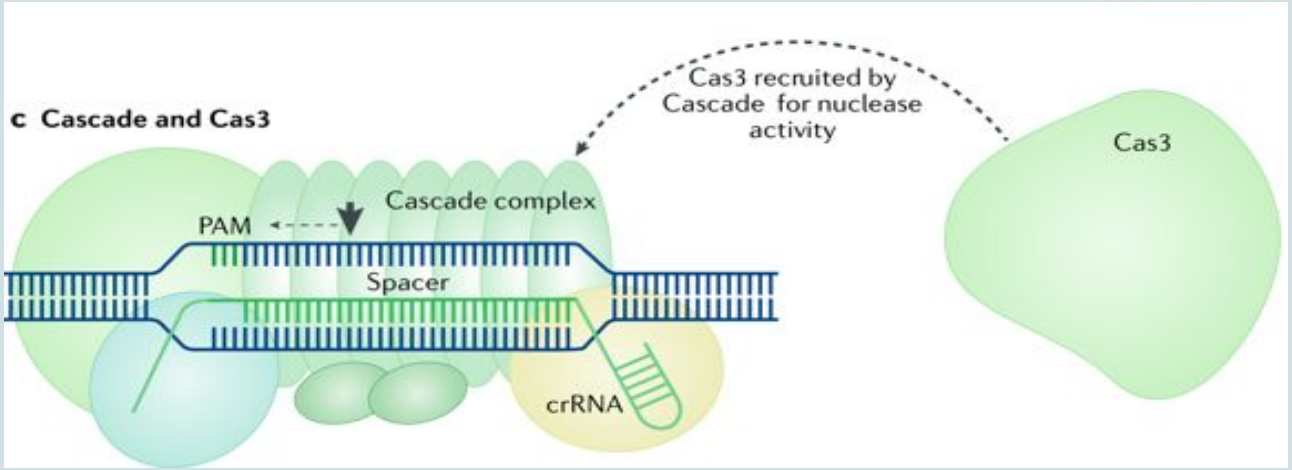
Analysis of Type I Cas3 Protein



Group 7:
Lauren Enriquez, Jiaxin Li, Anita Silver, and Huoran Yuan

Background and Project Scopes

Background:

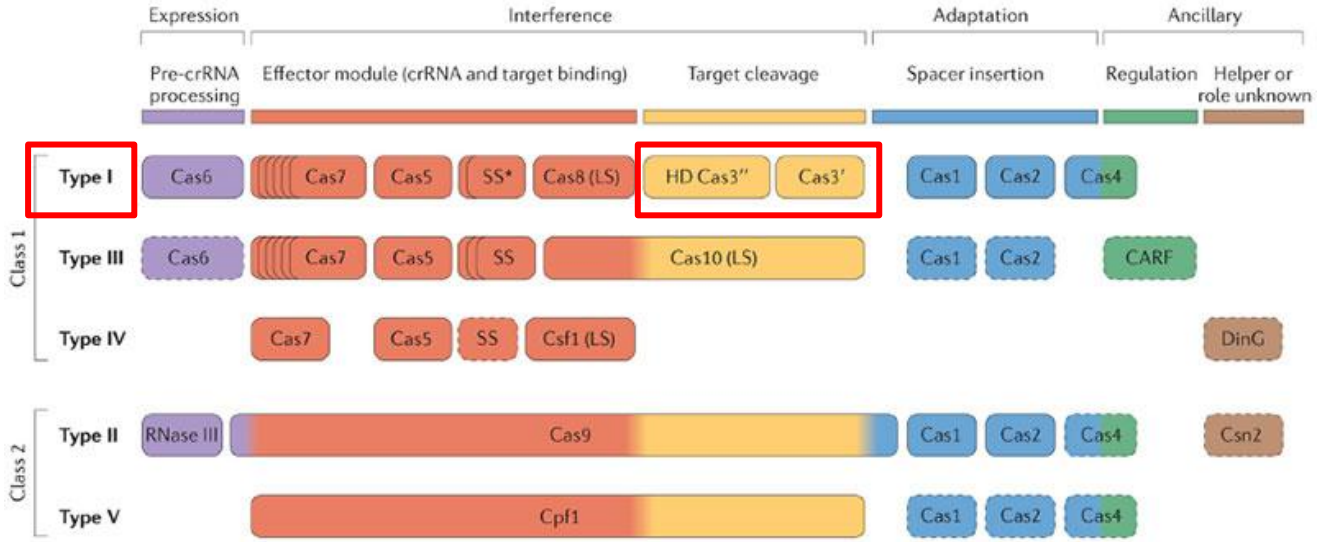
- CRISPR-Cas systems are commonly used in genome editing and diagnostic tools.
- *Salmonella* commonly uses the Type 1 CRISPR-Cas system, which include: Cas6, Cas7, Cas5, Cas8, Cas3, Cas1, Cas2, and Cas4.
- Cas3 (CRISPR-associated protein 3) is a key protein that is necessary for crRNA-guided interference of virus proliferation

