Evaluation of long read error correction software

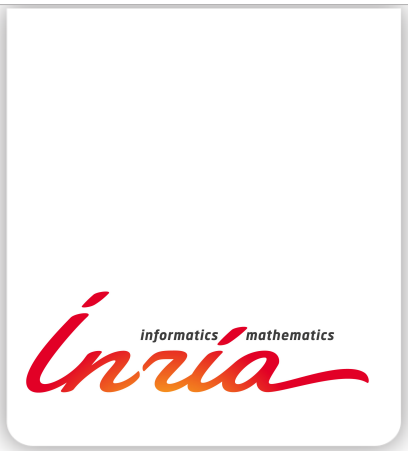
Laurent Bouri, Dominique Lavenier

Evaluation of long read error correction software

Laurent Bouri, Dominique Lavenier

RESEARCH REPORT
N° 9028
february 2017
Project-Team Genscale

ISSN 0249-6399 ISRN INRIA/RR--9028--FR+ENG



Evaluation of long read error correction software

Laurent Bouri*, Dominique Lavenier†

Project-Team Genscale

Research Report n° 9028 — february 2017 — 47 pages

Abstract: This report compares several read error correction software that use 3rd generation sequencing technology (long reads). The experimentations have been performed on several reference genomes and the results evaluated with QUAST. The long read error correctors that have been evaluated are: *LSC-2*, *Proovread*, *Ectools*, *Lordec*, *Nanocorr*, *Nas*, *Jabba*, *Pacbiotoca*, *Lorma* et *MHAP*. The first 8 software can merge long and short reads, while the last 3 software use only long reads.

Key-words: 3rd generation sequencing, long reads, correction

* CNRS Engineer/ France génomique
† CNRS Research Director, GenScale team leader

RESEARCH CENTRE
RENNES – BRETAGNE ATLANTIQUE

Campus universitaire de Beaulieu
35042 Rennes Cedex

Evaluation des logiciels de correction de lectures longues

Résumé : Ce rapport compare plusieurs programmes de correction d'erreurs de lectures (*reads*) issus de la technologie de séquençage de 3ème génération (longues lectures). Les expérimentations ont été menées sur plusieurs génomes de référence. Les logiciels de correction d'erreurs évalués sont : *LSC-2*, *Proovread*, *Ectools*, *Lordec*, *Nanocorr*, *Nas* , *Jabba*, *Pacbiotoca*, *Lorma* et *MHAP*. Les 8 premiers mixent longues et courtes lectures tandis que les 3 derniers n'utilisent que des longues lectures.

Mots-clés : séquençage de 3ème génération, lectures longues, correction

Contents

1 Introduction

1.1 Contexte

Third-technology sequencing brought by Pacbioscience or Oxford Nanopore is able to produce average long read length of more than 10,000 bp and thus can advantageously be used to improve the genome assembly. In fact, long reads span more repetitive elements and thus can produce more contiguous reconstruction of the genome. However, long reads have raw error rate ranging from 10% to 15%, requiring a preliminary stage of correction before the assembly process.

The available correction software are mainly based on two strategies :

- hybrid correction (both long reads and short reads are required) ;
- denovo correction (only long reads are required) ;

Hybrid correction uses short reads, such as Illumina, which have a much lower error rate, to correct long reads. The second strategy consists in aligning long reads against themselves.

1.2 Evaluated long read error correctors

1.2.1 Hybrid correctors (*Short and Long Reads*)

The evaluated hybrid correctors are listed in the table below. Some of them have the possibility of both recovering only the corrected regions ('trim' function) or the corrected and non-corrected regions ('untrim').

Correcteurs	trimmed reads	untrimmed reads
LSC 2 [al12a]	yes	yes
Pacbiotoca [al12b]	no	yes
Ectools [al14]	no	yes
Proovread [T14]	yes	yes
Lordec [Sal14]	yes	yes
Nanocorr [al15c]	no	yes
Nas [al15b]	no	yes
Jabba [al16a]	yes	no

1.2.2 De-novo correctors (*Long Read Only*)

The following de-novo correctors have been evaluated:

- Pacbiotoca [al12b]
- Lorma [al16b]
- MHAP (CANU) [al15a]

1.3 Method

1.3.1 Evaluation of long read error correction

In order to show the efficiency of long read error correction, several metrics reflecting long read quality can be calculated before and after assembly.

First, *BLASR* generates alignment of corrected long reads against the reference genome. From these alignments the following metrics are extracted:

- The average proportion of long read regions, aligned to the reference genome.
- The average number of match per long read and the percentage of identity of the aligned regions.
- The sum of corrected read length divided by the uncorrected read length. This gives an idea of the amount of sequences that has actually been corrected.

Then, the corrected long reads are assembled by *Smartdenovo*, a de-novo assembler giving good results, even for large genomes (> 100Mb) or with low coverage. Lastly, the QUAST (QUality ASsessment Tool) software [al13] evaluates the assembly by extracting several metrics:

- The number of contigs,
- Genome assembly length,
- N50,
- The fraction of the reference genome found among the contigs. This fraction is deduced by aligning the assembled genome against the reference genome using the MuMmer software [al99].

1.3.2 Datasets

The table below shows the 4 reference genomes and the datasets used for the various tests.

Genome	length (M bp)	Test	Minion	PacBio	Illumina
Acinetobacter	3.6 M	1	10x, 3.4K reads		211K reads
		2	20x, 10K reads		211K reads
E.Coli	4.6 M	3		10x, P4C2, 36K reads	16M reads
		4		100x, P4C2, 91K reads	16M reads
		5		10x, P6C4, 8.7K reads	16M reads
		6		100x, P6C4, 87K reads	16M reads
		7	20x, 22K reads		16M reads
S. Cerevisa	11.6 M	8		10x, P4C2, 26K reads	3.8M reads
		9		100x,P4C2, 261K reads	3.8M reads
		10	20x, 47K reads		3.8M reads
C. Elegans	100 M	11		10x, P6C4, 92K reads	55M reads
		12		100x, P6C4, 740K reads	55M reads

1.3.3 Hardware resources

Since most of the long read error correctors requires a large amount of hardware resources (CPU, RAM), their evaluation was done by submitting tasks to the GenOuest[1] platform cluster.

Correctors	Number of threads
LSC 2	8
PacbiotoCa	8
Ectools	1
Proovread	8
Lordec	8
Nanocorr	1
Nas	8
Jabba	8
MHAP	8
Lorma	8

Cluster node configuration:

- **Number of CPU**: 40
- **CPU frequency**: 2.6 GHz
- **Available RAM**: 256 GBytes

[1] GenOuest: https://www.genouest.org/

2 Hybrid correctors (*short and long reads*)

2.1 LSC-2

Introduction

LSC-2 is an hybrid corrector of long reads. Long reads and short reads are first compressed into homopolymers, then short reads are mapped to long reads with *Bowtie2*. Finally, the short read consensus replaces the long read sequences.

Website: `http://www.healthcare.uiowa.edu/labs/au/LSC/default.asp`

Installation

LSC-2 can be downloaded as pre-compiled binaries. Bowtie2 is required and must be installed.

Extraction of pre-compiled binaries files:

```
$ tar zxvf LSC-2.0.tar.gz
```

Input data

LSC-2 takes FASTA or FASTQ files as input.

Pipeline

runLSC script divides long read error correction into 5 steps:

- The sequences of long and the short reads are transformed by homopolymer compression so that each sequence of the same nucleotide is replaced by a single nucleotide of the same type.
- The short reads quality is checked. Indeed, some of these reads contain too much N letters or are too short.
- The short reads are aligned against the long ones with Bowtie2.
- The long reads are then modified according to the information provided by the short read consensus obtained with previous alignment.
- Once the correction points have been replaced by the corresponding short reads consensus, the rest of the compressed points are decompressed.

```
runLSC.py --long_reads LR.fa --short_reads SR.fa --specific_tempdir temp --output output_dir
```

- `longs_reads` : long read file.
- `short_reads` : short read file.
- `specific_tempdir` : folder containing temporary files (optional).
- `output` : final assembly folder.

