Instruction Manual for the Ray *de novo* genome assembler software

Sébastien Boisvert

February 4, 2011

Ray version 1.2.3

http://denovoassembler.sf.net

Reference to cite:

Sébastien Boisvert, François Laviolette & Jacques Corbeil.
Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies.
Journal of Computational Biology (Mary Ann Liebert, Inc. publishers, New York, U.S.A.).
November 2010, Volume 17, Issue 11, Pages 1519-1533.
doi:10.1089/cmb.2009.0238
http://dx.doi.org/doi:10.1089/cmb.2009.0238

Contents

1 Inputs

The input files for Ray contain sequences. The files must be formatted in one of the supported formats.

Table 1: File formats compatible with the Ray *de novo* genome assembler software

Format	Obligatory extension
Fasta format	.fasta
Fasta format, compressed with GNU zip (gzip)	.fasta.gz
Fasta format, compressed with bzip2	.fasta.bz2
Fastq format	.fastq
Fastq format, compressed with GNU zip (gzip)	.fastq.gz
Fastq format, compressed with bzip2	.fastq.bz2
Standard flowgram format	.sff

2 Parameters

2.1 k-mer length with -k

Ray builds a distributed catalog of all occuring k-mers in the reads and their reverse-complement. k must be greater or equal to 15 and lower or equal to 32.

2.2 Output prefix with -o

Output files are named according to the prefix provided by the option -o.

2.3 Single-end reads with -s

```
-s <sequencesFile>
```

2.4 Paired-end reads and mate-pair reads with -p

```
-p <leftSequencesFile> <rightSequencesFile> [ <averageFragmentLength> <standardDeviation> ]
```

Example for paired-end reads (the ends of DNA fragments):

```
-p s_200_1.fastq s_200_2.fastq
```

Example for mate-pair reads (owing to circularization and ligation, the order must be reverse):

```
-p s_20000_2.fastq s_20000_1.fastq
```

2.5 Paired-end reads and mate-pair reads with -i (interleaved sequences)

`-i <sequencesFile [ <averageFragmentLength> <standardDeviation> ]`

In the interleaved file (example is for a fasta file):

```
>200_1_1234/1
ATCGATCGATCGACTCAGACACGTACG
>200_1_1234/2
ACTGACGACGTACGACGTCATGCAACT
...
```

For mate-pair reads, the order must be reverse.

3 Output

3.1 Contiguous sequences

OutputPrefix.fasta contains contiguous sequences.

3.2 Paired-end and mate-pair libraries

OutputPrefix.LibraryLibraryNumber.txt contains the distribution of distances for paired-end and mate-pair libraries. One file per library.

3.3 Coverage distribution

OuputPrefix.CoverageDistribution.txt contains the k-mer coverage distribution.

3.4 Messages

OuputPrefix.ReceivedMessages.txt contains a matrix.
It contains the number of received messages for each MPI rank.
(MPI communication matrix; rows=destinations, columns=sources)

4 Example

4.1 Bacterial genome with paired-end and mate-pair short reads

The command:

```
mpirun -np 32 ~/Ray/trunk/code/Ray \
-p /home/boiseb01/nuccore/Large-Ecoli/200_1.fastq \
   /home/boiseb01/nuccore/Large-Ecoli/200_2.fastq \
```

```
-p /home/boiseb01/nuccore/Large-Ecoli/1000_1.fastq \
   /home/boiseb01/nuccore/Large-Ecoli/1000_2.fastq \
-p /home/boiseb01/nuccore/Large-Ecoli/4000_1.fastq \
   /home/boiseb01/nuccore/Large-Ecoli/4000_2.fastq \
-p /home/boiseb01/nuccore/Large-Ecoli/10000_1.fastq \
   /home/boiseb01/nuccore/Large-Ecoli/10000_2.fastq \
-o BacterialGenome | tee RayLog
```