Goals:

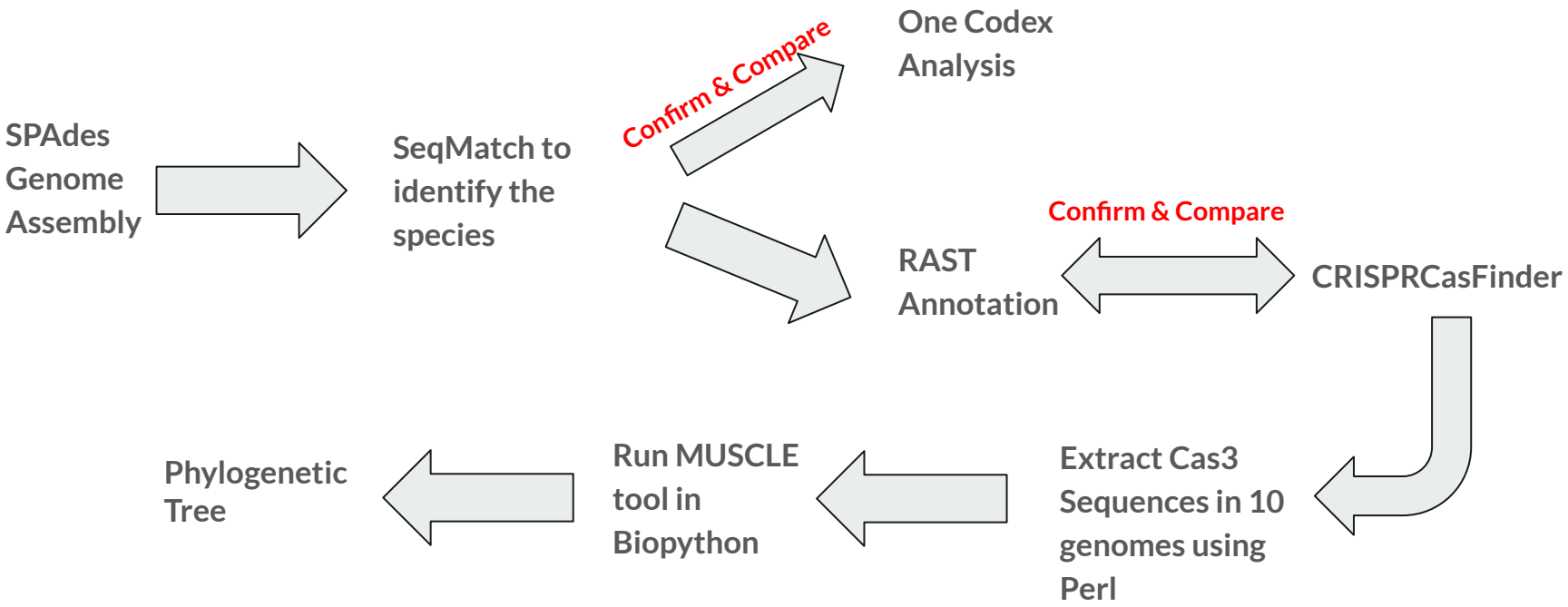
- Identify and analysis the unknown genome given (*Salmonella enterica*)
- Understand critical perspectives in gene-editing research
- Original goal: identify, tally, and compare Cas9 homologs including ISC genes within the transposons of Salmonella genomes.
- New Goal: Identify and compare Cas3 sequences in 10 sets of Salmonella genomes

Overview of CRISPR-Cas systems

- CRISPR-Cas systems in prokaryotic cells mediate an antiviral response that counteracts infection
- Cas3 is a target-degrading nuclease/helicase in Type I
- Cas3 coordinates binding, ATP-dependent translocation, and nuclease digestion of invader DNA.



Methods flowchart



Terminal Commands & Assembly Results

1. Unzip the sequences: $ gunzip SARA_7_S30_L004_R1_001.fastq.gz
2. SPAdes Terminal Command: $ spades.py -o /bigdata/FinalProject_groups/Group_7 -1 ./SARA_7_S30_L004_R1_001.fastq -2 ./SARA_7_S30_L004_R2_001.fastq -t 1