Encountered errors

Died at /LSC-2.0/bin/../utilities/explode_fasta.pl
ValueError: invalid literal for int() with base 10: ”

solution : Convert short reads from FASTQ format to FASTA format (AWK command line or Biopython Seq.IO module).

[bam_header_read] EOF marker is absent. The input is probably truncated.

solution : Install Bowtie2.

Output data

The corrected sequences are written into the ”corrected_read.fasta” file, while full_LR.fasta file contains concatenate uncorrected terminus sequences and corrected sequences. Both files are located in the final assembly folder.

2.2 PacBioToCA

Introduction

Pacbiotoca is an hybrid and de-novo corrector, taking as input long reads in FASTQ file format and short reads in a frg file.

Website: `http://wgs-assembler.sourceforge.net/wiki/index.php/PacBioToCA`

Installation

PacBioToCA can be downloaded as source code or pre-compiled binaries.

Code source compilation:

```
$ bzip2 -dc wgs-8.3rc2.tar.bz2 | tar -xf
$ cd wgs-8.3rc2
$ cd kmer && make install && cd ..
$ cd src && make && cd ..
$ cd ..
```

Extraction of pre-compiled binaries files:

```
$ bzip2 -dc wgs-8.3rc2-*.tar.bz2 | tar -xf
```

Input data

PacBioToCA requires as input an estimation of the genome length, a configuration file, the long reads in a FASTQ file format and a fragments file "frg" containing short reads in the case of an hybrid correction. Programs such as **fastatoCA** and **fastqtoCA** are available in the PBcR package to perform conversions from common formats such as FASTA or FASTQ. Converting the FASTA file to .frg requires not only a file containing the sequences but also a quality value FASTA file.

```
fastaToCA -l libraryname -s seq.fasta -q qlt.fasta > seq.frg
```

- `l` : library name
- `s` : short reads in FASTA file format
- `q` : quality values FASTA file

```
fastqToCA -libraryname LIB -technology pacbio-corrected -reads seq.fastq > seq.frg
```

- `libraryname` : library name
- `technology` : type of data (pacbio, illumina, 454,...)
- `reads` : short reads in FASTQ file format

Pipeline

PacBioToCA is able to perform hybrid or de-novo correction. In order to launch a de-novo error correction, simply do not provide the .frg file (short reads).

```
./pacBioToCA -threads 8 -libraryname name -s spec_file.spec -genomeSize <genome length>
             -fastq Long_reads.fastq short_reads.frg
```

- `libraryname` : prefix used for naming output files
- `s` : file containing specific options for correction
- `genomeSize` : genome length
- `fastq` : long reads to correct in FASTQ file format.

Output data

The corrected reads are stored in the two output files libraryname.fasta and libraryname.fastq.

2.3 Ectools

Introduction

Ectools is an hybrid corrector, taking as input unitigs from a short read assembly, and uses them to correct long reads.

Website: `https://github.com/jgurtowski/ectools`

Installation

Ectools requires the prior installation of Mummer and blastn. *ECtools* comes in the form of series of python scripts located in the *Ectools* github page:

`https://github.com/jgurtowski/ectools`

Import the folder containing *ECtools* scripts:

```
$ git clone https://github.com/jgurtowski/ectools.git
$ cd ectools
```

Input data

Ectools needs a contig file generated from short reads as well as a file including long reads.

Pipeline

Set the variable `ECTOOLS_HOME` with the path for input data:

```
$ ECTOOLS_HOME=/path/to/this/directory
```

Create a folder for long read error correction:

```
$ mkdir organism_correct
```

Create symbolic link to contig file:

```
$ cd organism_correct
$ ln -s /path/to/organims_contigs.fasta
```

Ectools authors recommend keeping reads larger than 1Kb. Then, partition long reads:

```
$ python ${ECTOOLS_HOME}/partition.py 20 500 pbreads.length_filtered.fa
```

Copy the bash script for correction to the main folder:

```
$ cp ${ECTOOLS_HOME}/correct.sh .
```

Modify the global variables at the beginning of the correct.sh script and then run the next command line to correct long reads:

```
$ for i in {0001..000N}; do cd $i; qsub -cwd -j y -t 1:M ../correct.sh; cd ..; done
```

- `N` : number of folders created by the "partition.py" script
- `M` : number of files in each folder.

Several steps are then performed to correct the long reads:

- Alignment of Pacbio reads against contigs with Nucmer.
- Selection of a set of contigs that covers best each long read to be corrected
- The "show-snps" program identifies the nucleotides that differ between contigs and long reads.
- Correction of long reads (Python script)

Output data

The corrected sequence files are written in the specified output folder: "p000N.cor.fa". Finally, merge files with a bash command:

```
$ cat p00*.cor.fa > organism.cor.fa
```

2.4 proovread

Introduction

Proovread is a de novo corrector, using a de-Bruijn graph constructed from long reads.

Website : `https://github.com/BioInf-Wuerzburg/proovread`

Installation

Proovread is available on Linux and requires NCBI Blast-2.2.24+, samtools-1.1+, Perl 5.10.1+ and the perl moduls Log::Log4perl and File::Which.

Code source compilation:

```
$ git clone --recursive https://github.com/BioInf-Wuerzburg/proovread
$ cd proovread/util/bwa
$ make
```

Input data

In order to not overload the processor and the memory, it is wise to divide the long reads data (FASTA, FASTQ) into several files:

```
$ SeqChunker -s 20M -o 0%03d pacbio_file
```

- `s` : file length
- `o` : output file name

Pipeline

Run the long read error correction process with the binary "proovread", located in the folder "bin", for every folders created in the previous step:

```
$ for i in {0001..000n}; do proovread -l $i -s /Path/to/short_reads {pre pb_$i; done
```

- `n` : Number of files generated by SeqChunker
- `l` : raw noisy long reads
- `s` : file containing accurate short reads
- `pre` : prefix used to name output file.

It is also possible to add unitigs with the argument "-unitigs".

Proovread corrects long reads in 2 steps:

1. The mapping of short reads on long reads is done by SHRIMP2, by adapting the score mode to consider that insertions are more frequent than deletions and that substitutions are rare events. Bowtie2 and bwa mem are also supported.

2. A consensus sequence is computed from these alignments.

Output data

The corrected sequences are written in the specified output folder (trimmed and untrimmed reads)

2.5 Nanocorr

Introduction

Nanocorr is a hybrid corrector designed for nanopore long reads using blastn to align short reads on long reads.

Website: `https://github.com/jgurtowski/nanocorr`

Installation

Nanocorr has been designed to work in a SGE or similar environment (GNU parallel). Note that the python module "h5py" requires the installation of the hdf5 library

Import nanocorr:

```
$ git clone https://github.com/jgurtowski/nanocorr
$ cd nanocorr
```

Create a virtual environment to install python dependencies:

```
$ virtualenv nanocorr_ve
$ source nanocorr_ve/bin/activate
```

python dependencies installation:

```
$ pip install git+https://github.com/cython/cython
$ export HDF5_DIR=/<chemin vers la librairie hdf5>
$ export LD_LIBRARY_PATH=$HDF5_DIR/lib:$LD_LIBRARY_PATH
$ export PATH=$HDF5_DIR/bin:$PATH
$ pip install numpy
$ pip install h5py
$ pip install git+https://github.com/jgurtowski/pbcore_python
$ pip install git+https://github.com/jgurtowski/pbdagcon_python
$ pip install git+https://github.com/jgurtowski/jbio
$ pip install git+https://github.com/jgurtowski/jptools
```

Install nanocorr:

```
$ python setup.py install
```

Input data

Nanocorr is supposed to correct only long reads from the nanopore technology (FASTA or FASTQ file format).

