Fall 2024 BIOINFORMATICS (BTEC-40220-001) Final Project

Evolutionary Analysis of Cas Protein Systems Using Bioinformatics, Comparative Sequence, Phylogenetic, and Domain Structure Analysis

Alec Pokrandt

Table of Contents

Abstract ..... Page 3
Introduction to CRISPR Systems and Cas9 ..... Page 3
Sequence Selection ..... Page 3
- Cas9 (Cleaves dsDNA) ..... Page 3
- Cas13 (Cleaves ssRNA) ..... Page 3

Jalview Annotations and Sequence Analysis ..... Page 4
- Bridge Helix ..... Page 4
- HNH Domain ..... Page 4
- RuvC Domain ..... Page 4
- PI Domain ..... Page 4

MEGA Phylogenetic Analysis.....Page 4
- Bootstrap Tree: Cas9s, Cas13s, and Cas1 ..... Page 5

Origins of Cas Proteins ..... Page 5
- Role of Transposable Elements ..... Page 6

Extremophile Variants of Cas9 ..... Page 8
- Planococcus Antarcticus ..... Page 9
- Bathymodiolus Septemdierum ..... Page 9
- Seonamhaeicola sp. S2-3 ..... Page 11

Taxa Distribution of Class 2 CRISPR Systems ..... Page 12
Cas9s and IscB Phylogenetic Analysis ..... Page 12
- Environmental Correlations ..... Page 14
- Taxonomic Breakdown ..... Page 14

Summary ..... Page 15
Future Studies .....Page 16
Data ..... Page 16
Citations ..... Page 15

Abstract:

Class 2 CRISPR systems have emerged as a powerful tool for genetic modification and engineering of genomes and transcriptomes. Class 2 systems are characterized by their single effector multidomain proteins. This all-in-one functionality is of particular interest as our findings indicate these systems occur predominantly in pathogenic and commensal bacteria **(1)**. We show unique adaptations in extremophile Class 2 Cas9 systems, and show evidence of Cas9 origins being the IS200/IS605 family of transposable elements. We present phylogenetic trees of these homologues, as well as Cas13, showing that the trees are inconsistent with traditional evolutionary mechanisms. We explore and annotate conserved functional regions of Cas9, and identify any outliers.

Introduction to CRISPR systems and Cas 9:

Clustered regularly interspaced short palindromic repeats (CRISPR) are prokaryotic immune systems. When a bacteria or archaea is infected by a bacteriophage, the systems integrate a specific portion of the phage genome into its own. If this prokaryote is later infected by the same phage, it is able to transcribe its stored sequence to produce a complimentary guide RNA (gRNA), complex that with CRISPR associated (Cas) proteins, and selectively target the invading complementary sequences (bacteriophage) for cleavage, effectively eliminating the threat. Class 2 CRISPR systems (Cas9 cleaves dsDNA, Cas12 cleaves ssDNA, and Cas13 cleaves ssRNA) are of particular interest because unlike other CRISPR systems, they are single proteins which can execute all the necessary functions for nucleic acid targeting and cleavage. Due to Cas9s programmable nature, it has been widely adapted for eukaryotic genome editing, debuting with a sickle cell disease cure earlier in late 2023.

Sequence Selection:

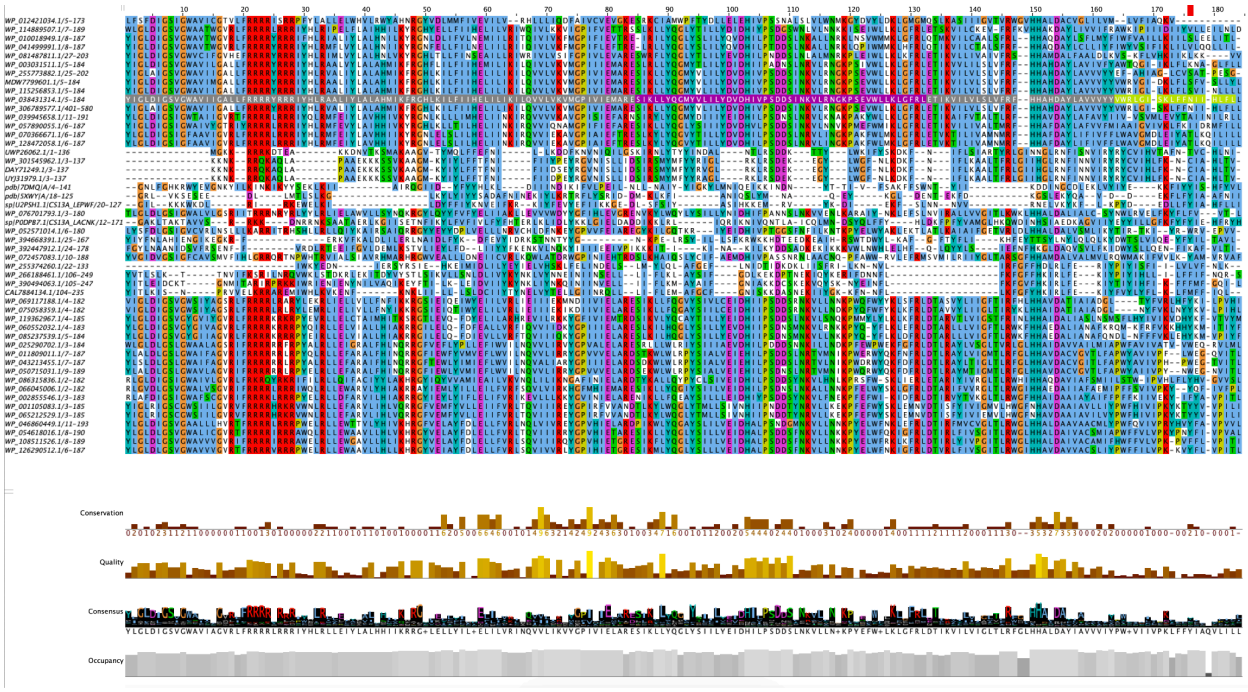
Cas9 (cleaves dsDNA): We selected *Streptococcus Pyogenes (Gene ID: 69900935)* to be our reference genome as it is one of the most extensively studied CRISPR-Cas proteins. Its sequence, structure, and function have been well documented, making it an ideal reference for aligning and comparing homologues. Using Interpro, we extracted the functional domains and executed a BLASTp to find homologs from diverse taxa from a broad range of environments. We also hand selected Cas9 sequences from extremophiles and symbionts for example, *Bathymodiolus septemdierum* thioautotrophic gill symbiont, which is found in deep-sea hydrothermal vent mussels, and *Planococcus antarcticus* which is found in cold environments, particularly Antarctic soil and water. We supplemented our data set with homologues from Gasiunas, G., Young, J.K., Karvelis, T. *et al.* These sequences formed the data set used for Cas9 homolog analysis with several tools. **(2)**

