Universitat Oberta de Catalunya

“Comparative genomic analysis of CRISPR/Cas systems in *Thermococcus* genomes”

Nombre Estudiante: **Marc Blanch Asensio.**

Plan de Estudios del Estudiante: **Máster en Bioinformática y Bioestadística.**

Área del trabajo final: **Microbiología, biotecnología y biología molecular**.

Nombre Consultor/a**: Paloma Pizarro Tobías.**

Nombre Profesor/a responsable de la asignatura**: Antoni Pérez Navarro.**

Fecha Entrega: **5/01/2021.**

FICHA DEL TRABAJO FINAL

Título del trabajo:	Comparative genomic analysis of CRISPR/Cas systems in Thermococcus genomes.
Nombre del autor:	Marc Blanch Asensio
Nombre del consultor/a:	Paloma Pizarro Tobías
Nombre del PRA:	Antoni Pérez Navarro
Fecha de entrega (mm/aaaa):	05/01/2021
Titulación::	Máster en Bioinformática y Bioestadística
Área del Trabajo Final:	Microbiología, biotecnología y biología molecular
Idioma del trabajo:	Inglés
Palabras clave	CRISPR/Cas, *Thermococcus*, structural features.

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados i conclusiones del trabajo.*

Los sistemas CRISPR/Cas son sistemas inmunes adaptativos extendidos en bacterias y arqueas que confieren protección contra ácidos nucleicos invasores. Sin embargo, más allá de su papel en procariotas, muchas aplicaciones se han derivado de estos sistemas, especialmente en el campo de la ingeniería del genoma. Tal ha sido el impacto de los sistemas CRISPR/Cas en este campo, que se consideran una tecnología revolucionaria. Este impacto ha impulsado el interés por identificar y caracterizar tales sistemas en procariotas. Aun así, los sistemas CRISPR/Cas de muchas procariotas, en particular en arqueas, siguen sin ser explorados.

En el presente estudio, los sistemas CRISPR/Cas de *Thermococcus* se caracterizan mediante un análisis genómico comparativo. *Thermococcus* es un género de aqueas que comprende microorganismos termófilos. Este género presenta interés industrial, principalmente debido a sus enzimas termoestables.

La caracterización de los sistemas identificados en *Thermococcus* reveló su presencia, diversidad, organización y elementos estructurales. Para dicha caracterización se ha requerido el uso de herramientas en línea, creadas especialmente para el análisis de los sistemas CRISPR/Cas, y de programas informáticos destinados a los análisis filogenéticos.

En líneas generales, se ha detectado una elevada presencia y diversidad de sistemas CRISPR/Cas en *Thermococcus*. Los análisis de los CRISPR loci, *cas* loci y sus elementos estructurales también sugirieron probables eventos de transferencia horizontal de genes entre las especies de *Thermococcus*. Además, se ha generado un marco de trabajo para detectar y caracterizar los sistemas CRISPR/Cas en los genomas de procariotas, que puede emplearse para otras especies o géneros.

Abstract (in English, 250 words or less):

CRISPR/Cas systems are adaptive immune systems widespread in bacteria and archaea that confer protection against invading nucleic acids. However, beyond its role in prokaryotes, many applications have derived from these systems, especially in the genome engineering field. Such has been the impact of CRISPR/Cas systems on this field, that they are considered a game-changing technology. This impact has triggered the interest in identifying and characterizing such systems in prokaryotes. Even so, the CRISPR/Cas systems of many prokaryotes, particularly in archaea, remain yet-to-be-explored.

In this study, the CRISPR/Cas systems of *Thermococcus* are characterized through a comparative genomic analysis. *Thermococcus* is an archaeal genus that comprises thermophilic microorganisms. *Thermococcus* are of interest for the industry, mainly owing to their thermostable enzymes.

The characterization of the systems identified in *Thermococcus* revealed their occurrence, diversity, organization and structural features. For said characterization, the use of online tools, specially created for CRISPR/Cas systems analyses, and software aimed at phylogenetic analyses have been required.

Overall, a high occurrence and diversity of CRISPR/Cas systems have been detected in *Thermococcus*. The CRISPR loci, *cas* loci and structural features analyses also suggested likely horizontal gene transfer events among *Thermococcus* species. Furthermore, a pipeline to detect and characterize CRISPR/Cas systems in prokaryotic genomes has been generated, which can be employed for other species or genera.

Índice

Lista de figuras

1. Introduction

1.1 Background.

The Clustered Regularly Interspaced Short Palindromic Repeats (**CRISPR**) along with the CRISPR-associated (**Cas**) proteins constitute the CRISPR/Cas systems [1]. These systems, which are encoded by nearly all archaea and about half of the bacteria, are adaptive immune systems that confer immunity against foreign nucleic acids such as plasmids and bacteriophages (hereinafter referred to as phages)[1] [2].

To be precise, CRISPR/Cas systems are composed of the CRISPR locus and the *cas* locus. The former, the CRISPR locus, is composed of direct repeats (DR) interspaced with certain sequences termed spacers. The length and sequence of DRs is normally the same within a particular CRISPR locus [3]. In contrast, spacers generally differ in length and sequence, and they are complementary to the nucleic acid of phages or plasmids (the spacer complementary sequence that is present in phages or plasmids is known as protospacer) [3]. Upstream the first DR there is an additional component of the CRISPR locus, the leader sequence. These leaders are AT-rich sequences key for the transcription initiation and the process of adaptation (explained below). These sequences differ in length among the CRISPR/Cas systems-encoding microorganisms, ranging from taround 50 base pairs (bp) to few hundred bp [4]. The latter, the *cas* locus, consists of typically several *cas* genes, which differ depending on the CRISPR/Cas system (**Fig. 1A**) [3].

The classification of CRISPR/Cas systems has varied throughout the years as new systems were discovered. Currently, CRISPR/Cas systems are divided into two classes and six types. Class I systems are characterized by the use of a complex of multiple Cas proteins, while class II systems employ a single Cas protein. Moreover, each class is composed of three different types. Class 1 comprises types I, III and IV, while class 2 includes types II, V and VI. Each type is constituted by different Cas and effector proteins (**Fig. 1B**) [2] [5].

The mechanism whereby CRISPR/Cas systems confer immunity relies on three well-distinguished stages [3].

- The first stage is known as **adaptation**. In this process, a sequence of the invading nucleic acid (i.e. protospacer) is incorporated in the CRISPR locus (note that this sequence is then named spacer once inserted in the CRISPR locus). New spacers are regularly inserted between the leader sequence and the first DR. Hence, for every newly acquired spacer, an additional DR is created (**Fig. 1A**) [6].

- The second stage, **expression** (also commonly-termed biogenesis), is based on i) the transcription of the CRISPR locus, thereby generating a long precursor CRISPR RNA (pre-crRNA), and ii) the transcription and subsequent translation of Cas and other effector proteins. Next, Cas proteins process the pre-crRNA into mature crRNAs. Each mature crRNA is based on a single spacer along with its adjacent DR, which is folded creating an RNA secondary structure. It is important to mention that the pre-crRNA processing also varies among the different CRISPR/Cas system types, and the proteins involved in the process differ as well (**Fig. 1B**) [7].

- Finally, the third and last stage, known as **interference**, takes place. Herein, the orchestrated action of the crRNA and Cas proteins allow the recognition and cleavage of target nucleic acids through the complementary binding of the spacers (i.e. crRNA) to their corresponding protospacers. Thus, new infections by the same phage or plasmids are prevented [8]. It is noteworthy mentioning that for this process to occur, it is normally required the protospacer adjacent motif (PAM), which is, by and large, a 2-6 bp-long sequence. Usually, the Cas proteins involved in the cleavage (i.e. nucleases), need to recognize the PAM sequence before cutting the invading sequence (**Fig. 1B**) [9].

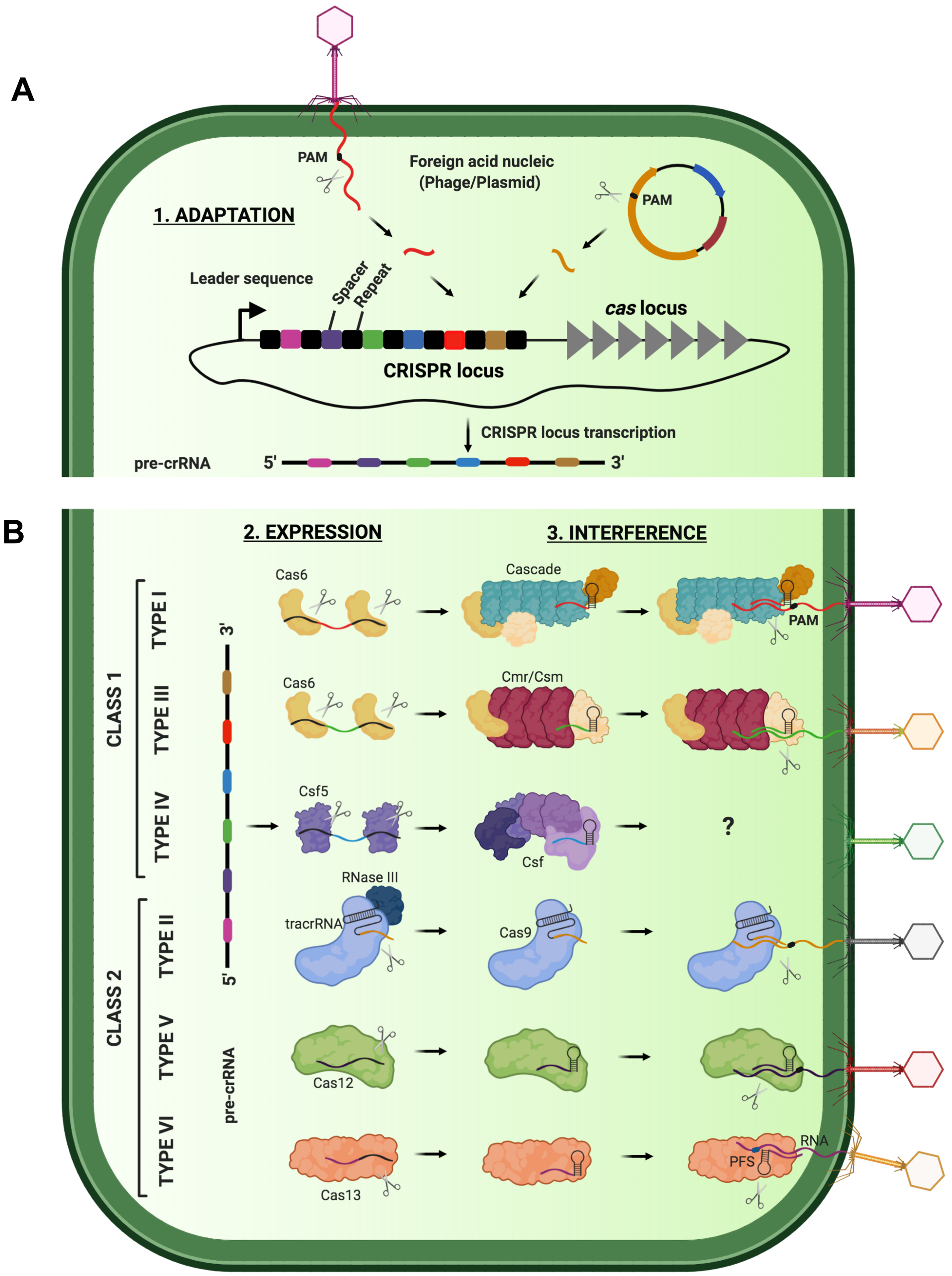


Figure 1. CRISPR/Cas systems constituents, stages and classification. **A**. The adaptation stage and the CRISPR locus and *cas* loci are shown. **B.** The stages of expression and interference are shown. Note that for each class and type, the proteins involved are typically different. For class 1 systems, a protein complex is assembled to process the pre-cRNA and cleave the invading acid nucleic. This complex varies among class 1 types. A complex called Cascade is characteristic of class 1 type I systems, whereas Cmr/Csm complex and Csf complex are characteristic of class 1 type III and class 1 type IV systems, respectively. On the other hand, for class 2 systems, just a single Cas protein participates in the processes. For class 1 type II systems, this protein is the Cas9 (an RNase III also participates), whereas Cas12 and Cas13 are the proteins from the class 2 type V and class 2 type VI systems, respectively [7] [8]. The interference stage for class 1 type IV systems remains unknown. For clarity purposes, this figure has been slightly simplified. For some systems, the interference stage is more complex, meaning that other non-Cas-protein components are involved. This figure has been created using Biorender [https://biorender.com/ (last visited on 21/12/2020)].

Furthermore, CRISPR/Cas systems play other roles beyond adaptive immunity. It has been reported that these systems are also involved in a wide range of mechanisms including microbial gene regulation and virulence, DNA repair, programmed cell death, dormancy, signal transduction, among others [10].

Over the last few years, the research and characterization of CRISPR/Cas systems have intensified exponentially due to the revolutionary application of such systems in the field of genome engineering. In 2012, scientists managed to adapt one CRISPR/Cas system from bacteria, *Streptococcus pyogenes* to be exact, to manipulate DNA from any organism at will. More precisely, they modified the Cas9 protein from the CRISPR/Cas class 2 type II system to cut a specific DNA sequence. Instead of the crRNA, the Cas 9 would use a single guide RNA, which would bind to the complementary target DNA sequence and facilitate the cleavage by Cas9. Through this cleavage, a target gene can be disrupted, obtaining a genetically modified organism in a certainly straightforward way. In addition, the scientists managed to make this procedure programmable, allowing to manipulate any desired sequence in virtually any organism [11]. Said scientists, Jennifer Doudna and Emmanuelle Charpentier, have been awarded with the 2020 Nobel Prize in Chemistry.

Over the past decade, numerous genome engineering technologies and applications have derived from CRISPR/Cas systems, beyond the adapted CRISPR/Cas9 system of S. *pyogenes*. For instance, CRISPR/Cas systems able to target RNA instead of DNA have been discovered and successfully programmed to manipulate desired RNA sequences [12]. Besides, Cas proteins have been modified, disrupting its cleavage activity while preserving its binding activity. Thus, nuclease-inactivated Cas proteins (dCas) were generated. These dCas proteins can interfere in gene regulation by binding themselves to promoter regions and hence repressing the gene transcription. This technique is known as CRISPR interference (CRISPRi) [13]. These are just some examples from the long list of technologies and applications that have been developed as a result of CRISPR/Cas systems research.

In short, these systems have already been and are being extremely useful for the scientific community. Scientists strongly believe in the potential of these systems to address a large number of diseases, especially genetic diseases. Many therapies based on CRISPR/Cas are already being tested in human clinical trials, pointing to promising results. That leads to the fact that many pharmaceutical and biotechnological companies have turned their attention to these game-changer systems. Therefore, there is plenty of interest in the characterization of CRISPR/Cas systems and the discovery of new ones as well [14].

A method to study CRISPR/Cas systems is via computer-based approaches. Bioinformatics has played a key role in the discovery of CRISPR/Cas systems in the genomes of prokaryotes, and also in identifying the functions they carry out. A myriad of bioinformatics online tools has been developed and tailored to extremely simply the tasks of identification and characterization of CRISPR/Cas systems harbored in prokaryotes genomes. These tools, as well as CRISPR/Cas-based databases, are freely available to the scientific community [15].

This study intends, through bioinformatics analysis, to characterize the CRISPR/Cas systems of a genus or species of microorganisms with medical or industrial interest that code for hitherto uncharacterized CRISPR/Cas systems. After a thorough literature search, *Thermococcus* was potentially identified as the candidate genus to focus on this study.

The genus *Thermococcus* belongs to the domain Archaea. Concisely, members of *Thermococcus* are thermophilic microorganisms that characteristically grow at temperatures between 60 to 105°C [16]. Not surprisingly, the enzymes of these thermophiles operate smoothly at high temperatures. This characteristic is of great importance for the industry since many industrial and biotechnological processes easily reach elevated temperatures. Thermophilic enzymes notwithstanding resist well denaturation and remain stable and functional [17]. In consequence, several *Thermococcus* enzymes are commonly used in industrial processes and reactions. For example, the enzyme Tk-SP from *Thermococcus kodakarensis*, which is a subtilisin homolog, presents enzymatic activity at high temperatures and in the presence of detergents. Moreover, this enzyme is capable of degrading the abnormal prion protein, which is responsible for transmissible spongiform encephalopathies [18]. Another enzyme, Tpa DNA polymerase from *Thermococcus pacificus*, exhibits polymerase chain reaction (PCR) applications [19]. Further research has shown that when this enzyme is fused to the SSs7d DNA binding protein from *Sulfolobus solfataricus*, it significantly enhances its PCR-related properties [20].

1.2 Research justification.

It is important to note that unidentified prokaryotes are continuously being discovered. And these newly discovered prokaryotes might code for CRISPR/Cas systems since roughly 90% of archaea and 50% of bacteria harbour such systems [3]. Furthermore, since CRISPR/Cas systems are highly widespread in prokaryotes, many systems have not been characterized through bioinformatics approaches yet. Altogether, this scientific field needs to be constantly investigated, and a study focusing on this field could be highly intriguing.

With regard to *Thermococcus*, only a couple of publications have focused on CRISPR/Cas systems of two *Thermococcus* species up to the present (January 2021). The first publication, published in 2013, characterized the previously unexplored CRISPR/Cas systems of *T. kodakarensis* [21]. The other study, published in 2016, revealed the structural features of Cas2 from *Thermococcus onnurineus* in type IV CRISPR/Cas system [22]. No analysis has nevertheless addressed all CRISPR/Cas systems encoded by *Thermococcus* species. Hence, the CRISPR/Cas systems in most *Thermococcus* remain undocumented thus far, which makes this genus suitable and ideal for this study.

This study aims to provide insightful information for better understanding the CRISPR/Cs systems in *Thermococcus*. Even more precisely, it it intended to determine the occurrence, diversity, organization and structural features of such systems, and ultimately try to infer to the evolutionary history of these systems in this archaeal genus.

1.3 Objectives

1.3.1 General objectives.

Two main general objectives were set for this study. The first objective is to appropriately and carefully select a group of microorganisms that allow the development of an interesting and relevant study in the chosen research field. Once identified the group, the second objective is to gain a comprehensive understanding of the CRISPR/Cas systems encoded by this particular group of microorganisms.

1.3.2 Specific objectives.

Regarding the second general objective, this objective has been further divided into several specific objectives. The specific objectives of this study are to:

- Determine the occurrence and diversity of the CRISPR/Cas systems in *Thermococcus*.

- Characterize the genomic architecture of the CRISPR/Cas systems in *Thermococcus* (CRISPR loci and Cas proteins).

- Characterize the structural features of the CRISPR/Cas systems in *Thermococcus* (leader sequences, DRs and spacers).

- Understand the phylogenetic relationships of the CRISPR/Cas systems in *Thermococcus.*

1.4 Planning

1.4.1 Tasks

Nine tasks have been established to ensure that the study is conducted in an orderly manner and that all the objectives are met. The tasks are listed in **Table 2**. In the Methodology section, all the methods and resources needed to conduct the tasks successfully are listed.

	TASK	DURATION (WEEKS)
Task 1	**Select the group of microorganisms.**	1
Task 2	**Identify and annotate the CRISPR loci.**	1.5
Task 3	**Identify and annotate the *cas* loci.**	1
Task 4	**Phylogenetic analysis of Cas proteins.**	1.5
Task 5	**Identify the leader sequences of the CRISPR loci.**	0.5
Task 6	**Reveal conserved regions within leader sequences.**	1
Task 7	**Analyse the CRISPR direct repeats.**	1
Task 8	**Predict the secondary structure of the direct repeats.**	0.5
Task 9	**Analyse and match the CRISPR spacers.**	1

Table 1. List of the tasks established for this study.

1.4.2 Timing

A Gantt chart has been created to manage the study conveniently and appropriately, and allow the visual representation of the workload over the period established to conduct the study (**Fig. 2**). The chart contains the established tasks, which are timely arranged. Besides, to make sure the study is conducted progressively over a fixed period, the coordinators of this Master defined several milestones, which have also been added to the chart.

The software Project Manager has been used to design the Gantt chart [https://www.projectmanager.com/ (21/12/2020)].

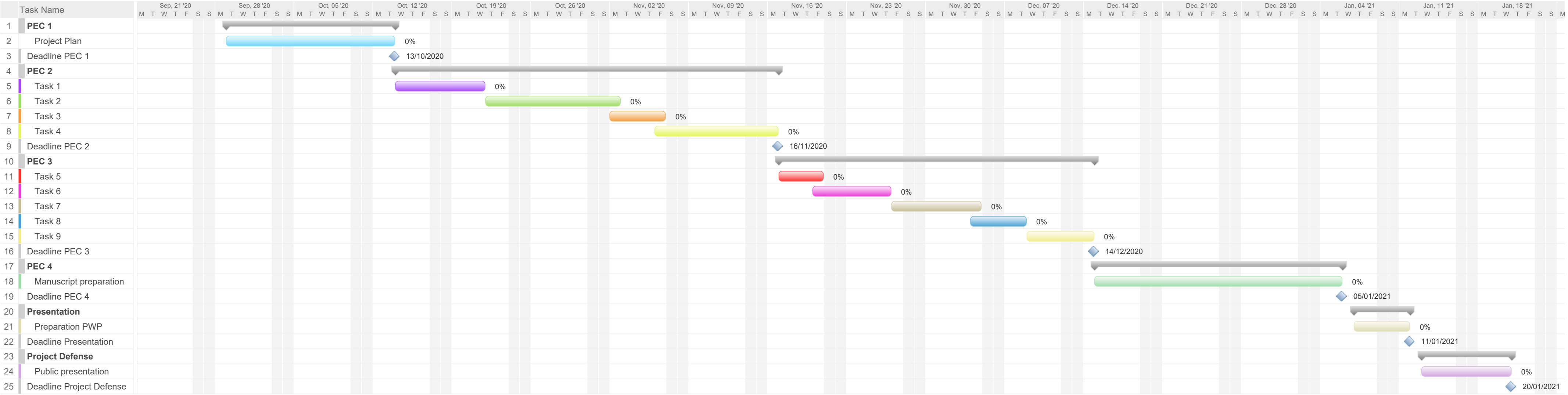


Figure 2. Gantt chart of the planning for this study.

1.5 Brief summary of products obtained.

The main product is the written report that follows. This report has been divided into the sections Methodology, Results, Discussion and Conclusions. A brief description of each of the previously mentioned sections is provided in section 1.6.

Apart from the report, a presentation has also been created to report the whole study orally. The presentation is based on the same parts as the report. This presentation will be used at the public defence of the study by the end of January.

Additionally, this study is aimed to be the basis for a scientific publication, as it is intended to publish the study in a scientific journal.

1.6 Brief description of the other chapters of the report.

As indicated, the Introduction section is concluded with a short description of the remaining parts of the report.

- **Methodology**. All the bioinformatics tools that have been used to conduct the study are listed and further explained. Several references that describe the use and implications of such tools have also been included and shortly discussed.

- **Results**. Mention of the findings obtained in the study, arranged in a logical sequence. Concise figures have been generated to confirm and better comprehend the findings.

- **Discussion**. Interpretation and description of the significance of the findings obtained, supported by other publications.

- **Conclusions**. The main points of the study have been highlighted in a synthesised way. These points are the most relevant aspects of the characterization of the CRISPR/Cas systems in *Thermococcus.* Also, further studies that could derive from this one have been suggested.

(The report also contains a part of **Abbreviations**, an **Appendix** and **References**).

2. Methodology

2.1 Selection of *Thermococcus.*

First of all, a list of the potential group of microorganisms with industrial or medical interest was made. Secondly, it was determined whether these groups code for CRISPR/Cas systems. To do so, the complete chromosome sequences of a couple of species of each group were retrieved from the National Center of Biotechnology Information (NCBI) genome database [https://www.ncbi.nlm.nih.gov/genome/ (10/12/2020)]. Then, a rapid analysis of the CRISPR/Cas systems was performed using CRISPRFinder, to get a quick overview of the systems present in these groups [https://crispr.i2bc.paris-saclay.fr/Server/ (21/12/2020)]. CRISPRFinder is a web tool that identifies CRISPR loci in given genomic sequences and provides information about its components [23]. Lastly, an extensive literature search was carried out to confirm that a similar study has not been conducted on the CRISPR/Cas systems of said potential groups. The search platforms used were PubMed and Google Scholar [https://pubmed.ncbi.nlm.nih.gov/ (14/12/2020)] [https://scholar.google.com/ (14/12/2020)].

2.2 *Thermococcus* genome sequences.

There are currently 22 *Thermococcus* species with their complete genomes available at NCBI genome database. Nevertheless, the genome of *Thermococcus* chitonophagus was not included in the analysis because a 2004 study suggested that this species was misclassified (basing on a 16 rRNA analysis), and it actually should belong to the genus *Pyrococcus* [24]. *Pyrococcus* belongs to the same family as *Thermococcus*, *Thermococcaceae*.

Hence, the genomes of all the remaining 21 *Thermococcus* species were retrieved from NCBI, and a comparative genomic analysis was performed via bioinformatics tools.

2.3 Bioinformatics analysis.

Firstly, all CRISPR/Cas systems present in these species were identified, and subsequently characterized, meaning that their occurrence, diversity and organization were determined. Secondly, the structural features of all CRISPR loci were analysed (i.e. the leader sequences, direct repeats (DRs) and spacers). Thirdly, phylogenetic analyses on several Cas proteins were performed.

2.3.1 Characterization of the occurrence, diversity and organization of CRISPR/Cas systems.

The identification of all CRISPR loci in *Thermococcus* genomes was performed using CRISPRFinder. All the identified loci were successively verified by CRISPRone, which is another web tool that predicts CRISPR/Cas systems in genomic sequences [https://omics.informatics.indiana.edu/CRISPRone/ (21/12/2020)]. The analysis was conducted only on the CRISPR loci confirmed by CRISPRFinder. The questionable CRISPR loci identified were not further analysed. After the identification, the structural features of all CRISPR loci were annotated, and also the *cas* loci adjacent to them, following manual curation (**Table 1**).

Be noted that by CRISPR/Cas systems, it has been considered all CRISPR loci that had a *cas* locus nearby (either upstream or downstream).