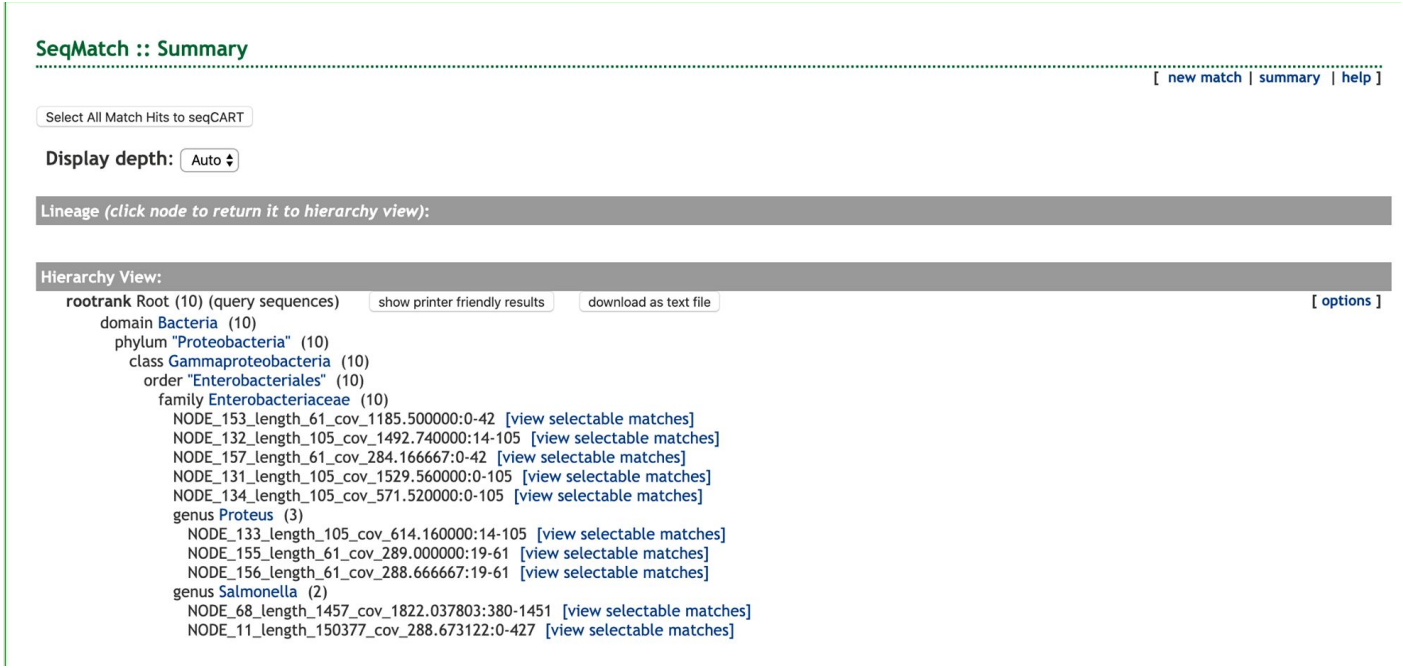
Contigs:

assembly_stats("./spades_outputs/contigs.fasta")
stats for ./spades_outputs/contigs.fasta
sum = 4912502, n = 166, ave = 29593.39, largest = 351593
N50 = 171259, n = 10
N60 = 129379, n = 14
N70 = 97600, n = 18
N80 = 81067, n = 24
N90 = 36944, n = 33
N100 = 56, n = 166
N_count = 0
Gaps = 0

Scaffolds:

assembly_stats("./spades_outputs/scaffolds.fasta")
stats for ./spades_outputs/scaffolds.fasta
sum = 4913156, n = 158, ave = 31095.92, largest = 370039
N50 = 177884, n = 9
N60 = 150377, n = 12
N70 = 113499, n = 15
N80 = 81302, n = 21
N90 = 38291, n = 29
N100 = 56, n = 158
N_count = 800
Gaps = 8

SeqMatch, RAST, One Codex Genome Analysis



SeqMatch :: Summary

[new match | summary | help]

Select All Match Hits to seqCART

Display depth: Auto

Lineage *(click node to return it to hierarchy view)*:

Hierarchy View:

rootrank Root (10) (query sequences) show printer friendly results download as text file [options]

- domain Bacteria (10)
  - phylum "Proteobacteria" (10)
    - class Gammaproteobacteria (10)
      - order "Enterobacteriales" (10)
        - family Enterobacteriaceae (10)
          - NODE_153_length_61_cov_1185.500000:0-42 [view selectable matches]
          - NODE_132_length_105_cov_1492.740000:14-105 [view selectable matches]
          - NODE_157_length_61_cov_284.166667:0-42 [view selectable matches]
          - NODE_131_length_105_cov_1529.560000:0-105 [view selectable matches]
          - NODE_134_length_105_cov_571.520000:0-105 [view selectable matches]
          - genus Proteus (3)
            - NODE_133_length_105_cov_614.160000:14-105 [view selectable matches]
            - NODE_155_length_61_cov_289.000000:19-61 [view selectable matches]
            - NODE_156_length_61_cov_288.666667:19-61 [view selectable matches]
          - genus Salmonella (2)
            - NODE_68_length_1457_cov_1822.037803:380-1451 [view selectable matches]
            - NODE_11_length_150377_cov_288.673122:0-427 [view selectable matches]

SeqMatch, RAST, One Codex Genome Analysis

Closest neighbors of Salmonella sp. (6666666.494818)