Cas13 (cleaves ssRNA): Cas13 homologues are classified into 4 broad

categories (a, b, c, and d) defined by their structures and functions. We indiscriminately selected the top 4 homologues of each subtype from NCBI. These sequences formed the data set used as an outgroup for Cas9 analysis.

Jalview

After using Clustal Omega to align our Cas9 and Cas13 sequences, we saw several conserved domains. Using Interpro, we imported spCas9 (our reference sequence) annotations in a .gff file into Jalview. Initially, this seemed promising as we had hypothesized that Cas9 and Cas13 had a common ancestor. Gaps have been manually removed in the following to better show homology of several domains at once. The divergent population in the center rows are the Cas13 homologues.



Annotations:

Bridge helix (Gray): Facilitates the interaction between the guide RNA and target DNA, playing a role in signal transduction during cleavage.

HNH domain (Pink): Cleaves the target DNA strand complementary to the guide RNA.

RuvC domain (Second Gray): Cleaves the non-target DNA strand during the double-strand break.

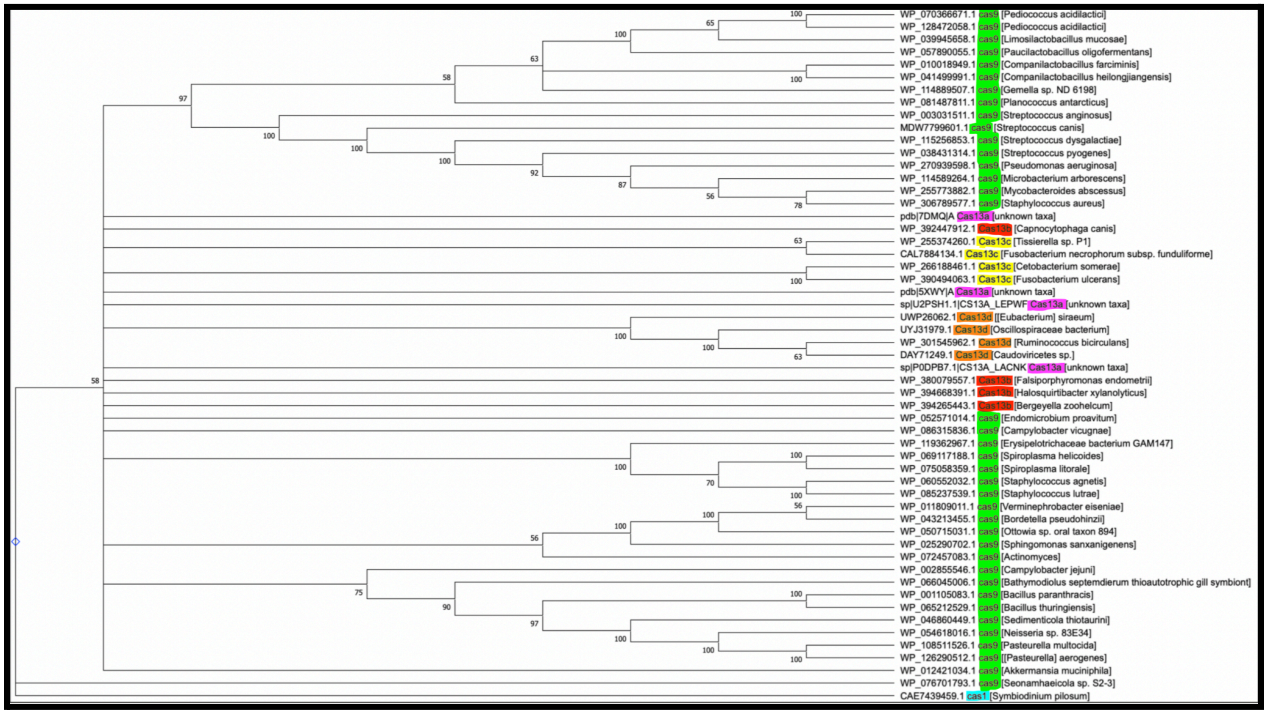
PI domain (Puke Yellow): Recognizes and binds the PAM sequence to ensure correct targeting of DNA.

MEGA Phylogenetic Tree

In Jalview, we observed significant variability in our Cas9 and Cas13 homologs so we chose to make a Maximum Likelihood (ML) bootstrap tree using the WAG model and SPR Heuristic. We did this because ML with WAG captures evolutionary patterns

more accurately than simpler models by accounting for the likelihood of specific amino acid changes based on observed evolutionary patterns. This is critical for homologs with high sequence divergence. We used SPR because it prunes subtrees and regrafts them in different locations to find topologies with higher likelihood scores. We found results incongruent with the idea of Cas9 and Cas13 having a common ancestor.

