**PacBioRead**

A screenshot of a cell phone

Description automatically generated

**Aim**

This tool performs a reference guided de novo assembly. It splits the reference genome in windows of size “Window size”, aligns the reads to this portion, and assembles them. Then it moves a number of bases equal to the “window step”, gets another “window size” number of bases, assembles them and so on until the entire genome is assembled. These local sub-assemblies are joined together and, if present, gaps are closed.

**Read quality filtering**

The assembly is performed only using reads with a threshold quality equal to “Min. quality”. And mapping for at least the 70% to the reference genome. This filter may reduce the number of total reads present in the dataset. For this reason, the reads statistics are calculated both at before and after filtering. The “Min quality” is set to 30 and produced good results. When few reads are available this value may be lowered, but otherwise I would leave it to this value.

**How to install it**

Just clone the repository in your home:

*git clone https://github.com/salvocamiolo/PacBioRead.git*

Then enter the downloaded folder and type:

*bash install\_linux.sh*

This will install all the dependencies and you can run the tool just calling the executable PacBioRead

**How to use it**

Just select a project (e.g. output) folder, the Pacbio reads in fastq format (it is important that the phred quality values are reported) and the reference genome. Window size and window step were good in the tests on HCMV reads. You can change the number of threads to speed it up. Finally hit the “Run button” to launch the assembly

**What does it produce?**

The tool produces many files that are left there for debug reason. What we are interested in are the following:

local\_assemblies.fasta: the fasta file with the sequences assembled in each window

scaffolds.fasta: the fasta file with the contigs joined together via their overlaps (e.g. window step < window size)

scaffold\_gapClosed.fasta: Due to repetitive sequences, sometimes it is not easy to find an overlap between all the contigs and the entire genome can not be reconstructed in one single scaffold. An internal algorithm search for reads that span the gap between two not overlapping but proximal contigs. If the scaffolds.fasta already contain one single sequence with no gap, then the scaffold\_gapClosed.fasta will be just a copy of scaffolds.fasta

finalAssembly.fasta: this is the main outuput file we are interested on. This is a fasta file obtained by applying a number of corrections on the previous scaffold\_gapClosed.fasta and represents the best prediction PacBioReads can do of the genome.