The nomenclature and classification of CRISPR/Cas systems used has been based on the classification suggested by Makarova *et al* [25].

2.3.2 Identification and conservation of leader sequences.

For each CRISPR locus, 160 bp upstream the first DR were selected as putative leader sequences. The CRISPR loci with a 3' → 5' orientation were reoriented using the online tool Reverse Complement from Sequence Manipulation Suite (SMS) [https://www.bioinformatics.org/sms2/rev_comp.html (14/12/2020)]. The leader sequences were aligned using ClustalX software to search for conserved regions within these sequences. ClustalX is a multiple sequence alignment method highly used to find conserved regions, prepare sequences for phylogenetic analysis, among others [26]. The results were also displayed with ClustalX. To complement the results, the GC content of each leader sequence was also calculated via Genomics %G~C Content Calculator, and compared the results to the GC genome content of the corresponding species [https://www.sciencebuddies.org/science-fair-projects/references/genomics-g-c-content-calculator (14/12/2020)]. The GC genome content was obtained at the NCBI genome database.

2.3.3 Direct Repeats analysis.

The consensus DR of every CRISPR locus found in *Thermococcus* genomes were retrieved from CRISPRFinder. The reverse complement sequences were obtained for the consensus DRs in which the orientation of the CRISPR locus was 3' → 5'. The reverse complement sequences were obtained through the Reverse Complement tool from SMS. A phylogenetic analysis was then performed on all the equally-oriented consensus DRs.

The alignment of the DRs was performed using the MUSCLE alignment algorithm. MUSCLE is another multiple sequence alignment method widely used for phylogenetic analysis [27] [https://www.ebi.ac.uk/Tools/msa/muscle/ (14/12/2020)]. Then, the software Molecular Evolutionary Genetics Analysis X (MEGAX) was used to build the phylogenetic tree. This well-established software provides tools to conduct comparative analysis of DNA and protein sequences [28]. The evolutionary history was inferred by using the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) method and Jukes-Cantor model calculating the bootstrap values of 500 samples. The

UPGMA method is frequently used to construct phylogenetic trees from a distance matrix. One characteristic of this method is that it makes the assumption that all lineages have the same evolutionary speed [29]. The UPGMA method was used instead of other methods such as the Neighbour-Joining because UPGMA is the method of choice for this kind of analysis [30]. Finally, the optimal UPGMA was exported and successively uploaded to the online tool iTOL. ITOL tool is commonly used to manage, annotate and display phylogenetic trees [https://itol.embl.de/ (14/12/2020)]. The final phylogenetic tree was displayed with midpoint rooting.

Several groups and subgroups were obtained in the phylogenetic analysis. The conservation of all the DRs from each group and subgroup was analysed using the online tool WebLogo [https://weblogo.berkeley.edu/logo.cgi (14/12/2020)]. WebLogo is a very handy tool to create sequence logos, which are graphical illustrations of the patterns within multiple sequence alignments [31]. To terminate the DR analysis, the most frequent consensus DR from each subgroup was selected, and its secondary structure was predicted via the online tool RNAFold [http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi (14/12/2020)]. RNAFold predicts the secondary structure of DNA or RNA sequences. Besides, this online tool provides a range of information related to the thermodynamics of the predicted structure [32].

2.3.4 Spacers analysis.

The spacers of every CRISPR locus found in *Thermococcus* genomes were retrieved from CRISPRFinder and submitted to analysis with CRISPRTarget [http://crispr.otago.ac.nz/CRISPRTarget/crispr_analysis.html (21/12/2020)]. The CRISPRTarget tool finds matches between spacers and target sequences to reveal the spacers identity [33]. The databases GenBank-Phage, RefSeq-Plasmid, IMGVR and RefSeq-Archea were selected as target databases. In parallel, to confirm the results obtained in CRISPRTarget, a Basic Local Alignment Search Tool (BLAST) search was conducted for all spacers [https://blast.ncbi.nlm.nih.gov/Blast.cgi (21/12/2020)]. BLAST detects local similarity between nucleotide or amino acid (aa) sequences by comparing said sequences with the sequences in a database. This tool also estimates the statistical significance of the matches obtained [34]. The database set was the Nucleotide collection (nr/nt) at NCBI, and the search was done using the default parameters. For both CRISPRTarget and BLAST, only matches showing at least 90% identity were considered.

The whole genome of the only two hitherto identified *Thermococcaceae* bacteriophages were subjected to a BLAST search: *Thermococcus prieurii* virus 1 (TPV1, 21592 bp), and *Pyrococcus* bacteriophage, *Pyrococcus abyssi* virus 1 (PAV1, 18098 bp) [35] [36].

2.3.5 Cas proteins phylogenetic analysis.

Phylogenetic analyses were performed on Cas1, Cas3 and Cas10 proteins. The genomic sequences of all the *cas1*, *cas3* and *cas10* genes included in the analyses were retrieved from NCBI (more information in the Results section). All genes with a 3'-5' orientation were re-orientated using the Reverse Complement tool from SMS. The alignments of the genomic sequences of *cas1*, *cas3* and *cas10* sequences were performed using the MUSCLE alignment algorithm [27]. Next, the software MEGAX was used to build the phylogenetic trees [28]. The evolutionary history was inferred by using the UPGMA method and Jukes-Cantor model calculating the bootstrap values of 500 samples [29]. UPGMA method is the preferred method for Cas proteins phylogenetic analysis [37] [38] [39].

The optimal UPGMA tree is shown for each Cas protein. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are illustrated next to the branches.

3. Results

3.1 Characterization of CRISPR/Cas systems.

The characterization of CRISPR/Cas systems in *Thermococcus* species has based on revealing their occurrence, diversity, organization and structural features.

3.1.1 Occurrence of CRIPSR/Cas systems.

Among the 21 genomes analysed, it was detected a certainly high occurrence of CRISPR loci. In total, 79 confirmed CRISPR loci were detected, and at least one CRISPR locus was observed in each *Thermococcus* species (**Fig. 3, Fig. 4, Appendix Table 1**). The number of CRISPR loci harbored widely ranges from one to ten loci per species. Up to five species harbor three (*T. gammatolerans, T. guaymasensis, T. kodakarensis, T. paralvinellae* and *T. profundus*) and four (*T. barophilus, T. eurythermalis, T. peptonophilus, T. piezophilus* and *T. radiotolerans*) CRISPR loci. Remarkably, *T. cleftensis* codes for ten CRISPR loci (**Fig. 3**).

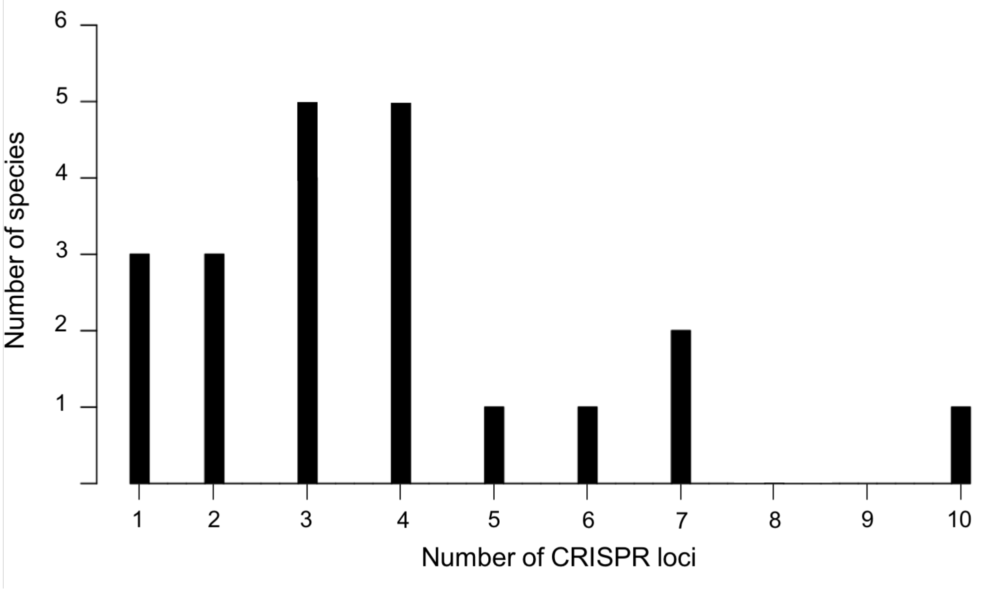
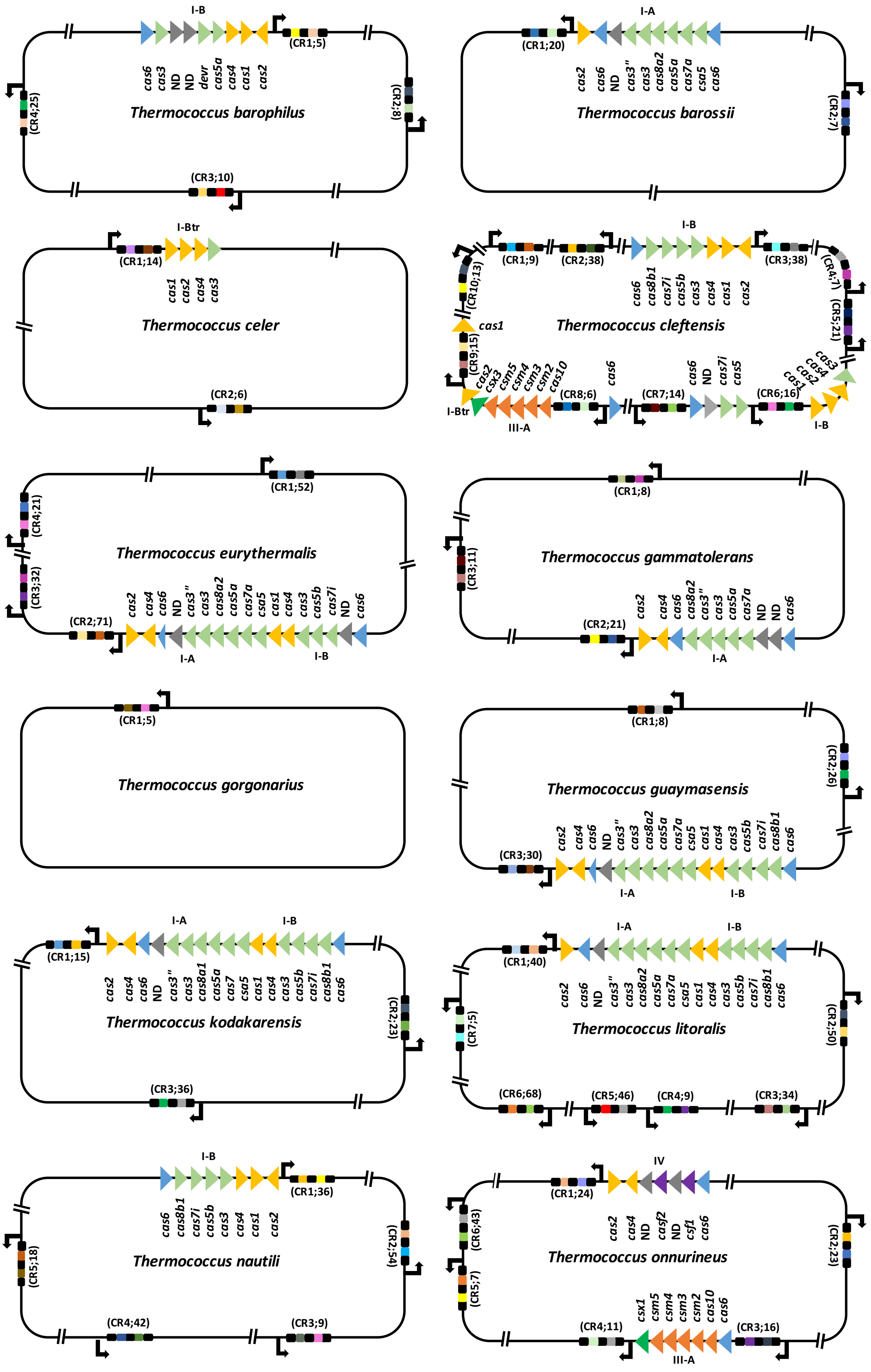


Figure 3. The number of CRISPR loci identified in the 22 fully-sequenced *Thermococcus* species.

Of these 79 CRISPR loci, 33 had adjacent *cas* genes. However, five of these cas loci were truncated. In addition, five *cas* loci were located between two CRISPR loci, which was considered to be just one CRISPR/Cas system (**Fig. 2, Table 1**).

Altogether, 23 complete CRISPR/Cas systems were confirmed among the 21 species. Only four species (*T. celer, T. gorgonarius, T. pacificus*, and *T. paralvinellae*) do not code for a complete CRISPR/Cas system (**Fig. 4, Appendix Table 1**).


I-B
(CR1;5)
cas6
cas3
ND
ND
devr
cas5a
cas4
cas1
cas2
(CR4;25)
(CR2;8)
Thermococcus barophilus
(CR3;10)
I-A
(CR1;20)
cas2
cas6
ND
cas3''
cas3
cas8a2
cas5a
cas7a
csa5
cas6
(CR2;7)
Thermococcus barossii
I-Btr
(CR1;14)
cas1
cas2
cas4
cas3
Thermococcus celer
(CR2;6)
I-B
(CR1;9)
(CR2;38)
(CR3;38)
(CR10;13)
(CR4;7)
(CR5;21)
cas6
cas8b1
cas7i
cas5b
cas3
cas4
cas1
cas2
cas1
Thermococcus cleftensis
(CR9;15)
cas2
csx3
csm5
csm4
csm3
csm2
cas10
cas6
cas6
ND
cas7i
cas5
cas3
cas4
cas2
cas1
(CR8;6)
(CR7;14)
(CR6;16)
I-Btr
III-A
I-B
(CR1;52)
(CR4;21)
Thermococcus eurythermalis
(CR3;32)
(CR2;71)
cas2
cas4
cas6
ND
cas3''
cas3
cas8a2
cas5a
cas7a
csa5
cas1
cas4
cas3
cas5b
cas7i
ND
cas6
I-A
I-B
(CR1;8)
(CR3;11)
Thermococcus gammatolerans
(CR2;21)
cas2
cas4
cas6
cas8a2
cas3''
cas3
cas5a
cas7a
ND
ND
cas6
I-A
(CR1;5)
Thermococcus gorgonarius
(CR1;8)
(CR2;26)
Thermococcus guaymasensis
(CR3;30)
cas2
cas4
cas6
ND
cas3''
cas3
cas8a2
cas5a
cas7a
csa5
cas1
cas4
cas3
cas5b
cas7i
cas8b1
cas6
I-A
I-B
I-A
I-B
(CR1;15)
cas2
cas4
cas6
ND
cas3''
cas3
cas8a1
cas5a
cas7
csa5
cas1
cas4
cas3
cas5b
cas7i
cas8b1
cas6
(CR2;23)
Thermococcus kodakarensis
(CR3;36)
I-A
I-B
(CR1;40)
cas2
cas6
ND
cas3''
cas3
cas8a2
cas5a
cas7a
csa5
cas1
cas4
cas3
cas5b
cas7i
cas8b1
cas6
(CR7;5)
Thermococcus litoralis
(CR2;50)
(CR6;68)
(CR5;46)
(CR4;9)
(CR3;34)
I-B
(CR1;36)
cas6
cas8b1
cas7i
cas5b
cas3
cas4
cas1
cas2
(CR5;18)
Thermococcus nautili
(CR2;54)
(CR4;42)
(CR3;9)
IV
(CR1;24)
cas2
cas4
ND
casf2
ND
csf1
cas6
(CR6;43)
Thermococcus onnurineus
(CR2;23)
(CR5;7)
csx1
csm5
csm4
csm3
csm2
cas10
cas6
(CR4;11)
(CR3;16)
III-A

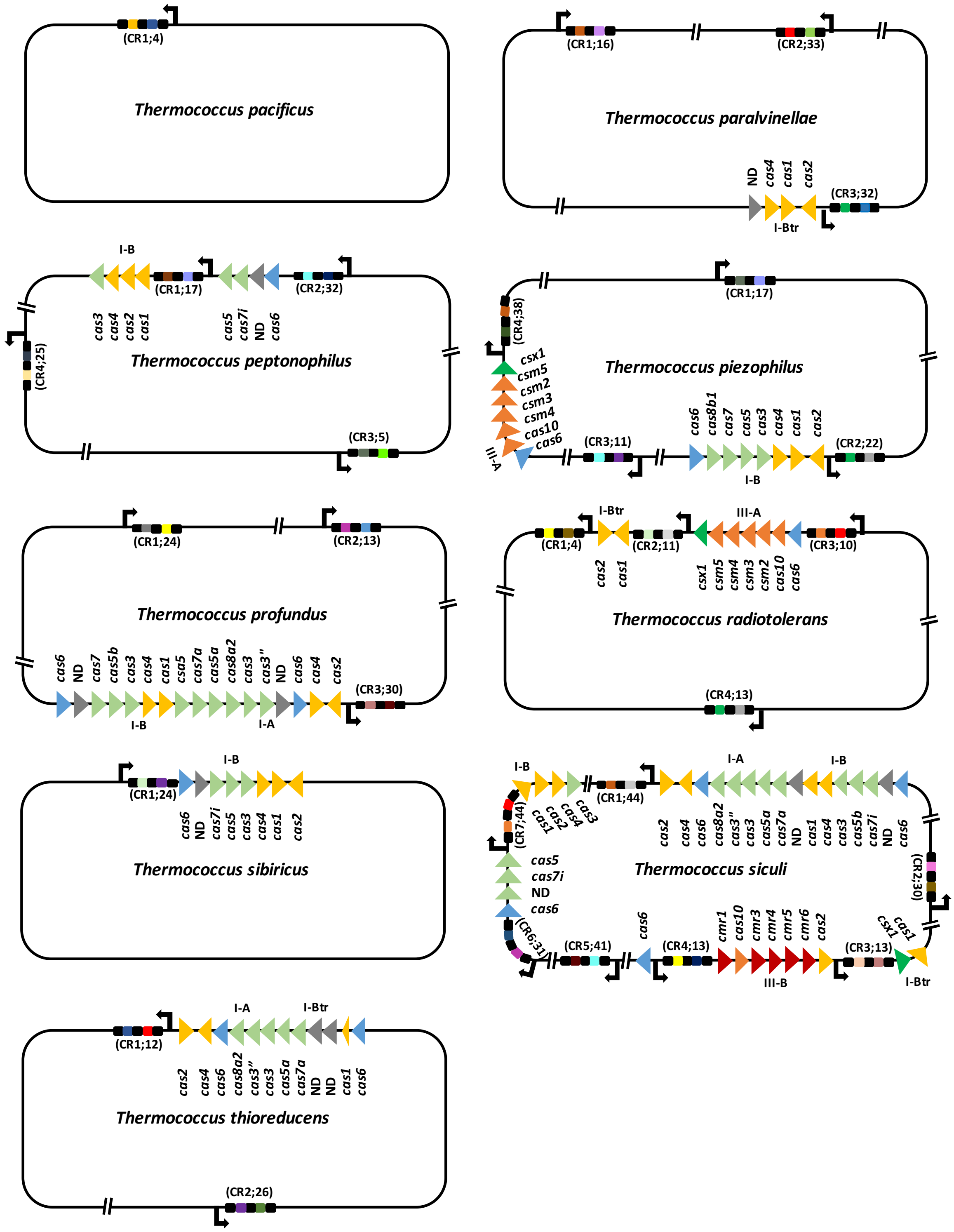


Figure 4. CRISPR and *cas* loci genomic architectures within *Thermococcus* genomes. For simplicity purposes, all CRISPR loci are based on three DRs (in black) and two spacers (coloured). The leader sequence of CRISPR loci is represented as a corner arrow. Both the CRISPR locus number and the number of spacers that the given locus are composed are indicated under each CRISPR locus. The *cas* genes are shown as arrowheads. Effector *cas* genes are shown in green, orange and red for type I, type III-A and type III-B systems, correspondingly. *Cas* genes involved in the informational module (*cas1*, *cas2* and *cas4*) are represented in yellow. C*as6* genes, which are the processing factors, are depicted in blue. C*sx1* genes, which are transcriptional regulators, are shown in light green. Finally, not determined genes (ND in short) are shown in grey. Narrower arrowheads correspond to partially annotated *cas* genes. Note that this illustration is not to scale.

3.1.2 Diversity of CRIPSR/Cas systems.

As regard to the diversity of CRISPR/Cas systems in *Thermococcus*, all class 1 CRISPR/Cas system types were detected (i.e. type I, III and IV) (**Fig. 4, Appendix Table 1**).

The type I systems were the most abundant. These were composed of eight subtype I-B, three subtype I-A and six subtype I-B/I-A systems. The latter are composed of two modules clustered together, one corresponding to type I-B *cas* genes and the other corresponding to type I-A *cas* genes. Type I-B is characterized by the presence of *cas8b1*, whereas type I-A codes for *cas8a1* [2] (**Fig. 4, Appendix Table 1**).

On the other hand, type III systems were less frequent than type I systems. Even so, the two type III subtypes known until now were found. Specifically, four subtype III-A and one type III-B are harbored in *Thermococcus* genomes (**Fig. 4, Appendix Table 1**).

Shockingly, only one type IV system was identified, in the species *T. onnurineus* (**Fig. 4, Appendix Table 1**). This system was already described in a previous study on this species [22]. A BLASTp search was performed on the type IV signature protein of *T. onnurineus*, Csf1. Surprisingly, proteins in *T. pacificus* (64.68 % of homology) and two unclassified *Thermococcus* species were similar to the Csf1 protein of *T. onnurineus.* Seemingly*, T. pacificus* contains the same *cas* locus as *T. onnurineus.* Yet, unlike *T. onnurineus,* the type IV *cas* locus in *T. pacificus* is not adjacent to a CRISPR locus. The Csf1 protein of *T. onnurineus* also shared a 25.33% homology with a protein in *Pyrodictium delaneyi.* The system is further discussed in the Discussion section.

In addition, it was inquired how many species harbor distinct complete CRISPR/Cas systems. Interestingly, the same amount of species, eight, code for two different CRISPR/Cas systems rather than one. Furthermore, one species, *T. siculi*, code for three distinct CRISPR/Cas systems (**Fig. 5**).

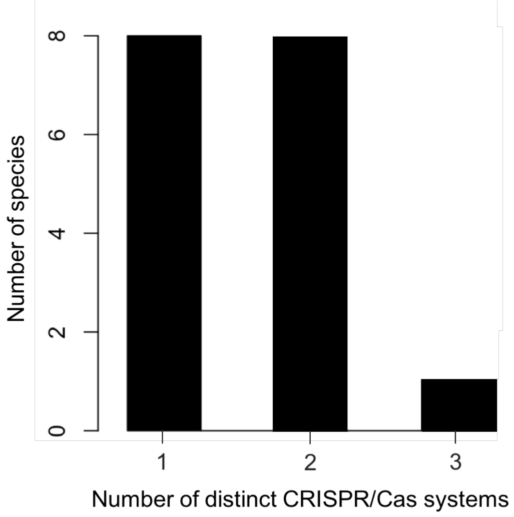


Figure 5. Number of different CRISPR/Cas systems harbored by *Thermococcus* species.

3.1.3 Organization of CRIPSR/Cas systems.

Intriguingly, type I-A systems were found more frequently concomitant to one type I-B (I-B/I-A system) than independently. When this type was not associated with type I-B systems, the *cas* loci presented the same organization, characterized by an inverted *cas2* and the absence of *cas1* (in *T. barossi* CR1, *T. gammatolerans* CR2 and *T. thioreducens* CR1) (**Fig. 4**).

Type I-A/I-B systems (found in *T. eurythermalis* CR2, *T. guaymasensis* CR3*, T. kodakarensis* CR1*, T. litoralis* CR1*, T. profundus* CR3 and *T. siculi* CR1) also happened to always have the same organization. Both modules were co-oriented, with one inverted *cas2* gene located at the beginning of the type I-A module, and one *cas1* gene found at the beginning of the type I-B module. However, the not-determined genes were not always the same for all these species (**Fig. 4**).

Type I-B systems, in contrast, presented different organizations when these were neither associated with a type I-A system nor with type III system. Two distinct organizations were found for these systems. One was mainly characterized by the presence of *cas* genes, all co-oriented, both upstream and downstream the CRISPR locus (*T. cleftensis* CR3, *T. peptonophilus* CR1 and *T. siculi* CR7). The other type I-B organization was characterized by an inverted *cas2* gene, as in type I-A systems, but these did encode for a *cas1* gene (*T. barophilus* CR1, *T. cleftensis* CR6, *T. nautili* CR1*, T. piezophilus* CR2 and *T. sibiricus* CR1) (**Fig. 4**).

Two truncated type I-B systems associated with other systems were found, one with a type III-A system (in *T. cleftensis* CR8), and one with a type I-A system (in *T. thioreducens* CR1). Additionally, three independent truncated type I-B systems, belonging to *T. celer* CR1, *T. radiotolerans* CR1 and *T. paralvinellae* CR3, were detected (**Fig. 4**).

Three independent type III-A systems were identified (*T. onnurineus* CR3, *T. piezophilus* CR4 and *T. radiotolerans* CR3). All these systems have the same organization, with all genes with the same orientation and order, being *csx1* the first one and *cas6* the last one of the locus. Only one type III-B system was detected (in *T. siculi* CR2), characterized by the presence of the *cmr* genes 1 to 6 (*cmr2* is also termed *cas10*) (**Fig. 4**).

As for the only type IV system found (*T. onnurineus* CR1), this was organized similarly to the type I-A systems. This system was characterized by an inverted *cas2* gene, and the absence of *cas1* gene (**Fig. 4**).

Finally, three partially annotated *cas* genes were detected (*cas6* of *T. eurythermalis* CR2 module 1, *cas6* of *T. guaymasensis* CR2 module 1, and *cas1* of *T. thioreducens* CR1) (**Fig. 4**).

3.2 Identification and conservation of leader sequences.

When retrieving the putative leader sequences, it was ensured that all these sequences had the same orientation, 5' → 3'. The CRISPR locus orientation was determined by recognizing the 3' terminal DR. This DR typically presents several single nucleotide polymorphisms when compared with the consensus DR [30].

