PERGA version 0.4
Reference Manual

License

PERGA is freely available to any scientist wishing to use it for non-commercial purposes.

Overview

A quick description of the program options are obtained by typing: perga -h or perga –help; perga_sf –h or perga_sf -help

PERGA is a Paired End Reads Guided Assembler that a novel sequence reads guided *de novo* assembly approach which adopts greedy-like prediction strategy for assembling reads to contigs and scaffolds. Instead of using single-end reads to construct contig, PERGA uses paired-end reads and different read overlap sizes from $O \geq O_{max}$ to O_{min} to resolve the gaps and branches. Moreover, by constructing a decision model using machine learning approach based on branch features, PERGA can determine the correct extension in 99.7% of cases. PERGA will also try to extend the contigs by all feasible nucleotides and determine if these multiple extensions due to sequencing errors or repeats by using looking ahead technology.

PERGA consists of two assembly phases: (1) assembly of reads and (2) assembly of contigs. The first phase is to assemble reads into contigs, and the second phase is to assemble contigs into scaffolds based on paired end reads.

Reads of Fasta and fastq format can be input into PERGA either in single or paired mode, and the output of contigs or scaffolds are multi-fasta format files.

Assembly of reads: assemble reads into contigs based on paired ends if the paired ends information is available. It can be carried out in single or paired mode. The final contigs are output into a multi-fasta file.

PERGA consists of three commands: “ht”, “assemble”, and “all”. The “ht” command is to construct the k-mer hash table based on k-mers; while the “assemble” command is to assemble reads into contigs. The “all” command means run the above two commands in turn. For example, when assemble *E.coli* genome based on reads in the file ERA000206_bix.fastq, we can execute PERGA by typing:

```
./perga all -k 25 -p 2 -f reads/ERA000206_bix.fastq -d outDir/ -o E.coli_ERA000206_bix -m 100
```

Alternatively, we can also execute “ht” or “assemble” command separately in turn, to execute the “ht” command:

```
./perga ht -k 25 -p 2 -f reads/ERA000206_bix.fastq -d outDir/ -o E.coli_ERA000206
```

And then execute the “assemble” command:

```
./perga assemble -g outDir/E.coli_ERA000206_bix_hashtable_K25_R70.bin -d outDir/ -o E.coli_ERA000206_bix -m 100
```

Assembly of contigs (scaffolding): this phase is based on paired ends. Contigs are ordered and oriented to construct scaffolds bases on paired ends, and the intra-scaffold gaps are filled, and finally, the scaffold sequences are output into a multi-fasta file. After the generating contigs of *E.coli* genome, we can construct scaffolds by:

```
./perga_sf -c outDir/E.coli_ERA000206_bix_contigs_K25_R70.fa -s 400 -m 100 -p 2 -f reads/ERA000206_bix.fastq -g yes -o E.coli_ERA000206_bix -d scafOutDir/
```

Usage: Assembly of reads (hash table construction)

A quick description of the program options is obtained by typing: perga -h or perga –help
Options are described as follows.

-k <INT> **k-mer size**

The k-mer size for assembly. Default is 21.

-r <INT> **read length cutoff**

Read length cutoff for reads. Reads longer than INT will be cut to INT for better base accuracy. Default is the read length.

-p <INT> **paired mode**

Paired end mode for read files. 0 - do not treat reads as paired (default); 1 - reads are paired with the first read in the first file, and the second read in the second file; 2 - reads are paired and are interleaved within a single file.

-f <files> **read files**

Read files. It is required for commands 'all' and 'ht'.

-q <INT> **single base quality threshold**

Single base quality threshold. Reads with single base quality < INT will be discarded. It is very useful for assembly with high reads coverage depth. Default is 2.

-d <dir> **output directory**

Output directory. All the result files are output into the directory. The default output directory is "./".

-o <String> **Prefix for output files**

The prefix of the output files.

-h (-help) **help information**

Show the help information and exit.

Usage: Assembly of reads (assembly)

A quick description of the program options is obtained by typing: perga -h or perga –help

Options are described as follows.

-g <file> **hash table file**

Hash table file for assembly. It is required for command 'assemble'.

-ins_len <float> **insert size for paired end library**

Insert size for paired end library. It will be used to assemble contigs in paired end mode.

-ins_sdev <float> **standard deviation for insert size**

Standard deviation of insert size of a paired end library. If not specified, this value is set to 0.15*insertSize if insert size is specified.

-d <dir> **output directory**

Output directory. All the result files are output into the directory. The default output directory is "./".

-o <String> **Prefix for output files**

The prefix of the output files.

-m <Int> **minimum size of the contigs to output**

The minimal size of contigs to output. Default value is 100 bp.

-h (-help) **help information**

Show the help information and exit.

Usage: Assembly of contigs (scaffolding)

A quick description of the program options is obtained by typing: perga_sf -h or perga_sf –help
Options are described as follows.

-c <file> **contigs file**

The input contigs file in fasta format. It is required to be specified.

-f <files> **read files**

Read files. It is required to be specified.

-p <INT> **paired mode**

Paired end mode for read files. 1 - reads are paired with the first read in the first file, and the second read in the second file (default); 2 - reads are paired and are interleaved within a single file.

-s <INT> **align region size at contig ends**

Region size of align region at contig ends. Default is 1000 bp.

-m <INT> **minimum size of the contigs for scaffolding**

Minimum size of contigs for scaffolding, and contigs with length < INT are discarded in scaffolding. Default is 100 bp.

-ins_len <float> **insert size for paired end library**

Insert size for paired end library. It will be used to scaffold contigs in paired end mode.

-ins_sdev <float> **standard deviation for insert size**

Standard deviation of insert size of a paired end library. If not specified, this value is set to 0.15*insertSize if insert size is specified.

-d <dir> **output directory**

Output directory. All the result files are output into the directory. The default output directory is "./".

-g <STR> **gap filling flag**

Gap filling flag that indicates whether the intra-scaffold gap are been filled or not: Y(es) - gap filling (default), N(o) - no gaps will be filled.

-o <String> **Prefix for output files**

The prefix of the output files.

-h (-help) **help information**

Show the help information and exit.

Suggestions, comments, bugs

ydwang@hit.edu.cn, zhuxiao.hit@gmail.com