Bootstrap tree: Cas9s, Cas13s, and Cas1



Pink (Cas13a); Red (Cas13b); Yellow (Cas13c); Orange (Cas13d); Green (Cas9); Blue (Cas1)

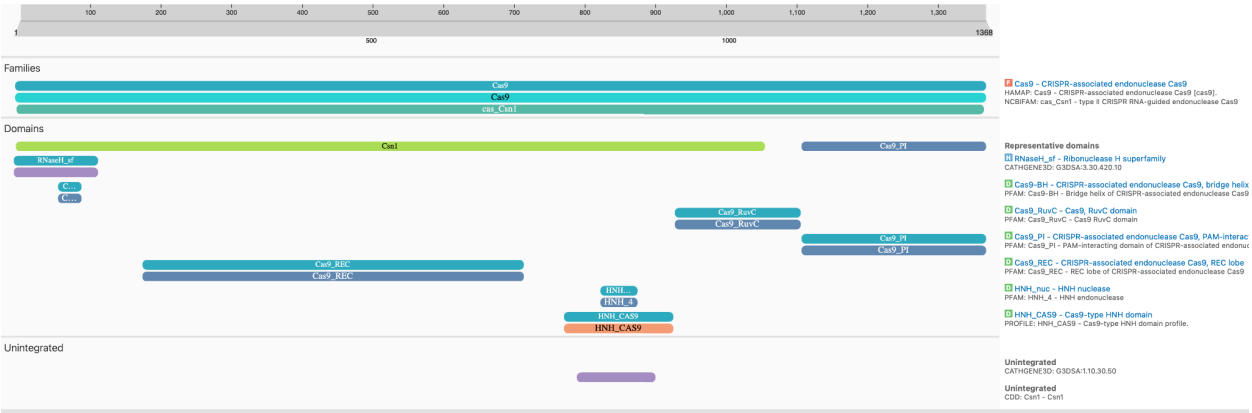
Here we begin to see data that is incongruent with the assumption that these systems all evolved from a common ancestor in the traditional sense. This 200 replicate bootstrap tree shows that an ancestral sequence evolved into Cas9 in 3 separate instances (convergent evolution). The same trend is much more striking in the Cas13 homologues, showing several instances of the same subtype (a, b, c, and d) of Cas13 having evolved separately. While most subtrees showed strong support, the overall topology was alarming. Based on this unexpected tree, we concluded there must have been a more complex evolutionary dynamic at work and began researching CRISPR's origins.

Cas Protein Origins

"The single over-arching theme of CRISPR-Cas evolution is the evolutionary entanglement between these systems of microbial adaptive immunity and various types of MGE. Strikingly, at least four unrelated MGE varieties have contributed to CRISPR-Cas evolution: (i) casposons that gave rise to the adaptation module, (ii) group

II introns that donated the RT to a distinct variety of type III adaptation modules, (iii) non-autonomous IS*605*-like transposons, the ancestors of type II and type V effectors, and (iv) a TA module that apparently contributed Cas2 …… The evolution of class 2 CRISPR-Cas systems clearly involved multiple acquisitions of ancestral MGE genes encoding nucleases that subsequently evolved into CRISPR-Cas effectors. In all likelihood, these genes were captured as a result of chance insertion of the respective MGE into pre-existing CRISPR-*cas* loci or next to orphan CRISPR arrays [43,137]. A puzzling aspect of this part of CRISPR-Cas evolution is the switch of the pre-crRNA processing from the Cas6-mediated mechanism characteristic of class 1 to the effector-catalysed and tracrRNA-dependent mechanisms in class 2. The tracrRNA which is required to recruit RNase III for processing apparently evolved on multiple occasions in different type II and type V systems [145], suggesting that the autonomous, effector-dependent processing is ancestral in class 2. The provenance of this mechanism remains an enigma on which the detailed study of the ancestral MGE-encoded nucleases might shed light." **(3)**

After exploring this research, and determining a Vertical Evolutionary perspective would not suffice, we seeked to explore the alignment and similarity between Cas9 sequences and these MGEs, transposons, and casposons. We turned to InterPro to find the domains in our spCas9 sequence.

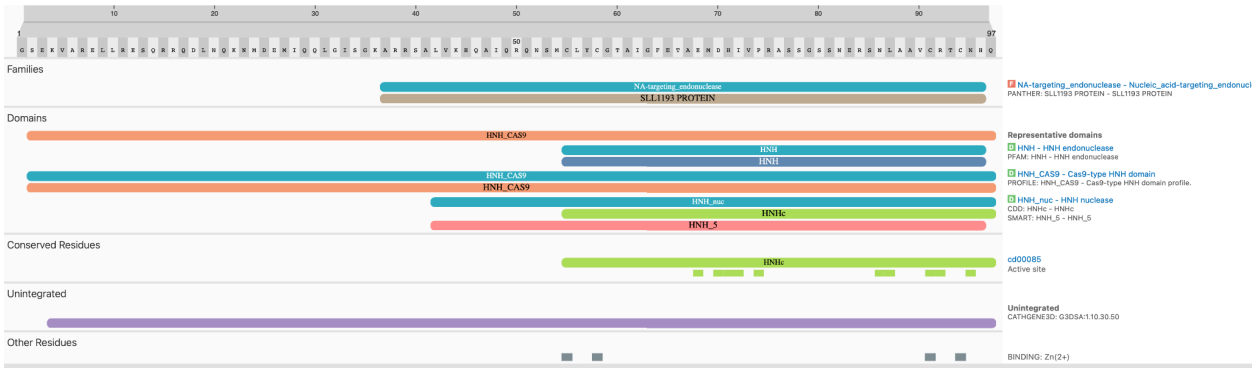


From this analysis, we can see evidence of the co-evolution arms-race explained by Koonin et al. The interesting hits were the HNH_4 and RNaseH superfamily. The RNaseH superfamily is abundant in a large set of proteins and taxa. While we can not explicitly prove that Cas9 gained this domain from a transposon, MGE, etc., our analysis at least shows that it is possible. The HNH-like domain is present in the following:

Bacteriophage Mu, transposase (IPR004189) - D
Tc1-like transposase, DDE domain (IPR038717) - D
Transposase, type 1 (IPR001888) - F
Transposase IS630-like (IPR047655) - F
IS481-like transposase (IPR047656) - F

Transposase IS3/IS150/IS904 (IPR050900) - F
Transposase_5 (IPR052338) - F
Lactococcus phage 712, M3 (IPR009773) - F
Retrotransposon Ty1/copia-like (IPR039537) - F

To explore the HNH in more depth, we manually extracted conserved domains from our Cas9 data set. Our intention was that by hand selecting short conserved sequences, InterPro's vision would be blurred because of the absence of taxa specific flanking regions. We queried with subsets of each conserved region from various taxa, for a total of 50 sequences ranging from 22-117 amino acids in length. Here, we show the most informative results.

