Exploratory analysis of CRISPR-Cas systems in Asgard archaea

Javier Marchena Bioinformatics project

Supervisor: Amelie Stein

Table of contents

1. Abstract

CRISPR-Cas systems provide adaptive immunity against invasive genetic elements in bacteria and archaea. Here, the recently discovered "Asgard" superphylum in archaea is searched for CRISPR-Cas systems. CRISPR protein profiles from other archaeal species were used to find CRISPR-related Asgard proteins through PSI-BLAST. The neighboring area of PSI-BLAST hits was explored with CRISPRCasFinder and the automatic genomic annotations, unveiling very diverse type I and type III CRISPR-Cas systems. Several systems seem to be fully functional, encompassing various Cas proteins and containing CRISPR arrays.

2. Introduction

CRISPR-Cas are adaptive immune systems that protect bacteria and archaea against viruses and other invasive genetic elements (Barrangou et al., 2007; Marraffini et al., 2008). The CRISPR-Cas machinery integrates fragments of virus or plasmid DNA into CRISPR arrays, where they become spacers between repeats, and uses processed transcripts of these spacers (crRNAs) as guides to detect and degrade the invading element (Hille et al., 2018). The CRISPR immune response consists of three consecutive stages: (1) adaptation (i.e., incorporation of new spacers into CRISPR arrays), (2) crRNA maturation from pre-CRISPR RNAs to mature crRNAs, and (3) interference, i.e., recognition and cleavage of the target DNA or RNA (Mohanraju et al., 2016; Amitai & Sorek, 2016). CRISPR–Cas systems are divided into two classes based on their effector complex composition. Class 1 systems employ multi-subunit effector complexes that consist of multiple Cas proteins, whereas in Class 2 systems, the effector is a single multi-domain protein (Makarova et al., Nat Rev Microbiol 2020). Inside each of these two classes, there is a subclassification into 3 CRISPR-Cas types (6 types in total), and each type is further ramified (**Figure 1**).

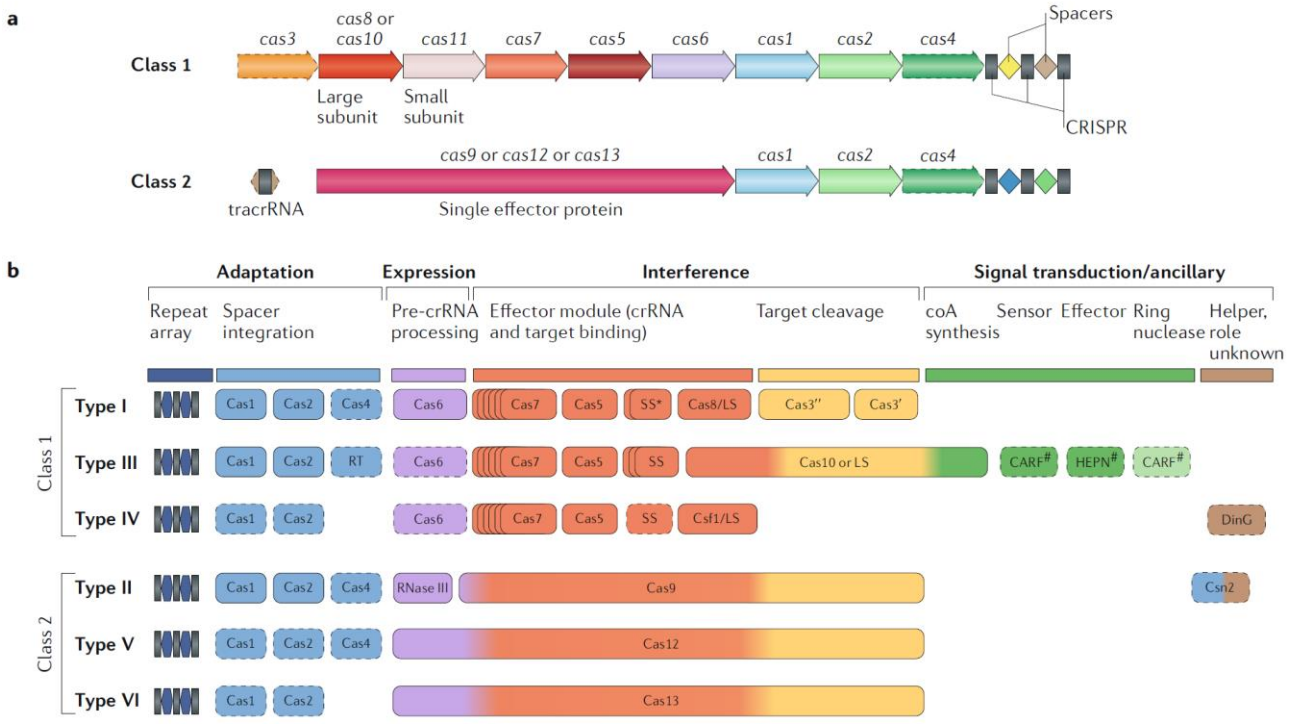

Figure 1. The two classes of CRISPR-Cas systems and their modular organization. **a)** General features of each class. **b)** Modular organization of each CRISPR-Cas class and type. Taken from Box 1 in Makarova et al., Nat Rev Microbiol 2020.

Since their discovery, CRISPR-Cas systems have been extensively sought and studied, mainly due to their potential as gene editing tools, but also in order to understand microbial defense mechanisms and physiology. Apart from adaptive immunity, CRISPR-Cas systems have been found to carry out a plethora of physiological functions such as gene regulation, DNA repair and cell dormancy (Westra et al., 2014). Thus, it is logical to analyze CRISPR-Cas systems of newly discovered microbes, in order to not only evaluate their potential as genome editing tools, but also to study their physiology.

The 'Asgard' superphylum in archaea is a recently discovered superphylum in archaea. The first phylum inside Asgard archaea, Thorarchaeota, was discovered on 2015 in a metagenomic study of deep marine sediments (Spang et al., 2015). Subsequently, four other phyla were discovered: Thorarchaeota (Seitz et al., 2016), Odinarchaeota and Heimdallarchaeota (Zaremba-Niedzwiedzka et al., 2017) and Helarchaeota (Seitz et al., 2019). Those phyla were grouped under the Asgard superphylum.

