the LoReTTA installation directory to your PATH or specify the complete installation directory path when launching the software. Ex:

```
/home/myUsername/Software/LoReTTa/LoReTTA
```

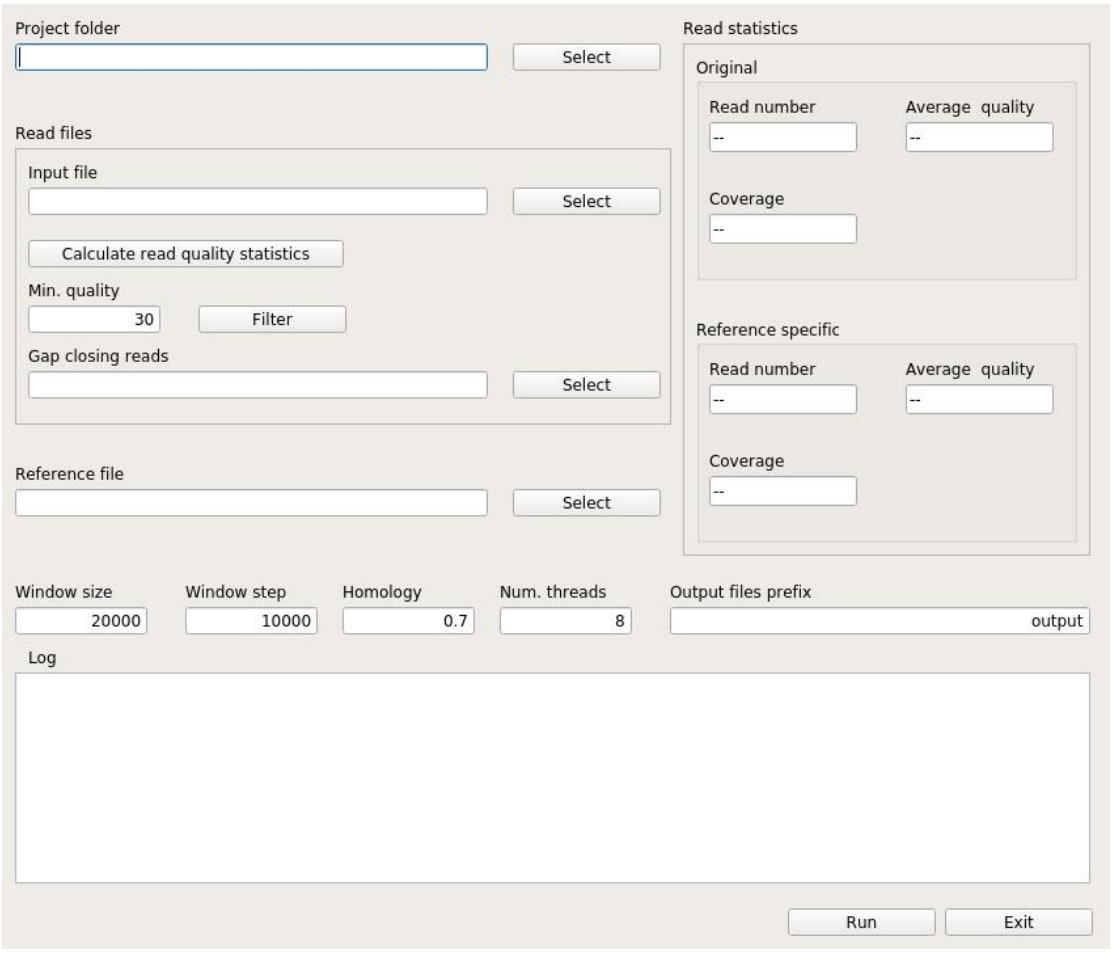
LoReTTA can be used under Linux 64bit and has been tested on Debian like systems such as Ubuntu and Linux Mint although it is expected to run under any Linux distributions.

The software works with fastq formatted read files and fasta formatted reference sequence files. It is important that none of the used folder or file names contain spaces to avoid crashes.

Please be aware that during intensive computations, such as the *de novo* assembly step, the LoReTTA graphic interface may become momentarily unresponsive (in some systems it may gray out) but the process will still proceed and each step will be promptly reported in the GUI log window as well as in the terminal used to run the LoReTTA command.

LoReTTA GUI

When launching LoReTTA the following window will open:



User defined parameters

Project folder

This is the folder where all the output files will be generated. Click the button "Select" on the right and browse your computer (new folders can be created at this stage). Specifying the project folder is compulsory to run any of the modules.

Input file
This is the fastq formatted input file containing the PacBio reads. To select the file just press “Select” on the right and browse your local files.

Min qual
This is the minimum Phred quality value used to filter the provided reads (see the “Read filtering” section below).

Gap closing reads
This is the fastq formatted file containing the reads that can be used to close gaps left after the *de novo* assembly step. The same file specified in “Input file” can be used here. Use the “Select” button on the right to select the file.

Reference file
This is the fasta formatted reference genome file. Use the “Select” button on the right to find it in your local folders.

Window size
The *de novo* assembly is performed by using a reference genome as a template. Such genome is split in portions and scanned using a sliding window approach. In this field you can specify the window size. The default value of 20,000 proved to be suitable for all the tested viral genomes.

Window step
This is the step by which the sliding window moves along the reference genome. The default value of 10,000 proved to be suitable for all the tested viral genomes.