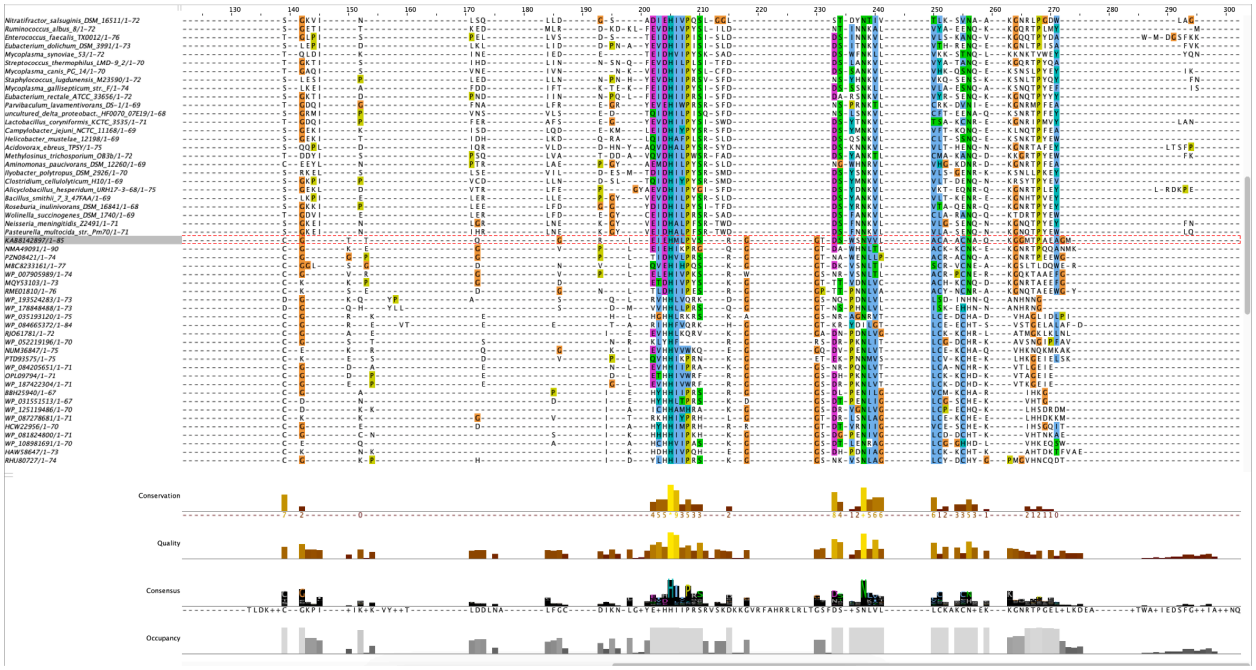


Again, this did not provide anything conclusive, but we did find that HNH endonuclease signatures can be found in the yeast intron 1 protein and in viral proteins. This HNH domain was likely inserted into the RuvC domain of the Cas9 protein, which is responsible for cleaving the non-target strand of DNA.

We obtained several hundred HNH domain sequences from Cas9 homologs and several hundred IscB HNH domains from the IS200/IS605 family of transposable elements. **(4)**

We combined a 171 sequence subset of these into a single FASTA file and aligned it with ClustalOmega.

The top half of the alignment with taxa names are Cas9 homolog HNH domains, and the bottom half with accession names only are a wide variety of HNH domains from IscB proteins of the IS200/IS605 family of transposable elements.
Using Jalview we could further visualize the alignments with the annotations.



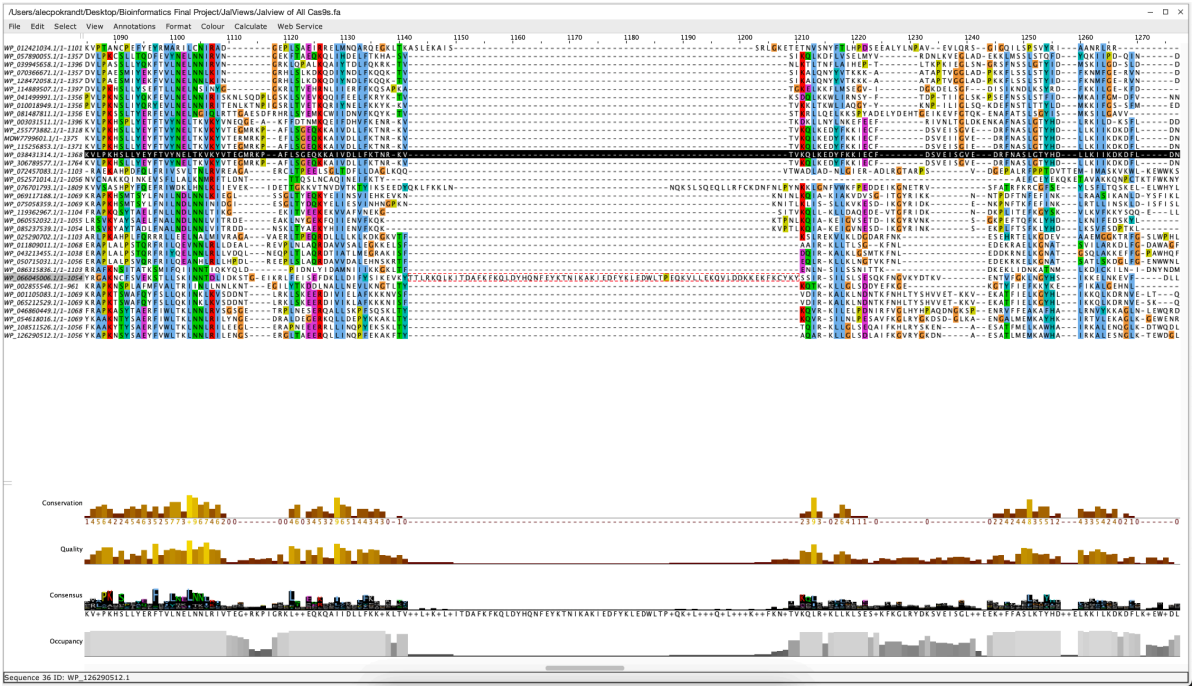
Extremophile Variants of Cas9:

Planococcus antarcticus- we found nothing interesting in the sequence.

Bathymodiolus septemdierum thioautotrophic gill symbiont

Aside from learning the hard way that Cas9 and other CRISPR systems did not evolve from the simple Darwinian dynamics, we found some outliers in our Cas9 homologs that stood out due to their unique living environments. Bathymodiolus septemdierum thioautotrophic gill symbiont which is found in deep-sea hydrothermal vent mussels has a unique feature in the REC lobe. **(5)**

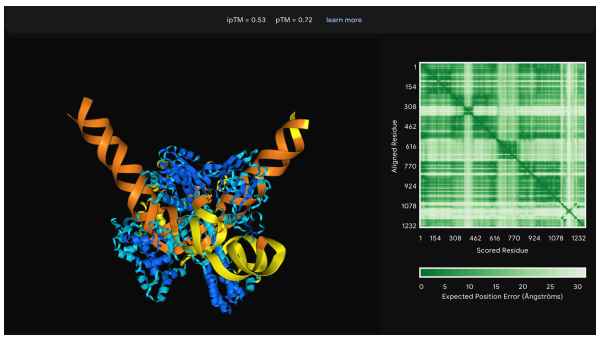
The REC (Recognition) lobe is crucial for the binding of the guide RNA and the formation of the RNA-DNA hybrid. This lobe helps position the target DNA in the active site and supports the stability of the entire complex. Due to the hyperthermic environment, perhaps this additional sequence (selected in red) plays an assisting role in thermostability of the Cas9-gRNA complex and the RNA-DNA hybrid. Our imported annotations for spCas9 are the black sequence which we used to determine what domain we were looking at.



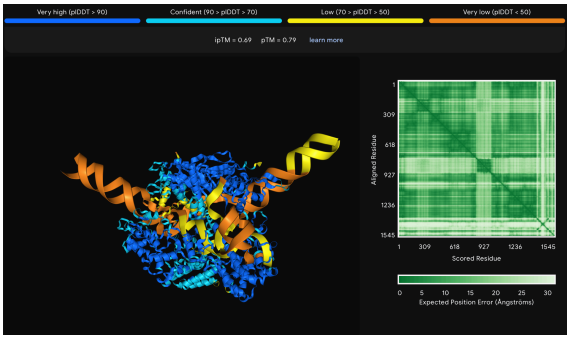
Our AlphaFold predictions showed effective complexing, but we note that there is no ability to control “simulation temperature”. However we were able to successfully model spCas9 and Bathymodiolus septemdierum thioautotrophic gill symbiont Cas9, each complexed with a gRNA targeting the human BCL11A gene. **(6)**

Here we show that both nucleases bind the gRNA, unzip the target DNA, and base pair the gRNAs with the target DNA successfully. We used the same seed in order to have comparable results.