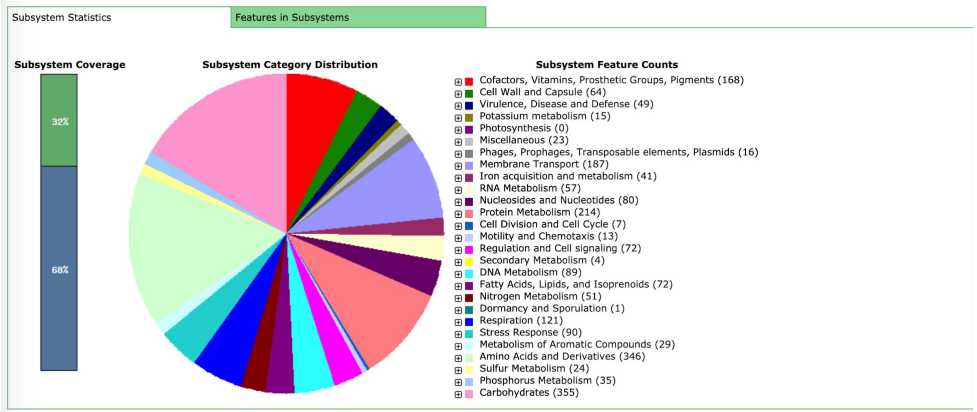
export to file | clear all filters

display 30 items per page

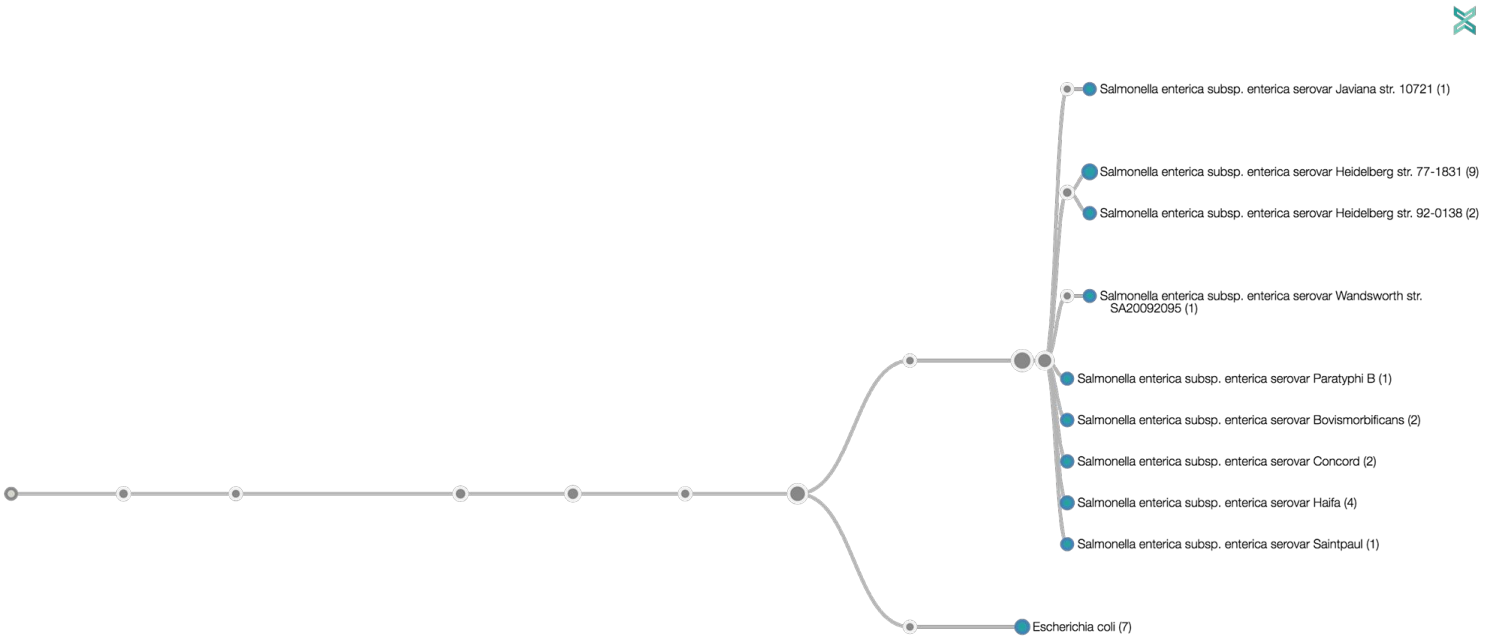
displaying 1 - 30 of 30

Genome ID	Score	Genome Name
209261.1	529	Salmonella enterica subsp. enterica serovar Typhi Ty2
198215.1	518	Shigella flexneri 2a str. 2457T
290339.8	467	Cronobacter sakazakii ATCC BAA-894
316407.3	465	Escherichia coli W3110
272620.3	463	Klebsiella pneumoniae MGH 78578
290338.6	453	Citrobacter koseri ATCC BAA-895
454166.6	439	Salmonella enterica subsp. enterica serovar Agona str. SL483
405955.9	438	Escherichia coli APEC O1
439843.6	424	Salmonella enterica subsp. enterica serovar Schwarzengrund str. CVM19633
511145.6	407	Escherichia coli str. K-12 substr. MG1655
300267.13	396	Shigella dysenteriae Sd197
99287.1	375	Salmonella typhimurium LT2
83333.1	367	Escherichia coli K12
399742.4	366	Enterobacter sp. 638
83334.1	361	Escherichia coli O157:H7
218491.3	345	Erwinia carotovora subsp. atroseptica SCRI1043
693216.3	344	Cronobacter turicensis z3032
243265.1	331	Photorhabdus luminescens subsp. laumondii TTO1
393305.7	312	Yersinia enterocolitica subsp. enterocolitica 8081
229193.1	312	Yersinia pestis biovar Medievalis str. 91001
218491.5	302	Pectobacterium atrosepticum SCRI1043
349968.3	300	Yersinia bercovieri ATCC 43970
273123.1	300	Yersinia pseudotuberculosis IP 32953
498217.4	280	Edwardsiella tarda EIB202
399741.3	271	Serratia proteamaculans 568
343509.6	269	Sodalis glossinidius str. 'morsitans'
706191.3	265	Pantoea ananatis LMG 20103
561229.3	262	Dickeya zeae Ech1591
590409.4	260	Dickeya dadantii Ech586
634503.3	238	Edwardsiella ictaluri 93-146

displaying 1 - 30 of 30



SeqMatch, RAST, One Codex Genome Analysis



CRISPRCasFinder

- Used the Fasta file of scaffolds to search for CRISPR-Cas genes in the Salmonella genomes
- CRISPRCasFinder is a program that can recognize the CRISPR-Cas genes in provided genome
- Found Cas1, Cas2, Cas,3, Cas5, Cas6, Cas7
- Orientation positive and negative, indicating the forward and reverse directions during DNA replication

Negative: Groups 8,9,10