Aligning the leader sequences of CRISPR loci, strictly conserved regions were identified among several alignment groups of leader sequences (**Fig. 6**). The most remarkable result was the high conservation of TATA boxes among these groups. Also, the first nucleotides (nt) upstream of the first DR were usually conserved between groups (**Fig. 6**).

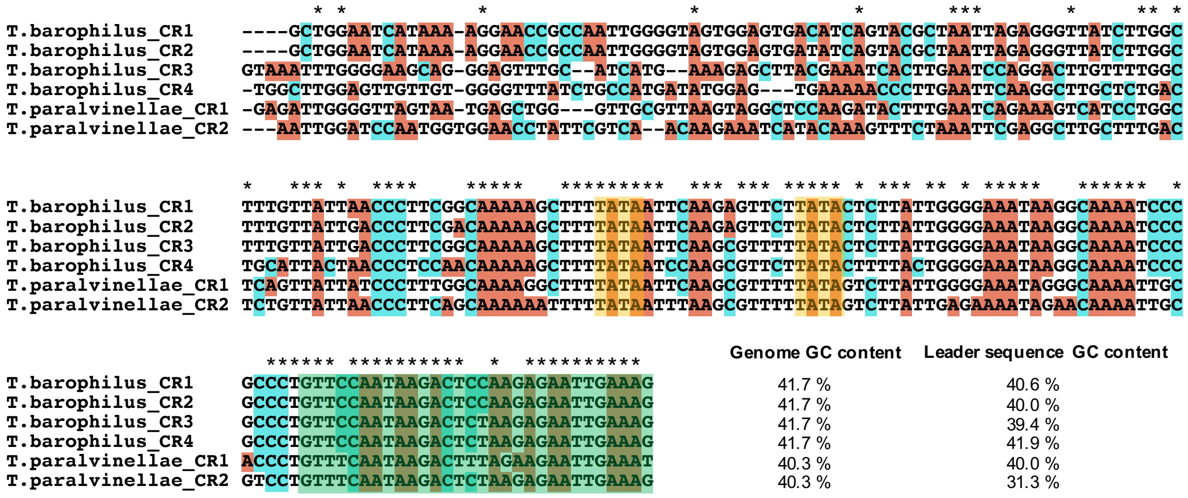
The leader sequences without conserved regions are not shown. Interestingly, all these sequences do not contain a TATA box in the vicinity of the first DR (160 nt upstream).

Furthermore, as already mentioned, another feature of the leader sequences is that they possess a rich content in AT. The GC content of all the analysed leader sequences was calculated and compared with the GC genome content of the corresponding *Thermococcus* species. The GC content from all leader sequences except one was lower than the GC genome content, indeed indicating richer AT regions in the leader sequences. This sequence belongs to *T. barophilus* CR4, and its GC content was slightly higher than the GC genome content, 41.7 % and 41.9 %, respectively. Several leader sequences with a high AT content were found. For instance, the leader sequences of *T. cleftensis* CR4, CR8 and CR9 have a GC content of 37.7 %, 38.9% and 37.7%, while the GC genome of this species is 55.8% (**Fig. 6**).

```
                                                                                     *       *******
T.cleftensis_CR5    ATCATTGTAAGGCTCTCTTCGGGTGAAA---------GCAGTCCTTTCGACGTTTCCCTATTCCTGTGCCGGCTTTATAA
T.cleftensis_CR6    ---AGTCCAAAGATCCTTCTGTGTGG------GCAACGGGGGCTTTAACACCTTTTTCATTTCGTGTGTCGGATTTATAA
T.cleftensis_CR7    --GTGGCCGAAAATGCCGAACCGGTGAAAAGAAAAATGGAGGTCT-------CTCCGCAGTTTTTGTGCCGGCTTTATAA
T.peptonophilus_CR1 ---------AATGTTTTGAAGTGTGTTGTCACAATTCCAAGCCCTTTGGTGGGTTCTTCGTTTTTGTGTCAGGTTTATAA
T.peptonophilus_CR2 --------AACCCAATTTGTGCATGAAATTGGGGTTAGAAAGTCTGTGAACGCTCCTCTAGTTTTGTGGCAAGTTTATAA
T.peptonophilus_CR4 -----ACTGAAGTTAATCTTATGAGGGACAG----TAGAAAGCCCCGGGATGCTCCTTTATTTTTGTGTTAGATTTATAA
T.radiotolerans_CR1 --------AGAGATGATGTAATGAAATTCTAGCCAAAAGAGATCCTAAAACGGAGTTCAACTCCAGCA-CGTATTTATAA
T.radiotolerans_CR2 --------AAGGATTATAACAGGGCAGCTACGAAATCCCAGTGAAGAGCAATCTCCGTTATCCTGGCACC-CCTTTATAA
T.siculi_CR6        ---------GCCAAATCCTGGAGTGCAATAATCCATAAAAGGCCTGGAAAAGCTTCTTTATTTCTGTGCCAGATTTATAA
T.siculi_CR7        ---------GGGTAGATCCATTGTTAGTCAACAAAATAAACTGTCCAACGCCCCACTTGATTTTTGTTCCACATTTATAA

                            *  *   **   *          **   *******   *       *  *            *
T.cleftensis_CR5    ATAGAATGCGGTGACACGGGTTTT-TAGAACAAGGTTTAAATAGCGGTGAAGAAAGAACTTAAAATCGAGAACCCGGGGA
T.cleftensis_CR6    ACCAACTGCGTTGCCACAGGTTTC-CAAGACAAAGTTTAAATAGGAGTTGAAAAACCAGATATTTTGAGGATCCGGGAAA
T.cleftensis_CR7    ATAGAATGCGGTGACACGGGTTTT-TAGAACAAGGTTTAAATAGCGGTGAAGGAAGAACTTAAAATCGAGAACCCGGGGA
T.peptonophilus_CR1 ATGGGGTGCAGTGCCACGGGTTCC-CAAAGAAAGCCTTAAATAGGAGTTGGGAAATCACTTTATTTGTCAAATGAAGGAG
T.peptonophilus_CR2 GTGGAGTGCAGTGCTACAGGTTTC-CGGAGAAAGTCTTAAATAGGAGTT-GGAAATCACTTTATTTGTCAAATGAAGGAG
T.peptonophilus_CR4 AGGACGTGCAGTGCTACGGGTTCC-CAAAGAAAGCCTTAAATAGGAGTTGGGGAATCACTTTGTTTGCCAATCAGAGAAA
T.radiotolerans_CR1 AGAAAAGCCAGTAACACGGGTTTCATGAAGAAATGCTTAAATAG-AGCAAGGAAAGAAGAGATATAACGAAACAGGACAA
T.radiotolerans_CR2 ATGAGTGCCGGTGCTACCGATTTC-CCAAGAAACAATTAAATAGAAGGAAGAAAATAATATTACACACCAAAAGAACTGA
T.siculi_CR6        ATCGGATACAGTGTTACAAGTCTC-CGGAAAAAGTCTTAAATAGGAGTTCGGAGATCACTTTACTATCCCATTAGGGGGA
T.siculi_CR7        ATAACTTGCAGTGCTACGGGTTCC-CGGGAAAAGTCTTAAATATAAGCTCAAACAGTACTTTATGTGTAAATCAAAGGAA

                            *  ***** **** **    * *********     Genome GC content   Leader sequence  GC content
T.cleftensis_CR5    ATTCCCGGAGCGTTTCCGTAGGACAGAATTGTGTGGAAAG        55.8 %                45.3 %
T.cleftensis_CR6    GTTGCCCGAGGGTTTCCGTAGGACAGAATTGTGTGGAAAG        55.8 %                44.1 %
T.cleftensis_CR7    ATTCCCGGAGCGTTTCCGTAGAACAGTGTTGTGTGGAAAG        55.8 %                46.6 %
T.peptonophilus_CR1 GAAAGTGAAGTGTTTCCGTAGAACGTAATCGTGTGGAAAG        51.7 %                41.0 %
T.peptonophilus_CR2 GAAAGTGAAGGGTTTCCGTAGAACGTAGTCGTGTGGAAAT        51.7 %                40.4 %
T.peptonophilus_CR4 AGAAGTGAAGAGTTTCCGTAGAACGTAGTCGTGTGGAAAG        51.7 %                41.0 %
T.radiotolerans_CR1 GAAACCGGAAAGTTTCGGTAGAACATAATTGTGTGGAAAC        55.6 %                36.6 %
T.radiotolerans_CR2 AAATCCTGAAAGTTTCCGTAGAACAGTATTGTGTGGAAAC        55.6 %                39.1 %
T.siculi_CR6        TAAAGTCAAGGGTTTCCGTAGGACATAGTTGTGTGGAAAG        55.0 %                39.8 %
T.siculi_CR7        AAAGAGTGAGAGTTTCCGTAGAACATGGTTGTGTGGAAAG        55.0 %                36.6 %
```

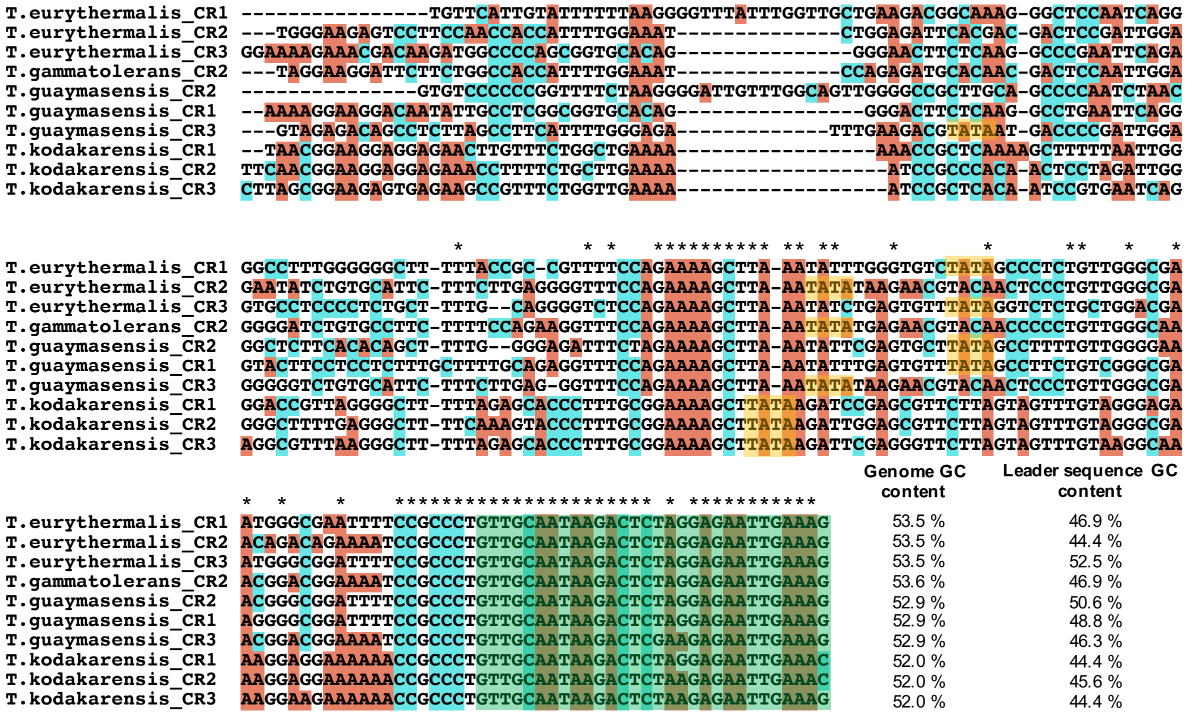
```
                                 * *        *    *  *   *  *** *       * ***        *      *
T.cleftensis_CR4 ----------GCTTTTTCTT------TCTTTTTCGGTTTTCTTACTATTTTGGATGGTTTGTGTTTGGAGTGTTTAGGGT
T.cleftensis_CR8 ----------GCTTTTTCTT------TCTTTTTCGGCTTTCTTACTAATTGAAGTGGTTTGTGTTGGGGGCGTTTGGAGT
T.cleftensis_CR9 ----------GCTTTTTCTT------TCTTTTTCGGTTTTCTTACTATTTTGGGTGGTTTGTGTTTGGAGTGTTTAGGGT
T.siculi_CR3     -----TCGGATTCCTCTTTTGATAGTTCTCCGTTTGATTTTCTACTACCGTCTGTGGTTCGGGAGAAGAGTAGAT-----
T.siculi_CR4     TGGGTTTGAAGCTCGCTCTAG-----CCCCAGTTTGGGCTTTTACAATTATGAATAGTTTAAACCAAGCCCGTTT-----

                           *     **** * *  *    * *** ** **** *****  *  *     * ***   **** **   *
T.cleftensis_CR4 TTTTGTTCTGGTGCTGTTTTCTGA-ACTTCTGGTAACCGCAAAATTTATATGGGGGTTCGGCGTTACTCTATTTGCTCAA
T.cleftensis_CR8 CTCCCTTCTAGGGTTGTTTTCTGA-ACTTCTGATAACCGCAAAATTTATATGGGAGTTTAACATTACTTTATTTGCCCGA
T.cleftensis_CR9 TTTTGTTCTGGTGCTGTTTTCTGA-ACTCCTGGTAACCGCAAAATTTATATAGGAGTTTAACATTACCCTATTTGCCCGA
T.siculi_CR3     --------AAGGGTCGTTTTTTGAGGCCTGAGCTAATCGAAAAACTTATAAGGAGGAAGATCGTTATTTTATTCGCCAAA
T.siculi_CR4     --------TCGTCTTATTTTCTAAGGCCCCAGCTAATCGAAAAACTTATAAGGAGGGAGAACATTATTTTATTCGCCAAA

                                                                     Genome GC    Leader sequence  GC
                   * ** ******* **************** * * *  *********       content          content
T.cleftensis_CR4 TAGGGCAAAAAAGTTAACCGTTTCAGAACCA-CATAATGTTTGGAAAC      55.8 %          37.7 %
T.cleftensis_CR8 AGGGGCAAAAAAGTTAACCGTTTCAGAACCA-CATGATGTTTGGAAAC      55.8 %          38.9 %
T.cleftensis_CR9 AGGGGCAAAAAAGTTAACCGTTTCAGAACCA-CATAATGTTTGGAAAC      55.8 %          37.4 %
T.siculi_CR3     AGGAGCGAAAAAGTGAACCGTTTCAGAACCAGCATAAGCTTTGGAAAC      55.0 %          41.0 %
T.siculi_CR4     AGGGGCGAAAAAGTGAACCGTTTCAGAACCAGCTTAAGCTTTGGAAAC      55.0 %          41.0 %
```

```
                               ***   **    **                *  *  * *          **    **  **
T.litoralis_CR1 -------------GAAAAAAGAGTAAGTTCAGGCGAGTGAGTGCTTTGAAGTCTTCTCTAGTGGCTTGGATTCGAGCTTT
T.litoralis_CR2 -------------AGAAAAGTAGTAAATTCAGGCGAGTGAGTGCTTTGAAGTCTTTTCTGAAAGCTGGAATTCAAGCCAG
T.litoralis_CR3 -------------AGAAAAGTAGTAGGTTCAGGTGAGTGAGTGCTTTGAAGTCTTTTCTAGTGGCTTGGATTCGAGCCTT
T.litoralis_CR5 -------------AGAAAAGTAGTAAATTCGGGCAAGTGAGTGCTTTGAAGTCTTTTCTAGTGGCTTGGATTCGAGCCTT
T.litoralis_CR6 -------------CAAAAAGGAGGAGTTTCAAGTGTGGTTGAGTTTCAAGGTTTCTCTCAAGGGCTCGAATTCGAGCTTT
T.sibiricus_CR1 GTTGTGAGAAAGCTGAAACATAGTTATTTTGGA------------TTAAACTCTTATGTGAGCACTTTATTTGAAGACTT

                   * **    ****** ****  *     ** * ******  *     *  * ** *   * * *  **  ****  **
T.litoralis_CR1 TTGAGGGGATTTTTATTGACCCTTTATGGAAAGGCTTATAAGATTTGGGCTTTCTAATTACTTTGTAGGGAGTTTAGAGG
T.litoralis_CR2 TTGAAGGGAGTTTTATTGACCCTTTGTGGAAAGGCTTATAAATTTCAAGCTTTCTAATTACTTTATAGGGAGTTTAGAGG
T.litoralis_CR3 CTTATGGCTGTTTTATTGACCCTTTGTGGAAAGGCTTATAAATTTCAAGCTCTCTAATAGTTTTGTAGGGGATTTAAAGG
T.litoralis_CR5 CTTATGGCTGTTTTATTGACCCTTTGTGGAAAGGCTTATAAATTTCAAGCTCTCTAATTACTTTATAGGGAGTTTAGAGG
T.litoralis_CR6 CTGAGGGGAGTTTTATTGACCCTTTGAAGAAAAGTTTATAAGATTCGGGCCTTCTAATTACTCTTTGAGGAGTTTAGAGG
T.sibiricus_CR1 TAGAGGGCTTATTTATT-ACCCCCTGCAATAAAGCTTATAAATTCTAAACTCTTTACTAGTTTTATAGGGAATTTAGGGG

                **** *** *** ***************** ***********     Genome GC content   Leader sequence  GC content
T.litoralis_CR1 AAAATTCGCCCCTGTTCCAATAAGACTTTAGAAGAATTGAAAG            43.1 %                40.0 %
T.litoralis_CR2 AAAATTCGCCCCGGTTCCAATAAGACTTTAAAAGAATTGAAAG            43.1 %                37.7 %
T.litoralis_CR3 AAAATTCGCCCCTGTTCCAATAAGACTTTAAAAGAATTGAAAT            43.1 %                39.4 %
T.litoralis_CR5 AAAATTCGTCCCTGTTCCAATAAGACTTTAAAAGAATTGAAAG            43.1 %                38.1 %
T.litoralis_CR6 AAAAATCGCCCCTGTTCCAATAAGACTTTAAAAGAATTGAAAG            43.1 %                41.3 %
T.sibiricus_CR1 AAAATTCGCCCCTGTTCCAATAAGACTTTAAAAGAATTGAAAG            40.2 %                33.1 %
```



```
                          * *           *                 *             *       ***       *     ** *
T.barophilus_CR1    ----GCTGGAATCATAAA-AGGAACCGCCAATTGGGGTAGTGGAGTGACATCAGTACGCTAATTAGAGGGTTATCTTGGC
T.barophilus_CR2    ----GCTGGAATCATAAA-AGGAACCGCCAATTGGGGTAGTGGAGTGATATCAGTACGCTAATTAGAGGGTTATCTTGGC
T.barophilus_CR3    GTAAATTTGGGGAAGCAG-GGAGTTTGC--ATCATG--AAAGAGCTTACGAAATCACTTGAATCCAGGACTTGTTTTGGC
T.barophilus_CR4    -TGGCTTGGAGTTGTTGT-GGGGTTTATCTGCCATGATATGGAG---TGAAAAACCCTTGAATTCAAGGCTTGCTCTGAC
T.paralvinellae_CR1 -GAGATTGGGGTTAGTAA-TGAGCTGG---GTTGCGTTAAGTAGGCTCCAAGATACTTTGAATCAGAAAGTCATCCTGGC
T.paralvinellae_CR2 ---AATTGGATCCAATGGTGGAACCTATTCGTCA--ACAAGAAATCATACAAAGTTTCTAAATTCGAGGCTTGCTTTGAC

                    *   *** *  ****    *****   *********  *** *** ***** * *** ** * *****   ******  *
T.barophilus_CR1    TTTGTTATTAACCCTTCGGCAAAAAGCTTTTATAATTCAAGAGTTCTTATACTCTTATTGGGGAAATAAGGCAAAATCCC
T.barophilus_CR2    TTTGTTATTGACCCTTCGACAAAAAGCTTTTATAATTCAAGAGTTCTTATACTCTTATTGGGGAAATAAGGCAAAATCCC
T.barophilus_CR3    TTTGTTATTGACCCTTCGGCAAAAAGCTTTTATAATTCAAGCGTTTTTATACTCTTATTGGGGAAATAAGGCAAAATCCC
T.barophilus_CR4    TGCATTACTAACCCTCCAACAAAAAGCTTTTATAATCCAAGCGTTCTTATACTTTTACTGGGGAAATAAGGCAAAATCCC
T.paralvinellae_CR1 TCAGTTATTATCCCTTTGGCAAAAGGCTTTTATAATTCAAGCGTTTTTATAGTCTTATTGGGGAAATAGGGCAAAATTGC
T.paralvinellae_CR2 TCTGTTATTAACCCTTCAGCAAAAAATTTTTATAATTTAAGCGTTTTTATAGTCTTATTGAGAAAATAGAACAAAATTGC

                      ****** **********  *  **********
T.barophilus_CR1    GCCCTGTTCCAATAAGACTCCAAGAGAATTGAAAG
T.barophilus_CR2    GCCCTGTTCCAATAAGACTCCAAGAGAATTGAAAG
T.barophilus_CR3    GCCCTGTTCCAATAAGACTCTAAGAGAATTGAAAG
T.barophilus_CR4    GCCCTGTTCCAATAAGACTCTAAGAGAATTGAAAG
T.paralvinellae_CR1 ACCCTGTTTCAATAAGACTTTAGAAGAATTGAAAT
T.paralvinellae_CR2 GTCCTGTTTCAATAAGACTCTAAGAGAATTGAAAG
```

	Genome GC content	Leader sequence GC content
T.barophilus_CR1	41.7 %	40.6 %
T.barophilus_CR2	41.7 %	40.0 %
T.barophilus_CR3	41.7 %	39.4 %
T.barophilus_CR4	41.7 %	41.9 %
T.paralvinellae_CR1	40.3 %	40.0 %
T.paralvinellae_CR2	40.3 %	31.3 %



```
T.eurythermalis_CR1 ----------------TGTTCATTGTATTTTTTAAGGGGTTTATTTGGTTGCTGAAGACGGCAAAG-GGCTCCAATCAGG
T.eurythermalis_CR2 ---TGGGAAGAGTCCTTCCAACCACCATTTTGGAAAT--------------CTGGAGATTCACGAC-GACTCCGATTGGA
T.eurythermalis_CR3 GGAAAAGAAACGACAAGATGGCCCCAGCGGTGCACAG---------------GGGAACTTCTCAAG-GCCCGAATTCAGA
T.gammatolerans_CR2 ---TAGGAAGGATTCTTCTGGCCACCATTTTGGAAAT--------------CCAGAGATGCACAAC-GACTCCAATTGGA
T.guaymasensis_CR2  ---------------GTGTCCCCCCGGTTTTCTAAGGGGATTGTTTGGCAGTTGGGGCCGCTTGCA-GCCCCAATCTAAC
T.guaymasensis_CR1  --AAAAGGAAGGACAATATTGCCTCGGCGGTGCACAG----------------GGGACTTCTCAAG-GCCTGAATTCAGG
T.guaymasensis_CR3  ---GTAGAGACAGCCTCTTAGCCTTCATTTTGGGAGA-------------TTTGAAGACGTATAAT-GACCCCGATTGGA
T.kodakarensis_CR1  --TAACGGAAGGAGGAGAACTTGTTTCTGGCTGAAAA-----------------AAACCGCTCAAAAGCTTTTTAATTGG
T.kodakarensis_CR2  TTCAACGGAAGGAGGAGAAACCTTTTCTGCTTGAAAA------------------ATCCGCCCACA-ACTCCTAGATTGG
T.kodakarensis_CR3  CTTAGCGGAAGAGTGAGAAGCCGTTTCTGGTTGAAAA------------------ATCCGCTCACA-ATCCGTGAATCAG

                                      *          * *   ********** ** **    *       *      **   *   *
T.eurythermalis_CR1 GGCCTTTGGGGGGCTT-TTTACCGC-CGTTTTCCAGAAAAGCTTA-AATATTTGGGTGTCTATAGCCCTCTGTTGGGCGA
T.eurythermalis_CR2 GAATATCTGTGCATTC-TTTCTTGAGGGGTTTCCAGAAAAGCTTA-AATATATAAGAACGTACAACTCCCTGTTGGGCGA
T.eurythermalis_CR3 GTGCCTCCCCTCTGCT-TTTG--CAGGGGTCTCCAGAAAAGCTTA-AATATCTGAGTGTTTATAGGTCTCTGCTGGACGA
T.gammatolerans_CR2 GGGGATCTGTGCCTTC-TTTTCCAGAAGGTTTCCAGAAAAGCTTA-AATATATGAGAACGTACAACCCCCTGTTGGGCAA
T.guaymasensis_CR2  GGCTCTTCACACAGCT-TTTG--GGGAGATTTCTAGAAAAGCTTA-AATATTCGAGTGCTTATAGCCTTTTGTTGGGGAA
T.guaymasensis_CR1  GTACTTCCTCCTCTTTGCTTTTGCAGAGGTTTCCAGAAAAGCTTA-AATATTTGAGTGTTTATAGCCTTCTGTCGGGCGA
T.guaymasensis_CR3  GGGGGTCTGTGCATTC-TTTCTTGAG-GGTTTCCAGAAAAGCTTA-AATATATAAGAACGTACAACTCCCTGTTGGGCGA
T.kodakarensis_CR1  GGACCGTTAGGGGCTT-TTTAGAGCACCCTTTGCGGAAAAGCTTATAAGATCCGAGCGTTCTTAGTAGTTTGTAGGGAGA
T.kodakarensis_CR2  GGGCTTTTGAGGGCTT-TTCAAAGTACCCTTTGCGGAAAAGCTTATAAGATTGGAGCGTTCTTAGTAGTTTGTAGGGCGA
T.kodakarensis_CR3  AGGCGTTTAAGGGCTT-TTTAGAGCACCCTTTGCGGAAAAGCTTATAAGATTCGAGGGTTCTTAGTAGTTTGTAAGGCAA

                    *  *    *    ********************** * ***********
T.eurythermalis_CR1 ATGGGCGAATTTTCCGCCCTGTTGCAATAAGACTCTAGGAGAATTGAAAG
T.eurythermalis_CR2 ACAGACAGAAAATCCGCCCTGTTGCAATAAGACTCTAGGAGAATTGAAAG
T.eurythermalis_CR3 ATGGGCGGATTTTCCGCCCTGTTGCAATAAGACTCTAGGAGAATTGAAAG
T.gammatolerans_CR2 ACGGACGGAAAATCCGCCCTGTTGCAATAAGACTCTAGGAGAATTGAAAG
T.guaymasensis_CR2  ACGGGCGGATTTTCCGCCCTGTTGCAATAAGACTCTAGGAGAATTGAAAG
T.guaymasensis_CR1  AGGGGCGGATTTTCCGCCCTGTTGCAATAAGACTCTAGGAGAATTGAAAG
T.guaymasensis_CR3  ACGGACGGAAAATCCGCCCTGTTGCAATAAGACTCGAAGAGAATTGAAAG
T.kodakarensis_CR1  AAGGAGGAAAAAACCGCCCTGTTGCAATAAGACTCTAGGAGAATTGAAAC
T.kodakarensis_CR2  AAGGAGGAAAAAACCGCCCTGTTGCAATAAGACTCTAAGAGAATTGAAAC
T.kodakarensis_CR3  AAGGAAGAAAAAACCGCCCTGTTGCAATAAGACTCTAAGAGAATTGAAAG
```