Thermophile Cas9



spCas9

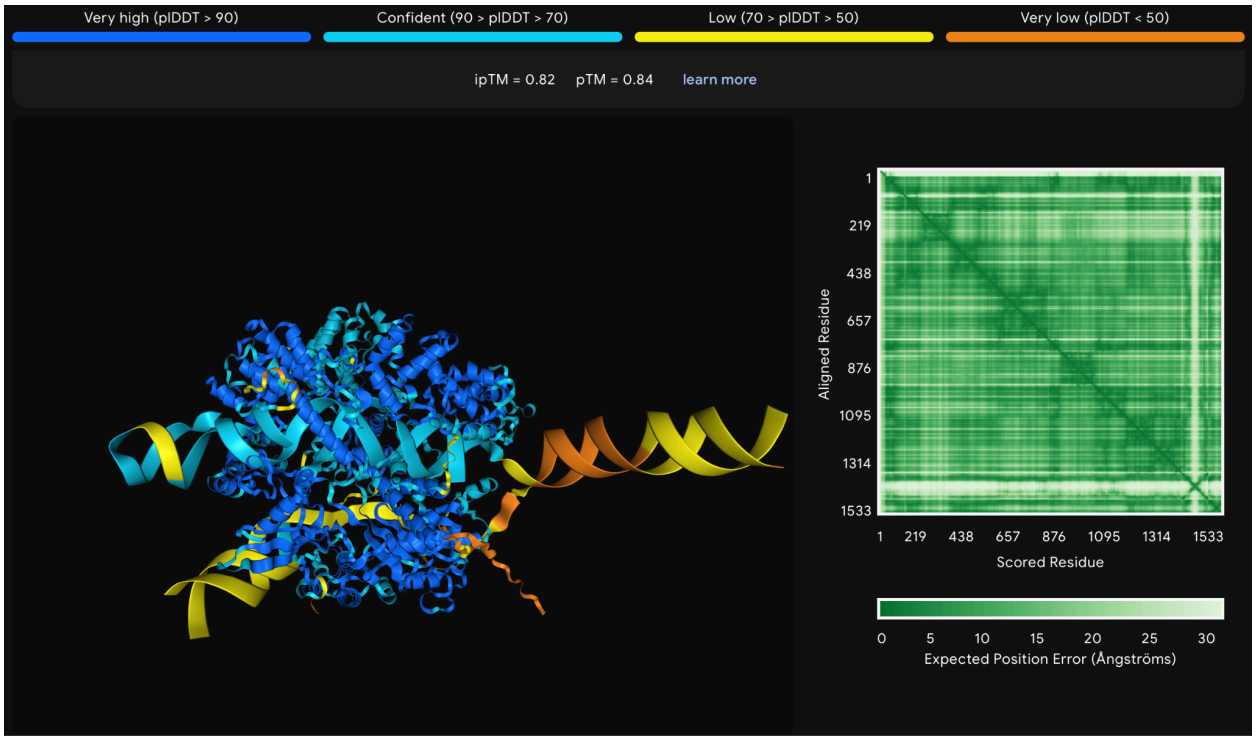


To elucidate if Bathymodiolus septemdierum thioautotrophic gill symbiont Cas9 has unique adaptations, we used ProtParam to understand the chemical composition of the thermophilic variant compared to our reference. These two homologs are quite different sizes, and have different ratios of positively charged amino acids and negatively charged amino acids. That being said, charged amino acids as a percent of the total residue count for each protein was exactly the same: 31.87%. Leaving this

aside, as we were more interested in the comparison of the instability index, we found the thermophilic homolog to be more stable as expected.
Bathymodiolus septemdierum thioautotrophic gill symbiont Cas9- The instability index (II) is computed to be 34.14
spCas9- The instability index (II) is computed to be 37.75

Planococcus antarcticus

We show below that *Planococcus antarcticus* which is found in cold environments, particularly Antarctic soil and water can form an RNA-DNA hybrid (the blue double helix) in AlphaFold3.



Seonamhaeicola sp. S2-3

Seonamhaeicola sp. S2-3 is a marine dwelling bacteria isolated from seawater collected off Jeju Island, South Korea. Out of all the Cas9 homologs we analyzed, it was the only one that had unique inserts in its HNH domain and its RuvC domain. While interesting, we were unable to determine why this is. It could be due to a unique source of its Cas9 sequence, or a distinct environmental pressure on the bacteria. **(7)**
Unique insert in the HNH domain:
>WP_076701793.1/954-1008 type II CRISPR RNA-guided endonuclease Cas9 [Seonamhaeicola sp. S2-3]
YNAIPNNKDHFERLKIVEDAAVTRNNFDKKFFEENEKLKNGNITKKNIEDILKKA

Unique insert in the RuvC domain:
>WP_076701793.1/1186-1245 type II CRISPR RNA-guided endonuclease Cas9 [Seonamhaeicola sp. S2-3]
WTELVAPRYMRLNKLIQPELFSDDSKEDKSCLFGRWQISKSGHQYFDCNLDKSIREKDES

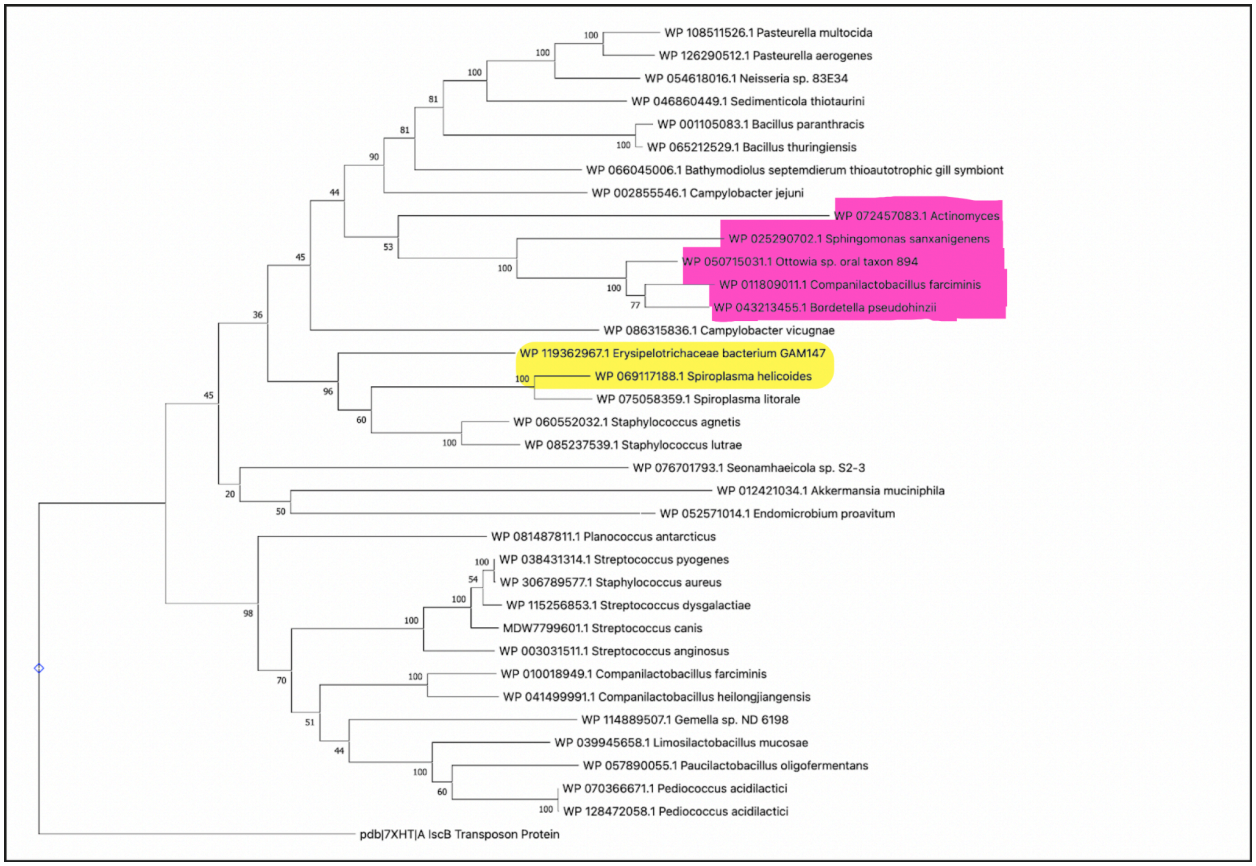
Taxa Distribution of Class 2 CRISPR systems