Asgard archaea associate with eukaryotes in phylogenomic analyses, and their genomes are enriched in eukaryotic proteins (Zaremba-Niedzwiedzka et al., 2017). Consequently, it has been hypothesized that an Asgard archaeon might have been the host cell from which eukaryotes emerged.

Until short ago, efforts to cultivate Asgard archaea were not successful, and therefore almost all genomic samples proceed from metagenomics analyses. These analyses are carried out by Next-Generation sequencing methods, usually Illumina sequencing technology, and therefore the resulting genomes are metagenome-assembled genomes (MAGs), represented by contigs that do not cover the full genome. Very recently, a representative of the Asgard superphylum was cultivated, providing the first complete Asgard genome (Imachi et al., 2020).

Apart from canonical, interferent CRISPR-Cas complexes, many derived, defective variants have been observed, which lack some previously considered essential elements (Westra et al., 2014). Many of these variants perform tasks other than defense against invasive elements, such as gene regulation and DNA repair.

Some examples of those variants have been found in CRISPR-Cas systems with HRAMP (Haloarchaeal Repeat Associated Mysterious Proteins) family proteins (Makarova et al., 2019). These are Haloarchaeal proteins resembling RAMP (Repeat Associated Mysterious Proteins) family proteins, which are characterized by extreme sequence diversity and by containing an RNA-recognition motif (RRM). HRAMPs contain Cas5 and Cas7 variants, a HRAMP-specific domain, and Deddy and HNH nucleases. Surprisingly, CRISPR-Cas systems with HRAMPs were found to contain no CRISPR arrays or other adaptation genes (Makarova et al., 2019). However, they are predicted to act as an RNA-based defense mechanism, with their effector mechanism still unknown.

The objective of the present project is to detect CRISPR-Cas systems in Asgard archaea and describe those systems. Starting from CRISPR-Cas related proteins in other archaeal species, CRISPR-Cas systems in Asgard archaea are identified and analyzed.

This study has a precedent (Makarova et al., CRISPR J. 2020). In that study, CRISPR-Cas systems in Asgard archaea were searched by similar methods, but some results differ, and the viewpoint of the result analysis also varies. Like in HRAMPs, ARAMPs (Asgard Repeat Associated Mysterious Proteins) were also found to lack CRISPR arrays or other adaptation genes. In addition, an

intriguing relationship was discovered between Cas1 proteins in Asgard archaea and TnpB-like nucleases.

3. Methods

arCOGs (Archaeal Clusters of Orthologous Genes, Makarova et al., 2015) were selected as the CRISPR-related proteins in other archaeal species. arCOGs group similar archaeal genes into clusters, annotate them and create multiple alignments for each cluster (also called arCOG profiles). The 95 arCOGs that were annotated to be CRISPR-related were deemed as "CRISPR arCOGs" and were the starting point for searching CRISPR-Cas systems in Asgard proteomes. The "CRISPR_arCOGs.ipynb" script removed all arCOGs that were not annotated as CRISPR-related, leaving only CRISPR arCOGs.

arCOGs are available at https://ftp.ncbi.nih.gov/pub/wolf/COGs/arCOG/ where arCOG profiles are accessible in the zip.aliar14.tgz file and annotations are accessible in the ar14.arCOGdef.tab file.

All proteins in Asgard archaea were downloaded from the NCBI Protein database, available at https://www.ncbi.nlm.nih.gov/protein/?term=txid1935183[Organism:exp. Perhaps due to its novelty, the Helarchaeota phylum is not present in the NCBI Protein database. Therefore, this study did not include proteins from the Helarchaeota phylum. The Asgard proteome was downloaded in fasta format and converted to a database by using the makeblastdb command. The resulting database contains 138500 proteins. Asgard genome assemblies were downloaded from the NCBI Assembly database, available at https://www.ncbi.nlm.nih.gov/assembly/?term=txid1935183[Organism:exp]. 136 genome assemblies were downloaded. As mentioned in the introduction, only 1 of those is a complete genome, while the rest are metagenome-assembled genomes.

Thereafter, arCOG profiles were employed to search that database by using PSI-BLAST (Altschul et al., 1997). This was carried out in the "PSI-BLAST.ipynb" script. In each PSI-BLAST search, one arCOG profile was used as query, with the Asgard proteome as database, a threshold E-value of 0.001, and otherwise default parameters. Each Asgard protein found by this procedure was called a hit.

Subsequently, all hit headers (lines starting with a > sign) were gathered in the "CRISPR_asgard_hits.txt" file. Hit headers gather the most important information about the hit: its Genbank ID, its phylum and/or species, and its definition (an automatic annotation of its function). This information was used to obtain the results illustrated in the "Asgard hits analysis" section.

Next, the neighborhood of each hit was analyzed. The 3 closest gene products (usually, proteins) on each side (5' and 3') were studied, for a total of 6 gene products around each hit. The SeqIO module of Biopython was employed to parse the genomes. This procedure was accomplished in the "hit_neighborhoods.ipynb" script.

Those 20000 bp around each hit are stored in the dnaseq_hits.fasta file, each header being the Genbank ID of the hit. Those DNA sequences were analyzed with CRISPRCasFinder (Couvin et al., 2018). Specifically, the CRISPRCasMeta online tool was used (available at https://crisprcas.i2bc.paris-saclay.fr/CrisprCasMeta/Index) with default parameters. Apart from CRISPR arrays, this tool also detects CRISPR-related proteins in general, and thus, its findings can be compared to the protein descriptions in the genome assemblies. The results can be found in the directory "Result analysis/crisprcasfinder_results". Those results can be viewed in a very user-friendly interface by uploading the result.json file to CRISPRCasFinder Viewer, available at https://crisprcas.i2bc.paris-saclay.fr/CrisprCasFinder/Viewer.

