M.Sc. Internship project

Title: Characterization of CRISPR transcripts through NGS sequence analysis

Supervisor : Daniel GAUTHERET (Professor)
Co-supervisors: Gilles VERGNAUD (Research Director), Christine POURCEL (scientist)

Location : I2BC. Department of Genome Biology, Institute for Integrative Biology of the Cell, CEA-CNRS-Université Paris Sud, Bat 400, 91405 Orsay.

Contact: daniel.gautheret@u-psud.fr, Tel: (+33) 1 69 82 62 38

Project summary

CRISPR elements confer bacteria and archaea a immune protection against virus. CRISPR contain a series of repeated DNA sequences separated by variable "spacers". The CRISPR genes are transcribed into long pre-RNAs which are then cleaved by proteins. While thousands of CRISPR loci have been characterized by bioinformatics, we still know very little about their transcripts. How abundant are they? What are their promoters and terminators? From which DNA strand are they produced? Thousands of bacterial RNA-seq libraries are now publicly available and can help us answer these questions.

Objectives:

1. We will set up a bioinformatics pipeline for scanning any bacterial RNA-seq library, identify reads mapping to CRISPRs and characterize all CRISPR RNAs.
2. Using this protocol on a large selection of libraries, we will compile a collection of expressed CRISPR loci, including the CRISPR RNA sequence, coordinates of transcribed regions, transcript abundance, strand and species/taxa.
3. We will use this result to train a machine learning classifier aiming at predicting the transcribed orientation of a CRISPR locus based on its genomic sequence.

Technical skills requested : Candidates should be in capacity to develop clean and reproducible Unix pipelines running on cluster or cloud infrastructures. They should understand the basic concepts of machine learning and hold a Bioinformatics M.Sc. level in R and/or Python programming. Finally they should be motivated by a unique opportunity to contribute to the understanding of CRISPR systems function and evolution.

Team: Our team is specialized in RNA bioinformatics and all its applications, from human health to bacterial/archaeal biology. The internship will be co-supervised by Gilles Vergnaud et Christine Pourcel, who are among the discoverers of the wonderful CRISPR system and authors of the highly cited CRISPRdb database.

References of team in connection to project :

1. Couvin D, Bernheim A, Toffano-Nioche C, Touchon M, Michalik J, Néron B, C Rocha EP, Vergnaud G, Gautheret D, Pourcel C. (2018) CRISPRCasFinder, an update of CRISPRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Res*. **46**:W246-W251.
2. Audoux J, Philippe N, Chikhi R, Salson M, Gallopin M, Gabriel M, Le Coz J, Commes T, Gautheret D. (2017) DE-kupl: Exhaustive capture of biological variation in RNA-seq data through k-mer decomposition. *Genome Biol.* **18**: 243.
3. Pain A, Ott A, Amine H, Rochat T, Bouloc P, Gautheret D. (2015) An Assessment of Bacterial Small RNA Target Prediction Programs. *RNA Biol.* **12**:509-13.
4. Toffano-Nioche C, Ott A, Crozat E, Nguyen AN, Zytnicki M, Leclerc F, Forterre P, Bouloc P, Gautheret D. (2013) RNA at 92°C - The non-coding transcriptome of the hyperthermophilic archaeon Pyrococcus abyssi. *RNA Biology* **10**:1211-20.
5. Toffano-Nioche C, Luo Y, Kuchly C, Wallon C, Steinbach D, Zytnicki M, Jacq A, Gautheret D. (2013) Detection of non-coding RNA in bacteria and archaea using the DETR'PROK Galaxy pipeline. *Methods*. **63**:60-5.

PhD continuation: to be discussed.

Duration : 6 months.

Monthly stipend : 540€