We initially set out to explore how Cas9 homologs in symbiotes were perhaps part of a distinct phylogenetic clade. Upon researching each bacteria's environment, it became apparent that the majority of our homologs were from commensal and pathogenic bacteria, even though we did not explicitly select for this in our initial sequence acquisition. The majority of taxa which we selected were obtained from intestinal or respiratory tracts of eukaryotes. Our findings are supported by Chylinski et al. “We also detected statistically significant over-representation of type II systems in host-associated (parasitic or commensal) bacteria (P = 4.3E−08)” ... “These observations suggest that environmental traits substantially contribute to the distribution of type II CRISPR-Cas systems among bacteria.”

They also provide additional evidence that supports the previously discussed evolutionary hypothesis emphasizing why Class 2 CRISPR systems may be over represented in commensal bacteria: “ In particular, horizontal gene transfer (HGT) of this system might have been favored in environments with diverse bacterial communities such as animal-associated microbiomes.” **(8)**

Cas9s and IscB Phylogentic analysis

Informed by our failed tree and subsequent research about the evolutionary origins of CRISPR systems, it became apparent that the transposable element would serve as a logical outgroup for phylogenetic analysis of the Cas9 sequences. This addresses the theory that this is likely the distant ancestor of the protein sequence. In addition, we can now analyze our tree to see that either these bacteria diverged at some point from a common ancestor, or that they gained the CRISPR sequences through HGT. We can further draw conclusions by analyzing the taxa’s lineages and evolutionary classification to see which are likely diverged species with diverged Cas9 proteins versus which are more likely to be bacteria with distant ancestors that more recently exchanged the Cas9 sequence through HGT.



Yellow

Erysipelotrichaceae bacterium GAM147 belongs to the family *Erysipelotrichaceae*, which is part of the Firmicutes phylum. *Spiroplasma helicoides* belongs to the genus Spiroplasma, within the Mollicutes class, which is derived from Firmicutes. Thus, the most recent common ancestor (MRCA) of these two species would be an ancestral bacterium within the Firmicutes phylum before the divergence of the Erysipelotrichaceae lineage and the Mollicutes lineage.

Erysipelotrichaceae bacterium GAM147 and Spiroplasma helicoides are bacteria that inhabit distinct environments. Members of the family *Erysipelotrichaceae* are commonly found in the gastrointestinal tracts of mammals, including humans and pigs. They are part of the normal gut microbiota and have been associated with various metabolic and inflammatory conditions. However, specific information about the habitat of *Erysipelotrichaceae bacterium GAM147* is limited. **(9)** *Spiroplasma helicoides* was isolated from the gut of a horsefly (*Tabanus abactor*) collected near Ardmore, Oklahoma, USA, in 1987. *Spiroplasma* species are often found in association with insects, where they can exist as commensals or pathogens. **(10)**

Based on the distinct habitats, we think it is most likely that these sequences diverged from their MRCA and did not exchange the genes through HGT.

Pink

Actinomyces: Primarily found in the oral cavity and gastrointestinal tract of humans and animals. They are part of the normal microbiota but can cause infections under certain conditions. **(11)**
Sphingomonas sanxanigenens: Isolated from soil environments, particularly cornfield soil in China. This species is known for producing extracellular biopolymers like sanxan. **(12)**
Ottowia sp. oral taxon 894: identified in the human oral cavity and considered part of the oral microbiome. **(13)**
Companilactobacillus farciminis: Originally isolated from marinated meat products. While some members of the Lactobacillaceae family are found in the gastrointestinal tracts of humans and animals, specific information about C. farciminis indicates its association with fermented foods. **(14)**
Bordetella pseudohinzii: Naturally colonizes the respiratory tracts of rodents, particularly mice and rats. It has been identified in laboratory animal colonies and can cause respiratory infections in these hosts. **(15)**

Similar Environments:

Actinomyces and Ottowia sp. oral taxon 894 are both found in the human oral cavity and are part of the oral microbiome. Companilactobacillus farciminis is associated with fermented meat products, indicating a presence in food-related environments rather than directly within the human gastrointestinal tract. Here we see a direct environmental correlation between these bacteria. Sphingomonas sanxanigenens was isolated from cornfield soil in China. Bordetella pseudohinzii is found in mouse and rat respiratory tracts. This perhaps indicates similar environments as humans and rats both eat corn which invariably has trace amounts of dirt on it.

Taxonomy:

These five bacterial genera —Actinomyces, Sphingomonas, Ottowia, Companilactobacillus, and Bordetella—belong to distinct taxonomic groups, indicating varied evolutionary relationships. Here's an overview:

Genus	Phylum	Family/Class
Actinomyces	Actinobacteria	Actinomycetaceae
Sphingomonas	Proteobacteria	Alphaproteobacteria
Ottowia	Proteobacteria	Betaproteobacteria
Companilactobacillus	Firmicutes	Lactobacillaceae
Bordetella	Proteobacteria	Betaproteobacteria

The phyla **Actinobacteria**, **Proteobacteria**, and **Firmicutes** represent major divisions within the bacterial domain, characterized by ancient evolutionary splits. Since

these phyla do not share a recent common ancestor beyond the origin of bacteria, their most recent common ancestor would trace back to the **Last Universal Common Ancestor (LUCA)**. Within the bacterial phylogenetic tree, *Actinomyces* belongs to the phylum **Actinobacteria*, while *Sphingomonas*, *Ottowia*, and *Bordetella* fall under the phylum **Proteobacteria*, further dividing into the classes **Alphaproteobacteria** and **Betaproteobacteria**. *Companilactobacillus*, on the other hand, is classified under the phylum **Firmicutes**. As such, the MRCA of these genera predates the divergence of these distinct bacterial phyla.