CAS-TypeIE

Type	CAS-TypeIE
Start	132,907
End	141,366

Gene name	Start	End	Orientation
Cas2_0_IE	132,907	133,200	-
Cas1_0_IE	133,200	134,120	-
Cas6_0_IE	134,117	134,767	-
Cas5_0_IE	134,749	135,495	-
Cas7_0_IE	135,506	136,564	-
Cse2_0_IE	136,578	137,138	-
Cse1_0_IE	137,135	138,691	-
Cas3_0_I	138,703	141,366	-

Positive: Groups 1,3,4,7,11,12,13

CAS-TypeIE

Type	CAS-TypeIE
Start	178,992
End	187,451

Gene name	Start	End	Orientation
Cas3_0_I	178,992	181,655	+
Cse1_0_IE	181,667	183,223	+
Cse2_0_IE	183,220	183,780	+
Cas7_0_IE	183,794	184,852	+
Cas5_0_IE	184,863	185,609	+
Cas6_0_IE	185,591	186,241	+
Cas1_0_IE	186,238	187,158	+
Cas2_0_IE	187,158	187,451	+

Isolating Cas3 Genes using Perl

- Perl code to identify, isolate, and copy the sequence encoding Cas3 in the fasta file “AllCasGenes.fasta”

```
print "Please enter your input file's adress:\n";
chomp(my $in = <STDIN>);
open(FNA, "<$in");

open(OF,
">>/Users/apple/desktop/AllCasGenes.fasta");

$genomename = "_Genome";
print("Please enter you genome number:\n");
chomp($nu = <STDIN>);
$genomename .= $nu;
$flag = 0;
```

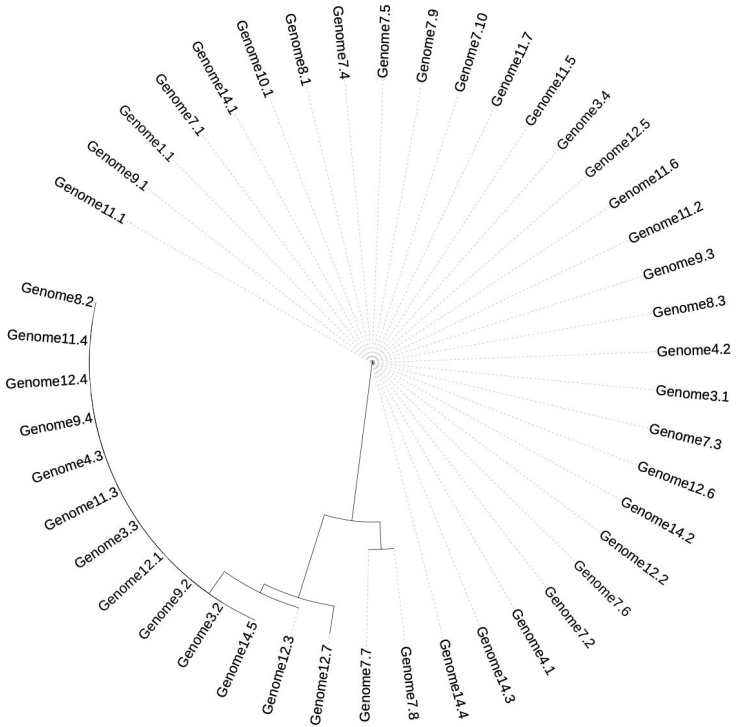
```
while($line = <FNA>){
        #print("1");
        if($flag == 1){
                if($line =~ />.*/){
                        if($line =~ /.*Cas3.*/){
                                $flag = 1;
                                chomp($line);
                                $line = $line . $genomename . "\n";
                                print OF "$line";
                        }else{
                                $flag = 0;
                        }
                }else{
                        print OF "$line";

                }
        }else{
                if($line =~ />.*/){
                        if($line =~ /.*Cas3.*/){
                                $flag = 1;
                                chomp($line);
                                $line = $line . $genomename . "\n";
                                print OF "$line";
                        }
                }
        }
}
close(FNA);
close(OF);
```