All plots were generated through the "Result_analysis.ipynb" script, using the Matplotlib library. Figure 7 (CRISPR-Cas systems in Asgard archaea) was created by using the GenomeDiagram module in Biopython.

4. Results

The procedure to find Asgard CRISPR-Cas systems consisted of selecting CRISPR-related proteins in other archaeal species (arCOGs), and searching through PSI-BLAST for similar proteins in Asgard archaea genomes. Each Asgard protein found by this procedure was called a hit. 505 CRISPR hits were found in the Asgard proteome.

4.1. Analysis of CRISPR hits

4.1.1. CRISPR hits by Asgard phylum

The Asgard superphylum contains 5 phyla: Heimdallarchaeota, Thorarchaeota, Odinarchaeota, Lokiarchaeota and Helarchaeota. Thus, it would be interesting to know how many hits correspond to each phylum. As mentioned in the Methods section, the Helarchaeota phylum was not included. It is also worth noting that some proteins did not indicate any association with a specific phylum. Those proteins were regarded as "Unclassified".

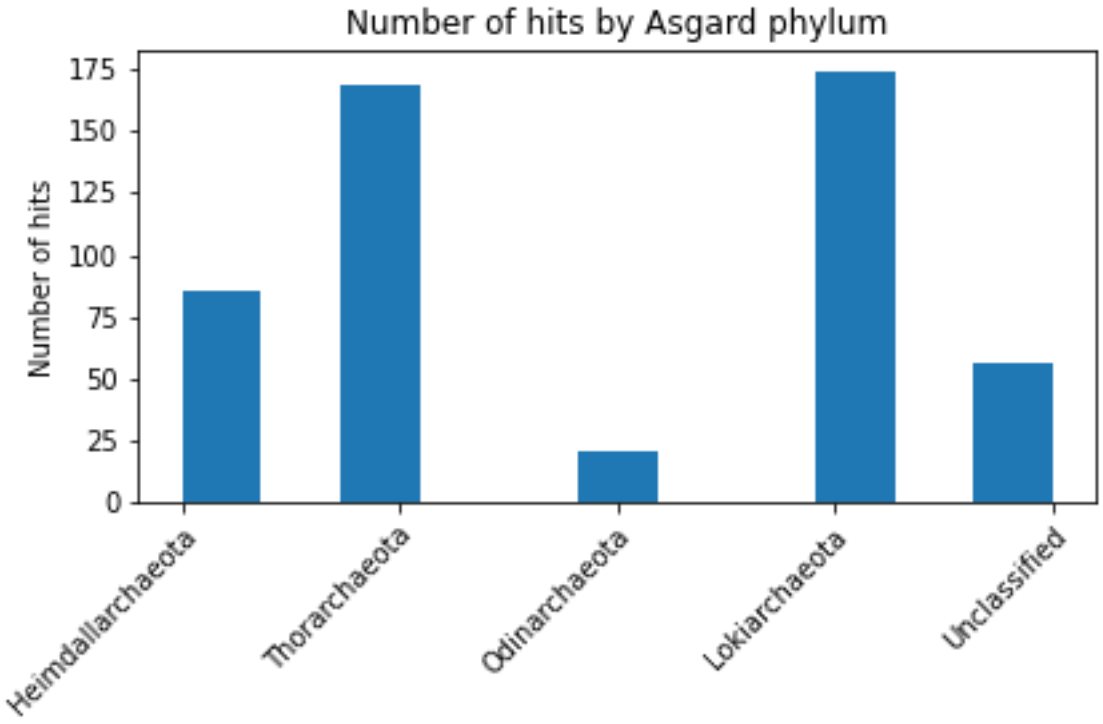

Figure 2. CRISPR hits by Asgard phylum.

The Asgard phyla with the highest number of CRISPR hits are Lokiarchaeota and Thorarchaeota, while the phylum with fewest CRISPR hits is Odinarchaeota (**Figure 2**).

However, this superficial analysis does not consider the fact that the diverse phyla have different representations in the Asgard protein database. For example, Lokiarchaeota is the most represented phylum, while Odinarchaeota is the least represented. This means that the likelihood of encountering a hit in Lokiarchaeota, even by pure chance, is much higher than in

Odinarchaeota. In order to eliminate this over-representation effect, the number of hits in each phylum was divided by the total number of proteins in each phylum. This normalization was called "CRISPR-Cas enrichment", because it indicates how much each phylum is enriched in CRISPR-related proteins.