The five taxa have a reasonable environmental association so they could have exchanged genes, or been exposed to similar transposable elements. Based on this information, we posit that these taxa gained their Cas9 sequences from HGT, or from a transposable element they were all being exposed to.

Summary

The idea that CRISPR systems evolved through HGT, and acquisition of sequences from transposable elements was completely new to me. As I'm specifically driving towards a career in genetic engineering, it was extremely insightful and useful to discover and learn more about CRISPR in this new light.

We initially set our to determine if there was a specific adaptation that had evolved in Cas9 sequences in symbiotes. During our research we found that Cas9 and other single effector molecule CRISPR systems are over represented in pathoge=nic and commensal bacteria. When we looked into the environments that the bacteria lived in that we had selected Cas9 sequences from, we found that the majority of them fit into this category. We selected most of our Cas9 sequences rather randomly so we essentially had a happy accident in being able to show the same trend and distribution.

As we explored our initial hypothesis that we could find unique adaptations in Cas9 effectors based on their lineage, which was further rooted in the assumption that all CRISPR systems diverged from a single common ancestor, we found that this was not reflected in our phylogenetic analyses. After exploring the existing research about the origins of CRISPR systems, and focusing on Type 2 systems, we took a new approach to comparing the sequences of Cas9. We successfully explored unique sequences in extremophile variants, and found supporting evidence for the existing research by identifying a clade of closely related Cas9 homologs that exist in extremely distantly related bacteria. We also compared the Cas9 homologs to the IS200/IS605 family of transposable elements, and used existing research to inform our questions.

Data Availability

Most of our data is not presented in this paper and is available in the attached documents.

Future Studies

Our study has illuminated to us many many areas to further research and explore. We left several interesting threads only explored on a surface level and they could all be explored in greater depth for further understanding. To further understand the evolutionary relationships of bacteria and the role of horizontal gene transfer, transposable elements, and phages in shaping the evolution of adaptive immune systems, future studies could focus on constructing and comparing phylogenetic trees of 16S rRNA and Cas9 gene sequences. These analyses should explore the similarities of the tree topologies to determine whether the Cas9 gene evolves in line with bacterial phylogeny or shows unique patterns indicative of horizontal gene transfer, transposable elements, and phages or lineage-specific adaptations. Comparative studies of Cas9 gene evolution across diverse bacterial lineages, contextualized with ecological and environmental factors, could shed light on the interplay between core genome evolution and the adaptive immune system in bacteria.

All of Our Data, Trees, Alignements, etc.

Bioinformatics Final Project

Citations

1. Chylinski K, Makarova KS, Charpentier E, Koonin EV. 2014. Classification and evolution of type II CRISPR-Cas systems. Nucleic Acids Res. 42:6091–6105.
2. Gasiunas G, Young JK, Karvelis T, Kazlauskas R, Urbaitis M, Siksnys D. 2020. A catalogue of biochemically diverse CRISPR-Cas9 orthologs. Nat Commun. 11:5512.
3. Koonin EV, Makarova KS. 2019. Origins and evolution of CRISPR-Cas systems. Philos Trans R Soc Lond B Biol Sci. 374:20180087.
4. Altae-Tran H, Kannan S, Demircioglu FE, Oshiro R, Nety CJ, Lawrence MM, Rouillon C, Bao S, He C, Zhang F. 2021. The widespread IS200/IS605 transposon family encodes diverse programmable RNA-guided endonucleases. Science. 374:57–65.
5. Patra AK, Perez M, Jang S, Won YJ. 2022. A regulatory hydrogenase gene cluster observed in the thioautotrophic symbiont of Bathymodiolus mussel in the East Pacific Rise. Sci Rep. 12:22232.
6. Jiang M, Ye Y, Li J. 2021. Core hairpin structure of SpCas9 sgRNA functions in a sequence- and spatial conformation–dependent manner. SLAS Technol. 26:66–77.
7. Kim DS, Chi WJ. 2020. Isolation and characterization of a new cellulase-producing marine bacterium, Seonamhaeicola sp. S2-3. Microbiol Biotechnol Lett. 48:539–546.

8. Chylinski K, Makarova KS, Charpentier E, Koonin EV. 2014. Classification and evolution of type II CRISPR-Cas systems. Nucleic Acids Res. 42:6091–6105.
9. Wu J, Liu M, Zhou M, Wu L, Yang H, Huang L, Chen C. 2021. Isolation and genomic characterization of five novel strains of Erysipelotrichaceae from commercial pigs. BMC Microbiol. 21:125.
10. Shen WY, Lo WS, Lai YC, Kuo CH. 2016. Complete genome sequence of Spiroplasma helicoides TABS-2^T (DSM 22551), a bacterium isolated from a horsefly (Tabanus abactor). Genome Announc. 4:e01201–16.
11. BiologyInsights Team. 2024. Actinomyces: structure, pathogenicity, and ecological impact. BiologyInsights.
12. https://www.dsmz.de/collection/catalogue/details/culture/DSM-19645
13. https://www.homd.org/taxa/tax_description?otid=894
14. *Companilactobacillus farciminis | DSM 20182, Lev II, CIP 102987, NCIMB 11954 | BacDiveID:6528*
15. *https://veteriankey.com/bordetella/*

Tool Citations

- PubMed
- Standard Protein BLAST
- Clustal Omega < EMBL-EBI
- AlphaFold Server
- Jalview
- MEGA Software
- Expasy - ProtParam