Homology
The PacBio reads are aligned to the reference genome during the early stages of the *de novo* assembly step. Only reads aligning in a proportion of their length equal to the Homology value are retained at this stage. The default value of 0.7 proved to be suitable when aligning reads to reference genome belonging to the same (or similar) strain. If a marked divergence is believed to exist between the reference genome and the genome from which the reads were originated, the Homology value can be decreased.

Num. threads
This is the maximum number of threads LoReTTA will use to perform the requested task.

Output file prefix
This is the prefix that will be used in all the output files generated by LoReTTA.

Read quality statistics

Before attempting a *de novo* assembly, LoReTTA can perform a quality test on the provided reads. After selecting a project folder, an input file and a reference file, this task can be performed by clicking on the "Calculate read quality statistics" button.

LoReTTA will calculate several statistics that will be reported in the "Read statistics" panel of the GUI. Initially, the number of reads, the average Phred quality and the coverage values will be calculated on the provided reads. The coverage value is calculated as the number of read bases divided by the reference genome length regardless of the number of reads originated from different organisms (e.g. host reads in clinical samples will be included in the computation at this stage). The reads are then aligned to the provided reference genome and the described statistics are calculated again only for the mapped reads. Such statistics will be also reported in a text file suffixed with the string "_reads_Quality_Statistics.log". A distribution plot will be also generated reporting the Phred quality distribution of the original input reads together with those mapped to the reference genome (file suffixed with the string "_qualityDist.png").

The read quality statistics module can be tested by selecting the human cytomegalovirus (HCMV) simulated reads contained in the file HCMV_reads.fastq as an input file and the HCMV_Reference_Genome.fasta (HCMV strain merlin) as a reference (provided in the testFiles folder within the LoReTTA distribution). Alternatively, the file HCMV_HSV_mix_reads.fastq can be used as input file. In this case a portion of the reads is originated *in silico* from a HSV-1 reference genome and will results in the differences observed between the "Original" and "Reference specific" sections of the "Read statistics" panel.

Read filtering

Once the Phred quality distribution has been computed, the user can filter the initial dataset using a Phred quality threshold (defined by the "Min qual" box in the GUI). In this process, the reads are scanned and all the bases that do not exceed such threshold are masked. Stretches of continuous unmasked bases are extracted from the original reads together with their Phred quality values and reported in a output file suffixed with the string "_hq_reads.fastq". Only sub-reads longer than 150 bases are retained at this stage.

***De novo* assembly**

After selecting the project folder, the input file, the gap closing file and the reference file, the *de novo* assembly can be performed by clicking the button "Run" in the LoReTTA window. All the remaining parameters can be either left with their default values or tuned by the user.

Each pipeline step is reported in the log window. After the process is completed a file suffixed with the string "_assembly.fasta" and containing the *de novo* assembled sequence is reported in the project folder.

The *de novo* assembly module can be tested using the human cytomegalovirus reads provided in the file HCMV_reads.fastq as both the input file and the gap closing file, and the HCMV reference genome (strain Merlin) provided in the HCMV_Reference_Genome.fasta file. Alternatively the file HCMV_HSV_mix_reads.fastq which contains approximately 20% of Human Simplex Virus 1 reads can be used. Since the human cytomegalovirus genome is used as a reference, only the HCMV reads will be mapped and extracted from the initial dataset prior the *de novo* assembly step. This results in the same assembled sequence previously produced from the dataset containing only HCMV reads.

LoReTTA_cl

This is the command line version of the *de novo* assembly module present in the LoReTTA GUI version. An help window can be invoked by typing:

```
./LoReTTA_cl -h
```

which produces the following output:

```
usage: LoReTTA_cl [-h] -i INPUTREADS -ref REFERENCE -o OUTPUTFOLDER -q QUALITY
                  -wsize WINDOWSIZE -wstep WINDOWSTEP -t THREADS -p PREFIX -cr
                  CLOSINGREADS -ho HOMOLOGY

Tool to perform reference guided de novo of PacBio reads

optional arguments:
  -h, --help            show this help message and exit
  -i INPUTREADS, --inputReads INPUTREADS
                        The Pacbio reads file in fastq format
  -ref REFERENCE, --reference REFERENCE
                        Reference file in fasta format
  -o OUTPUTFOLDER, --outputFolder OUTPUTFOLDER
                        The output folder name
  -q QUALITY, --quality QUALITY
                        Phred quality threshold
  -wsize WINDOWSIZE, --windowSize WINDOWSIZE
                        Window size
  -wstep WINDOWSTEP, --windowStep WINDOWSTEP
                        Window step
  -t THREADS, --threads THREADS
                        Number of threads
  -p PREFIX, --prefix PREFIX
                        Prefix of output files
  -cr CLOSINGREADS, --closingReads CLOSINGREADS
                        Fasta file with reads to be used to close the gaps
  -ho HOMOLOGY, --homology HOMOLOGY
                        The homology the target and the reference genome share
```

Please refer to the “User defined parameters” chapter for a description of the parameters that LoReTTA_cl can accept. Each parameter must be passed using the corresponding flags. As a way of example:

```
./LoReTTA_cl -i /testFiles/HCMV_reads.fastq
-ref ./testFiles/HCMV_Reference_Genome.fasta -o ../../testLoretta/ -q 30      -wsize
20000 -wstep 10000 -t 8 -p commandLineAssembly
-cr ./testFiles/HCMV_reads.fastq -ho 0.7
```