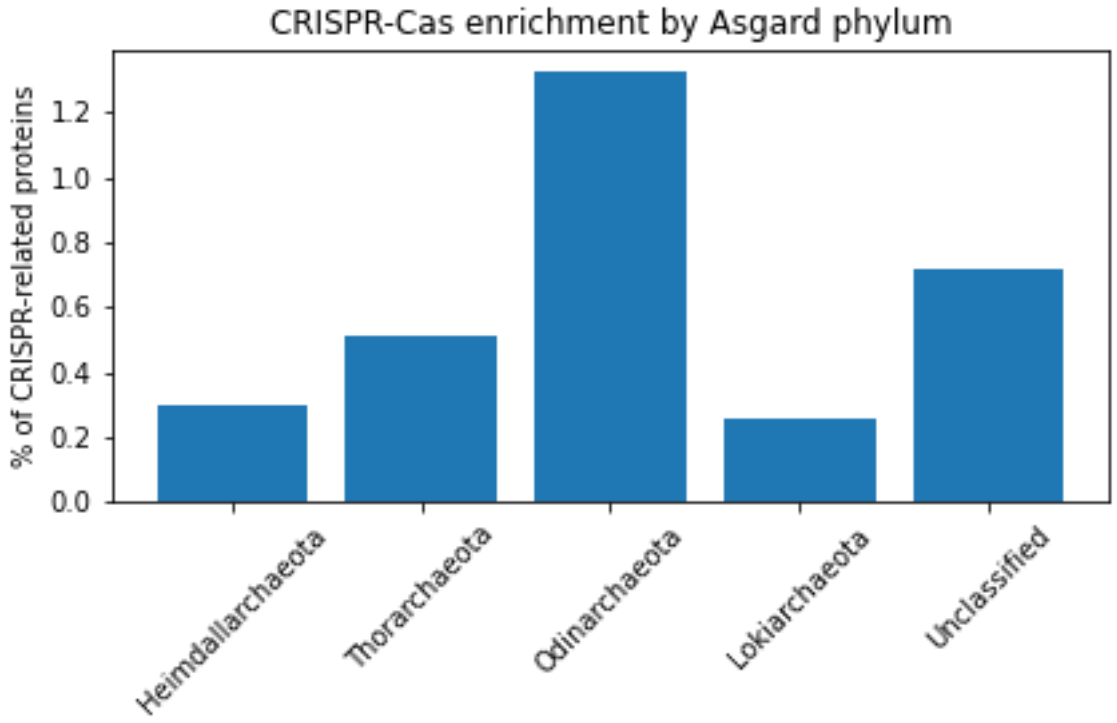

Figure 3. CRISPR-Cas enrichment by Asgard phylum.

The CRISPR-Cas enrichment representation in **Figure 3** reveals a very different outlook from that given by **Figure 2**. Odinarchaeota is the most CRISPR-enriched Asgard phylum, while Lokiarchaeota is the least CRISPR-enriched Asgard phylum. The enrichment of Odinarchaeota should possibly be judged cautiously, given the small sample size of Odinarchaeota proteins. There are only 1584 Odinarchaeota proteins out of all 138500 proteins.

4.1.2. CRISPR hits by category

Automatic annotations in genomes predict the type of protein that each gene product belongs to. Here, the predicted categories of Asgard hits are analyzed.

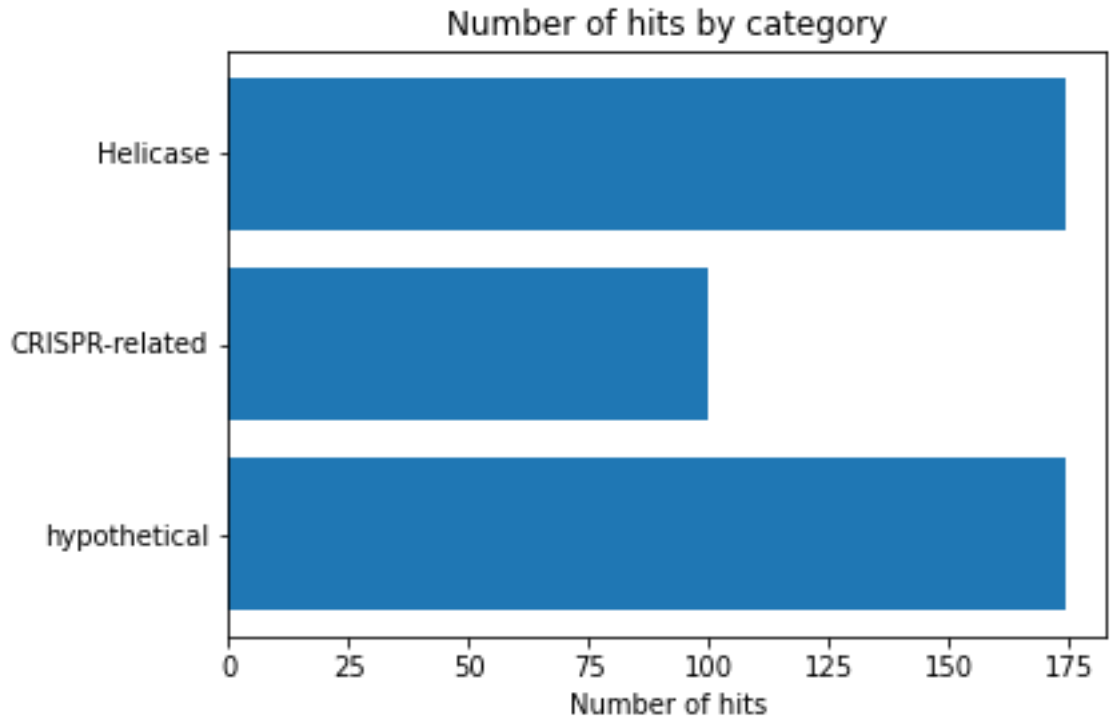

Figure 4. Histogram of Asgard CRISPR hits by category.

When automatic annotation tools could not predict the function of the protein, the label "hypothetical protein" is assigned. It is quite a common occurrence, and it does not necessarily mean that the protein is hypothetical and not real.

On the other hand, many Asgard hits were classified as Helicases. Most commonly, either DEAD/DEAH box helicases or ski2-type helicases. This could be explained in 2 ways: 1) these hits are false positives, that is, they are actually helicases that, due to their sequence similarity to CRISPR-related proteins, appeared in the PSI-BLAST searches, or 2) a misclassification by the annotation algorithm, because they are actually CRISPR-related proteins.

In any case, multiple hits are still classified as CRISPR-related, and many of those annotated as "hypothetical protein" could also be CRISPR-related.

4.1.3. CRISPR hits by CRISPR-type

Sometimes, even though not very often, in the protein type annotation, the CRISPR-type of CRISPR-related proteins is specified.