Pipeline

First, partition long reads into multiple files:

```
$  python partition.py 100 500 nanopore_reads.fa
```

Then, run the "nanocorr.py" script in order to start long read error correction:

```
$ qsub -cwd -v PATH,LD_LIBRARY_PATH -t 1:500 -j 500 \
-o nanocorr_out /path/to/nanocorr.py query.fa reference.fa
```

- `t` : Number of files in the folder containing the partitioned long reads
- `j` : Declares whether the standard error stream of the task will be merged with the standard output stream of the same task

Nanocorr launches blast to align short reads on Nanopore long reads and then uses a dynamic programming algorithm based on the longest increasing subsequence problem to select sets of short reads corresponding to each long read. The consensus reads are then calculated using *pbdagcon*

Output data

The corrected sequences is divided into several files "p000N.blast6.r.fasta" and need to be concatenated in a single file:

```
$ cat p*.blast6.r.fasta > output_file.fasta
```

Encountered errors

c.fatal error: hdf5.h: No such file or directory virtualenv

solution : installation of hdf5 library

2.6 LoRDEC

Introduction

LoRDEC is an hybrid corrector, using a de-Bruijn graph constructed from short reads to correct long reads.

Website : `http://www.atgc-montpellier.fr/lordec/`

Installation

LoRDEC is available on linux and requires Cmake 2.6+ and GCC 4.7+.

Import *LoRDEC* and the GATB library (`http://gatb-core.gforge.inria.fr/`) :

```
$ wget http://www.atgc-montpellier.fr/download/sources/lordec/LoRDEC-0.6.tar.gz
$ tar zxvf LoRDEC-0.6.tar.gz
$ cd LoRDEC-0.6
$ wget https://github.com/GATB/gatb-core/releases/download/v1.1.0/ \
  gatb-core-1.1.0-bin-Linux.tar.gz
$ tar zxvf gatb-core-1.1.0-bin-Linux.tar.gz
```

Modify the variable GATB_VER from the Makefile (1.1.0) Install *LoRDEC*

```
$ make
$ cd ..
```

Input data

LoRDEC requires short reads in FASTA or FASTQ file format and long reads in FASTA or FASTQ file format.

Pipeline

Run the long read error correction with the binary "lordec-correct":

```
$ lordec-correct -2 illumina.fasta -k 19 -s 3 -i pacbio.fasta \
  -o pacbio-corrected.fasta
```

- `2` : File of short reads.
- `k` : Size of the kmer used in the de-Bruijn graph
- `s` : Abundance threshold of a kmer to be considered correct
- `i` : Input file
- `o` : Output file

A series of steps is then performed in order to correct the long reads:

1. Construction of a de-Bruijn graph from the short reads
2. Suppression of k-mer with occurrence less than the s value
3. Choose an optimal path of the graph by calculating the edit distance between the path and a region of long read.

Output data

The corrected sequences will be in the output file indicated after the "-o" parameter. The output file in FASTA format contains long reads. Corrected sequences are defined by uppercase letters while uncorrected sequences appears as lowercase letters. *Lordec* offers the possibility to remove the uncorrected sequences at the beginning and at the end of the long reading or to keep only the corrected sequences.

```
$ lordec-trim -i fichier_reads.fasta -o fichier_trim.fasta
```

- `i` : corrected reads file
- `o` : output file

```
$ loredec-trim-split -i fichier_reads.fasta -o fichier_trim_split.fasta
```

- `i` : corrected reads file
- `o` : output file

2.7 Nas

Introduction

Nas adopts an hybrid approach to correct long reads from nanopore technology. The mapping of illumina sequences on nanopore reads will give long reads without errors called NaS (Nanopore Synthetic-long).

Website : `https://github.com/institut-de-genomique/NaS`

Installation

Nas is designed to run on linux or on grid environments (SGE,...). It can also be installed in a docker container (http://registry.hub.docker.com/u/rdbioseq/nas/). *Nas* requires the following dependencies:

- Shell tool GNU Parallel (http://www.gnu.org/software/parallel/) at least (22052015)
- Perl 5.8.0 or higher
- Perl graph library (http://search.cpan.org/ jhi/Graph/)
- Perl GetOpt module (http://search.cpan.org/dist/Getopt-Long/)
- Perl Set::IntSpan module (http://search.cpan.org/ swmcd/Set-IntSpan-1.19/IntSpan.pm)
- Newbler assembler v2.9 (available from 454 website)
- Blat binary (at least v35) accessible through your PATH environnment variable
- Last binary (at least 588) accessible through your PATH environnment variable

-*Nas* Installation:

```
$ git clone https://github.com/institut-de-genomique/NaS.git
```

Specify the paths to Last and Blast in the "NaS" binary located in the "NaSv2" folder

Input data

Nas is able to correct reads smaller than 60kb and whose genome is smaller than 30Mb

Pipeline

Run long read error correction:

```
$ ./NaS --fq1 short_reads1.fastq --fq2 short_reads2.fastq --nano Longs_reads.fa \
   --out NaS_output --mode sensitive --nb_proc 5
```

- `nano` : input file containing long reads
- `mode` : mapper used (fast=blat; sensitive=last)
- `nb_proc` : number of processor

Nas corrects long reads in two steps:

1. First, alignments are made between short reads and long reads (BLAT (fast mode) or LAST (sensitive mode)) in order to identify "seed-reads". The number of these "seed-reads" is then extended by looking for similar short reads in the initial set using comparisons.
2. Then, for each long reads, a microassembly is made from the short reads recruited using an Overlap-Layout-Consensus (Newbler) strategy.

Output data

The corrected sequences are written into the specified output folder, contained in the "NaS_hqctg_reads_final.fa" file.

2.8 Jabba

Introduction

Jabba is an hybrid corrector, taking as input a de-Bruijn graph resulting from an assembly of short reads and uses it to correct long reads.

Website : `https://github.com/gmiclotte/Jabba`

Installation

Jabba is available for Linux and requires Cmake 2.6+ et GCC 4.7+. It is possible to install *Karect* for the step of short read correction and *Brownie* for the construction of the de-Bruijn graph. Note that Jabba needs sparsehash to be installed.

Install karect:

```
$ git clone https://github.com/aminallam/karect.git
$ cd karect
$ make
$ cd ..
```

Install sparsehash

```
$ git clone https://github.com/sparsehash/sparsehash.git
$ ./configure {prefix=<chemin vers le dossiersparsehas>
$ make
```

Install brownie:

```
$ git clone https://github.com/jfostier/brownie.git
$ mkdir build
$ mkdir bin
$ cd build
$ cmake -DSPARSEHASH_INCLUDE_DIR=<chemin vers le dossiersparsehas/include>
```

Install sparsehash

```
$ git clone https://github.com/sparsehash/sparsehash.git
$ ./configure {prefix=<chemin vers le dossiersparsehas> -DMAXKMERLENGTH=75 ..
$ make brownie
$ cp src/brownie ../bin/brownie
$ cd ..
```

Installer Jabba:

```
$ git clone https://github.com/gmiclotte/Jabba.git
$ cd Jabba
$ ./compile.sh
$ cd ..
```

Input data

Jabba takes long reads in a FASTA or FASTQ file format and the de-Bruijn graph constructed by Brownie from the short reads, themselves corrected by a tool like *Karect*.

Pipeline

First, the authors recommend to start the process with *karect*:

```
$ mkdir karect_output
$ ./karect/karect -correct -matchtype=hamming -celltype=haploid \
  -inputfile=short_reads.fastq -resultdir=karect_output -tempdir=karect_output
```

Then run long read error correction as follow:

```
$ /Jabba/bin/jabba -o jabba_output -k 75 -brownie_output/DBGraph.fasta \
  -fastq long_reads.fastq
```

- `l` : minimum seed length
- `k` : kmer length used in Debruijn graph.

However, for each new tested dataset and the construction of a new de-Bruijn graph, the corrected reads must be redirected to a new output file. *Jabba* corrects the long reads by aligning them on a de-Bruijn graph using a "seed and extend" method where the "maximum exact matches" represent the seeds.