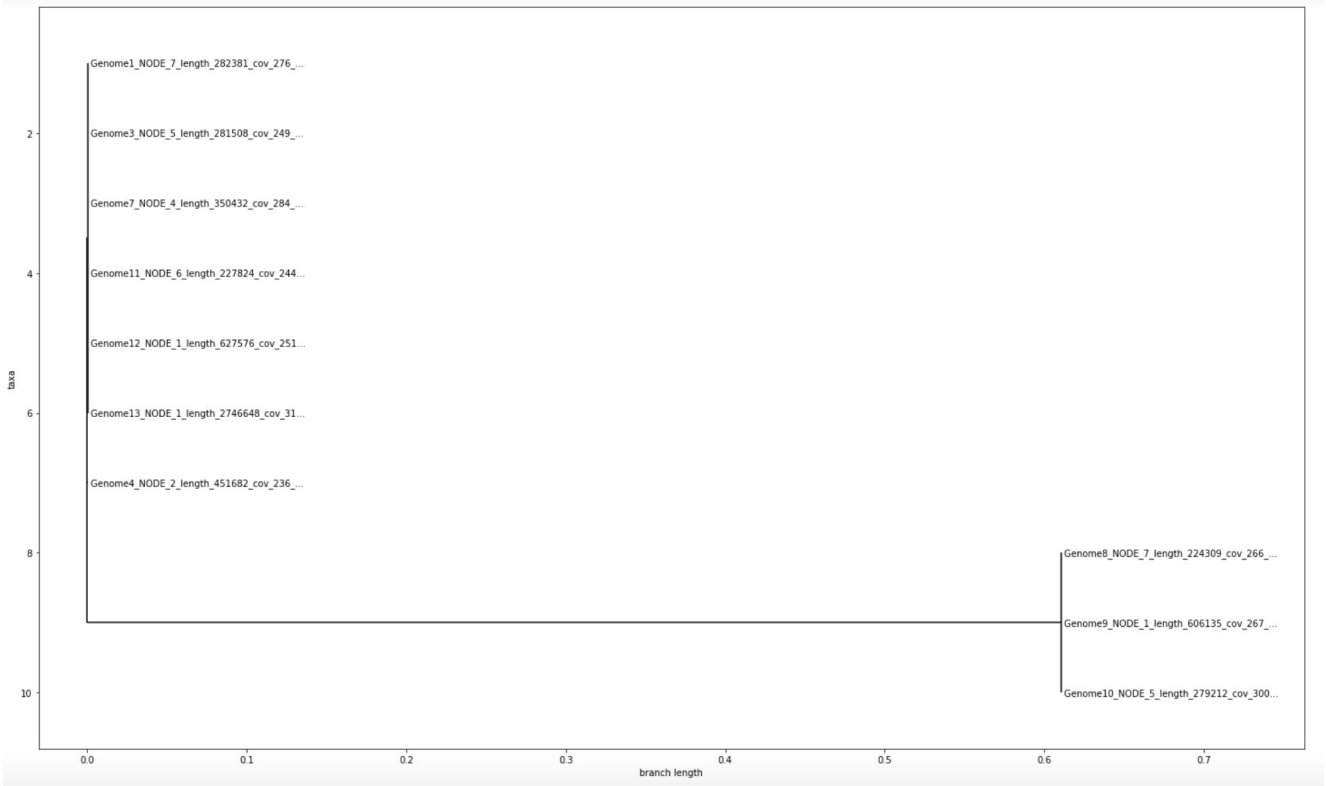
Build A Species Tree as Reference of the Cas3 genes clustering

- We chose 16SRNA as marker
- rna_hmm3.py
- MUSCLE

No obvious cluster found, all of the 10 strains we study are close to each other in evolutionary relationship.