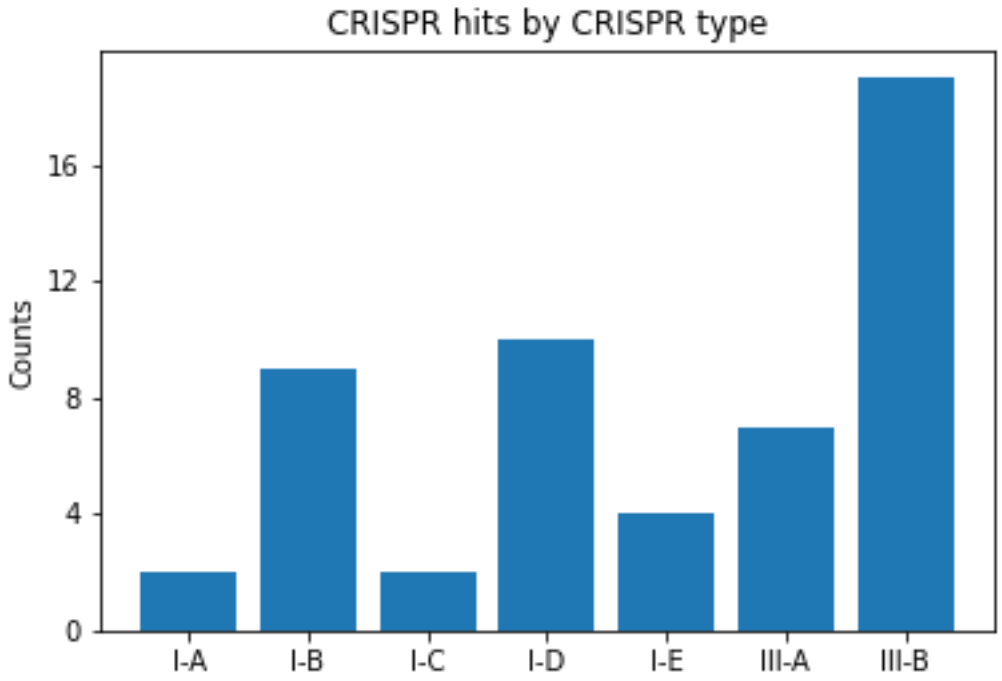

Figure 5. Histogram of Asgard CRISPR hits by CRISPR type.

Among Asgard CRISPR hits that reported the CRISPR type to which they belong, only types I and III were present, and in the same proportion (28 type III and 27 type I). Type III-B was the most widespread sub-type.

4.2. Hit neighborhood analysis

Another objective was to analyze the "neighborhood" of each hit, that is, to study what proteins or gene products are present in the genome region around each hit. For each hit, the 6 neighboring gene products were analyzed.

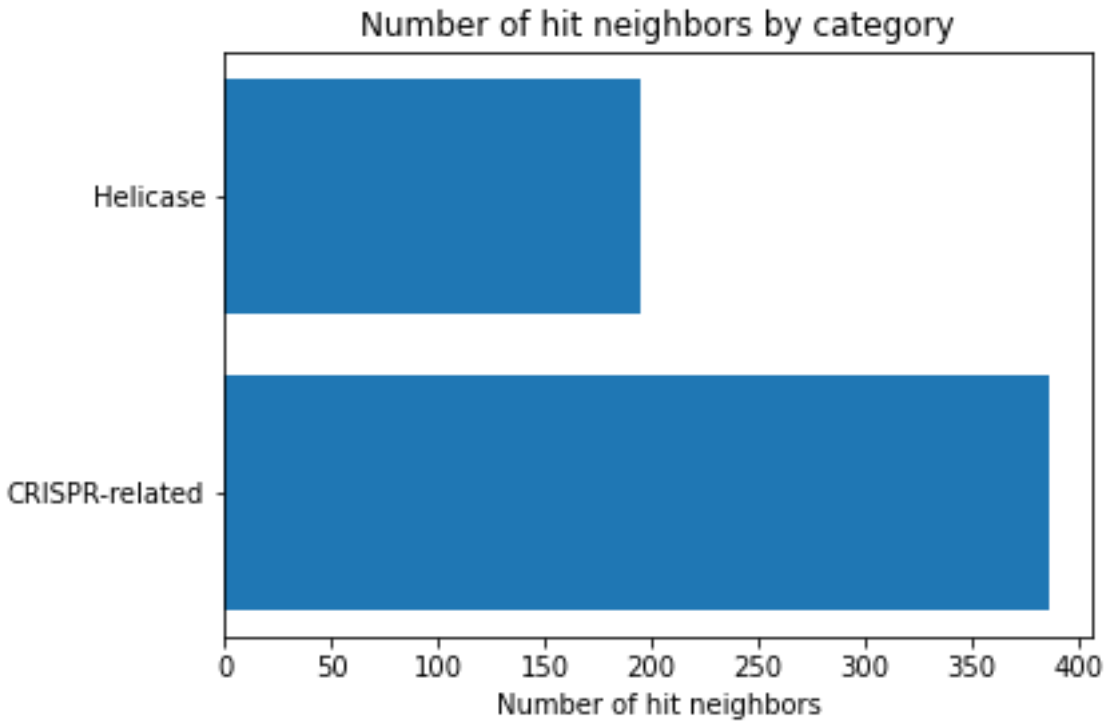

Figure 6. Histogram of Asgard CRISPR hit neighbors by CRISPR type.

In Figure 6, the "hypothetical" category was removed because there were many more of those cases. This is quite logical, because many of the neighbors will not be CRISPR-related or will not have a particularly clear function.

In this neighbor analysis by category, there are many more CRISPR-related proteins than predicted helicases, compared with the analysis of hits. This is logical, because it means that many of the neighbors are CRISPR-related proteins.

When looking at the "neighborhood_hits_products.txt" file, helicases are isolated, mostly surrounded by "hypothetical proteins". This signals that those helicases were perhaps false positives hits, that are not CRISPR-related, because otherwise some of those helicases would be surrounded by CRISPR-related proteins.

4.3. CRISPR-Cas systems in Asgard archaea

A final aim was to search for CRISPR arrays in the Asgard genome. As CRISPR arrays are expected to be in a zone nearby CRISPR-Cas systems, the 20000 bp around each hit (10000 bp on each side) were analyzed with CRISPRCasFinder.

Additionally, CRISPRCasFinder predicts and annotates CRISPR-related proteins. Therefore, this is a second source of CRISPR annotation, apart from the automatic annotations in the genome assemblies.

Various full CRISPR-Cas systems were detected in Asgard archaea. **Figure 7** shows the most relevant CRISPR-Cas systems in Asgard archaea.