Output dataset

The corrected sequences will be in the specified output folder, contained in a "Jabba-input_filename.fasta" file.

3 Denovo correctors (*Long Read only*)

3.1 LoRMA

Introduction

LoRMA is a de novo corrector using a de-Bruijn graph constructed from long reads to correct themselves.

Website : `https://www.cs.helsinki.fi/u/lmsalmel/LoRMA/`

Installation

LoRMA is available on Linux and requires Cmake 2.6+, GCC 4.7+ and Lordec.

Import *LoRDEC* and the GATB library:

```
$ wget http://www.atgc-montpellier.fr/download/sources/lordec/LoRDEC-0.6.tar.gz
$ tar zxvf LoRDEC-0.6.tar.gz
$ cd LoRDEC-0.6
$ wget https://github.com/GATB/gatb-core/releases/download/v1.1.0/ \
  gatb-core-1.1.0-bin-Linux.tar.gz
$ tar zxvf gatb-core-1.1.0-bin-Linux.tar.gz
```

Modify the GATB_VER variable from Makefile (1.1.0 here)

Install *LoRDEC*

```
$ make
$ cd ..
```

Install *loRMA*

```
$ wget https://www.cs.helsinki.fi/u/lmsalmel/LoRMA/LoRMA-0.3.tar.gz
$ tar zxvf LoRMA-0.3.tar.gz
$ cd LoRMA-0.3
$ mkdir build
$ cd build
$ cmake ..
$ make
```

Change the LORDECDIR variable in the lorma.sh file to specify the path to the *LoRDEC* installation folder.

Input data

LoRMA takes long reads in a FASTA or FASTQ file format as input.

Pipeline

LoRMA software correct long reads in two steps:

1. Iterative construction of a Dde-Bruijn graph from long reads, with an increasing value of the kmer size. Each node in the graph represents a genomic kmer, selected according to its abundance in the dataset of long reads. For each long read, identify possible paths in the graph between the solid kmer and then select the best path. Correct the regions of the long read.
2. Finally, use the graph to identify similar reads. Align these reads in order to build a consensus.

Output data

The corrected sequences are located in a file named "final.fasta"

3.2 MHAP (CANU)

Introduction

CANU includes a de novo correction module, using *MHAP* to align long reads against themselves and *pbdagcon* to correct the long reads by a consensus step.

Website : `https://github.com/marbl/canu`

Installation

Canu is designed to run on grid environments (LSF/PBS/Torque/Slrum/SGE are supported) and requires java 1.8 and gnuplot.

Install *Canu* :

```
$ git clone https://github.com/marbl/canu.git
$ cd canu/src
$ make -j <nombre de threads>
```

Input data

Canu Requires only long reads in FASTA or FASTQ format.

Pipeline

Run the long read error correction with the "canu" binary, located in the folder "canu/Linuxamd64/bin":

```
$ ./canu -correct -p ecoli -d ecoli genomeSize=4.8m -pacbio-raw longs-reads.fastq
```

- `genomeSize` : Genome size
- `pacbio-raw` : Type of long reads (pacbio-raw or naopore-raw)

It is also possible to isolate the corrected sequences with the "canu" binay:

```
$ ./canu -trim -p ecoli -d ecoli genomeSize=4.8m -pacbio-corrected corrected-reads.fastq
```

Canu corrects long reads in two steps:

1. At first, *MHAP* finds inaccurate overlaps between long reads.
2. *Pbdagcon* then builds a consensus sequence from these overlaps.

Encountered errors

canu failed with 'can't configure for SGE'

solution : Add the following arguments in the command line: gridEngineMemoryOption="-l h_vmem=MEMORY" gridEngineThreadsOption="-pe make THREADS"

Output data

The corrected sequences will be stored in multiple FASTA files concatenated in the output folder: "correction/2-correction/correction_outputs"

4 Evaluation of correction and assembly

4.1 Tested genomes

Genome name	Number of chromosome	genome size
Acinetobacter DP1	1 chromosome	3 976747 pb
Escherichia Coli K12 MG1655	1 chromosome	4 641652 pb
Saccharomyces Cerevisae W303	16 chromosomes	11 633571 pb
Caenorhabditis elegans	6 chromosomes	100 272607 pb

4.2 Evaluation methods

Long reads correction will be evaluated before and after assembly. First, a series of information is extracted either directly from the corrected long reads or from alignments performed with *BLASR* between long reads and the reference genome:

- **Size** : The sum of the length of the corrected long reads, divided by the length of uncorrected long reads.
- **Execution time** : Corrector execution time.
- **% mapped region** : Proportion of corrected reads that *BLASR* succeeded in mapping to the reference genome.
- **Mean number of match** : The average number of match per long read.
- **% identity** : Percentage of identity given by *BLASR* for the region of long read that could be aligned.

Then, the reference genomes allow to evaluate the quality of the different assemblies, based on the results of the metrics produced by the **QUAST** software. The different metrics are listed below:

- **# contigs (> 1000pb)** : Total number of contigs exceeding 1000 bp after assembly.
- **Total length** : Assembly length.
- **N50** : Length for which the collection of all contigs of this length or greater, covers at least half of the assembly.
- **Genome fraction (%)** : the percentage of aligned bases in the reference genome.

4.3 Datasets

Acinetobacter sp. adp1, Illumina: Internal data

Acinetobacter sp. adp1, Minion run5: http://www.genoscope.cns.fr/externe/nas/datasets/MinION/acineto/acineto_nanopore_2D_run5.fa.gz

Acinetobacter sp. adp1, Minion run6: http://www.genoscope.cns.fr/externe/nas/datasets/MinION/acineto/acineto_nanopore_2D_run6.fa.gz

Escherichia coli k-12, Illumina: ftp://webdata:webdata@ussd-ftp.illumina.com/Data/SequencingRuns/MG1655/MiSeq_Ecoli_MG1655_110721_PF_R1.fastq.gz
ftp://webdata:webdata@ussd-ftp.illumina.com/Data/SequencingRuns/MG1655/MiSeq_Ecoli_MG1655_110721_PF_R2.fastq.gz

Escherichia coli k-12, Minion: http://www.genoscope.cns.fr/externe/nas/datasets/MinION/ecoli/Ecoli_LomanAll2D.fa.gz

Escherichia coli k-12, Pacbio (p4c2): http://sourceforge.net/projects/wgs-assembler/files/wgs-assembler/wgs-8.0/datasets/escherichia_coli_k12_mg1655.m130404_014004_sidney_c100506902550000001823076808221337_s1_p0.1.fastq.xz
http://sourceforge.net/projects/wgs-assembler/files/wgs-assembler/wgs-8.0/datasets/escherichia_coli_k12_mg1655.m130404_014004_sidney_c100506902550000001823076808221337_s1_p0.2.fastq.xz
http://sourceforge.net/projects/wgs-assembler/files/wgs-assembler/wgs-8.0/datasets/escherichia_coli_k12_mg1655.m130404_014004_sidney_c100506902550000001823076808221337_s1_p0.3.fastq.xz

Escherichia coli k-12 Pacbio (p6c4): https://github.com/PacificBiosciences/DevNet/wiki/E.-coli-Bacterial-Assembly

Saccharomyces cerevisiae W303, Illumina: Accession number: SRR567755

Saccharomyces cerevisiae W303, Minion: http://www.genoscope.cns.fr/externe/nas/datasets/MinION/yeast/W303_ONT_Raw_reads_2D.fa.gz

Saccharomyces cerevisiae W303, Pacbio (p4c2): https://github.com/PacificBiosciences/DevNet/wiki/Saccharomyces-cerevisiae-W303-Assembly-Contigs

Caenorhabditis elegans, Illumina: Accession numbers: SRR065388, SRR065389, SRR065390

Caenorhabditis elegans, Pacbio (p6c4): http://datasets.pacb.com.s3.amazonaws.com/2014/c_elegans/list.html