	Genome GC content	Leader sequence GC content
T.eurythermalis_CR1	53.5 %	46.9 %
T.eurythermalis_CR2	53.5 %	44.4 %
T.eurythermalis_CR3	53.5 %	52.5 %
T.gammatolerans_CR2	53.6 %	46.9 %
T.guaymasensis_CR2	52.9 %	50.6 %
T.guaymasensis_CR1	52.9 %	48.8 %
T.guaymasensis_CR3	52.9 %	46.3 %
T.kodakarensis_CR1	52.0 %	44.4 %
T.kodakarensis_CR2	52.0 %	45.6 %
T.kodakarensis_CR3	52.0 %	44.4 %

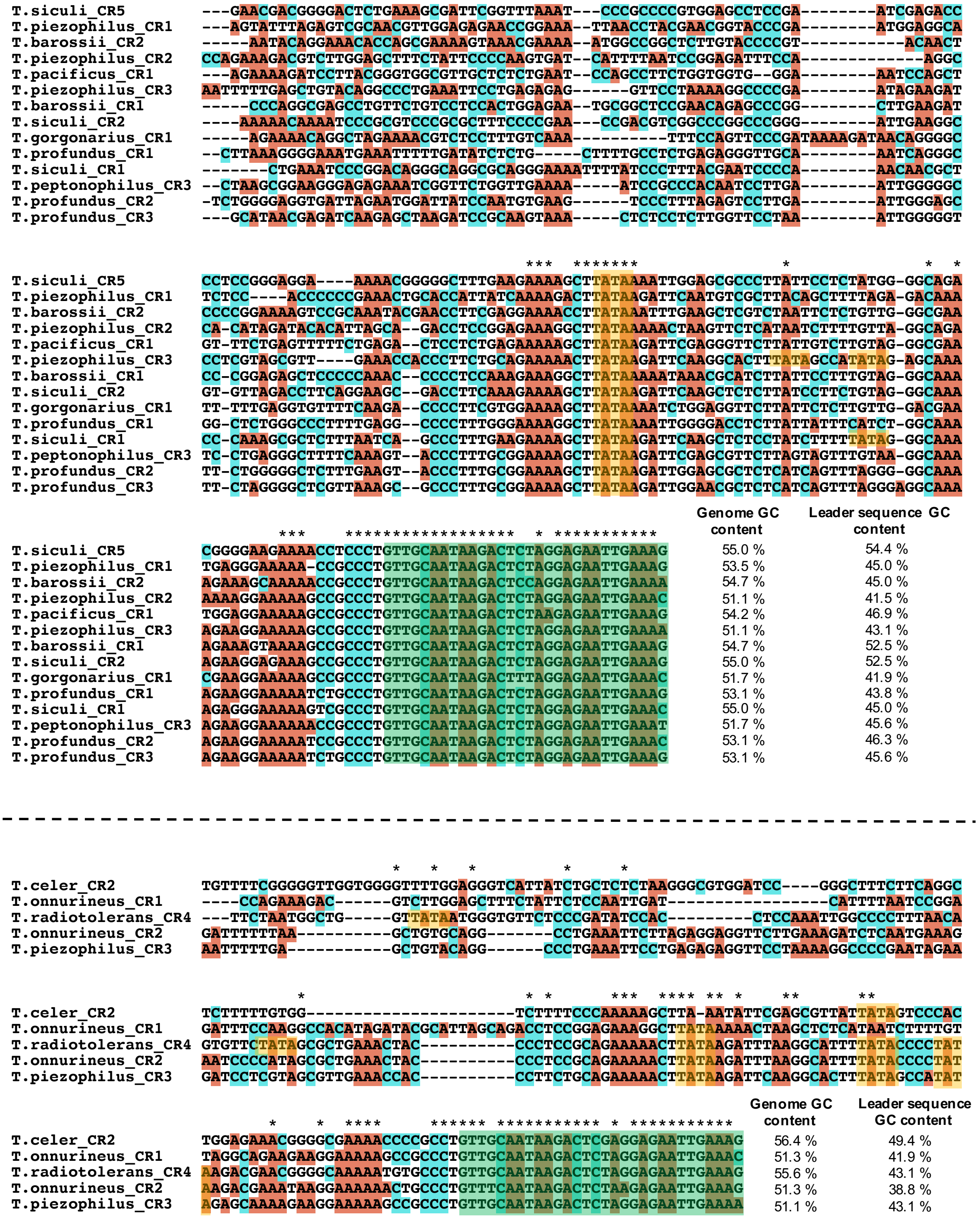

T.siculi_CR5 ---GAACGACGGGGACTCTGAAAGCGATTCGGTTTAAAT---CCCGCCCCGTGGAGCCTCCGA--------ATCGAGACC
T.piezophilus_CR1 ---AGTATTTAGAGTCGCAACGTTGGAGAGAACCGGAAA--TTAACCTACGAACGGTACCCGA--------ATGGAGGCA
T.barossii_CR2 -----AATACAGGAAACACCAGCGAAAAGTAAACGAAAA--ATGGCCGGCTCTTGTACCCCGT-----------ACAACT
T.piezophilus_CR2 CCAGAAAGACGTCTTGGAGCTTTCTATTCCCCAAGTGAT--CATTTTAATCCGGAGATTTCCA------------AGGC
T.pacificus_CR1 ---AGAAAAGATCCTTACGGGTGGCGTTGCTCTCTGAAT--CCAGCCTTCTGGTGGTG--GGA--------AATCCAGCT
T.piezophilus_CR3 AATTTTTGAGCTGTACAGGCCCTGAAATTCCTGAGAGAG------GTTCCTAAAAGGCCCCGA--------ATAGAAGAT
T.barossii_CR1 -----CCCAGGCGAGCCTGTTCTGTCCTCCACTGGAGAA--TGCGGCTCCGAACAGAGCCCGG--------CTTGAAGAT
T.siculi_CR2 ----AAAAACAAAATCCCGCGTCCCGCGCTTTCCCCGAA---CCGACGTCGGCCCGGCCCGGG--------ATTGAAGGC
T.gorgonarius_CR1 -----AGAAAACAGGCTAGAAAACGTCTCCTTTGTCAAA----------TTTCCAGTTCCCGATAAAAGATAACAGGGGC
T.profundus_CR1 --CTTAAAGGGGAAATGAAATTTTTGATATCTCTG-----CTTTTGCCTCTGAGAGGGTTGCA--------AATCAGGGC
T.siculi_CR1 -------CTGAAATCCCGGACAGGGCAGGCGCAGGGAAAATTTTATCCCTTTACGAATCCCCA--------AACAACGCT
T.peptonophilus_CR3 --CTAAGCGGAAGGGAGAGAAATCGGTTCTGGTTGAAAA-----ATCCGCCCACAATCCTTGA--------ATTGGGGGC
T.profundus_CR2 -TCTGGGGAGGTGATTAGAATGGATTATCCAATGTGAAG------TCCCTTTAGAGTCCTTGA--------ATTGGGAGC
T.profundus_CR3 ---GCATAACGAGATCAAGAGCTAAGATCCGCAAGTAAA-----CTCTCCTCTTGGTTCCTAA--------ATTGGGGGT
T.siculi_CR5 CCTCCGGGAGGA----AAAACGGGGGCTTTGAAGAAAAGCTTATAAAATTGGAGCGCCCTTATTCCTCTATGG-GGCAGA
T.piezophilus_CR1 TCTCC----ACCCCCCGAAACTGCACCATTATCAAAAGACTTATAAGATTCAATGTCGCTTACAGCTTTTAGA-GACAAA
T.barossii_CR2 CCCCGGAAAAGTCCGCAAATACGAACCTTCGAGGAAAACCTTATAAATTTGAAGCTCGTCTAATTCTCTGTTG-GGCGAA
T.piezophilus_CR2 CA-CATAGATACACATTAGCA--GACCTCCGGAGAAAGGCTTATAAAAACTAAGTTCTCATAATCTTTTGTTA-GGCAGA
T.pacificus_CR1 GT-TTCTGAGTTTTTCTGAGA--CTCCTCTGAGAAAAAGCTTATAAGATTCGAGGGTTCTTATTGTCTTGTAG-GGCGAA
T.piezophilus_CR3 CCTCGTAGCGTT----GAAACCACCCTTCTGCAGAAAAACTTATAAGATTCAAGGCACTTTATAGCCATATAG-AGCAAA
T.barossii_CR1 CC-CGGAGAGCTCCCCCAAAC--CCCCTCCAAAGAAAGGCTTATAAAAATAAACGCATCTTATTCCTTTGTAG-GGCAAA
T.siculi_CR2 GT-GTTAGACCTTCAGGAAGC---GACCTTCAAAGAAAAGCTTATAAGATTCAAGCTCTCTTATCCTTCTGTAG-GGCAAA
T.gorgonarius_CR1 TT-TTTGAGGTGTTTTCAAGA--CCCCTTCGTGGAAAAAGCTTATAAATCTGGAGGTTCTTATTCTCTTGTTG-GACGAA
T.profundus_CR1 GG-CTCTGGGCCCTTTTGAGG--CCCCTTTGGGAAAAGGCTTATAAAAATTGGGGACCTCTTATTATTTCATCT-GGCAAA
T.siculi_CR1 CC-CAAAGCGCTCTTTAATCA--GCCCTTTGAAGAAAAGCTTATAAGATTCAAGCTCTCCTATCTTTTTATAG-GGCAAA
T.peptonophilus_CR3 TC-CTGAGGGCTTTTCAAAGT--ACCCTTTGCGGAAAAGCTTATAAGATTCGAGCGTTCTTAGTAGTTTGTAA-GGCAAA
T.profundus_CR2 TT-CTGGGGGCTCTTTGAAGT--ACCCTTTGCGGAAAAGCTTATAAGATTGGAGCGCTCTCATCAGTTTAGGG-GGCAAA
T.profundus_CR3 TT-CTAGGGGCTCGTTAAAGC--GCCCTTTGCGGAAAAGCTTATAAGATTGGAACGCTCTCATCAGTTTAGGGAGGCAAA
Genome GC content
Leader sequence GC content
T.siculi_CR5 CGGGGAAGAAAACCTCCCTGTTGCAATAAGACTCTAGGAGAATTGAAAG 55.0 % 54.4 %
T.piezophilus_CR1 TGAGGGAAAAA-CCGCCCTGTTGCAATAAGACTCTAGGAGAATTGAAAG 53.5 % 45.0 %
T.barossii_CR2 AGAAAGCAAAAACCGCCCTGTTGCAATAAGACTCCAGGAGAATTGAAAA 54.7 % 45.0 %
T.piezophilus_CR2 AAAAGGAAAAAGCCGCCCTGTTGCAATAAGACTCTAGGAGAATTGAAAC 51.1 % 41.5 %
T.pacificus_CR1 TGGAGGAAAAAGCCGCCCTGTTGCAATAAGACTCTAAGAGAATTGAAAG 54.2 % 46.9 %
T.piezophilus_CR3 AGAAGGAAAAAGCCGCCCTGTTGCAATAAGACTCTAGGAGAATTGAAAA 51.1 % 43.1 %
T.barossii_CR1 AGAAAGTAAAAGCCGCCCTGTTGCAATAAGACTCTAGGAGAATTGAAAG 54.7 % 52.5 %
T.siculi_CR2 AGAAGGAGAAAGCCGCCCTGTTGCAATAAGACTCTAGGAGAATTGAAAG 55.0 % 52.5 %
T.gorgonarius_CR1 CGAAGGAAAAAGCCGCCCTGTTGCAATAAGACTTTAGGAGAATTGAAAC 51.7 % 41.9 %
T.profundus_CR1 AGAAGGAAAAATCTGCCCTGTTGCAATAAGACTCTAGGAGAATTGAAAG 53.1 % 43.8 %
T.siculi_CR1 AGAGGGAAAAAGTCGCCCTGTTGCAATAAGACTCTAGGAGAATTGAAAC 55.0 % 45.0 %
T.peptonophilus_CR3 AGAAGGAAAAAACCGCCCTGTTGCAATAAGACTCTAGGAGAATTGAAAT 51.7 % 45.6 %
T.profundus_CR2 AGAAGGAAAAATCCGCCCTGTTGCAATAAGACTCTAGGAGAATTGAAAC 53.1 % 46.3 %
T.profundus_CR3 AGAAGGAAAAATCTGCCCTGTTGCAATAAGACTCTAGGAGAATTGAAAG 53.1 % 45.6 %
T.celer_CR2 TGTTTTCGGGGGTTGGTGGGGTTTTGGAGGGTCATTATCTGCTCTCTAAGGGCGTGGATCC----GGGCTTTCTTCAGGC
T.onnurineus_CR1 ----CCAGAAAGAC------GTCTTGGAGCTTTCTATTCTCCAATTGAT----------------CATTTTAATCCGGA
T.radiotolerans_CR4 ---TTCTAATGGCTG-----GTTATAATGGGTGTTCTCCCGATATCCAC---------CTCCAAATTGGCCCCTTTAACA
T.onnurineus_CR2 GATTTTTTAA----------GCTGTGCAGG-------CCTGAAATTCTTAGAGGAGGTTCTTGAAAGATCTCAATGAAAG
T.piezophilus_CR3 AATTTTTGA-----------GCTGTACAGG------CCCTGAAATTCCTGAGAGAGGTTCCTAAAAGGCCCCGAATAGAA
T.celer_CR2 TCTTTTTGTGG---------------------TCTTTTCCCAAAAAGCTTA-AATATTCGAGCGTTATTATAGTCCCAC
T.onnurineus_CR1 GATTTCCAAGGCCACATAGATACGCATTAGCAGACCTCCGGAGAAAGGCTTATAAAAACTAAGCTCTCATAATCTTTTGT
T.radiotolerans_CR4 GTGTTCTATAGCGCTGAAACTAC----------CCCTCCGCAGAAAAACTTATAAGATTTAAGGCATTTTATACCCCTAT
T.onnurineus_CR2 AATCCCCATAGCGCTGAAACTAC----------CCCTCCGCAGAAAAACTTATAAGATTTAAGGCATTTTATACCCCTAT
T.piezophilus_CR3 GATCCTCGTAGCGTTGAAACCAC----------CCTTCTGCAGAAAAACTTATAAGATTCAAGGCACTTTATAGCCATAT
Genome GC content
Leader sequence GC content
T.celer_CR2 TGGAGAAACGGGGCGAAAACCCCGCCTGTTGCAATAAGACTCGAGGAGAATTGAAAG 56.4 % 49.4 %
T.onnurineus_CR1 TAGGCAGAAGAAGGAAAAACCGCCCTGTTGCAATAAGACTCTAGGAGAATTGAAAC 51.3 % 41.9 %
T.radiotolerans_CR4 AAGACGAACGGGGCAAAAATGTGCCCTGTTGCAATAAGACTCTAGGAGAATTGAAAG 55.6 % 43.1 %
T.onnurineus_CR2 AAGACGAAATAAGGAAAAAACTGCCCTGTTTCAATAAGACTCTAAGAGAATTGAAAG 51.3 % 38.8 %
T.piezophilus_CR3 AGAGCAAAAGAAGGAAAAAGCCGCCCTGTTGCAATAAGACTCTAGGAGAATTGAAAA 51.1 % 43.1 %

Figure 6. Putative leader sequences of CRISPR loci in *Thermococcus* genomes. 160-bp sequences upstream of each CRISPR locus were retrieved and subsequently aligned using ClustalX. Alignments are shown for groups of leader sequences that share conserved regions. Fully conserved regions are marked with an asterisk (*). DRs and TATA boxes are displayed in green and yellow, correspondingly.

3.3 Direct Repeats analysis.

The DRs in *Thermococcus* CRISPR loci range from 24 to 30 nt. DRs that differ in length and sequence are present in different CRISPR loci within the same *Thermococcus* species (**Appendix Table 1**). In other words, the DR consensus from different CRISPR loci in the same *Thermococcus* genome may not necessarily be the same, indicating that there is DRs diversity within the same genome.

A phylogenetic analysis with the 79 consensus DRs was performed. The phylogenetic tree classified the spacers into two groups. The group 1, blue-coloured in the tree, is composed of DRs of 24, 28 and 29 bp. This group was further divided into four subgroups, coloured in different blue shades, according to the similarity between the DRs. On the other hand, group 2, green-coloured in the tree, is composed of all DRs of 30 bp and two DRs of 29 bp, which have most likely suffered a nucleotide deletion at the 3'- terminus (*T. barossi* CR1 and *T. pacificus* CR1). This group was divided into seven subgroups, coloured with different green shades, based on the DRs similarity (**Fig. 7**).

Identical consensus DRs are widespread in distinct *Thermococcus* genomes. For instance, the most notable result was one consensus DR, GTTGCAATAAGACTCTAGGAGAATTGAAAG, that is present in up to seventeen different CRISPR loci (in ten distinct *Thermococcus* species). All these DRs are comprised in a cluster within the subgroup 2.7 (**Fig. 7**).

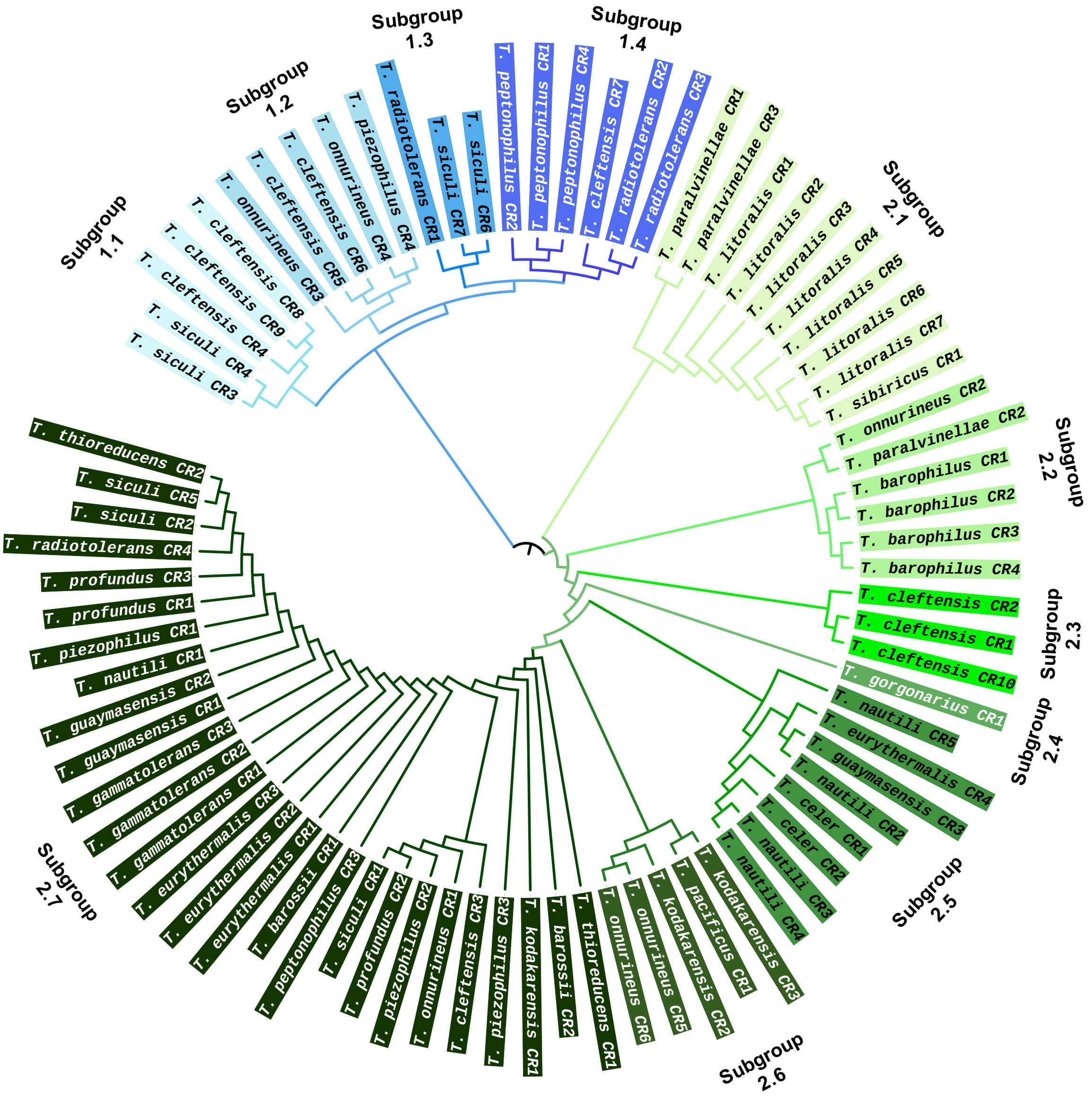


Figure 7. Phylogenetic tree for all consensus DRs within *Thermococcus* genomes. The CRISPR locus to which each DR belongs is shown after the species name. All consensus DRs were classified into two groups (coloured in blue and in green), and eleven different subgroups (coloured in different blue and green shades).

Next, the conservation of the DRs of these groups and subgroups was investigated by generating sequence logos using WebLogo. By and large, sequence logos provide more accurate and visual descriptions than consensus sequences. The DR from group 1 are more conserved than the DRs from group 2 (**Fig. 8**). The consensus DRs from Group 1 are composed of the conserved 5'-GTT terminus and the 3'-AGAATTGGAAA(C/G/T) terminus. By contrast, the consensus DRs from Group 2 are generally based on the conserved 5'-GTTT terminus and the 3'- GGAAA(C/G) terminus. The conservation of consensus DRs slightly varies between subgroups (**Fig. 8**).

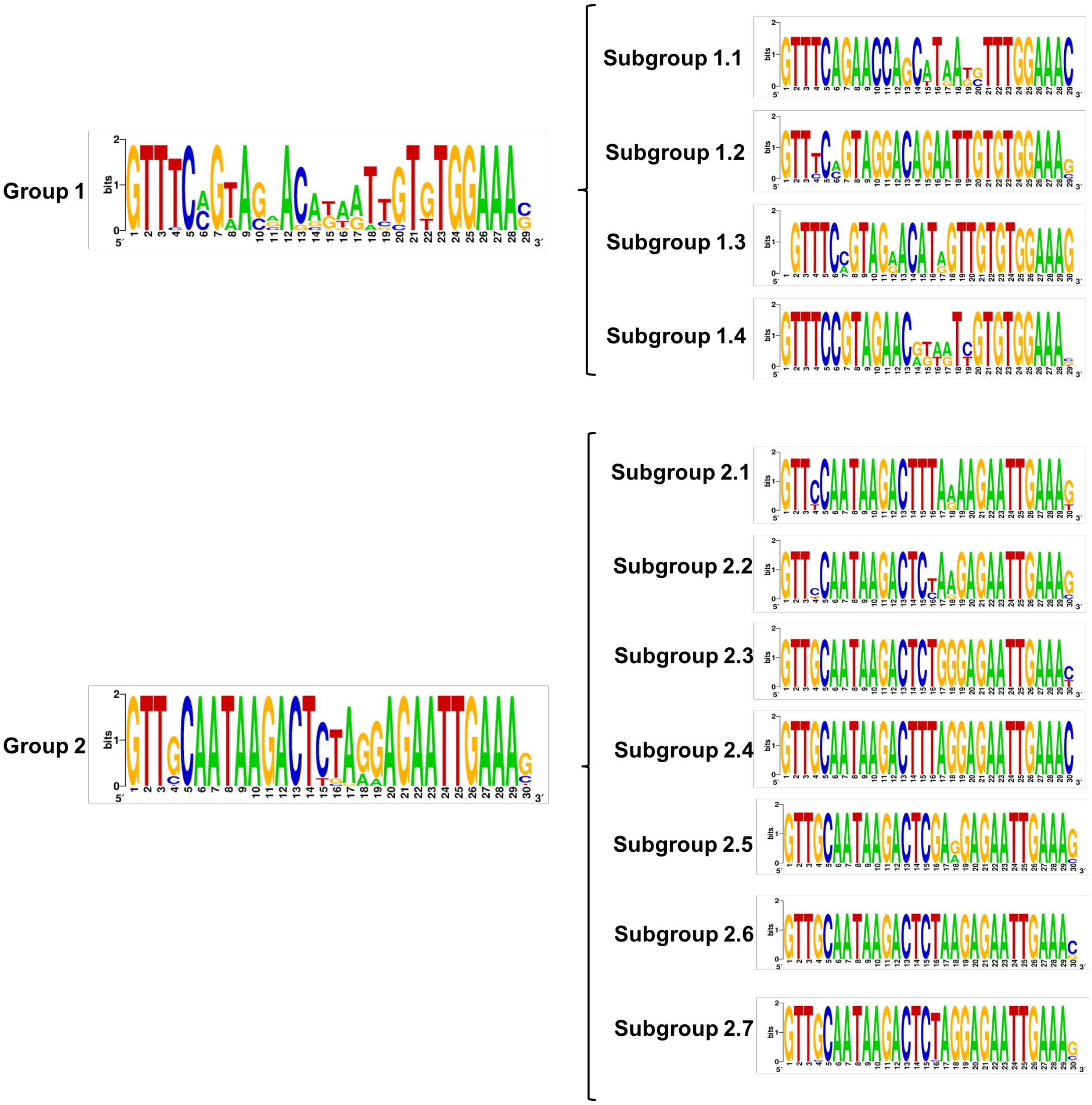


Figure 8. Conservation of the DRs within *Thermococcus* genomes. The conservation of the two groups and the eleven subgroups is shown. Sequence logo was generated using WebLogo. The relative frequency of the corresponding nucleotide at each position is displayed by the height of the letters.

