Analysis of Type I Cas3 Protein in *Salmonella enterica*

Group 7: Lauren Enriquez, Jiaxin Li, Anita Silver, and Huoran Yuan
Dec 04, 2019

Abstract
This project's original focus was to identify, tally, and compare Cas9 homologs including ISC genes within the transposons of *Salmonella* genomes. However, we discovered that *Salmonella* commonly uses the Type 1 CRISPR-Cas system. The components of which include: Cas6, Cas7, Cas5, Cas8, Cas3, Cas1, Cas2, and Cas4. Since our and other group's genomes did not contain the sequences that encode Cas9 proteins (and consequently, any Cas9 homologs), we chose to study the presence of Type 1 Cas3 protein in the *Salmonella* genomes. We chose to focus on the Cas3 protein because similarly to Cas9, it cuts phage DNA. However, "Cas3 is a target-degrading nuclease/helicase in Type I" and does not perform as "clean" of a cut as Cas9.[7] In this lab, we identify the location of CRISPR elements and Cas proteins. We then isolate Cas3 encoding sequences, generate a phylogenetic tree based on the MUSCLE alignment of these sequences, and determine the Cas3 gene evolutionary relationship among the 10 genomes.

Introduction
The prokaryotic adaptive immune system (CRISPR)–Cas in prokaryotic bacteria is one of the most widely used genome editing tools for mammalian genetics. Prokaryotes developed the CRISPR-Cas system to protect themselves from viral influences by building defense barriers through the prevention of adsorption, blocking of injection or degradation of the foreign nucleic acid.[9] The CRISPR-Cas genomic loci in different species of prokaryotic bacteria are highly complex and diverse. Investigating the CRISPR-Cas variants might help to unpuzzle the evolution of genomic and functional characterization of this immunity.[4] The knowledge of the evolution of cas proteins will be an important asset to the future of gene-editing research, and this will require studying non-model organisms. Since each CRISPR-Cas feature is extremely conserved in the Salmonella[8], we want to discover the biological meaning of their conservation and potential function other than immunogenic by investigating the Cas3 gene in Salmonella enterica. Cas3 (CRISPR-associated protein 3) is a key protein that is necessary for crRNA-guided interference of virus proliferation in CRISPR-Cas system.[9] "Cas3 coordinates binding, ATP-dependent translocation, and nuclease digestion of invader DNA."[10] Through analyzing the phylogenetic tree, we are investigating the role of Cas3 in the CRISPR system of *Salmonella enterica* and its mutation in different genomes.

Methods and Data
For our original analysis, we identified the location of CRISPR elements and Cas proteins in 10 genomes of *Salmonella*. We then isolated the Cas3 encoding sequences using Perl programming (**Appendix 3**) and generated a phylogenetic tree based on the MUSCLE alignment of these sequences. We analyzed 10 genomes of
Salmonella species, from the scaffold files from Groups 1,3,4,7,8,9,10,11,12, and 14.

Results and Discussion
1) **Assembly and Annotation**
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
```

The assembly-stat results were printed and analyzed in **Figure 1**.

The N50 for contigs was 171259. The N50 for scaffolds was 177884. Since scaffolds are created by chaining contigs together using additional information about the relative position and orientation of the contigs in the genome, their fragments are usually longer than those of contigs, therefore, the N50 is larger for scaffolds.

We plotted a histogram comparing the lengths of Contig to understand the distribution. The histogram is not uniformly distributed and it is closest to the Gaussian distribution with only one peak, indicating the contigs are mostly lengthed at 0~40,000 bps. Most of the contig fragments have similar lengths.

Additionally, we analyzed our genome using SeqMatch, RAST, and One Codex. This was to compare and confirm the results of our genome. From RAST analysis, we can conclude that our species is *Salmonella enterica* (**Figure 2**). Which means, the Proteus genus from SeqMatch could be due to contamination.

One Codex identifies microbial sequences using a "k-mer based" taxonomic classification algorithm through a web-based data platform, using a reference database that currently includes approximately 40,000 bacterial, viral, fungal, and protozoan genomes.[6]

Our genome is an isolated/low-complexity sample of *Salmonella enterica*. In the sample, 42.77% of reads (n=71) are specific to *Salmonella enterica*. Overall, 74.7% of 166 reads were classified using the One Codex database. An additional 6.63% of reads were classified, but are non-specific or host reads.

Our One Codex result (**Appendix 1 & 2**) confirmed that our species is *Salmonella enterica*. However, it showed the second source of DNA is from *Escherichia Coli*, not from genus *Proteus*. According to One Codex, 97.81% of the genome abundance matches the species *Salmonella enterica* and 2.19% belongs to species *Escherichia Coli*. There is an inconsistency in the two databases/ annotation tools. We would need further analysis to understand which genus by separating the *Salmonella* sequence and the other part of the sequence, for example using Bowtie2. However, this might be extremely hard since we don't have a reference sequence/database.

2) Cas gene identification

We ran all 10 genomes through CRISPRCasFinder, an online program that detects CRISPRs and cas genes. The program identifies the repeated sequence, the unique spacers, and the locations of cas proteins. The location of CRISPR arrays and cas gene clusters is organized in the order in which they lie on the scaffolds.fasta file. The program allows the user to view the data either online or in a downloaded file, with the Cas genes isolated in a fasta file. We then compared the CRISPR and cas results found by the program to those identified by RAST. This was to ensure that the online program was reliable. We concluded that the results from CRISPRCasFinder matched those perfectly with the results of RAST as illustrated in **Figure 4**.

Table 1 to the right shows the starting and ending base pair positions for the Cas3 gene for each genome. These results were determined from CRISPRCasFinder.

The location of the Cas3 protein within the nucleotide sequence was identified for the 10 genomes. Using Perl,
the sequences that encoded Cas3 were isolated and organized in a fasta file labeled "AllCasGenes.fasta". The code is provided in **Appendix 3**.

3) Phylogenetic analysis

The file "AllCasGenes.fasta" which contained the sequences that encoded Cas3 from the 10 genomes was input into MUSCLE. The output was an aligned file that was then used to construct a phylogenetic tree and saved as a newick file (**Figures 5 & 6**). The code used in the server's terminal is provided below:

```
muscle -in AllCasGenes.fasta -out AllCasGenes.aligned.fasta
fasttree -nt <AllCasGenes.aligned.fasta> CasTree.nwk
```

Additionally, we created another phylogenetic tree using the online tool iTOL (Interactive Tree of Life). We compare the outputs of these two trees in the result section.

There are two clusters in the phylogenetic tree. The first cluster contains the Cas3 genes from Groups 1, 3, 7, 11, 12, 13, and 4. The second cluster contains the Cas3 genes from Groups 8, 9, and 10. After further investigation of the CRISPRCasFinder data, we determined that the distinguishing factor that defines these clusters are in the orientation of the Cas3 gene. Cluster 1 has the Cas3 gene in the (+) orientation, indicating that the gene can be found on the forward DNA strand. Cluster 2 has the Cas3 gene in the (-) orientation, indicating that the gene can be found in the reverse DNA strand.

The branch length of cluster 2 was estimated to be 0.610517982 and that of cluster 1 was 0.000181676. Although Group 4's genome was placed in cluster 1, it is slightly different, with a branch length of 0.000187735. The length of all the sequences in each cluster was 2,664 base pairs.

Figure 7 displays the fraction of A or T and the fraction of G or C for each position in the Cas3 gene of each cluster in a graph. Our analysis showed that there was little to no transversion mutation in the Cas3 genes in each cluster. The similar GC content in the two clusters suggested the highly conserved characteristic of the Salmonella CRISPR-Cas system. There have been researchers speculating whether the *Salmonella* CRISPR-Cas system provides immunity against viral invasion and their data were similar to observations made with E. coli,[8] suggesting the CRISPR system does not "exhibit typical characteristics of an active immune defense system."[11] Therefore, our observation of the conserved Cas3 gene in the two clusters might suggest the immune defense system in Salmonella enterica is becoming less active.

Conclusion

From an unknown genome sequence, we successfully assembled and confirmed the species name to be *Salmonella enterica* via SeqMatch, RAST, and One Codex. However, we were not able to confirm the source of contamination to be *E. coli* or *Proteus* nor remove the contaminated sequence due to the lack of a database. We further searched and isolated the Cas3

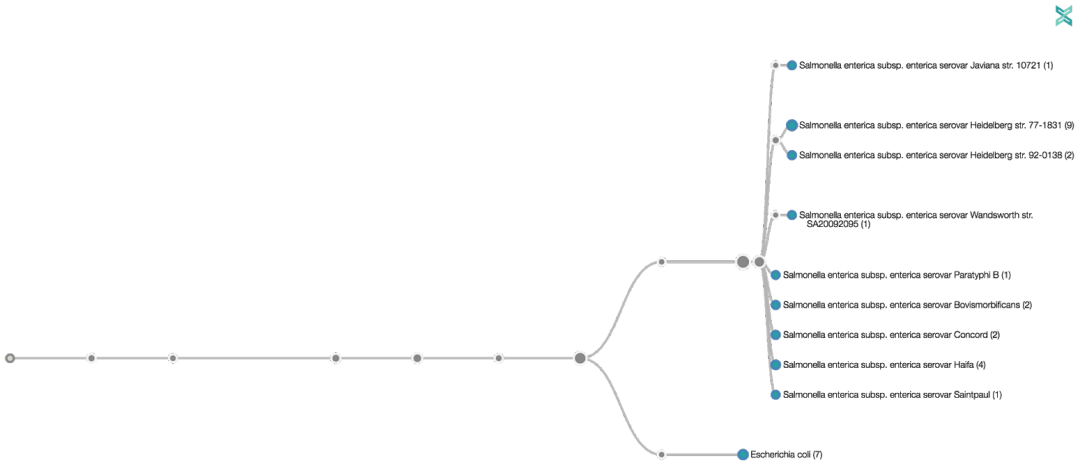
genes from 10 scaffold Fasta files, used them to create a Newick format and plotted a phylogenetic tree to find out the relationships between the Cas3 genes in the 10 genomes. From the phylogenetic tree by iTOL, we observed three branches, two of which had extremely low value in between. However, the tree generated in Biopython only showed two clusters. Therefore, we assumed there were only two main clusters for the 10 Cas3 genes. CRISPRCasFinder showed that the Cas genes in the two clusters had different orientations, suggesting that the DNA replication directions were different. We further analyzed the GC content of the two clusters by plotting a GC content against the sequence position graph in Biopython and no significant difference was observed in between.

We concluded that the Cas3 gene might be highly conserved in *Salmonella enterica* since no major difference was found in the 10 sets of data. In order to verify whether the CRISPR-Cas system is immune active, in addition to Cas3, we need to design further experiments to understand more of the Type I Cas genes: Cas6, Cas7, Cas5, Cas8, Cas1, Cas2, and Cas4 in *Salmonella enterica*.We would need to obtain a complete, uncontaminated *Salmonella enterica* genome, and repeat with a larger data set. Additionally, the cutting efficiency of Cas3 can be further investigated. This would require a combination of computational analysis along with wet lab experimentation. Determination of Cas3 efficiency in the *Salmonella enterica* genome would be beneficial in the research efforts to improve CRISPR genome editing tools.

Appendix

Organism Name	Rank	Tax ID	% of All Reads	# of Reads	# of Reads (w/ Children)	Est. Depth	Est. Abundance
Salmonella enterica	Species	28901	16.27	27	71	0.976 ×	97.81%
Escherichia coli	Species	562	4.22	7	7	0.022 ×	2.19%
Bacteria	Superkingdom	2	1.20	2	124	—	—
Enterobacteriaceae	Family	543	13.25	22	102	—	—
Escherichia	Genus	561	0.60	1	8	—	—
Salmonella	Genus	590	0.60	1	72	—	—
Salmonella enterica subsp. enterica serovar Heidelberg	No Rank	611	2.41	4	15	—	—
Proteobacteria	Phylum	1224	4.22	7	122	—	—
Gammaproteobacteria	Class	1236	6.02	10	115	—	—
Salmonella enterica subsp. enterica serovar Paratyphi B	No Rank	57045	0.60	1	1	—	—
Salmonella enterica subsp. enterica serovar Bovismorbificans	No Rank	58097	1.20	2	2	—	—
Salmonella enterica subsp. enterica	Subspecies	59201	10.24	17	44	—	—

Appendix 1 One Codex Genome Analysis Result



Appendix 2 One Codex Genome Analysis Species Tree