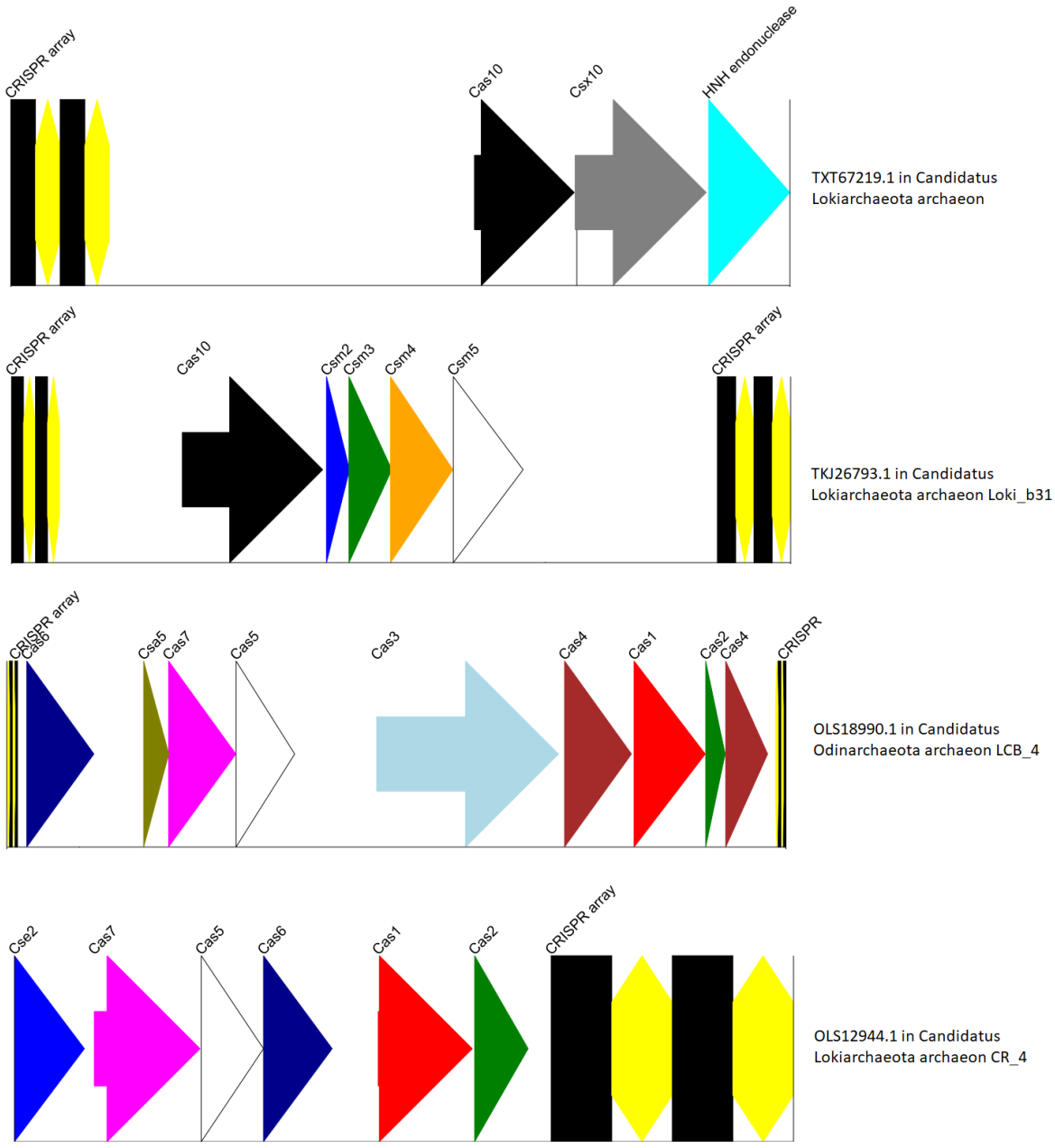

Figure 7. Main CRISPR-Cas systems in Asgard Archaea. In each system, the label is the protein ID of the leading Asgard hit in the system followed by the species in which the system is located. The annotations are mainly based in CRISPRCasFinder annotations, but they were occasionally combined with automatic annotations in the genome assemblies.

Each system is labeled by the protein ID of the "leading hit", which is the Asgard hit that appears the first in the system (most on the 5' side, on the left side of the image) and the species that harbors the system.

All these systems contain CRISPR arrays, either on one side or on both sides of the CRISPR proteins. In the two systems shown on top (TXT67219.1 and TKJ26793.1) in Figure 7, it could be argued that those CRISPR arrays are not adjacent to the rest of the CRISPR system. However, the presence of full CRISPR-Cas systems in their relative vecinities suggest the CRISPR arrays as being part of those systems.

The TXT67219.1 system in Candidatus Lokiarchaeota archaeon is a derived type III CRISPR-Cas system. The presence of the signature protein Cas10 indicates a type III system (Molina et al., 2020). Nonetheless, this system does not contain the typical Cmr or Csm complex proteins. Instead, it contains Csx10 and a HNH endonuclease. Possibly, in this system Cas10 recognizes the invading nucleic acid, while the HNH endonuclease could be chiefly responsible for the cleavage.

The TKJ26793.1 CRISPR-Cas system in Candidatus Lokiarchaeota archaeon Loki_b31 contains a full type III-A CRISPR-Cas system with Csm complex. It has the prototypical architecture of a Csm complex: a large Cas10 protein (also called Csm1) and the general 5-subunit Csm complex Csm1-Csm5 (Rouillon et al., 2013).

The OLS18990.1 system in Candidatus Odinarchaeota archaeon LCB_4 appears to be a prototypical type I-A CRISPR-Cas system (Makarova et al., Nat Rev Microbiol 2020). It contains the signature protein Csa5, surrounded by Cas7 and Cas5, which form the effector complex. The system also contains a Cas1-Cas2-Cas4 module, typical of this CRISPR sub-type. Only the characteristic protein Cas8a1 is missing for this system to be an absolutely paradigmatic type I-A CRISPR-Cas system. Regarding the CRISPR arrays, this system is a slightly unfortunate case in this regard, because the genomic contig is truncated in the middle of the CRISPR arrays. Therefore, probably the CRISPR arrays in this system are actually larger than depicted here. This system is fully located on the reverse strand.

