# User Manual

| | |
|---|---|
| **Date:** | **08 November 2011** |
| **Software Version:** | **2.0** |
| **Developed by:** | **Marten Boetzer and Walter Pirovano** |

What is SSPACE?
---------------

To start; SSPACE is not a de novo assembler, it is used after a pre-assembled run. SSPACE is a script to extend and scaffold pre-assembled contigs using a number of mate pairs or paired-end libraries. It uses Bowtie to map all the reads to the pre-assembled contigs. Unmapped reads are used for extending, if desired, the pre-assembled contigs with the SSAKE assembler. Again Bowtie is used to map the reads to the extended contigs. Positions and orientation of the reads are stored and used for scaffolding. If both reads of a pair are found within the allowed distance, they are used for scaffolding to determine the orientation, contig pairing and ordering of the contigs.

Why SSPACE?
---------------
SSPACE can be used for a number of reasons;

- A pre-assembly was performed with single reads, generating contigs. The user now has additional data, like mate pair data, and wants to use these data to extend and scaffold the contigs.

- A pre-assembly was performed with, for example, mate pair data of 1 kb insert size, generating contigs. The user now has additional data with larger insert size. The user wants to include the data for extending and scaffolding the contigs.

- A pre-assembly was performed with mate pair data on an assembler, generating contigs. The assembler, however, has no scaffolder. Inserting the contigs, along with the mate pair data, still can scaffold the contigs.

How to use SSPACE scaffolder?
---------------
SSPACE scaffolder comes with a number of files.

SSPACE_Basic_v2.0.pl
- Main program. Perl file with all the script for reading, extending, mapping and scaffolding.

README

- README file. Information about the process, input files/ parameter options, and output files.

MANUAL
- This file.

TUTORIAL
- Small tutorial on an E.coli dataset.

bin folder
- perl subscripts used by SSPACE.

Bowtie folder
- bowtie scripts for mapping the reads to the contigs

Dotlib folder
- Contains DotLib library for generating .dot file to visualize the scaffolds and its contigs

Example folder
- Example contigs, read TUTORIAL file.

Tools folder
- A number of useful tools including trimming, insert size estimation and conversion from .sam to .tab tools

To run the main script, type;

```
Perl SSPACE_Basic_v2.0.pl
Or
./SSPACE_Basic_v2.0.pl
```

This will print the options and parameters to the screen. Below is each parameter explained in detail.

The ‘-l’ library file:
---------------
The library file contains information about each library. The library file contains six columns, each separated by a space. An example of a library file is;

```
Lib1 file1.1.fasta file1.2.fasta 400 0.25 FR
Lib1 file2.1.fasta file2.2.fasta 400 0.25 FR
Lib2 file3.1.fastq file3.2.fastq 4000 0.5 RF
Lib2 TAB file4.tab 4000 0.5 RF
Lib3 TAB file5.tab 10000 0.5 RF
```

Each column is explained in more detail below;

Column 1:
---------
Name of the library. A short name to keep track of the names of the libraries. All temporary files and summary statistics are named by this library name. Libraries having same name are considered to be of same distance and deviation (column 4 and 5). In addition, these libraries with similar names are use for the same scaffolding iteration. Libraries should be sorted on distance, the first library is scaffolded first, followed by next libraries.

Column 2 & 3:
---------
Fasta or fastq files for both ends. For each paired read, one of the reads should be in the first file, and the other one in the second file. The paired reads are required to be on the same line. No naming convention of the reads is required, because names of the headers are not used in the protocol. Thus names of the headers shouldn't be the same and do not require any overlap of names like (…).x and (…).y, which is commonly used in assembly programs.

During the reading step, the sequences of both pairs are merged together and filtered for;

-Mapping. the filtered read pairs with only ACGT characters and no duplicates are used. The remaining merged read pairs are split to single reads, and are mapped to contigs.

-Extension. only unmapped single reads are used for extension of the pre-assembled contigs.

-Scaffolding. Besides the filtering of mate pairs containing "N" and non-ACGT character, duplicate mate pair sequences are also removed.

Concluding, each read should be larger than 16 (or the '–m' parameter if -x 1). If they are shorter, the program will simply omit them from the whole process.

If at the second column "TAB" (mind the capitals!) is set, the third column is considered as a Tabulated text file containing positions of read-pairs on contigs. The format is;

```
<ctg1>    <start1>   <end1>   <ctg2>    <start2>   <end2>
```

For example;

```
contig1    100   150   contig1    350   300
contig1    4000 4050 contig2     110   60
```

Some notes about TAB files;
1).
If Tab file is inserted;
- no filtering (-z option) of the contigs will be applied
- contigs will not be extended if -x option is set.

Both features can not be used, since otherwise the positions of the reads on the contigs are not correct.

2).
It is possible to include multiple different TAB libraries and combination of tab library with normal .fasta/.fastq files.

