README.md

Plewka\_et\_al\_PAIS

***Prediction of circular insertion sequences in the river mesocosm experiment***

This here is an operating procedure to test circularity for short mobile genetic elements (i.e. prokaryotic insertion sequences). It describes the workflow used in the publication.

**Contact**

You can contact me via Julia.plewka@uni-due if you have any questions as to the contents of this subfolder.

**LICENSE**

??

**Overview**

The described workflow has the following steps:

1. Splitting of the nucleotide sequence of the element in half and concatenation of the beginning of the first half to the end of the second half.
2. Mapping of the quality-controlled reads to the newly rearranged element using Bowtie2 (sensitive mode)
3. Removal of reads with more than five miss-matches.
4. Extraction of reads mapping across the region 50 bps before and 50 bps after the concatenation point.
5. Conversion of the SAM output into sorted BAM files (using samtools).
6. Calculation of the positional coverage depth with genomecov from bedtools (using the “bg” flags).

**Dependencies**

**Software**

* Bowtie2 (tested with version 2.5.4)
* Samtools (tested with version 1.21)
* Bedtools (tested with version 2.31.1)

Dependencies installation with mamba:

mamba create -n circularity\_test bioconda::bowtie2 bioconda::samtools bioconda::bedtools

**Remarks**

* Required inputs are:
  + The nucleotide sequence of the element in a single FASTA file (multi FASTA files are not permitted).
  + Quality-controlled raw reads, where the element was extracted from.
  + A tab separated BED file with the fasta header, bps position 50 bps before the midpoint of the sequence, bps position 50 bps after the midpoint of the sequence.
* Scripts work with both contigs and scaffolds.
* The quality controlled raw reads need to be separated into forward and reverse reads in FASTQ files. Note that Bowtie2 flags have to be adjusted if long-reads or interleaved files are mapped. The downstream analysis of the workflow can be used for both read types.
* The BED file can contain multiple entries of elements in separated rows.

**Workflow**

1. Activate conda environment

mamba activate circularity\_test or conda activate circularity\_test

1. Cut and rearrange nucleotide sequence with:

python3 01\_reorder\_firstSecHalf.py -i input.fasta -o output.fasta

1. Create bowtie2 index with:

bash 02\_mapping\_index.sh output.fasta

1. Map reads and filter for mapped region:

bash 03\_mapping.sh output.fasta fwd\_reads rev\_reads bed\_file

**Output**

In the end, a tab separated file with positional coverage information over the cut region (output.fasta\_filteredFLANKING\_CUT\_genomecov.txt) and a SAM file including only reads in the cut region are created (output.fasta\_filteredFLANKING\_CUT.sam.gz).