4.4 Test 1 : Acinetobacter sp, ADP1, run5; Minion 10x

Dataset:

- Oxford NanoPort reads (MinIon) corrected by Lordec (10x coverage) : 3427 reads
- Illumina reads (MiSeq) : 211219 reads of length 2x150pb
- Contigs generated by sparse assembler from Illumina reads (233 contigs)

Hybrid correctors

Metrics	LSC-2		PacBioToCA	Ectools	Proovread	
	trim	untrim			trim	untrim
Size	79.14	88.04	16.42	27.04	40.39	101.02
Execution time	5 342s	5 342s	1 084s	90s	5 214s	5214s
% mapped region	99.11	94.76	99.25	99.7	92.69	34
Mean number of match	10 135	10 731	1 904	12 417	6 442	11 239
% identity	87.49	86.68	93.07	99.6	99.93	98.37
contigs (>=1000pb)	29	34	7	17	20	18
N50	153 908	120 006	19 707	338 569	50 752	315 675
Total length	3 267 123	2 668 193	153 749	3 560 442	879 221	3 446 093
Genome fraction	15.19	13.62	0	96.75	24.42	92.54

Metrics	LoRDEC		Nanocorr	Nas	Jabba
	trim	untrim			
Size	79.91	101.82	72.53	96.1	82.9
Execution time	521s	521s	1 749s	22 500s	24.72s
% mapped region	99.81	91	99.33	99.45	88.8
Mean number of match	6 607	11 038	9 668	13 320	8 658
% identity	99.9	97.13	97.06	99.93	99.9
contigs (>=1000pb)	29	26	27	20	39
N50	178 054	164 126	127 495	211 411	90 115
Total length	3 369 809	3 341 599	3 329 760	3 203 245	2 809 569
Genome fraction	92.87	89.02	87.31	88.76	78.08

de-novo correctors

Metrics	PacBioToCA	MHAP (CANU)	LoRMA
Size	30.44	33.36	0.37
Execution time	1 037	120	243.84
% mapped region	99.25	99.6	99.6
Mean number of match	1 869	6 381	174
% identity	93.10	98.83	98.58
contigs (>=1000pb)	6	23	no data
N50	23 350	29 225	no data
Total length	132 365	628 069	no data
Genome fraction	0	0	no data

4.5 Test 2 : Acinetobacter sp, ADP1, run6; Minion 20x

Dataset:

- Oxford NanoPort reads (MinIon), corrected by lordec (20x coverage) : 10116 reads
- Illumina reads (MiSeq) : 211219 reads of length length 2x150pb
- Contigs generated by sparse assembler from Illumina reads (233 contigs)

Hybrid correctors

Metrics	LSC-2		PacBioToCA	Ectools	Proovread	
	trim	untrim			trim	untrim
Size	81.75	90.39	62.25	81.30	32.07	83.84
Execution time	8 935s	8 935s	4 088	120s	7 087s	7 087s
% mapped region	99.15	95.45	99.63	99.7	99.8	97.05
Mean number of match	10 005	10 561	1 227	12 436	4 559	11 831
% identity	88.73	87.88	99.93	99.6	99.95	98.99
contigs (>=1000pb)	1	2	47	1	29	1
N50	3 605 124	2 415 855	85 022	3 649 197	175 774	3 631 367
Total length	3 605 124	3 598 621	2 630 061	3 649 197	3 063 864	3 631 367
Genome fraction	25.82	28.88	73.07	98.79	84.98	97.21

Metrics	LoRDEC		Nanocorr	Nas	Jabba
	trim	untrim			
Size	82.28	102.02	76.14	98.33	84.4
Execution time	661s	661s	2 520s	91 920s	29s
% mapped region	99.82	92.51	99.32	99.32	99.8
Mean number of match	6 575	10 919	9 813	12 808	8 658
% identity	99.92	97.20	97.39	99.94	99.9
contigs (>=1000pb)	4	1	1	32	28
N50	1 113 078	3 622 997	3 647 606	116 593	110 934
Total length	3 601 490	3 622 997	3 647 606	3 230 285	2 648 543
Genome fraction	99.99	97.22	94	89.69	73.59

de-novo correctors

Metrics	PacBioToCA	MHAP (CANU)	LoRMA
Size	68.47	54.01	21.1
Execution time	6 745s	540s	670s
% mapped region	99.15	99.67	99.6
Mean number of match	7 297	8 631	189
% identity	93.10	99.49	98.13
contigs (>=1000pb)	14	16	no data
N50	2 020 356	1 892 568	no data
Total length	3 249 142	3 230 123	no data
Genome fraction	1.3	7.93	no data

4.6 Test 3 : Escherichia coli k-12, reads Pacbio 10x (P4-C2)

Dataset:

- Pacbio reads corrected by lordec (10x coverage) : 36355 reads
- Illumina reads (MiSeq) : 16759877 reads of length 2x150pb
- Contigs generated by sparse assembler from illumina reads (1876792 contigs)

hybrid correctors

Metrics	LSC-2		PacBioToCA	Ectools	Proovread	
	trim	untrim			trim	untrim
Size	86.11	95.40	80.38	80.39	79.18	106.9
Execution time	31 768s	31 768s	17 298s	160s	2 791s	2 791s
% mapped region	94.93	90.23	98.01	99.6	98.51	91.04
Mean number of match	4 266	4 433	1 596	6 188	3 634	4 631
% identity	95.47	94.35	99.85	99.3	99.82	97.64
contigs (>=1000pb)	71	66	40	61	65	77
N50	41 815	42 879	25 658	35 761	32 061	44 775
Total length	2 723 468	2 526 333	1 000 670	1 935 279	2 069 936	3 005 248
Genome fraction	17.56	50.05	21.55	41.52	44.40	63.44

Metrics	LoRDEC		Nanocorr	Nas	Jabba
	trim	untrim			
Size	73.24	94.67	80.95	–	80
Execution time	2 392s	2 392s	15 480s	–	267s
% mapped region	99.10	93.16	96.53	–	99.9
Mean number of match	1 034	2 517	4 150	–	4 803
% identity	99.53	94.82	97.95	–	99.9
contigs (>=1000pb)	1	90	67	–	77
N50	12 447	33 149	44 366	–	46 908
Total length	12 447	2 787 937	2 573 103	–	3 155 083
Genome fraction	0.26	53.04	53.78	–	68

The lack of correction by the *Nas* tool is due to an incompatibility with the illumina data.

de-novo correctors

Metrics	PacBioToCA	MHAP (CANU)	LoRMA
Size	36.54	32.07	1.04
Execution time	6 396s	2 280s	218s
% mapped region	99.66	99.6	99.6
Mean number of match	1973	3 898	335
% identity	99.03	99.72	99.8
contigs (>=1000pb)	10	20	no data
N50	19 702	22 729	no data
Total length	195 674	429 222	no data
Genome fraction	4.22	9.28	no data

4.7 Test 4 : Escherichia coli k-12, reads Pacbio 100x (P4-C2)

Dataset:

- Pacbio reads corrected by lordec (100x coverage) : 91394 reads
- Illumina reads (MiSeq) : 16759877 reads of length 2x150pb
- Contigs generated by sparse assembler Illumina reads (1876792 contigs)

hybrid correctors

Metrics	LSC-2		PacBioToCA	Ectools	Proovread	
	trim	untrim			trim	untrim
Size	85.12	95.91	29.79	55.5	81.26	94
Execution time	215 528s	215 528s	53 572s	312s	72 420s	72 420s
% mapped region	94.17	88.62	97.57	99.62	98.37	89.69
Mean number of match	4 124	4306	1 592	6 130	3 431	4 528
% identity	95.01	93.79	99.84	99.31	99.82	97.10
contigs (>=1000pb)	1	1	19	22	123.81	1
N50	4 683 796	4 692 163	406 673	692 587	4 653 196	4 665 748
Total length	4 683 796	4 692 163	4 608 578	4 523 355	4 653 196	4 665 748
Genome fraction	96.21	94.58	99.21	97.03	99.93	98.84

