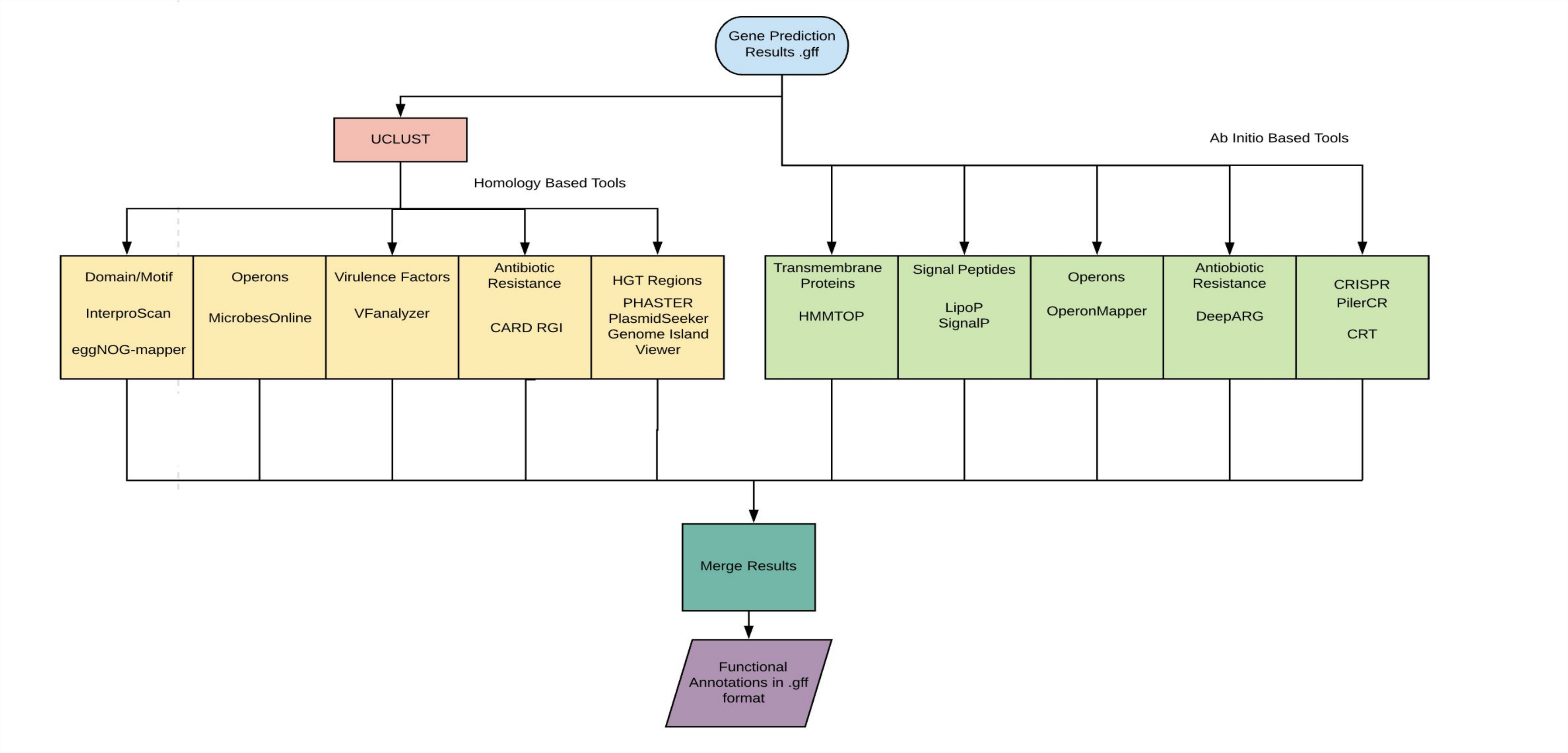
Functional Annotation Results

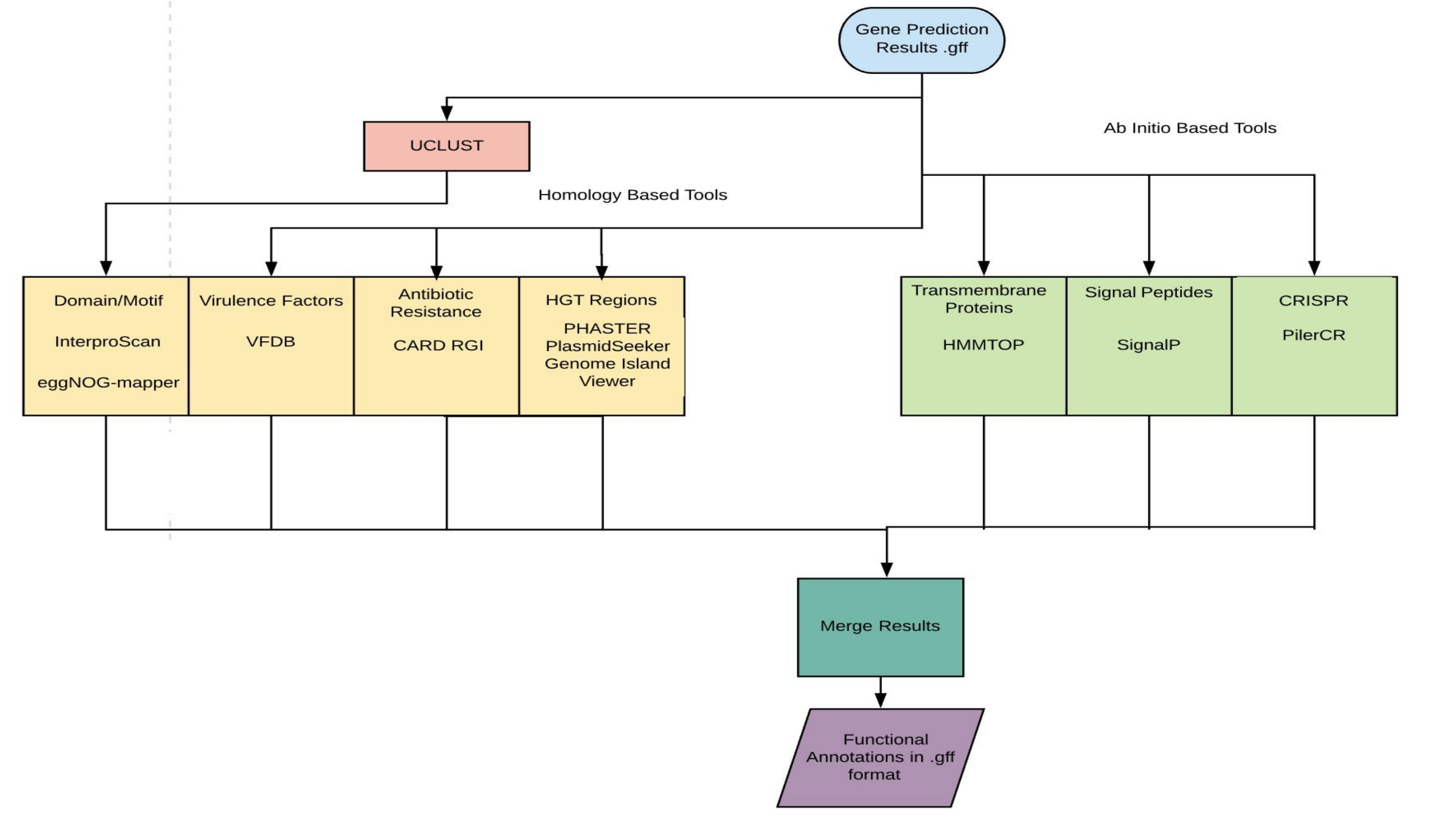
Team 3

Allison Rozanski, Gulay Bengu Ulukaya, Cheng Shen-Yi, Pallavi Misra

Initial Pipeline



Updated Pipeline



UCLUST[2]

Each cluster contains a single centroid sequence upon which the other sequences must have a certain sequence similarity to to be considered apart of the cluster [2]