Finally, the putative secondary RNA structure of the most representative consensus DR from each subgroup was examined. This RNA secondary structure is part of a functional crRNA, which is essential for the targeting of foreign nucleic acids [8]. After predicting the structures via RNAFold, different RNA secondary structures were observed among the subgroups (**Fig. 9**). The predicted structures of the subgroups from group 1 were nevertheless quite similar. Also, some predicted structures of the subgroups from group 2 were the same (subgroup 2.1 and 2.4, and subgroup 2.3 and 2.6) (**Fig. 9**).

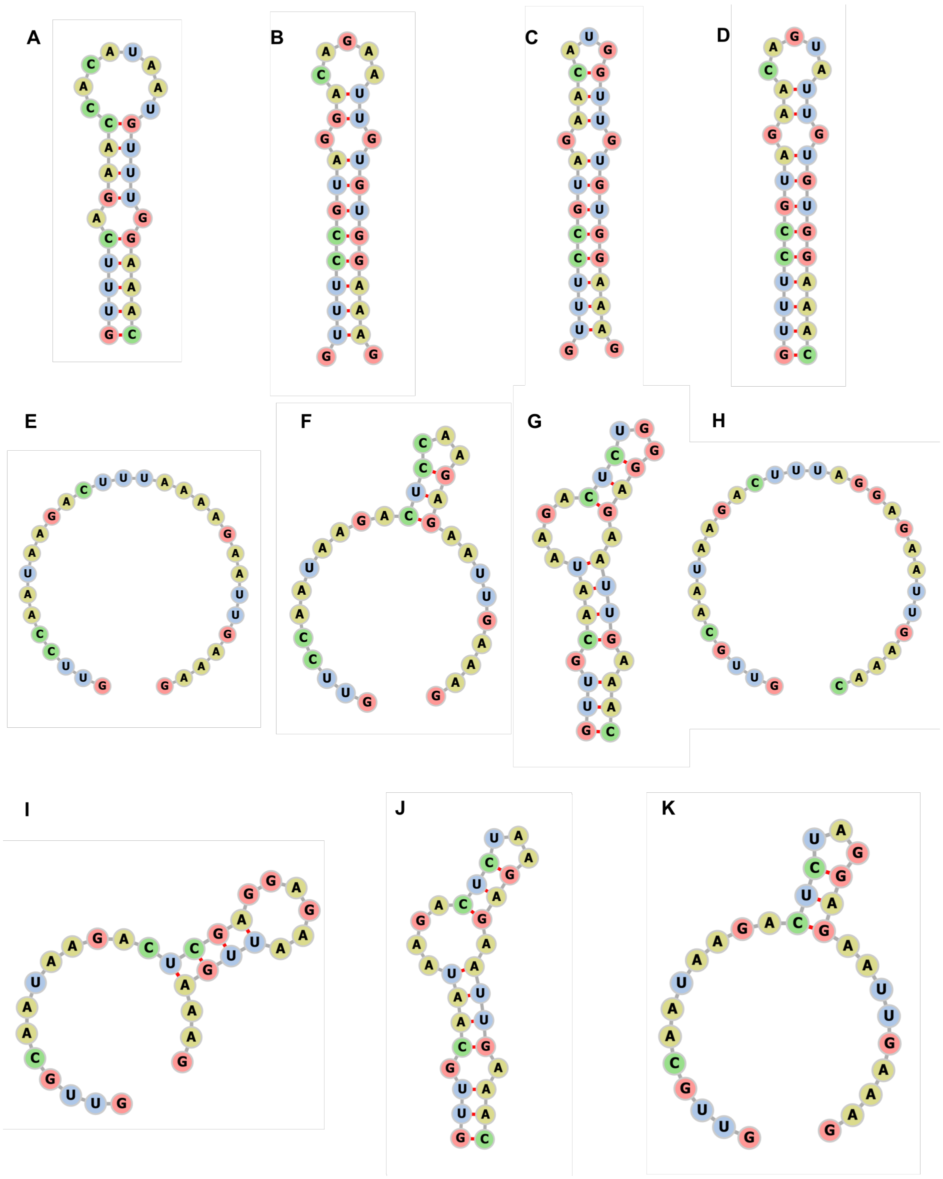


Figure 9. Secondary structures of the DRs predicted by RNAFold. The structures are coloured based on the sequence composition. **A.** Subgroup 1.1; structure corresponding to the consensus DR of *T. cleftensis* CR4 and CR9. **B.** Subgroup 1.2; structure corresponding to the consensus DR of *T. cleftensis* CR5 and CR6. **C.** Subgroup 1.3; structure corresponding to the consensus DR of *T. siculi* CR7. **D.** Subgroup 1.4; structure corresponding to the consensus DR of *T. radiotolerans* CR2 and CR3. **E.** Subgroup 2.1; structure corresponding to the consensus DR of *T. litoralis* CR2, CR3, CR4, CR5, CR6, CR7, and *T. sibiricus* CR1. **F.** Subgroup 2.2; structure corresponding to the consensus DR of *T. barophilus* CR1 and CR2. **G.** Subgroup 2.3; structure corresponding to the consensus DR of *T. cleftensis* CR1 and CR10. **H.** Subgroup 2.4; structure corresponding to the consensus DR of *T. gorgonarius* CR1. **I.** Subgroup 2.5; structure corresponding to the consensus DR of *T. nautili* CR3 and CR4, *T. celer* CR1 and CR2. **J.** Subgroup 2.6; structure corresponding to the consensus DR of *T. kodakarensis* CR2, *T. onnurineus* CR6 and CR6. **K.** Subgroup 2.7; structure corresponding to the consensus DR of *T. barossi* CR1, *T. eurythermalis* CR1, CR2 and CR3, *T. gammatolerans* CR1, CR2 and CR3, *T. guaymasensis* CR2, *T. nautili* CR1, *T. piezophilus* CR1, *T. profundus* CR1 and CR3, *T. radiotolerans* CR4, *T. siculi* CR2 and CR5, and *T. thioreducens* CR2.

3.4 Spacers analysis.

A total of 1800 spacers are contained in the 79 CRISPR loci identified in *Thermococcus* genomes. These spacers range from 25 (*T. piezophilus* CR3 spacer 11) to 70 (*T. radiotolerans* CR3 spacer 10) bp.

The number of spacers contained in the CRISPR loci of each CRISPR/Cas type system harbored in *Thermococcus* genomes was examined. I-A/I-B CRISPR/Cas systems, which are composed of one module corresponding to a I-A type and another module corresponding to an I-B type, harbored a higher average number of spacers, roughly 39. This elevated average is due to the presence of 71 spacers in one CRISPR locus (*T. eurythermalis* CR2; **Appendix Table 1**). In the second position are the I-B CRISPR/Cas systems, with an average of approximately 26 spacers (**Fig. 10**).

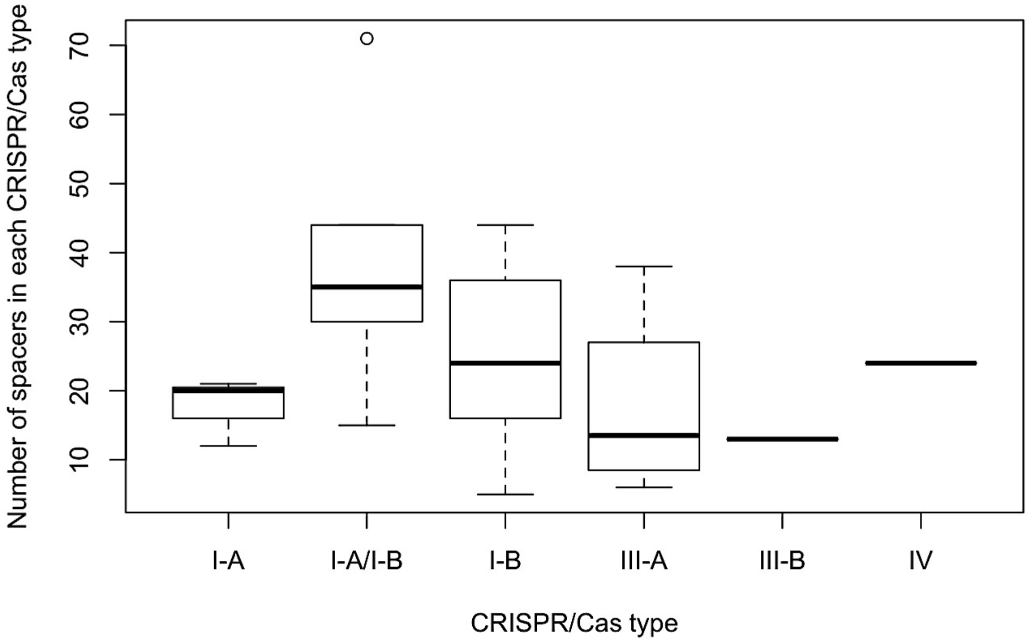


Figure 10. Boxplot representation of the number of spacers detected in the CRISPR loci of each CRISPR/Cas system identified in *Thermococcus*.

Regarding the CRISPRTarget and BLASTn results, 42 matches between *Thermococcus* spacers and protospacers were found (**Appendix Fig. 1, Appendix Table 2**). These 42 matches corresponded to only 32 different spacers.

No matches corresponded to TPV1 or any other bacteriophage. To confirm these results, the whole genomes of TPV1 and PAV1 were subjected to a BLASTn search. No matching sequences were found between *Thermococcus* spacers and the TPV1 and PAV1 genomes. Surprisingly, it was obtained TPV1 short sequences completely matching a region of the *tRNA-gly* gene from several *Thermococcus* species, including *T. onnurineus*, *T. nautili, T. paralvinellae* and two unclassified *Thermococcus* species.

The 42 matches found were spacers that targeted plasmids, *Thermococcus* genomes, and other archaea's genomes (**Appendix Fig. 1, Appendix Table 2**). More precisely, the results obtained were:

- Three spacers in *T. barophilus* and a spacer in *T. profundus* matching sequences from *P. yayanosii* CH, and a spacer in *T. nautili* matching a sequence from the archaea *Palaeococcus ferrophilus* DSM 13482.

- Eleven spacers targeting five different plasmids: pMETVU01 from *Methanocaldococcus vulcaniu,* pTN3 from *T. nautili*, and pT26-2, pAMT7, pIRI33 from unclassified *Thermococcus* species.

- Five spacers in *T. barophilus* MP matching sequences from the other *T. barophilus* strain, *T. barophilus* CH5.

- Nineteen spacers matching sequences from other *Thermococcus* species.

- Four self-targeting spacers, in *T. cleftensis*, *T. nautili*, *T. paralvinellae* and *T. siculi.* Self-targeting spacers match sequences from the same genome or plasmids present in the same cell.

Then the identity of all target sequences was investigated. Most of the sequences belong to hypothetical proteins, but known proteins were also identified. These proteins are: HTH domain-containing protein, DUF87 domain-containing protein, t26-26p, t26-16p, t26-15p, t26-14p, SAM pointed domain-containing ETS-like transcription factor, one integrase, S8 family peptidase, SPOUT family RNA methylase and tRNA (N(6)-L-threonylcarbamoyladenosine(37)-C(2))-methylthiotransferase (**Appendix Table 2**).

Another result that is noteworthy mentioning is that it was common to find spacers that were repeated in different CRISPR loci within the same *Thermococcus* genome.

Conclusive results regarding the identification of PAM sequences were not obtained. This matter is further discussed in the Discussion section.

3.5 Cas proteins phylogenetic analysis.

Phylogenetic analyses were performed on three different Cas proteins in order to infer the level of conservation and divergence of CRISPR/Cas systems in *Thermococcus*. The analyses were conducted for the Cas1, Cas3 and Cas10 proteins. These proteins were chosen because Cas1 is the most universally conserved protein in CRISPR/Cas systems, and Cas3 and Cas10 are the signature proteins of the type I and type III systems respectively [2] [30]. Hence, Cas1 proteins typically provide insight into the type of CRISPR/Cas systems, Cas3 proteins shed light on the sub-type of type I CRISPR/Cas systems, and Cas10 proteins inform about the subtype of type III CRISPR/Cas systems.

It was considered of interest to include Cas proteins encoded by *Pyrococcus* species in the analysis since it is the closest genus related to *Thermococcus.* Thereby*,* it could be obtained information about the phylogenetic relationships between these two genera concerning CRISPR/Cas systems. The diversity and organization of CRISPR/Cas systems in *Pyrococcus* was revealed in a publication from 2013 [40]. Yet, the genomes of two additional *Pyrococcus* species (*Pyrococcus kukulkanni* and *Pyrococcus chitonophagus*) became available after 2013*.* The CRISPR/Cas systems of these species were unknown, so their identification was done via the same procedure employed for *Thermococcus* genomes.

3.5.1 Cas1.

Nineteen Cas1 proteins belonging to 15 different *Thermococcus* species are depicted in the phylogenetic tree. Moreover, as previously mentioned, it was added in the analysis the Cas1 proteins from *Pyrococcus horikoshii, Pyrococcus furiosus, Pyrococcus yayanosii, Pyrococcus chitonophagus,* and *Pyrococcus kukulkanni.* Finally, the Cas1 proteins from the other major sub-types of type I CRISPR/Cas systems were included. More precisely, it was included the Cas1 proteins from the type I-A system of *Archaeoglobus fulgidus* [2], the type I-C system of *Moraxella bococuli* [41], the type I-D system of *Microcystis aeruginosa* [42], the type I-E system of *Escherichia coli* [2], and the type I-F system of *Yersinia pseudotuberculosis* [2]. Also, note that the partially annotated Cas1 protein from *T. thioreducens* was not included in the analysis (**Fig. 11**).

All *Thermococcus* and *Pyrococcus* Cas1 proteins cluster together at the top of the tree, in an independent cluster, which corresponds to type I-B systems. On the other hand, the Cas1 proteins form the rest of the type I systems are found at the bottom of the tree in different clusters. It is also noteworthy that the Cas1 proteins of *P. yayanosii* (CR5) and *P. chitonophagus* (CR2) are phylogenetically closer to *Thermococcus* species than to *Pyrococcus* species (**Fig. 11**).

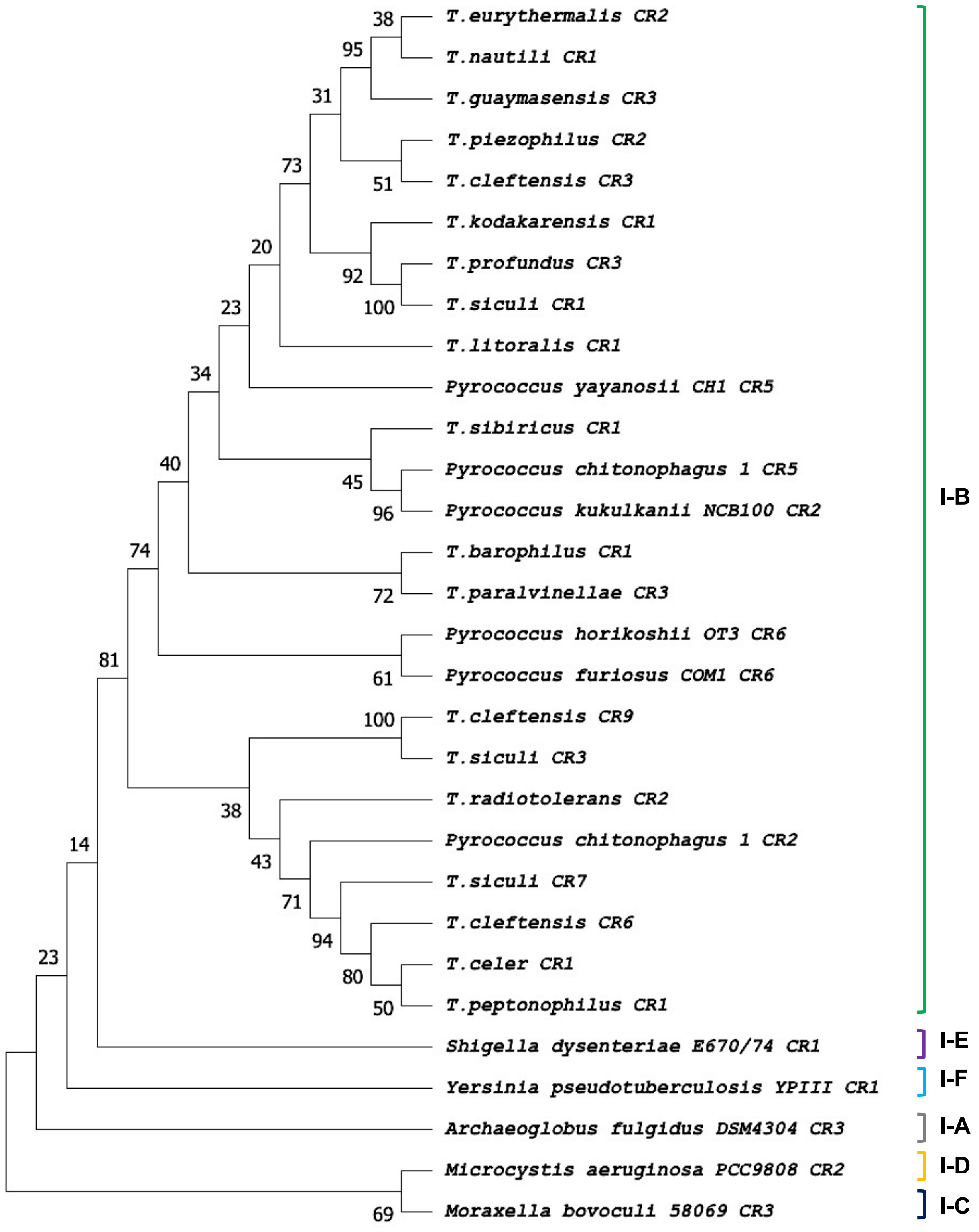


Figure 11. Phylogenetic tree for Cas1 proteins in *Thermococcus*. The CRISPR locus to which each Cas1 belongs is shown after the species name. Besides, representative Cas1 proteins from all six major sub-types of type I CRISPR/Cas systems, I-A to I-F, were included in the analysis.

3.5.2 Cas3.

Twenty-two Cas3 proteins belonging to 14 different *Thermococcus* species are depicted in the phylogenetic tree (**Fig. 12**). Besides, it was included in the analysis Cas3 proteins from *P. horikoshii, P. furiosus, P. yayanosii, P. chitonophagus* and *P. kukulkanni.*

The Cas3 proteins from *T. sibiricus* (CR1) and *T. profundus* (CR3.2) were excluded from the analysis because they presented a really low homology with other *Thermococcus* Cas3 proteins. BLASTx was used to find similar proteins to these two Cas3 proteins. Regarding the Cas3 protein of *T. profundus* (CR3.2), this protein showed a lower than 35% of homology to any other protein from the database. Surprisingly, this protein is more similar (34.87% of homology) to the Cas3 protein of *Dictyoglomus thermophilum* than to any other Cas3 protein encoded by *Thermococcus* (29.52% of homology with an unclassified *Thermococcus* species). The Cas3 protein of *T. sibiricus* (CR1), instead, possesses a relatively higher homology with Cas3 proteins encoded by *Thermococcus* (34.72 % of homology with an unclassified *Thermococcus* species). This protein is notwithstanding much more similar to Cas3 proteins from other microorganisms. For instance, it was obtained a 65.44% of homology with the Cas3 protein of Candidatus *Methanofastidiosum methylthiophilus,* and 56.91% of homology with Candidatus *Desulfofervidus auxilii.*

As for the results of the analysis, three type I-B clusters and two type I-A clusters were identified (**Fig. 12**).

Firstly, the genomic sequences of all type I-B Cas3 proteins were investigated. It was found that the proteins differ in length between the three clusters. The Cas3 proteins of the top type I-B cluster range from 725 aa (in *P. chitonophagus* CR5*, P. kukulkanni* CR2.2*, P. furiosus* CR6 and *P. horikoshii* CR6) to 759 aa (in *T. barophilus* CR1). The Cas3 proteins of the middle type I-B cluster are nevertheless shorter, ranging from 690 aa (in *P. chitonophagus* CR2) to 706 aa (in *T. cleftensis* CR6). Finally, the Cas3 proteins of the bottom type I-B cluster are the longest ones, 869 aa (in *T. eurythermalis* CR2.2) and 940 aa long (in *T. siculi* CR1.2). Then, the attention was turned to type I-A clusters, and similar results were obtained. The Cas3 proteins of the top type I-A cluster range from 651 aa (in *P. kukulkanni* CR2.1) to 678 aa (in *T. thioreducens* CR1). Instead, the Cas3 proteins of the bottom type I-A cluster range from 522 aa (in *T. guaymasensis* CR3.1 and *T. eurythermalis* CR2.1) to 534 aa (in *T. profundus* CR3.1). All type I-A Cas3 proteins are shorter than type I-B Cas3 proteins. These results confirm that there are distinct *cas* loci widespread in both *Thermococcus* and *Pyrococcus.*

In terms of phylogenetic relationships between both genera, it was also found that the type I-B Cas3 protein of *P. yayanosii* (CR5.2) is phylogenetically closer to *Thermococcus* species than to *Pyrococcus* species. The same finding was obtained for the type I-A Cas3 proteins of *P. chitonophagus* (CR2) and *P. kukulkanni (*CR2.1).

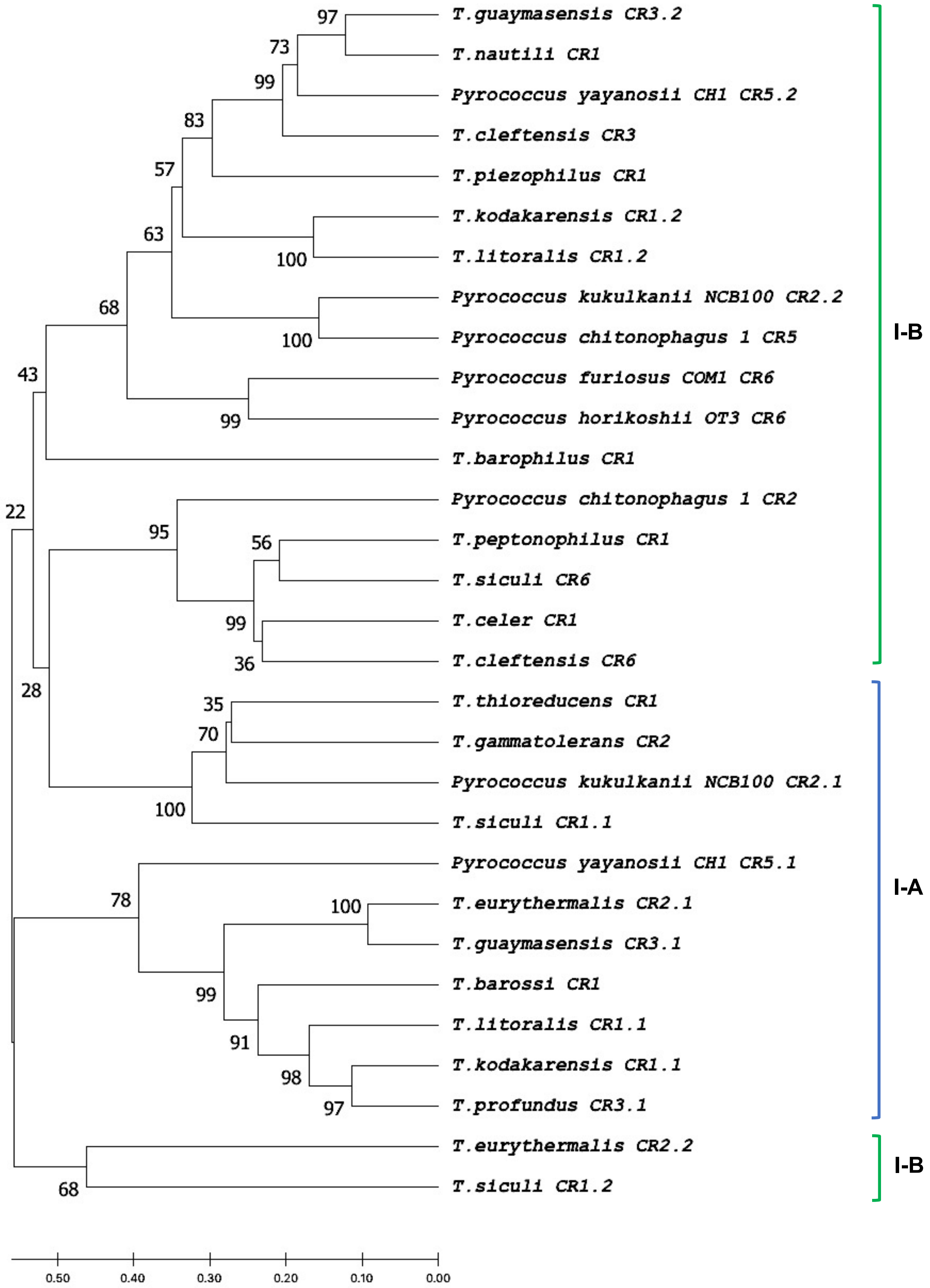


Figure 12. Phylogenetic tree for Cas3 proteins in *Thermococcus*. The CRISPR locus to which each Cas3 belongs is shown after the species name.

3.5.3 Cas10.

Five Cas10 proteins belonging to five different *Thermococcus* species are depicted in the phylogenetic tree. Furthermore, it was added in the analysis the Cas10 proteins from the type III CRISPR/Cas systems of *P. yayanosii, P. furiosus, P. horikoshii,* and *P. chitonophagus.* Lastly*,* the Cas10 proteins from both the type III-A system of *Thermococcus thermophilus* [43] and the type III-B system of *Archaeoglobus fulgidus* were included [2] (**Fig. 13**).

In the phylogenetic tree, type III-A Cas10 proteins cluster together at the top of the tree, whereas all type III-B Cas10 proteins are clustered at the bottom. An interesting finding is that the type III-A Cas10 protein of *P. yayanosii* (CR7) is phylogenetically closer to *Thermococcus* Cas10 proteins than to *Pyrococcus* Cas10 proteins. The same occurs with the type III-B Cas10 protein of *P. yayanosii* (CR4). This protein is phylogenetically more related to the type III-B Cas10 protein of *Thermococcus siculi* (CR4) than to *Pyrococcus* type III-B Cas10 proteins (**Fig. 13**).