The OLS12944.1 system in Candidatus Lokiarchaeota archaeon CR_4 is a type I-E CRISPR-Cas system. It contains the signature protein Cse2, as well as other proteins that can be part of type I-E systems, like Cas7, Cas5, Cas6, Cas1 and Cas2. The biggest deviation is the absence of the other signature protein Cse1, but that can be explained by the fact that there is an assembly gap in the genome before Cse2, from -2229 bp to -769 bp with respect to the start of Cse2.

Those 4 are the most relevant CRISPR-Cas systems in Asgard archaea, due to the completeness of the systems. However, there were some other possible CRISPR-Cas systems that were not included in Figure 7 due to their lack of completeness and in order to give precedence to the most relevant systems. RLI61793.1 in Candidatus Lokiarchaeota archaeon contains a CRISPR array, Cas1 and Cas2. However, no more proteins were identified by CRISPRCasFinder or the automatic genome annotations, and therefore the system seems slightly incomplete. OLS18600.1 in Candidatus Odinarchaeota archaeon LCB_4 only contains CRISPR arrays, but those CRISPR arrays are absolutely unequivocal, as they cover a span of 2000 bp and accommodate 29 spacers.

5. Discussion

CRISPR-Cas immunity is a key defense system against invading nucleic acids in bacteria and archaea. Apart from a defense mechanism, CRISPR-Cas immunity also regulates a wealth of physiological processes (Westra et al., 2014). Recently, CRISPR-Cas systems have gained attention owing to their promise as gene editing tools. In this study, CRISPR-Cas systems and CRISPR-related proteins in the Asgard superphylum of archaea are revealed and analyzed. Starting from CRISPR-related protein profiles in other archaea, CRISPR-related proteins in Asgard archaea were revealed through PSI-BLAST. Afterwards, neighboring regions of those Asgard hits were analyzed, uncovering CRISPR arrays and full CRISPR-Cas systems.

When analyzing metagenomic assemblies of recently discovered superphyla like Asgard archaea, two additional challenges arise: 1) the scarce genomic data available (there are 136 incomplete genomic assemblies of Asgard archaea) and 2) the small contig size in the assemblies. Out of the 4 systems described in Figure 7, 2 of them were considerably affected by the small contig size, which did not allow a complete overview and description of the system. Hence, there are presumably many more undiscovered CRISPR-Cas systems in Asgard archaea.

In this study, was the applied methodology ideal? The biggest question is perhaps whether the CRISPR-Cas system detection could be carried out simply by uploading the full Asgard genomes into CRISPRCasFinder. In this way, potentially more systems could be discovered. On the other hand, it is possible that many false positives would be discovered, making the CRISPRCasFinder result too lengthy to inspect it.

Due to time constraints, this option could not be properly investigated; however, a small exploration was performed. The full Asgard genome assemblies were submitted to CRISPRCasFinder as a fasta file. The results can be found under "Result Analysis/crisprcasfinder_full_asgard_genomes_result.json". The results were what might be expected: on the one hand, lengthy to examine due to the presence of many CRISPR arrays and clusters, with many that seem to be false positives, but on the other hand, various new CRISPR-Cas systems that were not discovered with the previous methodology. These new systems possibly contain proteins that are not so similar to other archaeal CRISPR-related proteins, and that might be why they were not detected during the PSI-BLAST. Therefore, the main methodology used in this study is more selective, as there seem to be less false positives, but fails to notice some CRISPR systems. In case that the user in more interested in finding many systems and does not object to having many false positives (most of them can be distinguished), this simple methodology of uploading genomes to CRISPRCasFinder should be superior. Admittedly, the reason why I did not perform this earlier is simply that I discovered the CRISPRCasFinder tool at an advanced stage in the project.

The CRISPR-related proteins found in this project are relatively distributed among the different Asgard phyla. However, the Odinarchaeota phylum is particularly enriched in CRISPR-related proteins (**Figure 3**), in spite of the relatively small number of proteins reported in this phylum. Therefore, in case of obtaining more genomic samples from this phylum in the future, its CRISPR-Cas systems are probably worth researching.

The discovery of CRISPR arrays in Asgard genomes is particularly remarkable, given that Makarova et al. (CRISPR J. 2020) affirmed that none of the CRISPR-related proteins in Asgard archaea were found adjacent to CRISPR arrays. The causes of this discrepancy are not clear. It is true that in that study they did not use exactly the same methods as I did in this project. For instance, they use the

minCED tool in order to search for CRISPR arrays, while I used CRISPRCasFinder. Nevertheless, the CRISPR arrays discovered in the current project are mostly unequivocal, long and containing many repeats and spacers, and therefore should definitely be detected by any tool. The most probable explanation is that they analyzed slightly different genomic sequences. In Supplementary Table 1 of that article, the data used in the study is shown. For example, they use the genome of Candidatus Odinarchaeota archaeon LCB_4, in which 1 CRISPR-Cas system is reported here. However, for some reason, they only report the use of 5 contigs, while there are many more in the genome. Possibly, they employed the BioSamples NCBI database, instead of the Assembly NCBI database I used. In any event, this difference in the genomic sequences analyzed seems to be the reason why in that article many CRISPR arrays were not spotted.

HNH endonucleses, like the one presented here in the TXT67219.1 system in Candidatus Lokiarchaeota archaeon (**Figure 7**), were reported to be very widespread in Haloarchaeal CRISPR-Cas systems (Makarova et al., 2019). Thus, certain Asgard CRISPR-Cas systems could be related to Haloarchaeal systems.

None of the DEAD Helicases or ski2-type helicases labeled as CRISPR Asgard hits (**Figure 4**) were found in any CRISPR-Cas system by CRISPRCasFinder or near other CRISPR-related proteins in automatic genomic annotations. Hence, almost certainly, those helicases were false positives, and not CRISPR-related proteins. In order to avoid the occurrence of these false positives, the E-value threshold in the PSI-BLAST search could be lowered. This, however, increases the risk of missing possible true hits.