```
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);
```

Appendix 3 Perl code to identify, isolate, and copy the sequence encoding Cas3 in the fasta file "AllCasGenes.fasta"

```
# PROCESSING: Calculates the G+C content for each position in Cluster 1
cluster1_GC = []
for i in range (len(align_map[1][1].seq)):
   sum = 0
   for x in align_map[1]:
      sum = sum + (GC(x.seq[i])/100)
   avg = sum / len(align_map[1])
   cluster1_GC.append(avg)

# PROCESSING: Calculates the G+C content for each position in Cluster 2
cluster2_GC = []
for i in range (len(align_map[2][1].seq)):
   sum = 0
   for x in align_map[2]:
      sum = sum + (GC(x.seq[i])/100)
   avg = sum / len(align_map[2])
   cluster2_GC.append(avg)

# PROCESSING: Calculates the A+T content for each position in Cluster 1
GCdata_1 = np.array(cluster1_GC)
ATdata_1 = 1 - GCdata_1

# PROCESSING: Calculates the A+T content for each position in Cluster 2
GCdata_2 = np.array(cluster2_GC)
ATdata_2 = 1 - GCdata_2
GC_vs_AT = {1:[GCdata_1, ATdata_1],
        2:[GCdata_2, ATdata_2]}
```

Appendix 4 This section of code calculates the percentage of A-T content for a given position in the sequences for each cluster. It then stores this information in the dictionary "GC_vs_AT".

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.

Written report

Submit a report (up to 5 pages, not included figures) along with your final presentation. This report should summarize your assembly results, your annotations, the methods that you used for your original analysis, and the findings of your analysis. Please be clear about the question you are trying to answer, how your chosen method will help you answer it, and any potential limitations of your results.