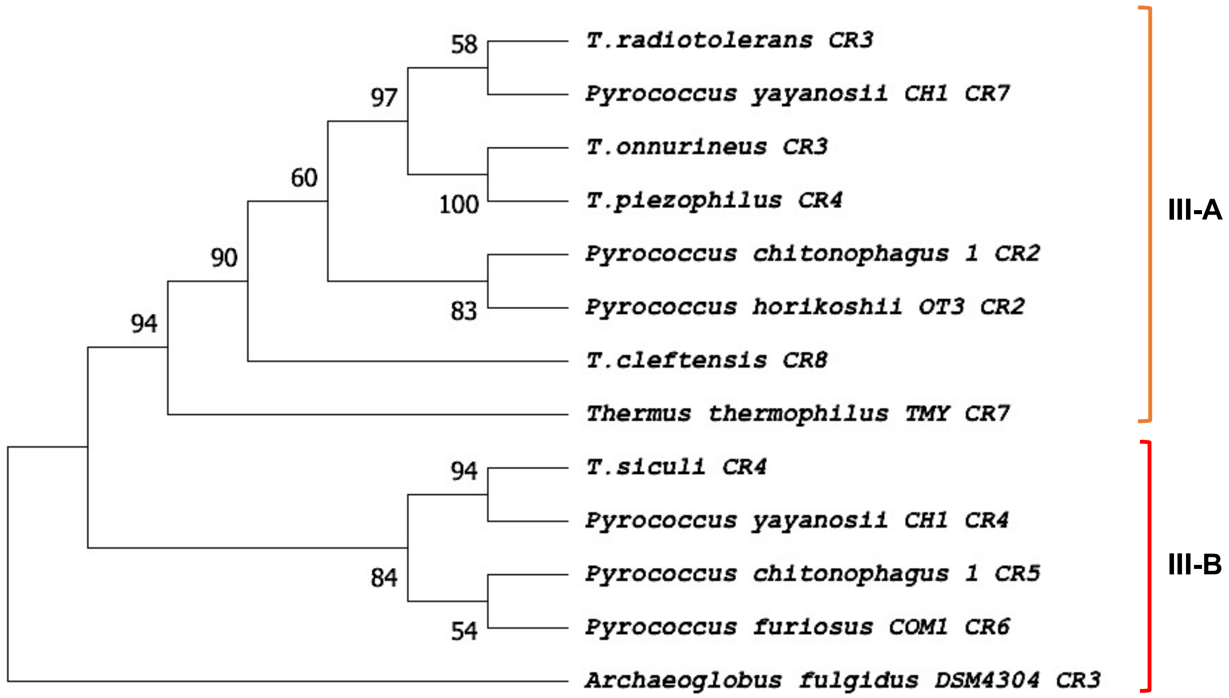

58
97
60
100
90
83
94
94
84
54
T.radiotolerans CR3
Pyrococcus yayanosii CH1 CR7
T.onnurineus CR3
T.piezophilus CR4
Pyrococcus chitonophagus 1 CR2
Pyrococcus horikoshii OT3 CR2
T.cleftensis CR8
Thermus thermophilus TMY CR7
T.siculi CR4
Pyrococcus yayanosii CH1 CR4
Pyrococcus chitonophagus 1 CR5
Pyrococcus furiosus COM1 CR6
Archaeoglobus fulgidus DSM4304 CR3
III-A
III-B

Figure 13*.* Phylogenetic tree for Cas10 proteins in *Thermococcus*. The CRISPR locus to which each Cas10 belongs is shown after the species name. In addition, representative Cas10 proteins from type III-A and type III-B CRISPR/Cas systems of other microorganisms were included in the analysis.

4. Discussion

The present study has identified and characterized the CRISPR/Cas systems in *Thermococcus*. A high occurrence and a wide diversity of CRISPR/Cas systems in *Thermococcus* genomes has been described. More accurately, the prevalence of CRISPR loci was really high given that 79 loci were identified in the 21 *Thermococcus* genomes analysed. Nevertheless, only 23 of these CRISPR loci have (non- truncated) cas locus/loci adjacent, so a total of 23 CRISPR/Cas systems are present in the *Thermococcus* species analysed. Not all the *Thermococcus* species encode for CRISPR/Cas systems, but an elevated number do. The prevalence of these systems in *Thermococcus* has been estimated to be around 81%. This figure is close to the estimated prevalence of CRISPR/Cas systems in archaea, which is thought to be nearly 90% [3].

Regarding the diversity, only class 1 CRISPR/Cas systems are present in *Thermococcus*, but the three types of systems that belong to this class (type I, III and IV) were identified. Moreover, two subtypes for both type I (A and B) and III (A and B) systems were detected. Similar bioinformatics analyses have been performed on other genera or species of archaea and especially bacteria, aiming to determine the occurrence and diversity of CRISPR/Cas systems in these species. For instance, computer-based approaches have studied the systems of *Pyrococcus* [40], *Mycobacterium* [44], *Salmonella* [45], *Streptomyces* [46], *Bifidobacterium* [37], *Klebsiella* [39], *Listeria monocytogenes* [47], *Bacteroides fragilis* [48], *Riemerella anatipestifer* [49], among others. However, such a rich diversity of CRISPR/Cas systems has not been described in these genera and species. This result confirms that the selection of *Thermococcus* as the target genus for this study was suitable.

The genomic organization of the *cas* loci happens to be pretty much alike in various *Thermococcus* species, suggesting that horizontal gene transfer (HGT) associated with *cas* loci among *Thermococcus* species has occurred. Another remarkable result was the identification of *cas* loci composed of a *cas* locus type I-A and a *cas* locus type I-B clustered together. These particular (non-truncated) *cas* loci were observed in six different *Thermococcus* species. I-A/I-B *cas* locus is highly uncommon, but it was previously described in *Pyrococcus*, the phylogenetically closest genus to *Thermococcus*. The occurrence, diversity and genomic organization of CRISPR/Cas systems in *Pyrococcus* was already described in a study from 2013. This study revealed type I (subtypes A and B) and III (subtypes A and B) systems in the genomes of six *Pyrococcus* species [40]. Furthermore, the genomic organization of the *cas* loci in *Pyrococcus* has turned out to be practically identical to the genomic organization of the *cas* loci observed in *Thermococcus.*

The *cas* loci similarity between *Thermococcus* and *Pyrococcus* led to investigate their phylogenetic relationships. The phylogenetic analysis revealed that some *Pyrococcus* Cas3 and Cas10 proteins are phylogenetically closer to *Thermococcus* Cas3 and Cas10 proteins than to other *Pyrococcus* Cas proteins. All these observations point out at likely HGT events between both genera. It is common knowledge that HGT involving CRISPR/Cas systems (or its loci independently) is recurrent among prokaryotes, and regularly this transfer is responsible for the rearrangement of the loci in the operons [50]. HGT mechanisms differ between archaea and bacteria. For example, despite the fact that classic bacterial HGT mechanisms (conjugation, transformation and transduction) have been described in particular archaea, alternative mechanisms have been identified for this domain as well. One of such mechanisms is the genetic material exchange via vesicles [51]. Even so, how genetic material is transferred among archaea is not well understood yet. HGT is nevertheless known to be incredibly frequent among archaea, and sometimes it takes place even among less-related populations [51].

One of the most shocking results was to identify only one type IV system, in *T. onnurineus*. A highly similar type IV *cas* locus is present in *T. pacificus*, but with no CRISPR locus nearby. Not surprisingly, no type IV systems are present in *Pyrococcus*, which makes the presence of type IV systems in the family *Thermococcaceae* extremely unusual. Type IV systems possibly are the least understood systems since several of its activities are still not clear. Furthermore, these systems lack several hallmark genes typical from other CRISPR/Cas systems, which further complicates its identification [52]. Yet, type IV systems are known to employ effector complexes rather than single nucleases, so they have been classified within class 1 CRISPR/Cas systems [53]. These systems are being investigated, and in 2019, using *Aromatoleum aromaticum* as a model, the processes of crRNA maturation and complex formation were elucidated [54]. Besides, little is known about the distribution of type IV systems in archaea. It could be of great interest to conduct an extensive phylogenetic analysis primarily aimed at determining how the type IV system from *T. onnurineus* has reached this *Thermococcus* species (also the *cas* loci in *T. pacificus*) and gaining insight into the occurrence of such systems in this domain.

The constituents of CRISPR loci (i.e. leader sequences, DRs and spacers) were also analysed. Conserved regions within leader sequences were common between certain *Thermococcus* species. A recent study has shown that conserved regions or motifs in the leader sequences of CRISPR loci are involved in the control of the spacers acquisition, which is a process of the adaptation stage, in type I-D systems [55]. Therefore, apart from regulating the transcription of CRISPR loci, other specific functions are being associated with the conserved motifs of leader sequences, highlighting the relevance for such conservation. Similarly, conservation among DRs was also observed. DRs were classified into two groups and eleven subgroups. Remarkably, the conservation of all the DRs classified into group 2 was high, suggesting that all these DRs probably arose from a single common DR ancestor. Both leader sequences and DRs observations support that HGT associated with CRISPR loci among *Thermococcus* species has most likely taken place.

Given that spacers play an imperative role in the interference stage, it was conducted a spacer analysis to reveal the identity of the protospacers. Unfortunately, this analysis could not gain insight into the phage exposure in *Thermococcus* because only one phage that infects *Thermococcus* has been hitherto described, TPV1. Even so, no spacers matching sequences from TPV1 were obtained. It would be interesting to unravel whether any spacer from the CRISPR loci of *T. prieurii* matches a TPV1 genomic sequence. The genome of this archaea is nevertheless not currently available. Hopefully, more efforts are made in the coming years to isolate and sequence new viruses from *Thermococcaceae.* Nonetheless, it should be noted that these archaea inhabit in deep-sea and terrestrial hot environments, which thoroughly complicates the collection of samples and its due analysis since scrupulous enrichment culture conditions are essential [56].

On the other hand, fortunately, several matches were obtained in the spacers analysis, 42 to be more precise. Eleven of these matches targeted five plasmids, four of such were *Thermococcus* plasmids. This indicates that, to some extent, there is more knowledge on plasmids than on phages for this species (and thermophilic archaea). Still, this knowledge is not extensive whatsoever. By the end of the 2000s and the beginning of the 2010s, there was a spark of interest in the isolation and characterization of mobile genetic elements (i.e. the mobilome) from *Thermococcus* and other thermophilic archaea. For instance, a study from 2013 characterized five newly discovered *Thermococcus* plasmids [57]. This spark was principally ignited due to the potential use of said plasmids for the development of new genetic tools for thermophilic archaea [58]. As a curiosity, *T. kodakarensis* has been established as the preferred model to study archaeal genetics, metabolism, biochemistry, among others. Therefore, both the genome and plasmids of this species have been subjected to a wide range of genetic manipulations [59]. Sadly, the isolation and characterization of the mobilome of *Thermococcus* happened to have ceased in the mid and late 2010s. This fact has hampered the mobilome characterization in *Thermococcus* and consequently the identification of spacers in *Thermococcus* CRISPR/Cas systems.

Most of the remaining matches corresponded to spacers that were identical or highly similar to sequences from distinct species or strains of *Thermococcus, Pyrococcus* and *Palaeococcus ferrophilus* (the latter belongs to *Thermococcaceae* family as well). Interestingly, four matches corresponded to self-targeting spacers. Hence, the ratio of self-targeting spacers is high (9.5%, 4/42). The incorporation of self-nucleic acid in CRISPR loci might result in autoimmunity, and it is a rather common phenomenon in prokaryotes [60]. Two of the self-targeted genes are chromosome-located and encode for a transfer RNA (tRNA) methylase and a tRNA methylthiotransferase. Both enzymes are involved in the modification of tRNA and potentially participate in various cellular processes [61][62]. Thus, the disruption of these genes would probably have a negative impact on the cells. However, cells are believed to evade CRISPR/Cas-based autoimmunity owing to mutations on *cas* genes, leader sequences, DRs, spacers, protospacers and PAM sequences [60]. For instance, a typical mechanism for phages to avoid CRSPR/Cas interference is through mutations at PAM sequences [63].

As mentioned in the Results section, no solid findings were obtained concerning the identification of PAM sequences. It is necessary to remark that the identity of only 32 out of 1800 spacers (0.0178%) could be revealed. Another fact that should be noted is that PAM sequences are only required for some specific CRISPR/Cas systems. Regarding class 1 systems, PAM sequences are only needed for type I systems. In such systems, PAM sequences are based on two or three conserved nucleotides located directly adjacent or one position after only one end of the protospacers [9]. This PAM sequence is vital for the recognition of the target DNA by the Cascade complex. In contrast, there is no evidence that PAM recognition is required for type III systems, whereas the interference stage is not well-understood for type IV systems yet [53]. As regard to class 2 systems, PAM sequences play an imperative role for type II and type IV systems [64], and a similar motif, called protospacer flanking site (also referred to as PFS) is typically involved in type VI systems [65]. Altogether, given the low percentage of spacers that could be identified and the absence of class 2 systems, it was impossible to obtain strong evidence for the identification of PAM sequences in *Thermococcus*.

Overall, this study has provided valuable information for the characterization of CRISPR/Cas systems in *Thermococcus* by revealing their occurrence, diversity, organization and structural features. The following section lists the concluding remarks of this study as well as suggests other studies/analyses that could derive from the present study. A brief final assessment of the study is also provided.

5. Conclusions

The main concluding remarks of this comparative genomics study are:

- At least one CRISPR locus is present in the genomes of all the *Thermococcus* species analysis. The number of CRISPR loci widely range from one to ten.

- *Cas* loci adjacent to CRISPR loci are less frequent. A total of 23 non-truncated cas loci are harbored in *Thermococcus* genomes, constituting 23 complete CRISPR/Cas systems.

- A high frequency of CRISPR/Cas occurrence has been found in *Thermococcus* (81%) as 17 out of the 21 genomes analysed code for at least one of said systems.

- CRISPR/Cas systems are vastly diverse in *Thermococcus*. All three class 1 types were identified (I, III and IV). Moreover, subtypes I-A and I-B were detected for both type I and III.

- Phylogenetic analyses confirmed that several distinct *cas* loci are widespread within *Thermococcus* genomes, and suggested that HGT events (of *cas* loci) have most likely occurred among *Thermococcus* species.

- HGT events probably also took place between *Pyrococcus* and *Thermococcus* since the genomic organization of *cas* loci is very similar in these genera. Besides, phylogenetic analyses confirmed that several *Pyrococcus* Cas proteins are phylogenetically closer to *Thermococcus* than to *Pyrococcus*.

- Leader sequences analysis revealed conserved regions within these sequences. TATA boxes could be identified for most of the leader sequences.

- DRs were classified into two groups and eleven subgroups according to sequence similarity. The RNA secondary structure of the DRs generally differed between the groups and was more similar within groups (for subgroups).

- The findings related to leader sequences and DRs hint at probable HGT events associated with CRISPR loci within *Thermococcus*.

- The history of mobilome and phage exposure was poorly understood in *Thermococcus*. No phage protospacers were identified. Only eleven protospacers belonging to five different plasmids could be revealed. The remaining of the protospacers belonged to *P. yayanosii*, *P. ferrophilus*, and *Thermococcus* species and strains. Four spacers were self-targeting spacers.

- No PAM sequences could be recognised owing to the low ratio of spacers that could be matched to protospacers (0.0178%).

Taken all together, it may be regarded as certain that the general objectives that were set when writing the proposal of this study have been met.

First of all, basing the study on *Thermococcus* has enabled the identification and understanding of the CRISPR/Cas systems that remained unknown for most of the *Thermococcus* which genome is publicly available to date. The fact that 21 genomes were available has also allowed inferring, to some extent, the evolutionary history of these systems in this archaeal genus. Furthermore, the phylogenetic relationships of both genera (concerning CRISPR/Cas systems) could be investigated as well because CRISPR/Cas-based documentation was already published for *Pyrococcus*.

On the other hand, the limited documentation of CRISPR/Cas systems in thermophilic archaea had a trade-off. Little is known for phages and mobile genetic elements in this kind of microorganisms as they dwell in extreme habitats. Therefore, the identification of protospacers was not as successful as desired. Hopefully, more research is focused on this field in the coming years and more extensive spacers analysis can be performed.

Further analyses and studies could derive from the findings obtained in this study. There are still several well-known *Thermococcus* species without its genome publicly available yet. Moreover, there is a long list of unclassified *Thermococcus*. Should more effort be made in the field of taxonomy, a much more detailed and extensive analysis could be carried out for this genus. Other genera phylogenetically close to *Thermococcus* (e.g. genera within the phylum *Euryarchaeota*) could be included in the analysis as well. That would provide a bigger picture of how CRISPR/Cas systems are widespread in this phylum.

This pipeline can be employed to analyse the CRISPR/Cas systems of other genera or species. The CRISPR/Cas systems of many archaea remain unexplored so far. Additionally, the occurrence of CRISPR/Cas systems in archaea is really high, and its diversity tends to be broad as well. This fact makes this domain particularly suitable for comparative genetics analyses.

A specific analysis focusing on the occurrence and distribution of type IV CRISPR/Cas systems in archaea could be an intriguing and promising study too. As stated above, these systems are not fully understood yet. Thus, a better comprehension of the type IV systems could develop further CRISPR/Cas-based applications of interest (e.g. in the field of genome engineering).

To conclude, it should be emphasized the vital importance of all the online tools that have been developed throughout the past two decades aiming to strongly facilitating the tasks of identifying and characterizing CRISPR/Cas systems in prokaryotes. That has allowed the authors of the study to select the online tools they worked with most comfortably. This selection was made while working on the study proposal. More precisely, a lot of time was devoted to carefully select the methodology of the study and to establish a logical and feasible planning. It was a time well spent as the authors have stuck to this initial planning during the study. Besides, no changes were implemented to ensure the success of the study.

6. Abbreviations

CRISPR: Clustered Regularly Interspaced Short Palindromic Repeats.

Cas: CRISPR-Associated.

DR: Direct Repeat.

BP: Base Pairs.

PAM: Protospacer Adjacent Motif.

CR: CRISPR locus (when associated with a number).

AA: Amino Acid.

NCBI: National Center of Biotechnology Information.

MEGAX: Molecular Evolutionary Genetics Analysis X.

UPGMA: Unweighted Pair Group Method with Arithmetic Mean.

BLAST: Basic Local Alignment Search Tool.

HGT: Horizontal Gene Tranfer.

7. References

[1] J. E. Garneau *et al.*, “The CRISPR/cas bacterial immune system cleaves bacteriophage and plasmid DNA,” *Nature*, 2010.

[2] K. S. Makarova *et al.*, “An updated evolutionary classification of CRISPR-Cas systems,” *Nat. Rev. Microbiol.*, 2015.

[3] F. Hille, H. Richter, S. P. Wong, M. Bratovič, S. Ressel, and E. Charpentier, “The Biology of CRISPR-Cas: Backward and Forward,” *Cell*, 2018.

[4] O. S. Alkhnbashi, S. A. Shah, R. A. Garrett, S. J. Saunders, F. Costa, and R. Backofen, “Characterizing leader sequences of CRISPR loci,” in *Bioinformatics*, 2016.

[5] E. V. Koonin and K. S. Makarova, “Origins and evolution of CRISPR-Cas systems,” *Philosophical Transactions of the Royal Society B: Biological Sciences*. 2019.

[6] S. H. Sternberg, H. Richter, E. Charpentier, and U. Qimron, “Adaptation in CRISPR-Cas Systems,” *Molecular Cell*. 2016.

[7] E. Charpentier, H. Richter, J. van der Oost, and M. F. White, “Biogenesis pathways of RNA guides in archaeal and bacterial CRISPR-Cas adaptive immunity,” *FEMS Microbiology Reviews*. 2015.

[8] N. Agrawal, P. V. N. Dasaradhi, A. Mohmmed, P. Malhotra, R. K. Bhatnagar, and S. K. Mukherjee, “RNA Interference: Biology, Mechanism, and Applications,” *Microbiol. Mol. Biol. Rev.*, 2003.

[9] F. J. M. Mojica, C. Díez-Villaseñor, J. García-Martínez, and C. Almendros, “Short motif sequences determine the targets of the prokaryotic CRISPR defence system,” *Microbiology*, 2009.

[10] G. Faure, K. S. Makarova, and E. V. Koonin, “CRISPR–Cas: Complex Functional Networks and Multiple Roles beyond Adaptive Immunity,” *Journal of Molecular Biology*. 2019.

[11] M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna, and E. Charpentier, “A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity,” *Science.*, 2012.

[12] D. B. T. Cox *et al.*, “RNA editing with CRISPR-Cas13,” *Science.*, 2017.

[13] M. H. Larson, L. A. Gilbert, X. Wang, W. A. Lim, J. S. Weissman, and L. S. Qi, “CRISPR interference (CRISPRi) for sequence-specific control of gene expression,” *Nat. Protoc.*, 2013.

[14] L. S. Tay, N. Palmer, R. Panwala, W. L. Chew, and P. Mali, “Translating CRISPR-Cas Therapeutics: Approaches and Challenges,” *CRISPR Journal*. 2020.

[15] O. S. Alkhnbashi, T. Meier, A. Mitrofanov, R. Backofen, and B. Voß, “CRISPR-Cas bioinformatics,” *Methods*. 2020.

[16] M. T. Price, H. Fullerton, and C. L. Moyer, “Biogeography and evolution of Thermococcus isolates from hydrothermal vent systems of the Pacific,” *Front. Microbiol.*, 2015.

[17] M. Á. Cabrera and J. M. Blamey, “Biotechnological applications of archaeal enzymes from extreme environments,” *Biological Research*. 2018.

[18] A. Hirata *et al.*, “Enzymatic activity of a subtilisin homolog, Tk-SP, from Thermococcus kodakarensis in detergents and its ability to degrade the abnormal prion protein,” *BMC Biotechnol.*, 2013.

[19] J. Il Lee, S. S. Cho, E. J. Kil, and S. T. Kwon, “Characterization and PCR application of a thermostable DNA polymerase from Thermococcus pacificus,” *Enzyme Microb. Technol.*, 2010.

[20] H. Ppyun, I. Kim, S. S. Cho, K. J. Seo, K. Yoon, and S. T. Kwon, “Improved PCR performance using mutant Tpa-S DNA polymerases from the hyperthermophilic archaeon Thermococcus pacificus,” *J. Biotechnol.*, 2012.

[21] J. R. Elmore *et al.*, “Programmable plasmid interference by the CRISPR-Cas system in Thermococcus kodakarensis,” *RNA Biol.*, 2013.

[22] T. Y. Jung *et al.*, “Structural features of Cas2 from Thermococcus onnurineus in CRISPR-cas system type IV,” *Protein Sci.*, 2016.

[23] I. Grissa, G. Vergnaud, and C. Pourcel, “CRISPRFinder: A web tool to identify clustered regularly interspaced short palindromic repeats,” *Nucleic Acids Res.*, 2007.

[24] E. Lepage *et al.*, “Molecular Diversity of New Thermococcales Isolates from a Single Area of Hydrothermal Deep-Sea Vents as Revealed by Randomly Amplified Polymorphic DNA Fingerprinting and 16S rRNA Gene Sequence Analysis,” *Appl. Environ. Microbiol.*, 2004.

[25] K. S. Makarova, Y. I. Wolf, and E. V. Koonin, “Classification and Nomenclature of CRISPR-Cas Systems: Where from Here?,” *Cris. J.*, 2018.

[26] J. D. Thompson, T. J. Gibson, and D. G. Higgins, “Multiple Sequence Alignment Using ClustalW and ClustalX,” *Curr. Protoc. Bioinforma.*, 2003.

[27] R. C. Edgar, “MUSCLE: A multiple sequence alignment method with reduced time and space complexity,” *BMC Bioinformatics*, 2004.

[28] S. Kumar, G. Stecher, M. Li, C. Knyaz, and K. Tamura, “MEGA X: Molecular evolutionary genetics analysis across computing platforms,” *Mol. Biol. Evol.*, 2018.

[29] J. O. Corliss, P. H. A. Sneath, and R. R. Sokal, “Numerical Taxonomy: The Principles and Practice of Numerical Classification,” *Trans. Am. Microsc. Soc.*, 1974.

[30] M. A. Nethery and R. Barrangou, “Predicting and visualizing features of CRISPR–Cas systems,” in *Methods in Enzymology*, 2019.

[31] G. E. Crooks, G. Hon, J. M. Chandonia, and S. E. Brenner, “WebLogo: A sequence logo generator,” *Genome Res.*, 2004.

[32] R. Lorenz *et al.*, “ViennaRNA Package 2.0,” *Algorithms Mol. Biol.*, 2011.

[33] A. Biswas, J. N. Gagnon, S. J. J. Brouns, P. C. Fineran, and C. M. Brown, “CRISPRTarget: Bioinformatic prediction and analysis of crRNA targets,” *RNA Biol.*, 2013.

[34] M. Johnson, I. Zaretskaya, Y. Raytselis, Y. Merezhuk, S. McGinnis, and T. L. Madden, “NCBI BLAST: a better web interface.,” *Nucleic Acids Res.*, 2008.

[35] A. Gorlas, E. V. Koonin, N. Bienvenu, D. Prieur, and C. Geslin, “TPV1, the first virus isolated from the hyperthermophilic genus Thermococcus,” *Environ. Microbiol.*, 2012.

[36] C. Geslin, M. Le Romancer, G. Erauso, M. Gaillard, G. Perrot, and D. Prieur, “PAV1, the first virus-like particle isolated from a hyperthermophilic euryarchaeote, ‘Pyrococcus abyssi,’” *J. Bacteriol.*, 2003.

[37] C. Hidalgo-Cantabrana, A. B. Crawley, B. Sanchez, and R. Barrangou, “Characterization and exploitation of CRISPR loci in Bifidobacterium longum,” *Front. Microbiol.*, 2017.

[38] A. E. Briner *et al.*, “Occurrence and diversity of CRISPR-Cas systems in the genus bifidobacterium,” *PLoS One*, 2015.

