

SRGA version 0.4

Reference Manual

License

SRGA 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: `SRGA -h` or `SRGA -help`; `SRGA_SF -h` or `SRGA_SF -help`

SRGA is a short read guided assembler that combined de Bruijn graph strategy and overlap graph strategy together. This program requires the reads to be all the same length, as Illumina reads are. And if you provided multiple files with different read lengths, SRGA will trim the 3' end of reads so that the reads are all the same length as the shortest reads in the files. And reads that are too long may impair the assembly performance. In such a case, SRGA will truncate the reads at 3' ends to be the same read length.

SRGA 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 SRGA 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.

SRGA consists of three commands: "graph", "assemble", and "all". The "graph" 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 `SRR001665.fastq`, we can execute SRGA by typing:

```
SRGA all -k 21 -r 34 -p 2 -f reads/SRR001665.fastq -d outDir/ -o E.coli_SRR001665 -m 100
```

Alternatively, we can also execute "graph" or "assemble" command separately in turn, to execute the "graph" command:

```
SRGA hashtable -k 21 -r 34 -p 2 -f reads/SRR001665.fastq -d outDir/ -o E.coli_SRR001665
```

And then execute the "assemble" command:

```
SRGA assemble -g outDir/E.coli_SRR001665_graph_K25_R34.bin -d outDir/ -o E.coli_SRR001665 -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:

```
SRGA_SF -c outDir/E.coli_SRR001665_contigs_K21_R34.fa -s 400 -m 100 -p 2 -f reads/SRR001665.fastq -g yes -o E.coli_SRR001665 -d scafOutDir/
```

Usage: Assembly of reads (hash table construction)

A quick description of the program options is obtained by typing: *SRGA -h* or *SRGA -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 necessary to be specified for commands 'all' and 'hashtable'.	
-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 <i>"/"</i> .	
-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: *SRGA -h* or *SRGA -help*
Options are described as follows.

-g <file>	graph 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 $0.15 * \text{insertSize}$ 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: `SRGA_SF -h` or `SRGA_SF -help`
Options are described as follows.

-c <file>	contigs file
The input contigs file in fasta format. It is necessary to be specified.	
-f <files>	read files
Read files. It is necessary 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