Phylogenetic Tree by MUSCLE



Two Clusters:

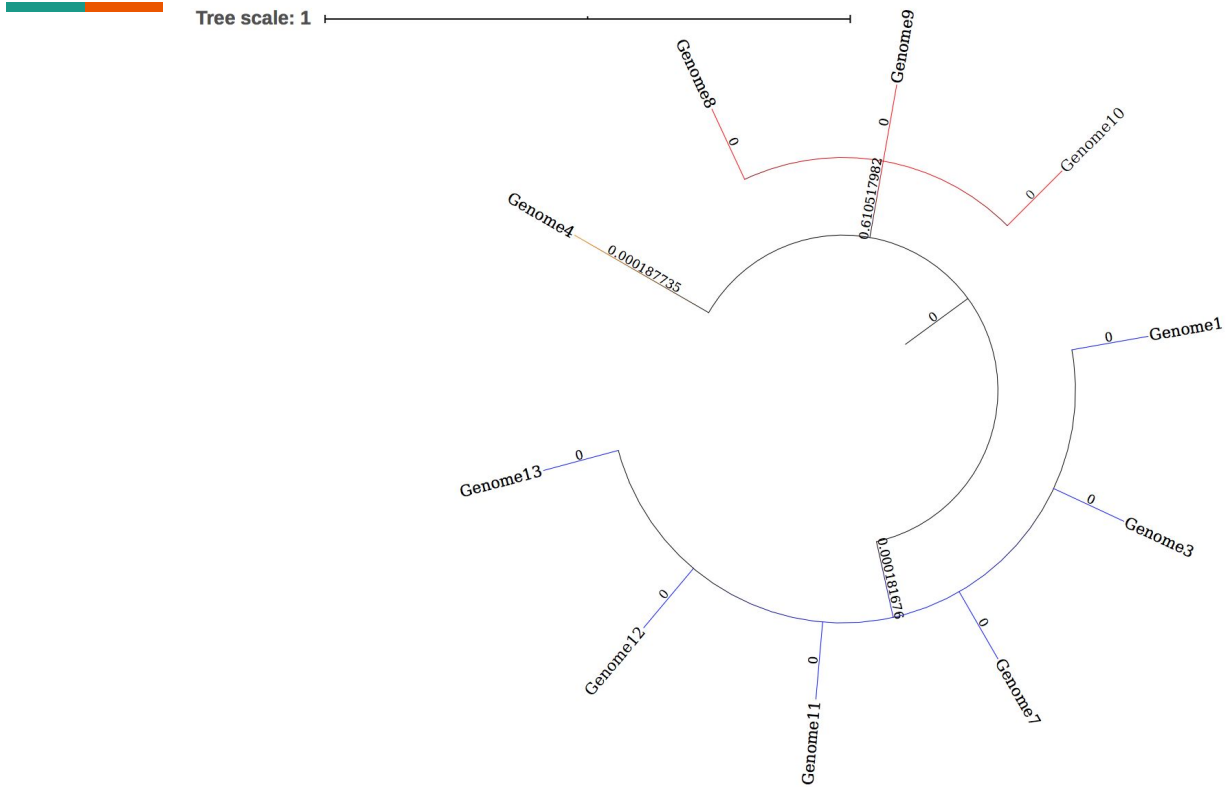
1) 7 Groups: 1, 3, 7, 11, 12, 13, and 4

   The larger cluster has Cas3 gene in the (+) direction

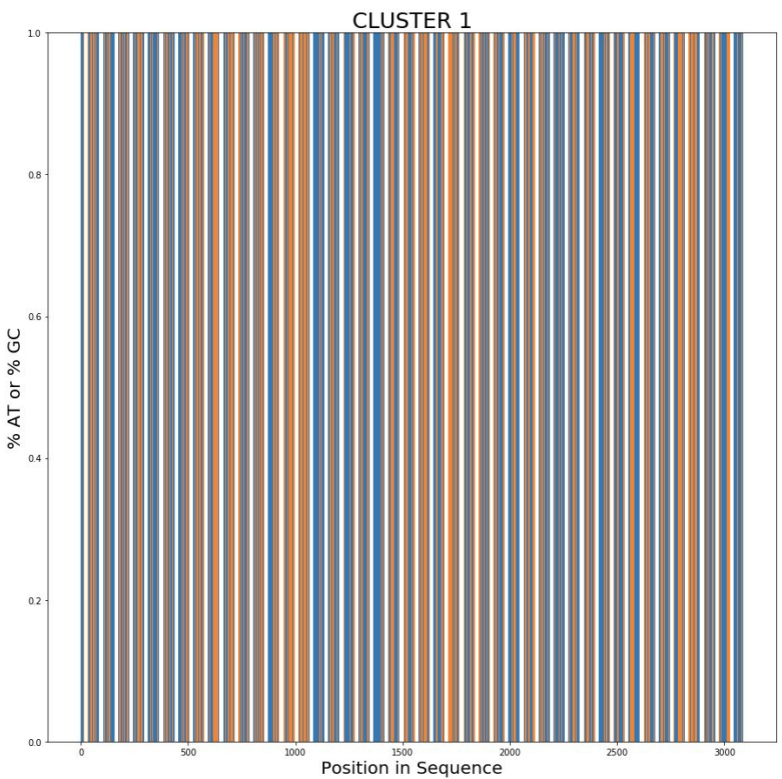
2) 3 Groups: 8,9, and 10

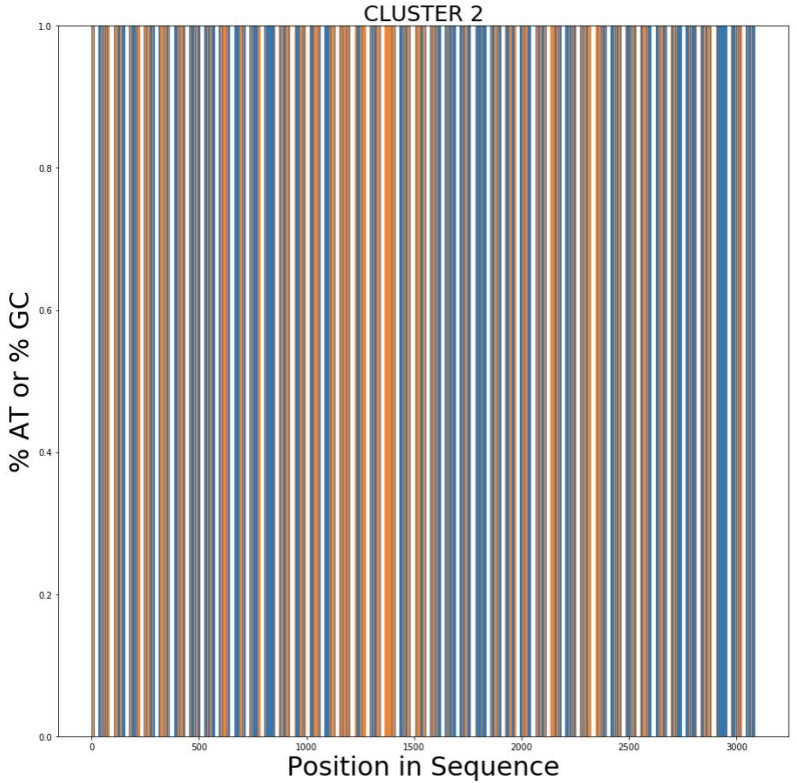
   The smaller cluster has Cas3 gene in the (-) direction

iTOL Phylogenetic Tree



GC% vs AT% Content





Conclusion and Future Steps

- No significant difference was observed between the two clusters
- The Salmonella CRISPR-Cas system might be highly conserved
- Confirm whether the CRISPR-Cas system exhibits typical characteristics of an active immune defense system

Future Steps:

- Obtain a complete, uncontaminated *Salmonella enterica* genome, and repeat with a larger data set
- Investigate the cutting efficiency of Cas3 protein
- Combination of computational analysis along with wet lab experimentation

References

1. Aach, J., Mali, P., & Church, G. M. (2014). CasFinder: Flexible algorithm for identifying specific Cas9 targets in genomes. doi: 10.1101/005074
2. Kapitonov, V. V.; Makarova, K. S.; Koonin, E. V. ISC, a Novel Group of Bacterial and Archaeal DNA Transposons That Encode Cas9 Homologs. *Journal of Bacteriology* 2016, 198 (5), 797–807.
3. Koonin, E. V., Makarova, K. S., & Zhang, F. (2017). Diversity, classification and evolution of CRISPR-Cas systems. Current Opinion in Microbiology, 37, 67–78. doi: 10.1016/j.mib.2017.05.008