Metrics	LoRDEC		Nanocorr	Nas	Jabba
	trim	untrim			
Size	78.97	96.48	79.03	–	77.7
Execution time	6 097s	6 097s	54 780s	–	296s
% mapped region	99.14	93.16	95.98	–	99.9
Mean number of match	2 188	4 250	3 987	–	4 752
% identity	99.79	96.14	97.75	–	99.99
contigs (>=1000pb)	28	1	1	–	47
N50	288 295	4 682 708	4 675 416	–	112 710
Total length	4 564 316	4 682 708	4 675 416	–	4 471 057
Genome fraction	98.23	96.69	97.87	–	96

The lack of correction by the *Nas* tool is due to an incompatibility with the illumina data.

de-novo correctors

Metrics	PacBioToCA	MHAP (CANU)	LoRMA
Size	38.63	29.55	79.6
Execution time	3 219s	3 180s	7 524s
% mapped region	99.19	96.01	99.7
Mean number of match	5 994	7 172	1 609
% identity	98.47	99.84	99.5
contigs (>=1000pb)	1	1	70
N50	4 624 650	22 729	86 784
Total length	4 624 650	4 647 081	4 109 490
Genome fraction	99.96	9.28	88.3

4.8 Test 5 : Escherichia coli k-12, reads Pacbio 10x (P6-C4)

Dataset:

- Pacbio reads corrected by lordec (couverture 10x) : 8746 reads
- Illumina reads (MiSeq) : 16759877 reads of length 2x150pb
- Contigs generated by sparse assembler from Illumina reads(1876792 contigs)

hybrid correctors

Metrics	LSC-2		PacBioToCA	Ectools	Proovread	
	trim	untrim			trim	untrim
Size	91.70	94.60	87.2	77.88	75.06	79.32
Execution time	110 418s	110 418s	45 542s	60s	12 840s	12 840s
% mapped region	97.23	95.03	98.34	99.53	99.34	93.56
Mean number of match	7 471	7569	2 756	10 144	7 054	9 121
% identity	97.30	96.93	99.67	99.2	99.85	98.96
contigs (>=1000pb)	6	7	–	20	7	7
N50	1491698	1 308 005	No data	525 979	1 486 401	1 259 739
Total length	4 676 830	4 677 980	No data	4 478 726	4 639 996	4 663 929
Genome fraction	97.15	96.2	No data	95.98	99.38	99.17

Metrics	LoRDEC		Nanocorr	Nas	Jabba
	trim	untrim			
Size	84.88	96.62	67.31	–	86.4
Execution time	603s	603s	20 460s	–	274s
% mapped region	98.84	93.64	96.87	–	99.8
Mean number of match	1 976	7 534	5 942	–	7 466
% identity	99.58	97.27	98.37	–	99.99
contigs (>=1000pb)	79	9	34	–	48
N50	31 940	1 207 236	215 605	–	99 142
Total length	2 323 363	4 650 012	4 443 229	–	4 103 224
Genome fraction	50.03	94.86	92.40	–	88.4

The lack of correction by the *Nas* tool is due to an incompatibility with the illumina data.

de-novo correctors

Metrics	PacBioToCA	MHAP (CANU)	LoRMA
Size	76.04	67.9	5.3
Execution time	2 738s	2 340s	323s
% mapped region	99.53	99.3	99.7
Mean number of match	5 336	7 396	283
% identity	99.18	99.3	99.68
contigs (>=1000pb)	57	50	–
N50	66 797	100 017	–
Total length	3 318 200	3 710 724	–
Genome fraction	71.62	80.13	–

The lack of results after the assembly of long reads corrected by *loRMA* is due to a low quantity of available long reads after the correction stage.

4.9 Test 6 : Escherichia coli k-12, reads Pacbio 100x (P6-C4)

Dataset:

- Pacbio reads corrected by lordec (100x coverage) : 87497 reads
- Illumina reads (MiSeq) : 16759877 reads of length 2x150pb
- Contigs generated by sparse assembler from Illumina reads (1876792 contigs)

Hybrid correctors

Metrics	LSC-2		PacBioToCA	Ectools	Proovread	
	trim	untrim			trim	untrim
Size	90.45	94.11	21.29	75.22	70.05	80.79
Execution time	8 634 43s	8 634 43s	69 305s	317s	161 940s	161 940s
% mapped region	96.61	93.94	97.64	99.6	99.16	94.02
Mean number of match	4 591	7711	3 391	9 824	6 580	9 221
% identity	96.68	96.23	99.88	99.2	99.83	99.08
contigs (>=1000pb)	47	21	15	12	1	1
N50	120585	368150	437 223	693 068	4 661 248	4 686 321
Total length	4 252 602	4 789 038	4 640 384	4 532 237	4 661 248	4 686 321
Genome fraction	82.98	95.94	99.42	97.18	99.98	99.30

Metrics	LoRDEC		Nanocorr	Nas	Jabba
	trim	untrim			
Size	82.72	96.43	67.55	–	84.1
Execution time	2 846s	2 846s	69 147s	–	304s
% mapped region	98.87	96.43	96.85	–	99.9
Mean number of match	1 896	7 664	6 184	–	7 670
% identity	99.57	96.78	98.04	–	99.9
contigs (>=1000pb)	39	12	1	–	44
N50	197 117	535 879	4 721 863	–	112 509
Total length	4 535 627	4 840 284	4 721 863	–	4 314 638
Genome fraction	97.54	95.90	96.06	–	92.9

The lack of correction of the *Nas* tool is due to an incompatibility of the illumina data.

de-novo correctors

Metrics	PacBioToCA	MHAP (CANU)	LoRMA
Size	30.20	21.16	83.9
Execution time	5 120s	7 380s	18 923s
% mapped region	97.68	99.6	99.8
Mean number of match	12 387	6 381	1 351
% identity	98.60	98.8	99.71
contigs (>=1000pb)	1	1	27
N50	4 646 758	4 646 370	306 808
Total length	4 646 758	4 646 370	4 597 457
Genome fraction	99.97	99.99	98.9

4.10 Test 7 : Escherichia coli k-12, reads Minion 20x

Dataset:

- Minion reads corrected by lordec (20x coverage) : 22270 reads
- Illumina reads (MiSeq) : 16759877 reads of length 2x150pb
- Contigs generated by sparse assembler from Illumina reads(1876792 contigs)

hybrid correctors

Metrics	LSC-2		PacBioToCA	Ectools	Proovread	
	trim	untrim			trim	untrim
Size	88.49	99.87	90.34	80.39	94.3	100.9
Execution time	35 918s	35 918s	40 138s	160s	3 365s	22 140s
% mapped region	99.16	95.76	98.39	99.9	99.78	95.47
Mean number of match	5 428	5933	1 749	7 838	4 277	5 881
% identity	88.35	86.86	99.79	99.8	99.91	97.96
contigs (>=1000pb)	5	6	36	25	2	1
N50	1 313 431	1 083 813	172 755	210 664	4 526 234	4 671 320
Total length	4 645 507	4 699 525	4 510 878	4 474 215	4 661 072	4 671 320
Genome fraction	17.56	17.29	97.20	94.97	99.97	96.31

Metrics	LoRDEC		Nanocorr	Nas	Jabba
	trim	untrim			
Size	87.54	101.08	92.60	–	94.8
Execution time	406s	406s	34 200s	–	272s
% mapped region	99.26	94.67	98.46	–	99.9
Mean number of match	1 719	5 763	5 505	–	5 699
% identity	99.63	95.14	96.49	–	99.9
contigs (>=1000pb)	114	2	1	–	51
N50	32 904	4 650 531	4 673 315	–	105 676
Total length	3 246 379	4 707 245	4 673 315	–	4 423 700
Genome fraction	69.91	91.15	93.60	–	95.3

The lack of correction of the *Nas* tool is due to an incompatibility of the illumina data.

de-novo correctors

Metrics	PacBioToCA	MHAP (CANU)	LoRMA
Size	78.96	76.99	13.4
Execution time	6 396s	720s	238s
% mapped region	99.21	99.69	99.6
Mean number of match	4 923	5 577	164
% identity	92.76	99.24	98.53
contigs (>=1000pb)	9	1	–
N50	642 915	4 396 819	–
Total length	4 321 437	4 396 819	–
Genome fraction	0.26	0.32	–

The lack of results after the assembly of long reads corrected by *loRMA* is due to a low quantity of available long reads after the correction stage.

