**ACLS Master Thesis Genomics: tutorial Nurdzane Memeti**

CRISPR systems are known to be involved in the defense of bacteria against phages and plasmids. The aim of your master work is to develop an analysis method that can analyze CRISPR repeat regions on a non-Darwinian way, modelling evolutionary steps in the development of these systems. We will use the two track modules to guide you towards the topic of your Master work. This analysis method will be developed based on the work done in the EGSB group on *Erwinia* spp. but will be tested on different *Enterobacteriaceae* for its function.

The Track Module 1 Genomics tutorial gave you a solid base in the literature on CRISPR systems in *Enterobacteriaceae* and especially *Erwinia* spp. (Introduction of your Master work). In Track Module 2 Genomics you will start with data analysis of the different *Erwinia amylovora* genomes.

**Proposed content**:

* Introduction: reduced text from TM1 and TM2, dependent on content. More literature to be added based on the different genomes.
* Methods: fitting to what will be reached (see below).
* Results: currently, there are (more than) 140 genomes of *E. amylovora* isolates available, grouping in five main groups as given in literature (Zeng 2018, Parcey 2020). The general question is how much divergence is present between the individual members of each group, and to program the specific tool for the identification and analysis of the CRISPR repeat regions in each genome. Another question is on how to visualize the results best.
* Discussion: can one discriminate, based on the CRISPR repeat regions, different groups, and if so, how would a proposed evolutionary path in *E. amylovora* look like?

**Methods**:

An analysis program for CRISPR repeat regions as an example *E. amylovora* by considering following points.

1. Firstly, integrate the CRISPRFinder resp. CRISPRcasFinder.
2. Get the inputs from CRISPRFinder program
3. Several checkpoints have to be performed such as: check repeats from CRISPR finder and assign to CRR1 – CRR3. Is the repeat in the right order? Does it start at leader via direction as in paper Rezzonico 2011?
4. Initialize a CRR CODE, because each spacer needs a specific identification number (strain specific identifier.)
5. Compare internal database with previous spacer sequences and allocate a CRR identifier (note: up to <4 SNP is considered similar sequence (see paper of Rezzonico 2011)) New spacers get a new CRR identifier in the database.
6. Output generation: sequence strain specific identifier with insert/deletion in CRR list (1000, 2000, 3000) with colored blocks

ideally: evolutionary aspects. In which order in and out.

**Time to invest**: 30 ETCS, 900 working hours, 107 working days, 21,4 working weeks for fulltime

**Bi-Weekly discussion:** with supervisor (set dates in advance).

**Early corrections of the report:** November

**Deadline**: 10. January 2022