Circular alignment viewer for bacterial plasmids and antibiotic resistance genes

Programmierprojekt 2018

Background

Bacterial genomes often contain circular DNA sequences. These so called plasmids facilitate horizontal gene transfer and often contribute to antibiotic resistances in bacteria. Genome viewers are interactive tools to display genome sequences, gene annotations and aligned sequencing reads. Most conventional genome viewers (IGV, Tablet) are unable to display circular alignment visualizations. Recent 3rd generation technologies (Nanopore, PacBio) produce long reads that overlap the "gap" inserted into linear representations. A circular genome viewer can therefore help to explore and analyze alignments on circular DNA. This can be used to finish assemblies and identify whether a DNA sequence (contig) is a circular plasmid or part of the main chromosomal DNA.

This tool should be able to generate, analyze and visualize the alignment. In a first step it will process the input data, align the sequence reads to the genome contigs and try to circularize the contigs based on the read alignment. Later we will add interactivity and additional statistics. If no difficulties arise up to this point, further functions can be implemented. The tool should be able to produce publication ready figures.

Features

- Generate alignments in BAM format
- Identify circular DNA (using reads aligning to both ends of the linear plasmid sequence)
- Circular visualization of plasmids and reads
- Interactivity: Zoom, Rotate, Highlight, extract Reads
- Add gene annotations in .gff Format
- Identify antibiotic resistance genes and highlight them
- Add sequence statistics: Coverage, GC content, alignment identity
- Blast selected regions against NCBI or EBI-Ensmbl database
- [optional] Identify wrong assemblies, circularize plasmids incorrectly integrated into the genome sequence, determine correct plasmid length (i.e. trim overlapping ends, combine multiple contigs)
- [optional] Find origin of replication (oriC) and rotate plasmid

- [optional] Identify structural variations in the plasmids using the aligned reads

Implementation

- OS: Linux (Windows and MacOS optional)
- Source Code management: Github
- Task assignment / Discussions: Github
- **Preferred** language: Java with GUI in JavaFX
- Alternative: Interactive Web interface (d3.js) with backend in Python
- C++/Qt

First steps

- Understand the aim of the application. What are plasmids? How can long-read alignments be used for circularization of plasmids after assembly
- Try out different Genome Viewers (sample input files are attached)
- Investigate useful libraries (i.e. BAM-File parsing)
- Create Proof of Concept application for the chosen Language/Framework with simple circular visualization.
- Assign tasks.

Literature

Take a look at existing genome viewers to gain an overview of their layout and functions.

- **IGV** (Integrated Genome Viewer): widely used tool with many functions. http://software.broadinstitute.org/software/igv/
- **Tablet:** well-designed user interface. https://ics.hutton.ac.uk/tablet/
- SAM/BAM File specifications: https://samtools.github.io/hts-specs/SAMv1.pdf
- Pubmed: Find publications about plasmids, genome assembly and Nanopore sequencing (or many other scientific studies): https://www.ncbi.nlm.nih.gov/pubmed/