4.11 Test 8 : Saccharomyces cerevisiae W303, reads Pacbio 10x (P4-C2)

Dataset:

- Pacbio reads corrected by Lordec (10x coverage) : 26196 reads
- Illumina reads (HiSeq) : 3815678 reads of length 2x100pb
- Contigs generated by sparse assembler from Illumina reads (10055 contigs)

hybrid correctors

Metrics	LSC-2		PacBioToCA	Ectools	Proovread	
	trim	untrim			trim	untrim
Size	40.46	92.41	21.49	51.98	44.23	99.45
Execution time	77 167s	77 167s	11 172	147s	28 260s	28 260s
% mapped region	83.06	76.99	97.83	95.5	86.74	91.04
Mean number of match	4 154	4533	639	6 904	2 187	4 631
% identity	92.45	91.42	99.6	99.4	99.40	97.64
contigs (>=1000pb)	175	169	–	173	79	173
N50	52 123	48 098	–	51 046	25 102	66 148
Total length	7 677 266	7 234 422	–	7 828 809	1 908 619	8 766 345
Genome fraction	19.81	47.53	–	66.49	16	69.16

Metrics	LoRDEC		Nanocorr	Nas	Jabba
	trim	untrim			
Size	68.65	99.91	42.41	–	9.2
Execution time	1 546s	1 546s	8 932s	–	122s
% mapped region	90.91	76	91.23	–	99.1
Mean number of match	1 830	4 497	2 842	–	574
% identity	99.44	95.72	97.25	–	99.9
contigs (>=1000pb)	87	174	70	–	–
N50	20 734	60 876	24 708	–	–
Total length	1 820 797	11 633 571	1 744 892	–	–
Genome fraction	15.07	66.95	13.89	–	–

The lack of correction of the *Nas* tool is due to an incompatibility of the illumina data. The lack of results after the assembly of long reads corrected by *PacBioToCA* ou *Jabba* is due to a low quantity of available long reads after the correction stage.

de-novo correctors

Metrics	PacBioToCA	MHAP (CANU)	LoRMA
Size	16.87	13.93	6.17
Execution time	3 039s	6 660s	849s
% mapped region	93.54	81.14	60.4
Mean number of match	1562	3 553	202
% identity	92.76	97.9	93.17
contigs (>=1000pb)	6	7	no data
N50	17 647	24 044	no data
Total length	91 430	158 844	no data
Genome fraction	0.74	1.21	no data

The lack of results after the assembly of long reads corrected by *loRMA* is due to a low quantity of available long reads after the correction stage.

4.12 Test 9 : Saccharomyces cerevisiae W303, reads Pacbio 100x (P4-C2)

Dataset:

- Pacbio reads corrected by lordec (100x coverage) : 261964 reads
- Illumina reads (HiSeq) : 3815678 reads of length 2x100pb
- Contigs generated by sparse assembler from Illumina reads (10055 contigs)

hybrid correctors

Metrics	LSC-2		PacBioToCA	Ectools	Proovread	
	trim	untrim			trim	untrim
Size	83.24	93.84	15.53	55.70	49.56	96.94
Execution time	943 079s	943 079s	33 179s	1 325s	200 040s	200 040s
% mapped region	85.33	80.45	97.46	95.35	85.97	79.78
Mean number of match	4 287	4531	802	6 719	2 402	4 441
% identity	93.74	92.82	99.64	99.47	99.40	96.07
contigs (>=1000pb)	23	21	236	60	71	20
N50	780551	781 007	40 462	340 500	135 059	818 767
Total length	1 2182 386	12 181 497	8 340 302	11 628 184	7 019 792	12 178 144
Genome fraction	89.69	86.40	57.66	96.74	57.66	95.05

Metrics	LoRDEC		Nanocorr	Nas	Jabba
	trim	untrim			
Size	76.77	97.66	87.31	–	11.6
Execution time	31 024s	31 024s	143 100s	–	213s
% mapped region	90.04	78.84	90.35	–	98.9
Mean number of match	1 869	4 413	3 227	–	576
% identity	99.41	96.39	97.54	–	99.9
contigs (>=1000pb)	183	20	24	–	–
N50	87 654	820 758	755 939	–	–
Total length	11 479 142	12 209 277	12 281 380	–	–
Genome fraction	92.70	95.26	97.32	–	–

The lack of correction of the *Nas* tool is due to an incompatibility of the illumina data.

de-novo correctors

Metrics	PacBioToCA	MHAP (CANU)	LoRMA
Size	29.21	23.9	70.3
Execution time	10 916s	9 360s	4 1036s
% mapped region	91.49	86.03	94.4
Mean number of match	6 923	8 327	1 461
% identity	97.83	99.14	99.5
contigs (>=1000pb)	23	24	210
N50	771921	776764	64372
Total length	12 143 815	12 284 076	9 946 745
Genome fraction	97.99	98.53	84.9

4.13 Test 10 : Saccharomyces cerevisiae W303, reads Minion 20x

Dataset:

- Minion reads corrected by lordec (20x coverage) : 47027 reads
- Illumina reads (HiSeq) : 3815678 reads of length 2x100pb
- Contigs generated by sparse assembler from Illumina reads (10055 contigs)

hybrid correctors

Metrics	LSC-2		PacBioToCA	Ectools	Proovread	
	trim	untrim			trim	untrim
Size	36.32	55.86	4.60	32.75	23.81	100.85
Execution time	13 686s	13 686s	10 866s	120s	47 020s	47 020s
% mapped region	92.23	85.42	98.74	96.4	87.43	78.46
Mean number of match	4 225	5921	216	7 537	1 349	5 943
% identity	79.67	77.40	–	99.3	99.40	90.12
contigs (>=1000pb)	76	15	–	140	6	131
N50	25 411	23 975	–	100 030	16 107	114 531
Total length	18 341 73	336 512	–	10 456 921	91 934	11 046 053
Genome fraction	0.15	0.034	–	88.48	0.55	80.16

Metrics	LoRDEC		Nanocorr	Nas	Jabba
	trim	untrim			
Size	40.45	101.31	56.14	–	4.7
Execution time	2 877s	2 877s	18 450s	–	140s
% mapped region	94.64	78.74	95.96	–	99.6
Mean number of match	1 269	6 062	6 542	–	579
% identity	99.64	90.36	99.22	–	99.9
contigs (>=1000pb)	15	156	81	–	–
N50	16 772	84 777	200 230	–	–
Total length	244 645	10 456 229	11 362 908	–	–
Genome fraction	1.85	77.93	95.74	–	–

The lack of correction of the *Nas* tool is due to an incompatibility of the illumina data. The lack of results after the assembly of long reads corrected by *PacBioToCA* ou *Jabba* is due to a low quantity of available long reads after the correction stage.

de-novo correctors

Metrics	PacBioToCA	MHAP (CANU)	LoRMA
Size	5.02	–	–
Execution time	4 908s	–	–
% mapped region	89.78	–	–
Mean number of match	986	–	–
% identity	88.58	–	–
contigs (>=1000pb)	1	–	–
RR n° 9028 N50	12991	–	–
Total length	12 991	–	–
Genome fraction	0.003	–	–

The lack of results after the assembly of long reads corrected by *PacBioToCA* ou *Jabba* is due to a low quantity of available long reads after the correction stage.

