GapCloser Manual

Introduction

The GapCloser is designed to close the gaps emerging during the scaffolding process by SOAPdenovo or other assembler, using the abundant pair relationships of short reads.

System Requirement

GapCloser aims for large plant and animal genomes, although it also works well on bacteria and fungi genomes.

The memory use is mainly related to the read number, the number of unique Kmers in the reads and the gap number and size in scaffolds.

The time consumption depends on gap number, gap size and read number.

Given the assembly of YH genome (genome size ~ 3G) as an example, the peak memory used by GapCloser was about 200G and the time cost was about 1 days.

Command Line Options

1. A typical command line:

GapCloser –b config_file –a scaffold_file –o output_file

2. Parameters:

GapCloser [options]

-a <string> input scaffold file name, required.

-b <string> input library info file name, required.

-o <string> output file name, required.

-l <int> maximum read length (<=155), default=100.

-p <int> overlap param(<=31), default=25.

-t <int> thread number, default=1.

-h -? output help information.Input files:

- The format of configure file is the same as the configure file for SOAPdenovo.
- The format of input scaffold file should be FASTA.

3. Output File:

- One file (named by -o) contains scaffold sequences with some gaps filled.
- *.fill

  It describes the information of the gaps in the scaffolds. The first column is the starting position of a gap in the output sequence. The second is the end position. The third and fourth are the length of sequences extending separately from the left and right boundaries of a gap. The status of the gaps can be seen from the fifth column of the file. If a gap was finished through the overlapping of Kmers, the flag was set to 1, otherwise 0. The sixth column tells the length of gap sequence with relative high quality. The seventh column is the original gap size. The eighth column is the final gap size. If the gap was closed (the value of the fifth column is 1), the value is the length of gap sequence. Otherwise, the value is equal to the value of the seventh column or 1 bp longer than it when the value of the seventh column is 1.

FAQ

1. What pair ends should be used for gap filling?

   GapCloser mainly uses read pairs with short and medium insert sizes, although the long insert PE reads (>2K) may also help.

   It is recommended that the reads be corrected before gap filling, both for the consideration of memory usage and accuracy of gap sequences produced at this stage.

2. How to set maximal read length (default 100)?

   The maximal read length can be set by parameter ‘l’.

3. What’s the sequence quality produced during gap filling?

   The sequence quality is statistically lower than that of the sequences on both sides of the gaps.