CRISPR-Cas systems in Asgard archaea display a broad diversity. **Figure 5** shows that Asgard archaea harbor many different sub-types of type I and type III CRISPR-Cas systems. Accordingly, **Figure 7** also reveals a broad diversity, as the 4 described systems all belong to different subtypes: I-A, I-E, III-A and a derived system. This broad diversity was already noted by Makarova et al. (CRISPR J. 2020), where the diversity was reported to be so high that Cas proteins in Asgard archaea were divided into three sub-categories.

6. Future perspectives

If more time had been available, this study could have been expanded in countless ways. To start with, the reported Cas proteins could be analyzed in order to study how similar or different they

are to Cas proteins from other organisms. Phylogenomic analysis could be carried out so as to explain possible evolutionary pathways that produced these Cas proteins. Derived CRISPR-Cas systems could be analyzed as a way to try to explain their function and mechanism.

CRISPR-Cas systems in Asgard archaea prove to be a fruitful field of study, with a very broad diversity, and harboring fully functional as well as derived CRISPR systems.

7. References

References are ordered alphabetically. In case there is more than one reference by a same author, references are ordered by date, from most recent to oldest. The APA format was used.

In the text, articles are referenced by the name of the first author followed by et al., or two authors if there were 2 or less authors. The year of the publication is mentioned, and in case there was more than one publication by the same author in the same year, the journal is also indicated.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research, 25(17), 3389-3402.

Amitai, G., & Sorek, R. (2016). CRISPR–Cas adaptation: insights into the mechanism of action. Nature Reviews Microbiology, 14(2), 67.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., ... & Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. Science, 315(5819), 1709-1712.

Couvin, D., Bernheim, A., Toffano-Nioche, C., Touchon, M., Michalik, J., Néron, B., ... & Pourcel, C. (2018). CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. Nucleic acids research, 46(W1), W246-W251.

Hille, F., Richter, H., Wong, S. P., Bratovič, M., Ressel, S., & Charpentier, E. (2018). The biology of CRISPR-Cas: backward and forward. Cell, 172(6), 1239-1259.

Imachi, H., Nobu, M. K., Nakahara, N., Morono, Y., Ogawara, M., Takaki, Y., ... & Takai, K. (2020). Isolation of an archaeon at the prokaryote–eukaryote interface. Nature, 577(7791), 519-525.

Makarova, K. S., Wolf, Y. I., Shmakov, S. A., Liu, Y., Li, M., & Koonin, E. V. (2020). Unprecedented Diversity of Unique CRISPR-Cas-Related Systems and Cas1 Homologs in Asgard Archaea. The CRISPR Journal, 3(3), 156-163.

Makarova, K. S., Wolf, Y. I., Iranzo, J., Shmakov, S. A., Alkhnbashi, O. S., Brouns, S. J., ... & Koonin, E. V. (2020). Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. Nature Reviews Microbiology, 18(2), 67-83.

Makarova, K. S., Karamycheva, S., Shah, S. A., Vestergaard, G., Garrett, R. A., & Koonin, E. V. (2019). Predicted highly derived class 1 CRISPR-Cas system in Haloarchaea containing diverged Cas5 and Cas7 homologs but no CRISPR array. FEMS microbiology letters, 366(7), fnz079.

Makarova, K. S., Wolf, Y. I., & Koonin, E. V. (2015). Archaeal clusters of orthologous genes (arCOGs): an update and application for analysis of shared features between Thermococcales, Methanococcales, and Methanobacteriales. Life, 5(1), 818-840.

Makarova, K. S., Zhang, F., & Koonin, E. V. (2017). SnapShot: class 1 CRISPR-Cas systems. Cell, 168(5), 946-946.

Marraffini, L. A., & Sontheimer, E. J. (2008). CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. Science, 322(5909), 1843-1845.

Mohanraju, P., Makarova, K. S., Zetsche, B., Zhang, F., Koonin, E. V., & Van der Oost, J. (2016). Diverse evolutionary roots and mechanistic variations of the CRISPR-Cas systems. Science, 353(6299).

Molina, R., Sofos, N., & Montoya, G. (2020). Structural basis of CRISPR-Cas Type III prokaryotic defence systems. Current Opinion in Structural Biology, 65, 119-129.

Rouillon, C., Zhou, M., Zhang, J., Politis, A., Beilsten-Edmands, V., Cannone, G., ... & White, M. F. (2013). Structure of the CRISPR interference complex CSM reveals key similarities with cascade. Molecular cell, 52(1), 124-134.

Seitz, K. W., Dombrowski, N., Eme, L., Spang, A., Lombard, J., Sieber, J. R., ... & Baker, B. J. (2019). Asgard archaea capable of anaerobic hydrocarbon cycling. Nature communications, 10(1), 1-11.

Seitz, K. W., Lazar, C. S., Hinrichs, K. U., Teske, A. P., & Baker, B. J. (2016). Genomic reconstruction of a novel, deeply branched sediment archaeal phylum with pathways for acetogenesis and sulfur reduction. The ISME journal, 10(7), 1696-1705.

Spang, A., Saw, J. H., Jørgensen, S. L., Zaremba-Niedzwiedzka, K., Martijn, J., Lind, A. E., ... & Ettema, T. J. (2015). Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature, 521(7551), 173-179.

Westra, E. R., Buckling, A., & Fineran, P. C. (2014). CRISPR–Cas systems: beyond adaptive immunity. Nature reviews Microbiology, 12(5), 317-326.

Zaremba-Niedzwiedzka, K., Caceres, E. F., Saw, J. H., Bäckström, D., Juzokaite, L., Vancaester, E., ... & Ettema, T. J. (2017). Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature, 541(7637), 353-358.

8. Code and data availability

All code, as well as the most important data files, is available at the Github repository https://github.com/bqn767/CRISPR_Asgard.