Tutorial

Date:	**10 July 2012**
Software Version:	**1.10**
Developed by:	**Marten Boetzer and Walter Pirovano**

GapFiller v1.10 Marten Boetzer - Walter Pirovano, Jul 2012
email: walter.pirovano@baseclear.com

Citation

If you use GapFiller in a scientific publication, please cite:
Boetzer, M. and Pirovano, W., Towards (almost) closed genomes with GapFiller, Genome Biology 13(6), 2012

License
----

getting started
============================

Download and obtaining data

Download the E.coli 200bp paired-end read data from
http://www.ebi.ac.uk/ena/data/view/SRR001665

Download both .fastq files and store them on disk. For this example, we place the two datasets in the 'example' folder which is present in the GapFiller .zip file.

In the 'example' folder, a scaffold dataset is present. This file is generated by assembling the E.coli 200bp paired-end dataset with ABYSS and consecutively scaffolded with SSPACE using the paired-end dataset of SRR001665.
In the same folder, a library file is available. This library file describes the datasets and distances between the reads for each library. Multiple libraries are allowed, here only one is described. Each column per library should be seperated by a space. In the file 'library.txt' a library is available for the E.coli 200bp paired-end dataset. Each column is described below;

-First column contains a desired name, here we take 'ecoli'.
-Second and third columns are the datasets.
-Fourth column is the mean distance between the paired reads, which is 215.
-Fifth column is a deviation of the mean distance that is allowed, here we take 0.25. This means that any pair having a distance between 150 and 250 is allowed.
-The final column indicates the orientation of the paired-reads. Here we set this to 'FR', since the pairs are orientated as --><--.

Before running GapFiller, set your current working directory to the 'example' directory(cd (path_to_GapFiller)/example) were both the library file and the scaffold sequences are stored.

Running GapFiller

We run GapFiller on the SSPACE scaffolds using the following parameters;

```
perl (path_to_SSPACE)/GapFiller.pl -l libraries.txt -s
SSPACE_scaffolds.fa -m 30 -o 2 -r 0.7 -n 10 -d 50 -t 10 -T 1 -i 1 -b test
```

This means that we use the paired files in 'libraries.txt' to fill the scaffolds of 'SSPACE_scaffolds.fa'. For information about the parameters used, see the README file. A number of files and folders are generated;

One file is the 'test.filled.final.txt'. This file gives useful information about the gapclosing, including the original gapsize, the number of nucleotides that are closed and whether the gap is closed or not.
The file 'test.closed.evidence.txt' contains detailed information about the gapfilling process of each gap, e.g. how the extension was performed.
The file 'test.summaryfile.txt' contains a summarized information about the analyses, e.g. parameters used and number of gaps that are closed.
The final file is the 'test.gapfilled.final.fa', which contains the final gapclosed sequences.