3).
The contigs in the TAB files are required to be the same as the names in the inserted contig file (-s option). Names of the contigs are splitted on spaces, so a contig name like '>contig1 cov300' will be 'contig1'. 'contig1' should thus be the name of the contig in the TAB file.

See the README for more information about how the TAB file works.
See the TUTORIAL on how to convert a SAM or BAM file to a .tab file.

Column 4 & 5:
---------
The fourth column represents the expected/observed inserted size between paired reads. The fifth column represents the minimum allowed error. A combination of both means e.g. that with an expected insert size

of 4000 and 0.5 error, the distance can have an error of 4000 * 0.5 = 2000 in either direction. Thus pairs between 2000 and 6000 distance are valid pairs.

Column 6:
---------
The final column indicates the orientation of the paired-reads. Orientations can be: FF, FR, RF or RR. Where the F stands for --> orientation, and R for <-- orientation. Orientation of FR thus means that the pairs are: --><--

MAIN PARAMETERS:

The '-s' contigs fasta file
---------------
The '–s' contigs file should be in a .fasta format. The headers are used to trace back the original contigs on the final scaffold fasta file. Therefore, names of the headers should not be too complex. A naming of ">contig11" or ">11", should be fine. Otherwise, headers of the final scaffold fasta file will be too large and hard to read.
Contigs having a non-ACGT character like "." or "N" are not discarded. They are used for extension, mapping and building scaffolds. However, contigs having such character at either end of the sequence, could fail for proper contig extension and read mapping.

The '-x' contig extension option
---------------
Indicate whether to do extension or not. If set to 1, contigs are tried to be extended using the unmapped sequences. If set to 0, no extension is performed.

EXTENSION PARAMETERS:

The '–m' minimum overlap
---------------
Minimum number of overlapping bases of the reads with the contig during overhang consensus build up. Higher '-m' values lead to more accurate contigs at the cost of decreased contiguity. We suggest to take a value close to the largest read length. For example, for a library with 36bp reads, we suggest to use a -m value between 32 and 35 for reliable

contig extension. For more information, see the SSPACE README file or the SSAKE paper/poster.

The -o number of reads
---------------
Minimum number of reads needed to call a base during an extension, also known as base coverage. The higher the ‘-o’, the more reads are considered for an extension, increasing the reliability of the extension.

The ‘-t’ trimming option
---------------
Trims up to ‘-t’ base(s) on the contig end when all possibilities have been exhausted for an extension. See SSAKE help files for information.

The ‘-u’ unpaired single reads
---------------
Unpaired reads can be inserted for contig extension. These reads are not used for scaffolding.

The ‘-r’ minimal base ratio
---------------
Minimum base ratio used to accept a overhang consensus base. Higher '-r' value lead to more accurate contig extension.

SCAFFOLDING PARAMETERS:

The ‘-k’ minimal links and ‘-a’ maximum link ratio
---------------
Two parameters control scaffolding (-k and -a). The -k option specifies the minimum number of links (read pairs) a valid contig pair must have to be considered. The -a option specifies the maximum ratio between the best two contig pairs for a given contig being extended. For more information see the .readme file or the poster of SSAKE.

The ‘-n’ contig overlap
---------------
Minimum overlap required between contigs to merge adjacent contigs in a scaffold. Overlaps in the final output are shown in lower-case characters.

The '-z' minimal contig
---------------
Minimal contig size to use for scaffolding. Contigs below this value are not used for scaffolding and are filtered out. Larger contigs produce more reliable scaffolds and also the amount of scaffolds is vastly reduced. Smaller contigs (< 100bp) are likely to be repeated elements and can stop the extension of the scaffold due to exceeding the -a parameter.

BOWTIE MAPPING PARAMETERS:

The '-g' maximum gaps
---------------
Maximum allowed gaps for Bowtie, this parameter is used both at mapping during extension and mapping during scaffolding. This option corresponds to the -v option in Bowtie. We strongly recommend using no gaps, since this will slow down the process and can decrease the reliability of the scaffolds. We only suggest to increase this parameter when large reads are used, e.g. Roche 454 data or Illumina 100bp.

The '-T' number of threads
---------------
Number of search threads for mapping reads to the contigs with Bowtie. See the Bowtie (http://bowtie-bio.sourceforge.net/manual.shtml) for more information.

ADDITIONAL PARAMETERS:

The '-p' plot option
---------------
Indicate whether to generate a .dot file for visualisation of the produced scaffolds.

The '-b' prefix base name
---------------
All files start with the '-b' prefix to allow for multiple runs on the same folder without overwriting the results.

The '-v' verbose option

---------------

Indicate whether to run in verbose mode or not. If set, detailed information about the contig extension and contig pairing process is printed on the screen.

Additional information about the input, output and general process of the script can be found in the README file.