Introduction

The evolution of CRISPR–cas loci, which encode adaptive immune systems in archaea and bacteria, involves rapid changes, in particular numerous rearrangements of the locus architecture and horizontal transfer of complete loci or individual modules. These dynamics complicate straightforward phylogenetic classification. Studying the evolution of these systems and quantifying them in the bacterial species will help us to further our understanding in regulating the microflora or regulating the metabolism within the different systems.

Objective

Our goal is to develop a pipeline for metagenome mining of CRISPR-associated (CAS) proteins or CRISPR loci and study their evolution and distribution across the phylogenetic tree (analysis of signature protein families)

Progress

So far we have developed a pipeline which involves:

Steps to explore cas proteins in bacterial genomes (This is a test run, in future we want to explore them in metagenomic datasets).
(As we proceed in this project, We understood about memory issues and low harddisk space)

0) ***Download complete bacterial genomes and associated plasmid sequences from NCBI***

For this project, it is downloaded like this:
(https://www.ncbi.nlm.nih.gov/genome/doc/ftpfaq/#downloadservice)
to download genome FASTA sequence for all RefSeq bacterial complete genome assemblies:

- Start with an "all[filter]" query on Assembly (https://www.ncbi.nlm.nih.gov/assembly)
- Select "Bacteria" from the "Organism group" facet in the left-hand sidebar
- Select "Complete genome" from the "Assembly level" facet in the left-hand sidebar
- Click on the "Download Assemblies" button to open the download menu
- Leave "Source database" set to RefSeq
- Select "Genomic FASTA" from the "File type" menu
- Wait for the "calculating size..." message to be replaced by an estimated size
- Click Download, you may get a pop-up window asking if/where you want to save the genome_assemblies.tar archive file
- After the download has finished, expand the tar archive
- The resulting folder named "genome_assemblies" will contain:

a report.txt file that provides a summary of what was downloaded

a folder named like "ncbi-genomes-YYYY-MM-DD", where YYYY-MM-DD is the date of the download, containing:
a README.txt file
an md5checksums.txt file
many data files with names like *_genomic.fna.gz, in which the first part of the name is the assembly accession followed by the assembly name

1) ***annotate bacterial genome (make faa files), then make protein database using prokka***

2) ***using hmmbuild: build a profile HMM from an alignment (256 cas protein alignment files was used)***

3) ***Using hmmsearch: search a sequence database with a profile HMM***

4) ***Now we have all significant hits' ids from hmmsearch and we need to extract correspond sequences from database for them (This step have done using esl-sfetch)***

5) ***Then we clustered the sequences using psi-cd-hit***
6) used centroids to do multiple sequence alignment using Muscle

7) ***Then maximum-likelihood tree was build using Iq-tree command line***

We annotate the trees using R scripts and prokaryotic description which is downloaded from NCBI database (leaves colored by family level).

The trees are huge and interpreting the tree is not easy.
Then we decided to BLAST all proteins that we have from hmmsearch results against nr database. Significant BLAST matches were scored using cutoff of 1e-5 for the E-value. Then BLAST results were filtered based on 50% similarities. Finally we made trees again. In this step we need to interpret and compare the trees.

Technical problems

There is an issue of memory on the local computer. The entire pipeline is set up as a bash script standalone within the docker container. It can work with multiple metagenomes and run the steps. But there are no docker privileges (as discussed earlier)

Now the pipeline is running in personal computer on Arriam lab's metagenome data

1. Metagenomic reads -> assembly
2. assembly -> taxonomic annotation in CAT
3. assembly + taxonomic annotation -> binning in SolidBin
4. binns (MAGs) -> taxonomic annotation in BAT
5. annotated MAGs -> search for CAS-associated proteins in each MAG individually; use a MAG's taxonomy as a reference for all extracted CAS-associated proteins

1. Construct gene trees for each family and find incongruence between individual gene trees; you can do gene-tree topology clustering. If several gene trees have the same topology, you can argue that they have been inherited/transferred together [This is the evolution of a CRISPR/CAS system within a single lineage]
2. Compare across MAGs: what types of CRISPR/CAS systems there are [This is evolution of CRISPR/CAS a whole: how these systems arise and diversify across the phylogenetic tree of bacteria]

1) We can draw a scheme for your pipeline and describe all dependencies for each step. This will help us workout, what is feasible;

2) I can help you implement a couple of steps, so that you could carry on yourself having our cooperative work as a reference;

3)I've finalized my ideas concerning the two types of evolutionary analysis and I thinks I've figured out a couple of ways to proceed with the second one.