[39] J. Shen, L. Lv, X. Wang, Z. Xiu, and G. Chen, “Comparative analysis of CRISPR-Cas systems in Klebsiella genomes,” *J. Basic Microbiol.*, 2017.

[40] C. Norais, A. Moisan, C. Gaspin, and B. Clouet-d’Orval, “Diversity of CRISPR systems in the euryarchaeal Pyrococcales,” *RNA Biology*. 2013.

[41] N. D. Marino *et al.*, “Discovery of widespread type i and type v CRISPR-Cas inhibitors,” *Science.*, 2018.

[42] C. Yang, F. Lin, Q. Li, T. Li, and J. Zhao, “Comparative genomics reveals diversified CRISPR-Cas systems of globally distributed Microcystis aeruginosa, a freshwater bloom-forming cyanobacterium,” *Front. Microbiol.*, 2015.

[43] R. H. J. Staals *et al.*, “RNA Targeting by the Type III-A CRISPR-Cas Csm Complex of Thermus thermophilus,” *Mol. Cell*, 2014.

[44] L. He, X. Fan, and J. Xie, “Comparative genomic structures of Mycobacterium CRISPR-Cas,” *J. Cell. Biochem.*, 2012.

[45] N. Shariat, R. E. Timme, J. B. Pettengill, R. Barrangou, and E. G. Dudley, “Characterization and evolution of Salmonella CRISPR-Cas systems,” *Microbiol. (United Kingdom)*, 2015.

[46] J. Zhang, X. Li, Z. Deng, and H. Y. Ou, “Comparative Analysis of CRISPR Loci Found in Streptomyces Genome Sequences,” *Interdiscip. Sci. Comput. Life Sci.*, 2018.

[47] H. Di, L. Ye, H. Yan, H. Meng, S. Yamasak, and L. Shi, “Comparative analysis of CRISPR loci in different Listeria monocytogenes lineages,” *Biochem. Biophys. Res. Commun.*, 2014.

[48] M. Tajkarimi and H. M. Wexler, “CRISPR-Cas systems in bacteroides fragilis, an important pathobiont in the human gut microbiome,” *Front. Microbiol.*, 2017.

[49] D. K. Zhu *et al.*, “Comparative genomic analysis identifies structural features of CRISPR-Cas systems in Riemerella anatipestifer,” *BMC Genomics*, 2016.

[50] K. S. Makarova *et al.*, “Evolution and classification of the CRISPR-Cas systems,” *Nature Reviews Microbiology*. 2011.

[51] A. Wagner *et al.*, “Mechanisms of gene flow in archaea,” *Nature Reviews Microbiology*. 2017.

[52] R. Pinilla-Redondo *et al.*, “Type IV CRISPR-Cas systems are highly diverse and involved in competition between plasmids,” *Nucleic Acids Res.*, 2020.

[53] Y. Li and N. Peng, “Endogenous CRISPR-cas system-based genome editing and antimicrobials: Review and prospects,” *Frontiers in Microbiology*. 2019.

[54] A. Özcan *et al.*, “Type IV CRISPR RNA processing and effector complex formation in Aromatoleum aromaticum,” *Nat. Microbiol.*, 2019.

[55] S. N. Kieper, C. Almendros, and S. J. J. Brouns, “Conserved motifs in the CRISPR leader sequence control spacer acquisition levels in Type I-D CRISPR-Cas systems,” *FEMS Microbiol. Lett.*, 2019.

[56] C. Geslin, M. Le Romancer, M. Gaillard, G. Erauso, and D. Prieur, “Observation of virus-like particles in high temperature enrichment cultures from deep-sea hydrothermal vents,” *Res. Microbiol.*, 2003.

[57] M. Krupovic, M. Gonnet, W. Ben Hania, P. Forterre, and G. Erauso, “Insights into Dynamics of Mobile Genetic Elements in Hyperthermophilic Environments from Five New Thermococcus Plasmids,” *PLoS One*, 2013.

[58] H. Atomi, T. Imanaka, and T. Fukui, “Overview of the genetic tools in the

Archaea," *Front. Microbiol.*, 2012.

[59] T. H. Hileman and T. J. Santangelo, "Genetics techniques for Thermococcus kodakarensis," *Front. Microbiol.*, 2012.

[60] A. Stern, L. Keren, O. Wurtzel, G. Amitai, and R. Sorek, "Self-targeting by CRISPR: Gene regulation or autoimmunity?," *Trends in Genetics*. 2010.

[61] S. Arragain *et al.*, "Identification of eukaryotic and prokaryotic methylthiotransferase for biosynthesis of 2-methylthio-N6-threonylcarbamoyladenosine in tRNA," *J. Biol. Chem.*, 2010.

[62] H. Hori, "Transfer RNA methyltransferases with a SpoU-TrmD (SPOUT) fold and their modified nucleosides in tRNA," *Biomolecules*. 2017.

[63] H. Deveau *et al.*, "Phage response to CRISPR-encoded resistance in Streptococcus thermophilus," *J. Bacteriol.*, 2008.

[64] S. Shmakov *et al.*, "Discovery and Functional Characterization of Diverse Class 2 CRISPR-Cas Systems," *Mol. Cell*, 2015.

[65] O. O. Abudayyeh *et al.*, "C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector," *Science*, 2016.

8. Appendix

The Appendix Table 1 is based on all the annotated CRISPR/Cas systems in *Thermococcus* genomes.

The Appendix Figure 1 contains the results obtained via CRISPRTarget for the identification of spacers.

The Appendix Table 2 contains all the information retrieved from the CRISPRTarget results.

Appendix Table 1.

***Thermococcus* species**	**Strain**	**CRISPR locus**	**CRISPR Type**	**Loci Location**	**CRISPR DR sequences**	**DR length**	**Spacers**	***cas* genes**
T. barophilus	MP	1	I-B	398320:398694	GTTCCAATAAGACTCCAAGAGAATTGAAAG	30	5	Yes
		2	-	579476:580049	GTTCCAATAAGACTCCAAGAGAATTGAAAG	30	8	No
		3	-	1334169:1334873	GTTCCAATAAGACTCTAAGAGAATTGAAAG	30	10	No
		4	-	1530305:1532069	GTTCCAATAAGACTCTAAGAGAATTGAAAG	30	25	No
T. barossii	SHCK-94	1	I-A	407181:408549	GTTGCAATAAGACTCTAGGAGAATTGAAAG	30	20	Yes
		2	-	859421:859919	GTTGCAATAAGACTCCAGGAGAATTGAAA	29	7	No
T. celer	Vu-13	1	I-Btr	529498:530473	GTTTCCGTAGAACGGTATCGTGTGGAAAG	29	14	Yes
		2	-	1567558:1567995	GTTGCAATAAGACTCGAGGAGAATTGAAAG	30	6	No
T. cleftensis	1	1	-	48:682	GTTGCAATAAGACTCTGGGAGAATTGAAAC	30	9	No
		2	-	100515:103079	GTTGCAATAAGACTCTGGGAGAATTGAAAT	30	38	No
		3	I-B	347447:350029	GTTGCAATAAGACTCTAGGAGAATTGAAAC	30	38	Yes
		4	-	1709530:1710076	GTTTCAGAACCACATAATGTTTGGAAAC	28	7	No
		5	-	1722315:1723753	GTTTCCGTAGGACAGAATTGTGTGGAAAG	29	21	No
		6	I-B	1739835:1740928	GTTTCCGTAGGACAGAATTGTGTGGAAAG	29	16	Yes
		7	Ad.	1745414:1746369	GTTTCCGTAGAACAGTGTTGTGTGGAAAG	29	14	Yes
		8	III-A	1868115:1868602	GTTTCAGAACCACATGATGTTTGGAAAC	28	6	Yes
		9	I-Btr	1877271:1878429	GTTTCAGAACCACATAATGTTTGGAAAC	28	15	Yes
		10	-	1949366:1950270	GTTGCAATAAGACTCTGGGAGAATTGAAAC	30	13	No
T. eurythermalis	A501	1	-	446433:449955	GTTGCAATAAGACTCTAGGAGAATTGAAAG	30	52	No
		2	I-A/I-B	1220049:1224832	GTTGCAATAAGACTCTAGGAGAATTGAAAG	30	71	Yes
		3	-	1581454:1583630	GTTGCAATAAGACTCTAGGAGAATTGAAAG	30	32	No

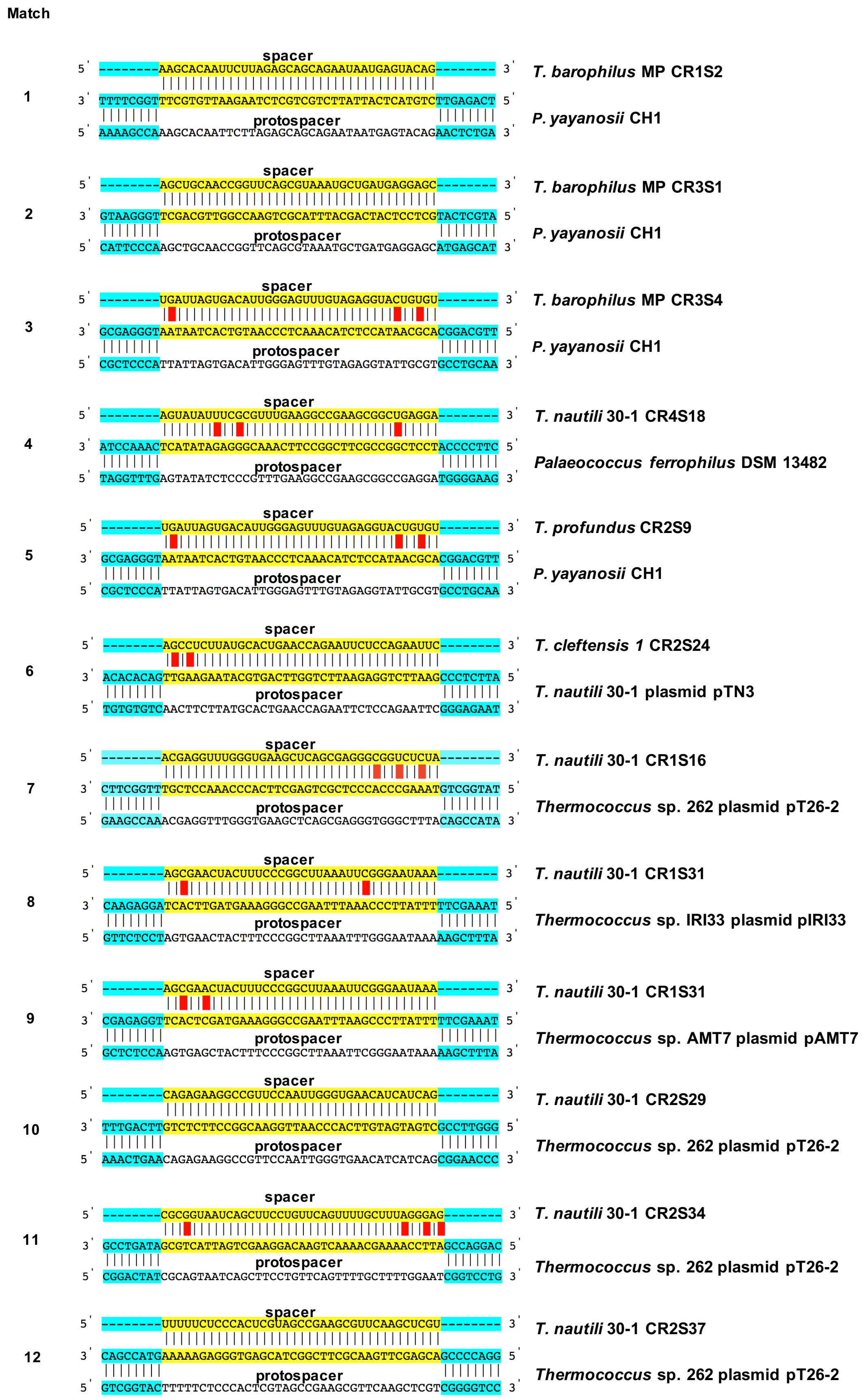
		4	-	1720540:1721978	GTTGCAATAAGACTCGAAGAGAATTGAAAG	30	21	No
T. gammatolerans	EJ3	1	-	208373:208937	GTTGCAATAAGACTCTAGGAGAATTGAAAG	30	8	No
		2	I-A	1221278:1222718	GTTGCAATAAGACTCTAGGAGAATTGAAAG	30	21	Yes
		3	-	1857022:1857814	GTTGCAATAAGACTCTAGGAGAATTGAAAG	30	11	No
T. gorganarius	W-12	1	-	157193:157558	GTTGCAATAAGACTTTAGGAGAATTGAAAC	30	5	No
T. guaymasensis	DSM11113	1	-	457101:457669	GTTGCAATAAGACTCTAGGAGAATTGAAAG	30	8	No
		2	-	604299:606074	GTTGCAATAAGACTCTAGGAGAATTGAAAG	30	26	No
		3	I-A/I-B	1798198:1800241	GTTGCAATAAGACTCGAAGAGAATTGAAAG	30	30	Yes
T. kodakarensis	KOD1	1	I-A/I-B	373033:374072	GTTGCAATAAGACTCTAGGAGAATTGAAAT	30	15	Yes
		2	-	468981:470566	GTTGCAATAAGACTCTAAGAGAATTGAAAC	30	23	No
		3	-	833495:835984	GTTGCAATAAGACTCTAAGAGAATTGAAAG	30	36	No
T. litoralis	DSM5473	1	I-A/I-B	29:2779	GTTCCAATAAGACTTTAGAAGAATTGAAAG	30	40	Yes
		2	-	208872:212274	GTTCCAATAAGACTTTAAAAGAATTGAAAG	30	50	No
		3	-	406450:408814	GTTCCAATAAGACTTTAAAAGAATTGAAAG	30	34	No
		4	-	860433:861075	GTTCCAATAAGACTTTAAAAGAATTGAAAG	30	9	No
		5	-	861993:865173	GTTCCAATAAGACTTTAAAAGAATTGAAAG	30	46	No
		6	-	1226231:1230894	GTTCCAATAAGACTTTAAAAGAATTGAAAG	30	68	No
		7	-	2214800:2215162	GTTCCAATAAGACTTTAAAAGAATTGAAAG	30	5	No
T. nautili	30-1	1	I-B	182361:184798	GTTGCAATAAGACTCTAGGAGAATTGAAAG	30	36	Yes
		2	-	451542:455179	GTTGCAATAAGACTCGAAGAGAATTGAAAG	30	54	No
		3	-	678193:678832	GTTGCAATAAGACTCGAGGAGAATTGAAAG	30	9	No
		4	-	705938:708770	GTTGCAATAAGACTCGAGGAGAATTGAAAG	30	42	No
		5	-	1461759:1462999	GTTGCAATAAGACTCGAGGAGAATTGAAAC	30	18	No
T. onnurineus	NA1	1	IV	294116:295760	GTTGCAATAAGACTCTAGGAGAATTGAAAC	30	24	Yes
		2	-	728608:730178	GTTTCAATAAGACTCTAAGAGAATTGAAAG	30	23	No
		3	III-A	818816:819904	GTTCCAGTAGGACAGAATTGTGTGGAAAG	29	16	Yes
		4	Ad.	828033:828820	GTTTCAGTAGGACAGAATTGTGTGGAAA	28	11	Yes
		5	-	994457:994969	GTTGCAATAAGACTCTAAGAGAATTGAAAC	30	7	No

		6	-	995057:997976	GTTGCAATAAGACTCTAAGAGAATTGAAAC	30	43	No
T. pacificus	P-4	1	-	1564300:1564593	GTTGCAATAAGACTCTAAGAGAATTGAAA	29	4	No
T. paralvinellae	ES1	1	-	553728:554840	GTTTCAATAAGACTTTAGAAGAATTGAAAT	30	16	No
		2	-	855389:857656	GTTTCAATAAGACTCTAAGAGAATTGAAAG	30	33	No
		3	I-Btr	954068:956265	GTTTCAATAAGACTTTAGAAGAATTGAAAT	30	32	Yes
T. peptonophilus	OG-1	1	I-B	43143:44313	GTTTCCGTAGAACGTAATCGTGTGGAAAG	29	17	Yes
		2	Ad.	48598:50778	GTTTCCGTAGAACGTAGTCGTGTGGAAAT	29	32	Yes
		3	-	445657:446033	GTTGCAATAAGACTCTAGGAGAATTGAAA	29	5	No
		4	-	1654477:1656204	GTTTCCGTAGAACGTAGTCGTGTGGAAAG	29	25	No
T. piezophilus	CDGS	1	-	147812:148984	GTTGCAATAAGACTCTAGGAGAATTGAAAG	30	17	No
		2	I-B	1383163:1384693	GTTGCAATAAGACTCTAGGAGAATTGAAAC	30	22	Yes
		3	-	1845174:1845926	GTTGCAATAAGACTCTAGGAGAATTGAAAA	30	11	No
		4	III-A	1926393:187	GTTTCAGTAGGACAGAATTGTGTGGAAAC	29	38	Yes
T. profundus	DT5432	1	-	149437:151081	GTTGCAATAAGACTCTAGGAGAATTGAAAG	30	24	No
		2	-	522704:523627	GTTGCAATAAGACTCTAGGAGAATTGAAAC	30	13	No
		3	I-A/I-B	1104751:1106801	GTTGCAATAAGACTCTAGGAGAATTGAAAG	30	30	Yes
T. radiotolerans	EJ2	1	Ad.	270715:271036	TGTTTCAGTAGAACATAGTTGTGT	24	4	Yes
		2	I-Btr	272808:273627	GTTTCCGTAGAACAGTATTGTGTGGAAAC	29	11	Yes
		3	III-A	282186:282967	GTTTCCGTAGAACAGTATTGTGTGGAAAC	29	10	Yes
		4	-	465077:465975	GTTGCAATAAGACTCTAGGAGAATTGAAAG	30	13	No
T. sibiricus	MM739	1	I-B	1328782:1330412	GTTCCAATAAGACTTTAAAAGAATTGAAAG	30	24	Yes
T. siculi	RG-20	1	I-A/I-B	1129384:1132361	GTTGCAATAAGACTCTAGGAGAATTGAAAC	30	44	Yes
		2	-	1500605:1502652	GTTGCAATAAGACTCTAGGAGAATTGAAAG	30	30	No
		3	I-Btr	1573978:1574997	GTTTCAGAACCAGCATAAGCTTTGGAAAC	29	13	Yes
		4	III-B	1584773:1585800	GTTTCAGAACCAGCTTAAGCTTTGGAAAC	29	13	Yes
		5	-	1651711:1654492	GTTGCAATAAGACTCTAGGAGAATTGAAAG	30	41	No
		6	Ad.	1726662:1728763	GTTTCCGTAGGACATAGTTGTGTGGAAAG	29	31	Yes
		7	I-B	1733089:1736042	GTTTCCGTAGAACATGGTTGTGTGGAAAG	29	44	Yes

T. thioreducens	OGL-20P	1	I-A/I-Btr	684432:685277	GTTCCAATAAGACTCTAGGAGAATTGAAAG	30	12	Yes
		2	-	1555442:1557222	GTTGCAATAAGACTCTAGGAGAATTGAAAG	30	26	No

Appendix Table 1. CRISPR/Cas systems in *Thermococcus* genomes. Ad is short for adjacent, referring to CRISPR loci that are adjacent to other CRISPR/Cas systems. Tr (I-Btr) is short for truncated, referring to Type-I systems that are not complete, so they lack several *cas* genes typical from these systems.

Appendix Figure 1.


Match
1
spacer
5' --------AAGCACAAUUCUUAGAGCAGCAGAAUAAUGAGUACAG-------- 3'
3' TTTTCGGTTTCGTGTTAAGAATCTCGTCGTCTTATTACTCATGTCTTGAGACT 5'
protospacer
5' AAAAGCCAAAGCACAATTCTTAGAGCAGCAGAATAATGAGTACAGAACTCTGA 3'
T. barophilus MP CR1S2
P. yayanosii CH1
2
spacer
5' --------AGCUGCAACCGGUUCAGCGUAAAUGCUGAUGAGGAGC-------- 3'
3' GTAAGGGTTCGACGTTGGCCAAGTCGCATTTACGACTACTCCTCGTACTCGTA 5'
protospacer
5' CATTCCCAAGCTGCAACCGGTTCAGCGTAAATGCTGATGAGGAGCATGAGCAT 3'
T. barophilus MP CR3S1
P. yayanosii CH1
3
spacer
5' --------UGAUUAGUGACAUUGGGAGUUUGUAGAGGUACUGUGU-------- 3'
3' GCGAGGGTAATAATCACTGTAACCCTCAAACATCTCCATAACGCACGGACGTT 5'
protospacer
5' CGCTCCCATTATTAGTGACATTGGGAGTTTGTAGAGGTATTGCGTGCCTGCAA 3'
T. barophilus MP CR3S4
P. yayanosii CH1
4
spacer
5' --------AGUAUAUUUCGCGUUUGAAGGCCGAAGCGGCUGAGGA-------- 3'
3' ATCCAAACTCATATAGAGGGCAAACTTCCGGCTTCGCCGGCTCCTACCCCTTC 5'
protospacer
5' TAGGTTTGAGTATATCTCCCGTTTGAAGGCCGAAGCGGCCGAGGATGGGGAAG 3'
T. nautili 30-1 CR4S18
Palaeococcus ferrophilus DSM 13482
5
spacer
5' --------UGAUUAGUGACAUUGGGAGUUUGUAGAGGUACUGUGU-------- 3'
3' GCGAGGGTAATAATCACTGTAACCCTCAAACATCTCCATAACGCACGGACGTT 5'
protospacer
5' CGCTCCCATTATTAGTGACATTGGGAGTTTGTAGAGGTATTGCGTGCCTGCAA 3'
T. profundus CR2S9
P. yayanosii CH1
6
spacer
5' --------AGCCUCUUAUGCACUGAACCAGAAUUCUCCAGAAUUC-------- 3'
3' ACACACAGTTGAAGAATACGTGACTTGGTCTTAAGAGGTCTTAAGCCCTCTTA 5'
protospacer
5' TGTGTGTCAACTTCTTATGCACTGAACCAGAATTCTCCAGAATTCGGGAGAAT 3'
T. cleftensis 1 CR2S24
T. nautili 30-1 plasmid pTN3
7
spacer
5' --------ACGAGGUUUGGGUGAAGCUCAGCGAGGGCGGUCUCUA-------- 3'
3' CTTCGGTTTGCTCCAAACCCACTTCGAGTCGCTCCCACCCGAAATGTCGGTAT 5'
protospacer
5' GAAGCCAAACGAGGTTTGGGTGAAGCTCAGCGAGGGTGGGCTTTACAGCCATA 3'
T. nautili 30-1 CR1S16
Thermococcus sp. 262 plasmid pT26-2
8
spacer
5' --------AGCGAACUACUUUCCCGGCUUAAAUUCGGGAAUAAA-------- 3'
3' CAAGAGGATCACTTGATGAAAGGGCCGAATTTAAACCCTTATTTTTCGAAAT 5'
protospacer
5' GTTCTCCTAGTGAACTACTTTCCCGGCTTAAATTTGGGAATAAAAAGCTTTA 3'
T. nautili 30-1 CR1S31
Thermococcus sp. IRI33 plasmid pIRI33
9
spacer
5' --------AGCGAACUACUUUCCCGGCUUAAAUUCGGGAAUAAA-------- 3'
3' CGAGAGGTTCACTCGATGAAAGGGCCGAATTTAAGCCCTTATTTTTCGAAAT 5'
protospacer
5' GCTCTCCAAGTGAGCTACTTTCCCGGCTTAAATTCGGGAATAAAAAGCTTTA 3'
T. nautili 30-1 CR1S31
Thermococcus sp. AMT7 plasmid pAMT7
10
spacer
5' --------CAGAGAAGGCCGUUCCAAUUGGGUGAACAUCAUCAG-------- 3'
3' TTTGACTTGTCTCTTCCGGCAAGGTTAACCCACTTGTAGTAGTCGCCTTGGG 5'
protospacer
5' AAACTGAACAGAGAAGGCCGTTCCAATTGGGTGAACATCATCAGCGGAACCC 3'
T. nautili 30-1 CR2S29
Thermococcus sp. 262 plasmid pT26-2
11
spacer
5' --------CGCGGUAAUCAGCUUCCUGUUCAGUUUUGCUUUAGGGAG-------- 3'
3' GCCTGATAGCGTCATTAGTCGAAGGACAAGTCAAAACGAAAACCTTAGCCAGGAC 5'
protospacer
5' CGGACTATCGCAGTAATCAGCTTCCTGTTCAGTTTTGCTTTTGGAATCGGTCCTG 3'
T. nautili 30-1 CR2S34
Thermococcus sp. 262 plasmid pT26-2
12
spacer
5' --------UUUUUCUCCCACUCGUAGCCGAAGCGUUCAAGCUCGU-------- 3'
3' CAGCCATGAAAAAGAGGGTGAGCATCGGCTTCGCAAGTTCGAGCAGCCCCAGG 5'
protospacer
5' GTCGGTACTTTTTCTCCCACTCGTAGCCGAAGCGTTCAAGCTCGTCGGGGTCC 3'
T. nautili 30-1 CR2S37
Thermococcus sp. 262 plasmid pT26-2

13

```
                         spacer
5' --------CUCAUUGAUGAGGUUCAGAACCUCCUCCAGCACCCCGA-------- 3'
           ■||||||||||||||||||||||||||||||||||■||
3' AGAAGAAGCAGTAACTACTCCAAGTCTTGGAGGAGGTCGTGGGACTACCGGCCT 5'
   ||||||||             protospacer              ||||||||
5' TCTTCTTCGTCATTGATGAGGTTCAGAACCTCCTCCAGCACCCTGATGGCCGGA 3'
```

T. nautili **30-1 CR2S38**

Thermococcus **sp. 262 plasmid pT26-2**

14

```
                         spacer
5' --------ACUCCUGAAGUAGCGGUCAACUUUCAUCAUAUCAUUGCU-------- 3'
           |||||||||||||||||||||||||||||||||||||■■
3' GTCGGTTATGAGGACTTCATCGCCAGTTGAAAGTAGTATAGTAACACCGAGGTGA 5'
   ||||||||              protospacer              ||||||||
5' CAGCCAATACTCCTGAAGTAGCGGTCAACTTTCATCATATCATTGTGGCTCCACT 3'
```