4.14 Test 11 : Caenorhabditis elegans, reads Pacbio 10x (P6-C4)

Dataset:

- Pacbio reads corrected by lordec (10x coverage) : 92597 reads
- Illumina reads (MiSeq) : 55070232 reads of length 2x150pb
- Contigs generated by sparse assembler from Illumina reads (1022387 contigs)

hybrid correctors

Metrics	LSC-2		PacBioToCA	Ectools	Proovread	
	trim	untrim			trim	untrim
Size	–	–	–	58.51	76.86	84.85
Execution time	–	–	–	2 166s	1 197 640s	1 197 640
% mapped region	–	–	–	99.20	98.3	93.61
Mean number of match	–	–	–	9 054	5 809	10 629
% identity	–	–	–	98.70	98.98	98
contigs (>=1000pb)	–	–	–	718	1 111	808
N50	–	–	–	67 819	58 024	138 446
Total length	–	–	–	42 539 808	56 155 537	87 425 140
Genome fraction	–	–	–	42.11	55.60	85.75

Metrics	LoRDEC		Nanocorr	Nas	Jabba
	trim	untrim			
Size	99.15	99.15	–	–	2.6
Execution time	9 155	9 155	–	–	1 502
% mapped region	99.34	94.59	–	–	99.5
Mean number of match	154	9 870	–	–	2 842
% identity	98.19	91.17	–	–	99.9
contigs (>=1000pb)	–	784	–	–	2
N50	–	151 698	–	–	33 717
Total length	–	91 256 859	–	–	57 816
Genome fraction	–	13.16	–	–	0.058

The lack of results after the assembly of long reads corrected by *PacBioToCA* is due to a low quantity of available long reads after the correction stage. The lack of read correction by *LSC-2* and *Nanocorr* is due to an excessive execution time. The lack of correction of the *Nas* tool is due to an incompatibility of the illumina data.

de-novo correctors

Metrics	PacBioToCA	MHAP (CANU)	LoRMA
Size	66.83	58.8	0.46
Execution time	153 269s	7 560s	4 870s
% mapped region	99.25	98.44	99.1
Mean number of match	3 876	7 778	473
% identity	98.22	98.31	99.4
contigs (>=1000pb)	796	1 062	–
RR n° 9028 N50	36 996	40 476	–
Total length	27 782 956	40 777 446	–
Genome fraction	27.59	40.5	–

The lack of results after the assembly of long reads corrected by *loRMA* is due to a low quantity of available long reads after the correction stage.

4.15 Test 12 : Caenorhabditis elegans, reads Pacbio 100x (P6-C4)

Dataset:

- Pacbio corrected by lordec (100x coverage) : 740776 reads
- reads Illumina reads (MiSeq) : 55070232 reads of length 2x150pb
- Contigs generated by sparse assembler from Illumina reads (1022387 contigs)

hybrid correctors

Metrics	LSC-2		PacBioToCA	Ectools	Proovread	
	trim	untrim			trim	untrim
Size	–	–	–	55.71	–	–
Execution time	–	–	–	44 340s	–	–
% mapped region	–	–	–	99.22	–	–
Mean number of match	–	–	–	8 711	–	–
% identity	–	–	–	98.60	–	–
contigs (>=1000pb)	–	–	–	618	–	–
N50	–	–	–	157 992	–	–
Total length	–	–	–	65 473 014	–	–
Genome fraction	–	–	–	60.28	–	–

Metrics	LoRDEC		Nanocorr	Nas	Jabba
	trim	untrim			
Size	81.94	97.90	–	–	40.4
Execution time	149 977s	149 977s	–	–	3 028s
% mapped region	97.90	94.59	–	–	99.2
Mean number of match	797	9 870	–	–	2 842
% identity	98.72	91.17	–	–	99.9
contigs (>=1000pb)	1 202	90	–	–	50
N50	25 551	2 249 996	–	–	28 845
Total length	27 660 721	107 109 382	–	–	1 360 729
Genome fraction	25.36	85.75	–	–	1.3

The lack of results after the assembly of long reads corrected by *PacBioToCA* is due to a low quantity of available long reads after the correction stage. The lack of read correction by *LSC-2* et *Nanocorr* is due to an excessive execution time. The lack of read correction by *Nas* tool is due to an incompatibility of the illumina data.

de-novo correctors

Metrics	PacBioToCA	MHAP (CANU)	LoRMA
Size	0.24	40.77	20.1
Execution time	12 559s	171 360s	59 621s
% mapped region	98.64	94.51	99.65
Mean number of match	7 479	14 654	465
% identity	96.69	98.65	99.7
contigs (>=1000pb)	–	129	3
RR n° 9028 N50	–	1 904 749	13 510
Total length	–	104 538 176	45 760
Genome fraction	–	99.38	0.046

The lack of results after the assembly of long reads corrected by *PacBioToCA* is due to a low quantity of available long reads after the correction stage.

References

[al99] Delcher AL et al. "Alignment of whole genomes". In: *Nucleic Acids Res. 1999 Jun 1;27(11):2369-76. PMID:10325427 ; PMCID:PMC148804* (1999).

[al12a] Kin Fai Au et al. "Improving PacBio Long Read Accuracy by Short Read Alignment". In: *http://dx.doi.org/10.1371/journal.pone.0046679* (2012).

[al12b] Sergey Koren et al. "Hybrid error correction and de novo assembly of single-molecule sequencing reads". In: *Nature Biotechnology 30(7):693-700* (2012).

[al13] Alexey Gurevich et al. "QUAST: quality assessment tool for genome assemblies". In: *Bioinformatics 29(8), 1072-1075.* (2013).

[al14] Hayan Lee et al. "Error correction and assembly complexity of single molecule sequencing reads". In: *http://dx.doi.org/10.1101/006395* (2014).

[Sal14] Rivals E Salmela L. "LoRDEC: accurate and efficient long read error correction". In: *Bioinformatics. 30(24):3506-14. doi: 10.1093/bioinformatics/btu538* (2014).

[T14] Hackl T. "proovread: large-scale high-accuracy PacBio correction through iterative short read consensus". In: *Bioinformatics. 30(21):3004-11. doi: 10.1093/bioinformatics/btu392* (2014).

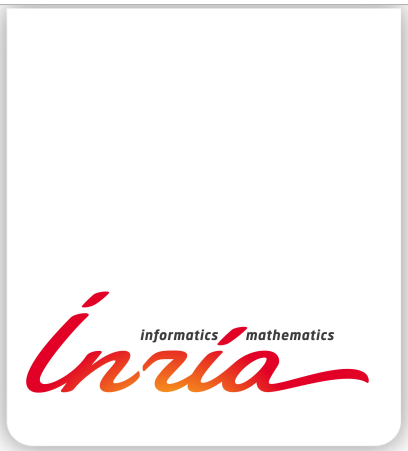
[al15a] Konstantin Berlin at al. "Assembling Large Genomes with Single-Molecule Sequencing and Locality Sensitive Hashing". In: *Nature Biotechnology doi: 10.1038/nbt.3238* (2015).

[al15b] Mohammed-Amin Madoui et al. "Genome assembly using Nanopore-guided long and error-free DNA reads". In: *BMC Genomics 16:327; doi: 10.1186/s12864-015-1519-z* (2015).

[al15c] Sara Goodwin et al. "Oxford Nanopore Sequencing and de novo Assembly of a Eukaryotic Genome". In: *Genome Research doi: 10.1101/gr.191395.115* (2015).

[al16a] Giles Miclotte at al. "Jabba: hybrid error correction for long sequencing reads". In: *Algorithms for Molecular Biology 11:10; doi: 10.1186/s13015-016-0075-7* (2016).

[al16b] Leena Salmela et al. "Accurate selfcorrection of errors in long reads using de Bruijn graphs". In: *arXiv:1604.02233* (2016).



RESEARCH CENTRE
RENNES – BRETAGNE ATLANTIQUE

Campus universitaire de Beaulieu
35042 Rennes Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399