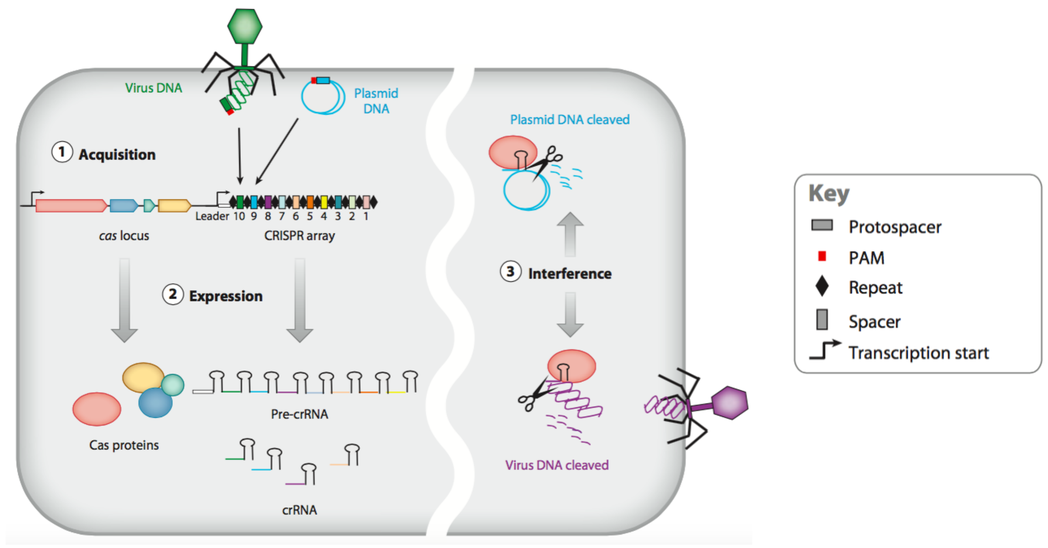
From CRISPR to Prophage: A Bioinformatics Pipeline

A project by: Hamed Borhani, Mohamedhakim Elakhrass, Yi Ming Gan, Cedric Lood

Background: CRISPR-Cas System



The Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) and CRISPR-associated (Cas) proteins comprises a multistep process by which specific small fragments of foreign nucleic acids are first recognized as being foreign and incorporated into host CRISPR region between short DNA repeats.

Subsequently, these fragments, also known as spacers are expressed as CRISPR RNA (crRNA). crRNA in conjunction with host Cas proteins, are used as a surveillance and adaptive immune system by which incoming foreign nucleic acids are recognized and destroyed.

Objectives

The objectives of this project were to look for a possible prophage that is able to escape the known CRISPR-Cas system and to quantify biological dark matter indirectly.

Intermezzo: Reproducibility

The project included the use of 10 different third-party softwares, a code base of over a thousand lines of code (bash and python), and different libraries. In order for everyone in the group to run the different pipelines, and to enable reproducibility of the results, a standard Linux configuration was created using Vagrant and Ansible. The code can be found on Github along with complete documentation on the repository of the project.

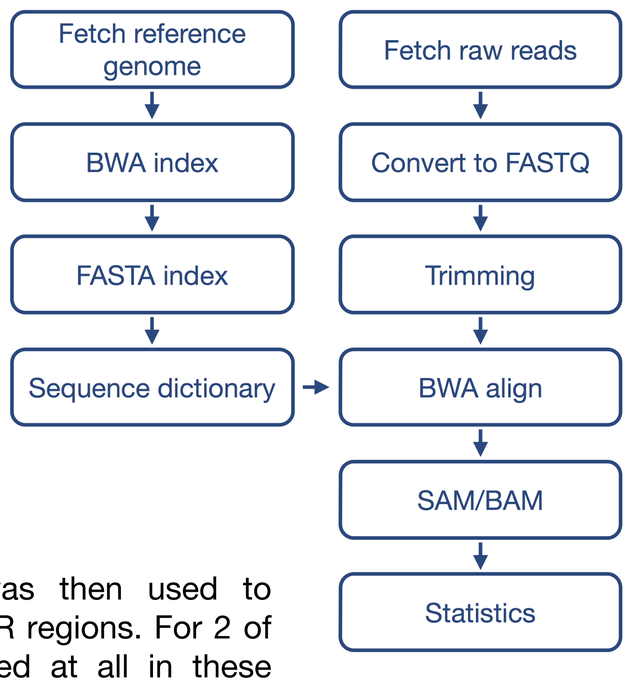
Repository address: *https://github.com/Milt0n/TheOmicians*

CRISPR 101

Three softwares that can locate CRISPR regions in genomes were investigated: CRT, PILER-CR, and CRISPRFinder. They returned highly similar results in terms of array detection for the following bacteria:

- Akkermansia Muciniphila
- Bacteroides Fragilis
- Faecalibacterium Prausnitzii
- Methanobrevibacter Smithii

Next, a Sam/Bam pipeline was developed to map raw reads from resequencing projects for the 4 bacteria. The pipeline consisted of 2 scripts, one that fetched the reference genome (NCBI Genome db) and processed it, and another one that fetched the raw reads (NCBI SRA db) and produced the Sam/Bam files.