4. Makarova KS, Grishin NV, Shabalina SA, Wolf YI, Koonin EV: A putative RNAinterference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. Biol Direct 2006, 1:7.
5. Koonin EV, Makarova KS. Origins and evolution of CRISPR-Cas systems. Philos Trans R Soc Lond B Biol Sci. 2019;374(1772):20180087. doi:10.1098/rstb.2018.0087
6. One Codex: A Sensitive and Accurate Data Platform for Genomic Microbial Identification Samuel S. Minot, Niklas Krumm, Nicholas B. Greenfield bioRxiv 027607; doi: https://doi.org/10.1101/027607
7. https://www.sinobiological.com/cas-proteins.html
8. Shariat, N. (2015). Characterization and evolution of Salmonella CRISPR-Cas systems. Microbiology, 161(2), 374–386. doi: 10.1099/mic.0.000005
9. Sinkunas, T., Gasiunas, G., Fremaux, C., Barrangou, R., Horvath, P., & Siksnys, V. (2011). Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. The EMBO Journal, 30(7), 1335–1342. doi: 10.1038/emboj.2011.41
10. Gong, B., Shin, M., Sun, J., Jung, C.-H., Bolt, E. L., Oost, J. V. D., & Kim, J.-S. (2014). Molecular insights into DNA interference by CRISPR-associated nuclease-helicase Cas3. Proceedings of the National Academy of Sciences, 111(46), 16359–16364. doi: 10.1073/pnas.1410806111
11. Touchon, M., Charpentier, S., Clermont, O., Rocha, E. P., Denamur, E. & Branger, C. (2011). CRISPR distribution within the Escherichia coli species is not suggestive of immunity-associated diversifying selection. J Bacteriol 193, 2460–2467.

Thank You