T. nautili **30-1 CR3S9**

T. nautili **30-1 plasmid pTN3**

15

```
                         spacer
5' --------UCAAGCGCUGGAGGGUGUGAUAUGGCUGAGAACGCGU-------- 3'
           ||||||||||||||||||||||||||||||■||||||
3' GTTGGGGAAGTTCGCGACCTCCCACACTATACCGACTCCTGCGCATAGTTTGG 5'
   ||||||||             protospacer             ||||||||
5' CAACCCCTTCAAGCGCTGGAGGGTGTGATATGGCTGAGGACGCGTATCAAACC 3'
```

T. nautili **30-1 CR4S33**

Thermococcus **sp. 262 plasmid pT26-2**

16

```
                         spacer
5' --------GCGGUGGCUAGGCUCAUGCACUUGGGUCUAGUGAAAA-------- 3'
           |||||||||||||||||||||||||||||||||||||
3' ATTGATGCCGCCACCGATCCGAGTACGTGAACCCAGATCACTTTTACCCTCTA 5'
   ||||||||             protospacer             ||||||||
5' TAACTACGGCGGTGGCTAGGCTCATGCACTTGGGTCTAGTGAAAATGGGAGAT 3'
```

T. thioreducens **OGL-20P CR1S8**

Methanocaldococcus vulcanius
M7 plasmid pMETVU01

17

```
                         spacer
5' --------AAGCACAAUUCUUAGAGCAGCAGAAUAAUGAGUACAG-------- 3'
           |||||||||||||||||||||||||■|||||||||||
3' TTTTCGGTTTCGTGTTAAGAATCTCGTCGTCTTGTTACTCATGTCCTGAGACT 5'
   ||||||||             protospacer             ||||||||
5' AAAAGCCAAAGCACAATTCTTAGAGCAGCAGAACAATGAGTACAGGACTCTGA 3'
```

T. barophilus **MP CR1S2**

T. barophilus **CH5**

18

```
                         spacer
5' --------UUGCAAGAAUAUCUCCAAGGUUCGCAACUCCGCUCAG-------- 3'
           |||■■||||||||||||||||||||||||||||||||
3' GCAACCTAAACAGTCTTATAGAGGTTCCAAGCGTTGAGGCGAGTCACCGTTAC 5'
   ||||||||             protospacer             ||||||||
5' CGTTGGATTTGTCAGAATATCTCCAAGGTTCGCAACTCCGCTCAGTGGCAATG 3'
```

T. barophilus **MP CR2S4**

T. barophilus **CH5**

19

```
                         spacer
5' --------UGGAAUCGGCUGCUGCUCAAGGAACUCAAUCAAAGAUUCA-------- 3'
           ■||■||■|||||||||||||||||||||||||||||||||
3' TAGGAATGGCCCTATCCGACGACGAGTTCCTTGAGTTAGTTTCTAAGTTCTCCTTC 5'
   ||||||||              protospacer               ||||||||
5' ATCCTTACCGGGATAGGCTGCTGCTCAAGGAACTCAATCAAAGATTCAAGAGGAAG 3'
```

T. barophilus **MP CR2S8**

T. barophilus **CH5**

20

```
                         spacer
5' --------GCUUUUUCAGCUGUUCAUUCUCUUGCUUCAGCUUUGC-------- 3'
           |||||||||||||||||||||||||||||||||||||
3' AAGTTGTACGAAAAAGTCGACAAGTAAGAGAACGAAGTCGAAACGTAAGAGAA 5'
   ||||||||             protospacer             ||||||||
5' TTCAACATGCTTTTTCAGCTGTTCATTCTCTTGCTTCAGCTTTGCATTCTCTT 3'
```

T. barophilus **MP CR3S4**

T. barophilus **CH5**

21

```
                         spacer
5' --------UGCUCGAAGUAUUCAAUACAACAUUUGUCAAUAUAACAG-------- 3'
           |||||||||||||||||||||||||||||||||||||||
3' TCAATGGAACGAGCTTCATAAGTTATGTTGTAAACAGTTATATTGTCACAACTCC 5'
   ||||||||              protospacer              ||||||||
5' AGTTACCTTGCTCGAAGTATTCAATACAACATTTGTCAATATAACAGTGTTGAGG 3'
```

T. barophilus **MP CR4S21**

T. barophilus **CH5**

22

```
                         spacer
5' --------UCCAGGUCGAGAAGCUCGAAGACGGCAAGAUAGUCG-------- 3'
           |■||||||||||||||||||||||||||||||||||
3' TCTTTGTAACGTCCAGCTCTTCGAGCTTCTGCCGTTCTATCAGCAGTTCCAC 5'
   ||||||||            protospacer             ||||||||
5' AGAAACATTGCAGGTCGAGAAGCTCGAAGACGGCAAGATAGTCGTCAAGGTG 3'
```

T. barophilus **MP CR3S9**

T. thioreducens **OGL-20P**

23

```
                         spacer
5' --------AUGCCAACGUCAGCAUGGAGGACGUUCUCAAGGCGA-------- 3'
           ||||||||||||||||||||||||||||||||||||
3' TCGCCGGCTACGGTTGCAGTCGTACCTCCTGCAAGAGTTCCGCTATTTACTC 5'
   ||||||||            protospacer             ||||||||
5' AGCGGCCGATGCCAACGTCAGCATGGAGGACGTTCTCAAGGCGATAAATGAG 3'
```

T. cleftensis **1 CR3S22**

Thermococcus **sp. JdF3**

24

```
                         spacer
5' --------AUGCCAACGUCAGCAUGGAGGACGUUCUCAAGGCGA-------- 3'
           ||||||||||||||||||||||||||||||||||||
3' TCGCCGGCTACGGTTGCAGTCGTACCTCCTGCAAGAGTTCCGCTATTTACTC 5'
   ||||||||            protospacer             ||||||||
5' AGCGGCCGATGCCAACGTCAGCATGGAGGACGTTCTCAAGGCGATAAATGAG 3'
```

T. cleftensis **1 CR3S22**

Thermococcus **sp. MV11**

25

```
                          spacer
5' --------AGUCGCGAACAUCGAGAUGGAAGACGAGAAAAUCAAG-------- 3'
           |||||||||||||||||||||||||||||||||||||
3' TAACTTTCTCAGCGCTTGTAGCTCTACCTTCTGCTCTTTTAGTTCCAACGTTA 5'
   ||||||||             protospacer             ||||||||
5' ATTGAAAGAGTCGCGAACATCGAGATGGAAGACGAGAAAATCAAGGTTGCAAT 3'
```

T. eurythermalis **A501 CR1S52**

T. guaymasensis **DSM 11113**

26

```
                          spacer
5' --------AGGGGGUGGUAAACGUGCGAAGAAGAGGUAGGGGCUU-------- 3'
           |||||||||||||||||||||||||||||||||||||
3' TTCCCGGCTCCCCCACCATTTGCACGCTTCTTCTCCATCCCCGAAAAAACGCC 5'
   ||||||||             protospacer             ||||||||
5' AAGGGCCGAGGGGGTGGTAAACGTGCGAAGAAGAGGTAGGGGCTTTTTTGCGG 3'
```

T. eurythermalis **A501 CR2S21**

Thermococcus **sp. EXT12c**

27

```
                          spacer
5' --------GAAUGGCGAUGAAGAGCAGGCGAAGAUUGACGCUCAAA-------- 3'
           ||||||||||||||||||||||||||||||||||||||
3' TTCCGAAACTTACCGCTACTTCTCGTCCGCTTCTAACTGCGAGTTTTCCCTGTC 5'
   ||||||||             protospacer              ||||||||
5' AAGGCTTTGAATGGCGATGAAGAGCAGGCGAAGATTGACGCTCAAAAGGGACAG 3'
```

T. kodakarensis **KOD1 CR1S6**

T. litoralis **DSM 5473**

28

```
                          spacer
5' --------AAAGUAUGGCCCUGUGAAGAGGAUUGGCUUUGAGCUA-------- 3'
           |||||||||||||||||||||||||||||||||||||
3' GCTGGGGCTTTCATACCGGGACACTTCTCCTAACCGAAACTCGATGGGATGCT 5'
   ||||||||             protospacer             ||||||||
5' CGACCCCGAAAGTATGGCCCTGTGAAGAGGATTGGCTTTGAGCTACCCTACGA 3'
```

T. nautili **30-1 CR1S15**

T. guaymasensis **DSM 11113**

29

```
                          spacer
5' --------CCAGCUGUGGAGCUUCGCCCUCCAACCGCCUUUCUUG-------- 3'
           ||||||||||||||||||||||||||||||■■|■|||
3' TAACTTTAGGTCGACACCTCGAAGCGGGAGGTTGGCGGGGAAAACTCTCAGAG 5'
   ||||||||             protospacer             ||||||||
5' ATTGAAATCCAGCTGTGGAGCTTCGCCCTCCAACCGCCCCTTTGAGAGTCTC 3'
```

T. nautili **30-1 CR2S9**

T. guaymasensis **DSM 11113**

30

```
                          spacer
5' --------AGUAUAUUUCGCGUUUGAAGGCCGAAGCGGCUGAGGA-------- 3'
           |■|||||||||||||||||||||||||||||■|||||
3' ATCTAAACTTATATAAAGCGCAAACTTCCGGCTTCGCCGGCTCCTACCGCTTC 5'
   ||||||||             protospacer             ||||||||
5' TAGATTTGAATATATTTCGCGTTTGAAGGCCGAAGCGGCCGAGGATGGCGAAG 3'
```

T. nautili **30-1 CR4S18**

T. thioreducens **OGL-20P**

31

```
                          spacer
5' --------UUUUCUUUCCUUUGGAGAGCUCACGGAUUGGGUACC-------- 3'
           ||||||||||||||||||||||||||||||||||||
3' TTTTGGGGAAAAGAAAGGAAACCTCTCGAGTGCCTAACCCATGGACCGATAC 5'
   ||||||||            protospacer             ||||||||
5' AAAACCCCTTTTCTTTCCTTTGGAGAGCTCACGGATTGGGTACCTGGCTATG 3'
```

T. nautili **30-1 CR4S23**

Thermococcus **sp. SY113**

32

```
                          spacer
5' --------CGGUCUCUGGCGUCAGCGUGCCGUGCUCCUGGAGAA-------- 3'
           |||||||||||||||||||||||||||||||■||||
3' GTAGAACTGCCAGAGACCGCAGTCGCACGGCACGAGGACATCTTAGTCGCGC 5'
   ||||||||            protospacer             ||||||||
5' CATCTTGACGGTCTCTGGCGTCAGCGTGCCGTGCTCCTGTAGAATCAGCGCG 3'
```

T. peptonophilus **OG-1 CR4S13**

T. thioreducens **OGL-20P**

33

```
                          spacer
5' --------CAGUACCGCUGGCAUUCGUAGUGCUCGUUGAACUCUUCGA-------- 3'
           |■||■|||||||||||||||||||||||||||||||||||
3' TTACCGTAGCCACGGCGACCGTAAGCATCACGAGCAACTTGAGAAGCTATTCGAGC 5'
   ||||||||              protospacer               ||||||||
5' AATGGCATCGGTGCCGCTGGCATTCGTAGTGCTCGTTGAACTCTTCGATAAGCTCG 3'
```

T. peptonophilus **OG-1 CR4S18**

T. kodakarensis **KOD1**

34

```
                          spacer
5' --------GGACGAGUUUGCGCUUAAUAAGUCGAAAUUCAUUAAU-------- 3'
           |||■|||■|||||||||||||||||||||||||||||
3' CGTCAATACCTCCTCGAACGCGAATTATTCAGCTTTAAGTAATTATTTGAGGA 5'
   ||||||||             protospacer             ||||||||
5' GCAGTTATGGAGGAGCTTGCGCTTAATAAGTCGAAATTCATTAATAAACTCCT 3'
```

T. piezophilus **CDGS CR3S2**

T. barophilus **CH5**

35

```
                          spacer
5' --------UGAUUAGUGACAUUGGGAGUUUGUAGAGGUACUGUGU-------- 3'
           |■|||||||||||||||||||||||||||||■|||||
3' ACGAGGGTAATAATCACTGTAACCCTCAAACATCTCCATAACACACGGACGTT 5'
   ||||||||             protospacer             ||||||||
5' TGCTCCCATTATTAGTGACATTGGGAGTTTGTAGAGGTATTGTGTGCCTGCAA 3'
```

T. profundus **DT5432 CR2S9**

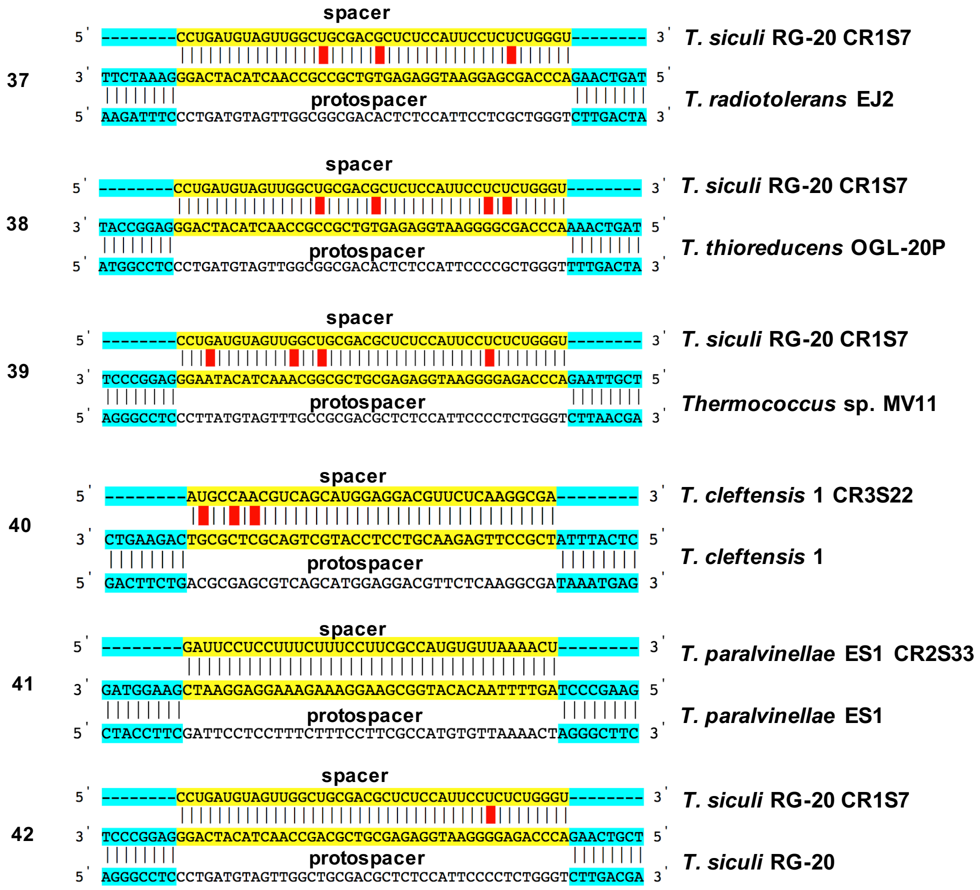
T. barophilus **CH5**

36

```
                          spacer
5' --------CCUGAUGUAGUUGGCUGCGACGCUCUCCAUUCCUCUCUGGGU-------- 3'
           |■|||||||||||||■|||||■||||||||||||||||||||
3' TCCCTGAGGAACTACATCAACCGCCGCTGTGAGAGGTAAGGAGAGACCCAGAATTGAT 5'
   ||||||||               protospacer                ||||||||
5' AGGGACTCCTTGATGTAGTTGGCGGCGACACTCTCCATTCCTCTCTGGGTCTTAACTA 3'
```

T. siculi **RG-20 CR1S7**

Thermococcus **sp. 5-4**


37
spacer
5' --------CCUGAUGUAGUUGGCUGCGACGCUCUCCAUUCCUCUCUGGGU-------- 3'
T. siculi RG-20 CR1S7
3' TTCTAAAGGGACTACATCAACCGCCGCTGTGAGAGGTAAGGAGCGACCCAGAACTGAT 5'
protospacer
T. radiotolerans EJ2
5' AAGATTTCCCTGATGTAGTTGGCGGCGACACTCTCCATTCCTCGCTGGGTCTTGACTA 3'
38
spacer
5' --------CCUGAUGUAGUUGGCUGCGACGCUCUCCAUUCCUCUCUGGGU-------- 3'
T. siculi RG-20 CR1S7
3' TACCGGAGGGACTACATCAACCGCCGCTGTGAGAGGTAAGGGGCGACCCAAAACTGAT 5'
protospacer
T. thioreducens OGL-20P
5' ATGGCCTCCCTGATGTAGTTGGCGGCGACACTCTCCATTCCCCGCTGGGTTTTGACTA 3'
39
spacer
5' --------CCUGAUGUAGUUGGCUGCGACGCUCUCCAUUCCUCUCUGGGU-------- 3'
T. siculi RG-20 CR1S7
3' TCCCGGAGGGAATACATCAAACGGCGCTGCGAGAGGTAAGGGGAGACCCAGAATTGCT 5'
protospacer
Thermococcus sp. MV11
5' AGGGCCTCCCTTATGTAGTTTGCCGCGACGCTCTCCATTCCCCTCTGGGTCTTAACGA 3'
40
spacer
5' --------AUGCCAACGUCAGCAUGGAGGACGUUCUCAAGGCGA-------- 3'
T. cleftensis 1 CR3S22
3' CTGAAGACTGCGCTCGCAGTCGTACCTCCTGCAAGAGTTCCGCTATTTACTC 5'
protospacer
T. cleftensis 1
5' GACTTCTGACGCGAGCGTCAGCATGGAGGACGTTCTCAAGGCGATAAATGAG 3'
41
spacer
5' --------GAUUCCUCCUUUCUUUCCUUCGCCAUGUGUUAAAACU-------- 3'
T. paralvinellae ES1 CR2S33
3' GATGGAAGCTAAGGAGGAAAGAAAGGAAGCGGTACACAATTTTGATCCCGAAG 5'
protospacer
T. paralvinellae ES1
5' CTACCTTCGATTCCTCCTTTCTTTCCTTCGCCATGTGTTAAAACTAGGGCTTC 3'
42
spacer
5' --------CCUGAUGUAGUUGGCUGCGACGCUCUCCAUUCCUCUCUGGGU-------- 3'
T. siculi RG-20 CR1S7
3' TCCCGGAGGGACTACATCAACCGACGCTGCGAGAGGTAAGGGGAGACCCAGAACTGCT 5'
protospacer
T. siculi RG-20
5' AGGGCCTCCCTGATGTAGTTGGCTGCGACGCTCTCCATTCCCCTCTGGGTCTTGACGA 3'

Appendix Figure 1. CRISPRTarget outputs. All 42 spacers matching identified protospacers are shown. The red boxes correspond to the non-complementary base pairing.

Appendix Table 2.

Match	Targeting	Spacer from	Protospacer from	Strand	Match identity	Target protein	GenBank record
1	Different genus	*T. barophilus* MP CR1S2	*P. yayanosii* CH1	-	100%	Hypothetical protein	WP_013906152.1
2	Different genus	*T. barophilus* MP CR3S1	*P. yayanosii* CH1	+	100%	Hypothetical protein	WP_013906143.1
3	Different genus	*T. barophilus* MP CR3S4	*P. yayanosii* CH1	+	94.6%	Hypothetical protein	WP_013906152.1
4	Different genus	*T. nautili* 30-1 CR4S18	*Palaeococcus ferrophilus* DSM 13482	-	91.9%	HTH domain-containing protein	WP_048148968.1
5	Different genus	*T. profundus* DT5432 CR2S9	*P. yayanosii* CH1	+	91.9%	DUF87 domain-containing protein	WP_013906153.1
6	Plasmid	*T. cleftensis* 1 CR2S24	*T. nautili* 30-1 plasmid pTN3	-	94.6%	Hypothetical protein	YP_008619350.1
7	Plasmid	*T. nautili* 30-1 CR1S16	*Thermococcus* sp. 262 plasmid pT26-2	-	91.9%	t26-26p	YP_003603619.1
8	Plasmid	*T. nautili* 30-1 CR1S31	*Thermococcus* sp. IRI33 plasmid pIRI33	+	94.4%	Hypothetical protein	WP_192964463.1
9	Plasmid	*T. nautili* 30-1 CR1S31	*Thermococcus* sp. AMT7 plasmid pAMT7	+	94.4%	Hypothetical protein	WP_192964925.1
10	Plasmid	*T. nautili* 30-1 CR2S29	*Thermococcus* sp. 262 plasmid pT26-2	-	100%	t26-14p	YP_003603607.1
11	Plasmid	*T. nautili* 30-1 CR2S34	*Thermococcus* sp. 262 plasmid pT26-2	+	90%	t26-16p	YP_003603609.1
12	Plasmid	*T. nautili* 30-1 CR2S37	*Thermococcus* sp. 262 plasmid pT26-2	+	100%	t26-14p	YP_003603607.1
13	Plasmid	*T. nautili* 30-1 CR2S38	*Thermococcus* sp. 262 plasmid pT26-2	-	94.7%	t26-14p	YP_003603607.1
14	Plasmid and self-targeting	*T. nautili* 30-1 CR3S9	*T. nautili* 30-1 plasmid pTN3	-	94.9%	SAM pointed domain-containing ETS-like transcription factor	YP_008619366.1

15	Plasmid	*T. nautili* 30-1 CR4S33	*Thermococcus* sp. 262 plasmid pT26-2	-	97.3%	t26-15p	YP_003603608.1
16	Plasmid	*T. thioreducens* OGL-20P CR1S8	*Methanocaldococcus vulcanius* M7 plasmid pMETVU01	-	100%	Hypothetical protein	WP_012819840.1
17	Different *Th*. strain	*T. barophilus* MP CR1S2	*T. barophilus* CH5	-	97.3%	Hypothetical protein	WP_056934646.1
18	Different *Th*. strain	*T. barophilus* MP CR2S4	*T. barophilus* CH5	+	94.6%	Hypothetical protein	WP_056934637.1
19	Different *Th*. strain	*T. barophilus* MP CR2S8	*T. barophilus* CH5	+	92.5%	Hypothetical protein	WP_056934660.1
20	Different *Th*. strain	*T. barophilus* MP CR3S4	*T. barophilus* CH5	+	100%	Hypothetical protein	WP_056934646.1
21	Different *Th*. strain	*T. barophilus* MP CR4S21	*T. barophilus* CH5	-	100%	Hypothetical protein	WP_056934637.1
22	Different *Th*. species	*T. barophilus* MP CR3S9	*T. thioreducens* OGL-20P	-	97.2%	Hypothetical protein	WP_055428441.1
23	Different *Th*. species	*T. cleftensis* 1 CR3S22	*Thermococcus* sp. JdF3	+	100%	Hypothetical protein	WP_167903915.1
24	Different *Th*. species	*T. cleftensis* 1 CR3S22	*Thermococcus* sp. MV11	+	100%	Hypothetical protein	WP_167774253.1
25	Different *Th*. species	*T. eurythermalis* A501 CR1S51	*T. guaymasensis* DSM 11113	-	100%	Hypothetical protein	WP_062372725.1
26	Different *Th*. species	*T. eurythermalis* A501 CR2S21	*Thermococcus* sp. EXT12c	-	100%	Hypothetical protein	WP_099209180.1
27	Different *Th*. species	*T. kodakarensis* KOD1 CR1S6	*T. litoralis* DSM 5473	-	100%	Hypothetical protein	WP_004069311.1
28	Different *Th*. species	*T. nautili* 30-1 CR1S15	*T. guaymasensis* DSM 11113	-	100%	Hypothetical protein	WP_062370217.1
29	Different *Th*. species	*T. nautili* 30-1 CR2S9	*T. guaymasensis* DSM 11113	+	91.9%	Hypothetical protein	WP_062370182.1

30	Different *Th*. species	*T. nautili* 30-1 CR4S18	*T. thioreducens* OGL-20P	-	94.6%	Hypothetical protein	WP_143597836.1
31	Different *Th*. species	*T. nautili* 30-1 CR4S23	*Thermococcus* sp. SY113	+	100%	Integrase	WP_148882593.1
32	Different *Th*. species	*T. peptonophilus* OG-1 CR4S13	*T. thioreducens* OGL-20P	-	97.2%	S8 family peptidase	WP_055428439.1
33	Different *Th*. species	*T. peptonophilus* OG-1 CR4S18	*T. kodakarensis* KOD1	+	95%	Hypothetical protein	WP_011249342.1
34	Different *Th*. species	*T. piezophilus* CDGS CR3S2	*T. barophilus* CH5	-	94.6%	-	-
35	Different *Th*. species	*T. profundus* DT5432 CR2S9	*T. barophilus* CH5	-	94.6%	DUF87 domain-containing protein	WP_056934647.1
36	Different *Th*. species	*T. siculi* RG-20 CR1S7	*Thermococcus* sp. 5-4	+	92.9%	SPOUT family RNA methylase	WP_088180241.1
37	Different *Th*. species	*T. siculi* RG-20 CR1S7	*T. radiotolerans* EJ2	-	92.9%	SPOUT family RNA methylase	WP_088866378.1
38	Different *Th*. species	*T. siculi* RG-20 CR1S7	*T. thioreducens* OGL-20P	-	90.5%	SPOUT family RNA methylase	WP_055428680.1
39	Different *Th*. species	*T. siculi* RG-20 CR1S7	*Thermococcus* sp. MV11	+	90.5%	SPOUT family RNA methylase	WP_167773488.1
40	Self-targeting	*T. cleftensis* 1 CR3S22	*T. cleftensis* 1	+	91.7%	Hypothetical protein	WP_014788367.1
41	Self-targeting	*T. paralvinellae* ES1 CR2S33	*T. paralvinellae* ES1	+	100%	tRNA (N(6)-L-threonylcarbamoyladenosine(37)-C(2))-methylthiotransferase	WP_042681706.1
42	Self-targeting	*T. siculi* RG-20 CR1S7	*T. siculi* RG-20	+	97.6%	SPOUT family RNA methylase	WP_088856279.1

Appendix Table 2. Annotated CRISPRTarget outputs. The table provides information about the matches between spacers and protospacers.