The Integrated Genome Viewer was then used to visualize the mappings in the CRISPR regions. For 2 of the bacteria, no reads were mapped at all in these genomic regions. It is hypothesized that this is a result of the high dynamicity of the CRISPR regions where spacers can change rapidly.

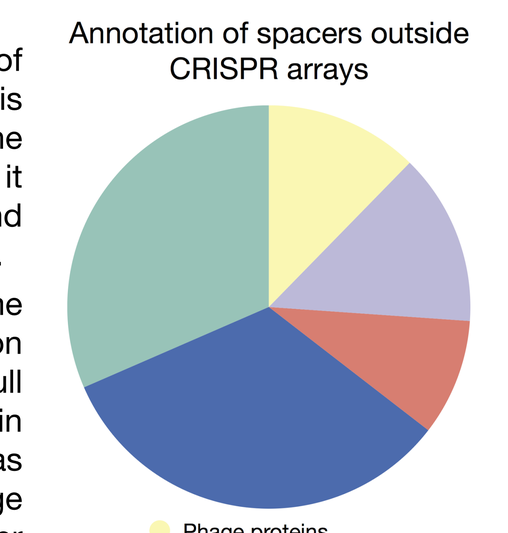
Finally, a search for spacers that would be repeated in the genome outside of the CRISPR regions (herein referred as “hits”) was performed. No results were found, so a decision was made to scale up the pipeline and perform an analysis on many more genomes (see next section).

CRISPR pipeline

The first step of the pipeline was fetching about 5200 genomes from NCBI database using Entrez API and storing each as a local file. Corresponding CRISPRs were then obtained and stored in one file per input by feeding the fetched genomes into the PILER-CR program in the second step.

After that the spacer sequences were extracted from PILER-CR reports and then each genome was searched in order to find a hit. Found hits were filtered to reduce false positive results in next step. Finally annotations for these found hits were obtained from NCBI using ENTREZ in the last step.

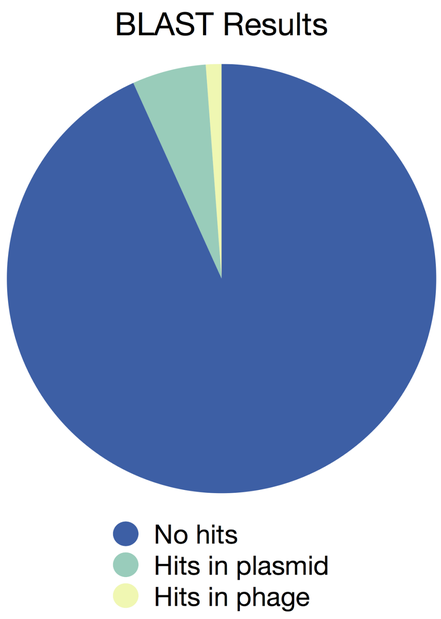
A local nucleotide BLAST was also run using all spacers found by PILER-CR program in the second step as the query in order to gain an insight into the spacers’ possible sources.



Results and discussion

The pipeline identified multiple areas outside of the CRISPR region where spacers occur. This was done using a script that matched the spacers back to the genome of the bacteria it was found in. The original goal was to try and identify bacteriophage genomes using spacers.

Some of these spacers matched to the genome of the bacteria where further investigation suggested the possibility of it being a full bacteriophage genome. These were identified in bacteria that contain the Type I or II CRISPR-Cas systems which do not allow bacteriophage genome integration. This must be further investigated because bacteria with Type I or II CRISPR-Cas system should not have hits.



76127 spacers were identified in the 5167 genomes run through this pipeline. All spacers were blasted to the NCBI nucleotide database filtering for bacteriophages and plasmids, 5615 blast hits were identified. From the 5615 blast hits 80% were located in plasmids and 20% were located in phages. This shows that only 7% of spacers match to a known sequence and 93% are considered biological dark matter.

This pipeline could further be applied to metagenomics experiments. The microbiome and virome of an individual could be sequenced, then the pipeline could be applied to the microbiome data. The spacers that are acquired from the microbiome data can be compared to the virome to see how prevalent certain bacteriophages are.

Authors contributions

Ming and Hakim: Interpretation of biological results
Cedric and Hamed: Pipeline development

CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res. 2007 May 31
Complete references can be found on the dedicated page on the github repository: https://github.com/Milt0n/TheOmicians/blob/master/references.md