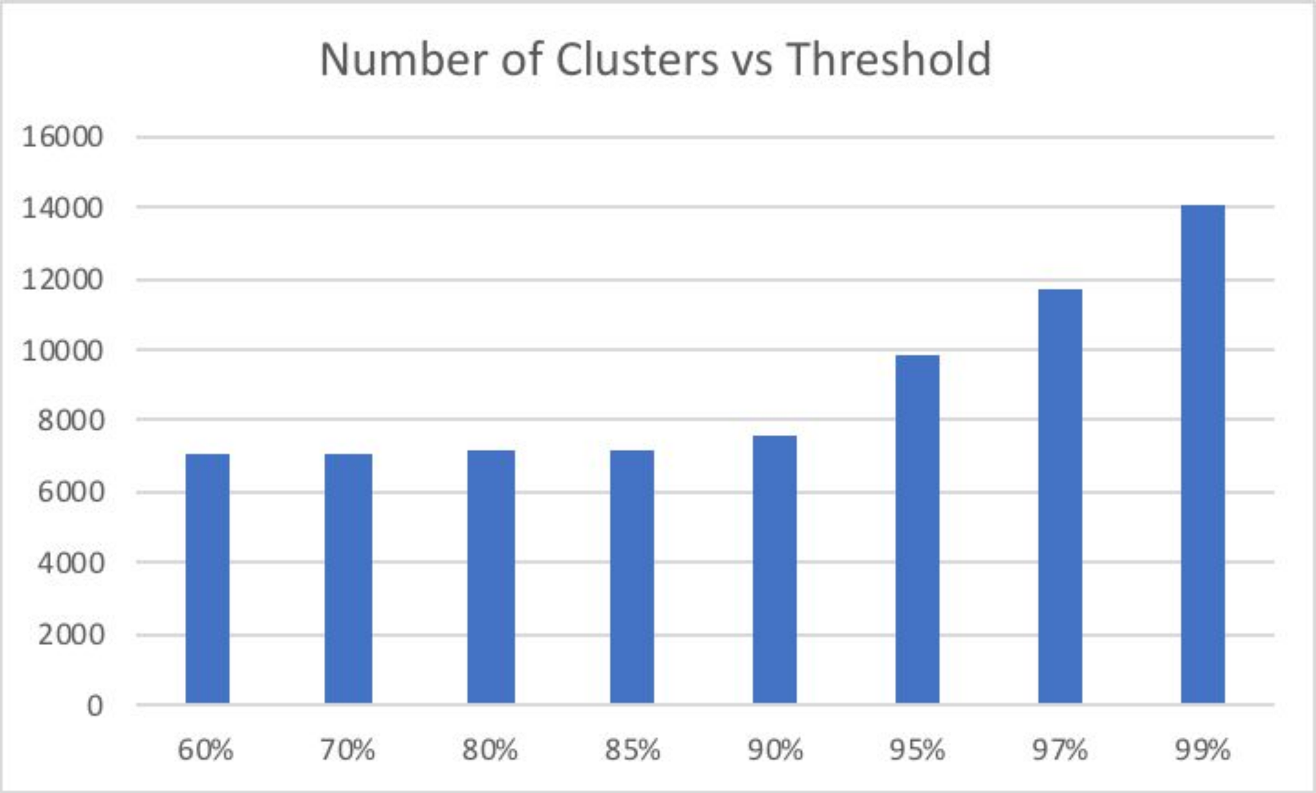
The threshold limit can be thought of as the radius of the cluster

Identities are determined by global alignment
Used a 97% identity cluster threshold: obtains larger average cluster sizes and relatively low amount of singletons: 1%

Although 95% identity had larger average cluster sizes we wanted a higher degree of specificity in our results.

Identity Threshold	Clusters	Singletons	Avg Cluster Size
99%	14,010	976	11
97%	11,720	459	14.2
95%	9,871	322	15.6

UCLUST: Cluster Frequencies



eggNOG Mapper[1]

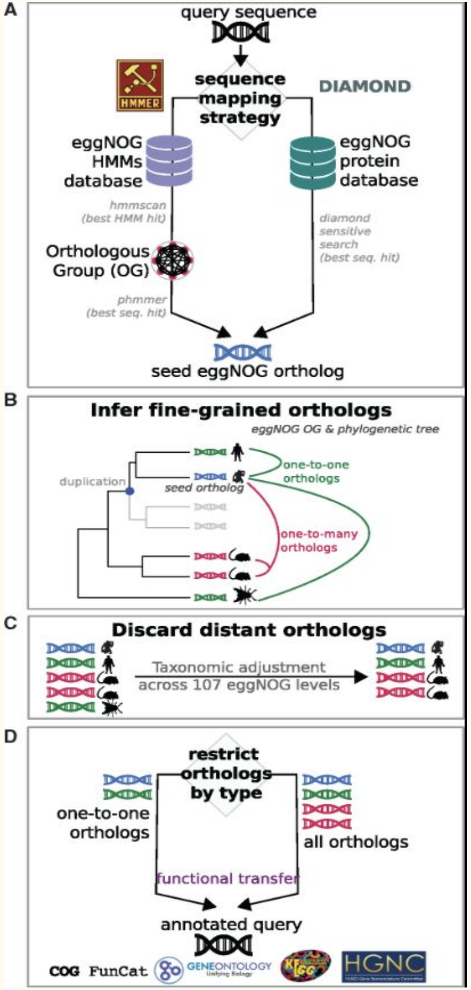
Download eggNOG databases:

```
download_eggnog_data.py bact
```

Run eggNOG:

- DIAMOND vs. HMMER
- “recommended for very large data sets such as metagenomes, as well as for annotating organisms with close relatives among the species covered by eggNOG”
- Because the eggNOG database covers close relatives of Listeria monocytogenes and due to the size of our analysis even after clustering we decided to utilize DIAMOND

```
emapper.py -i ../USEARCH/All_Centroid.fasta -m diamond
--translate -d bact -o eggNog_annotations
```

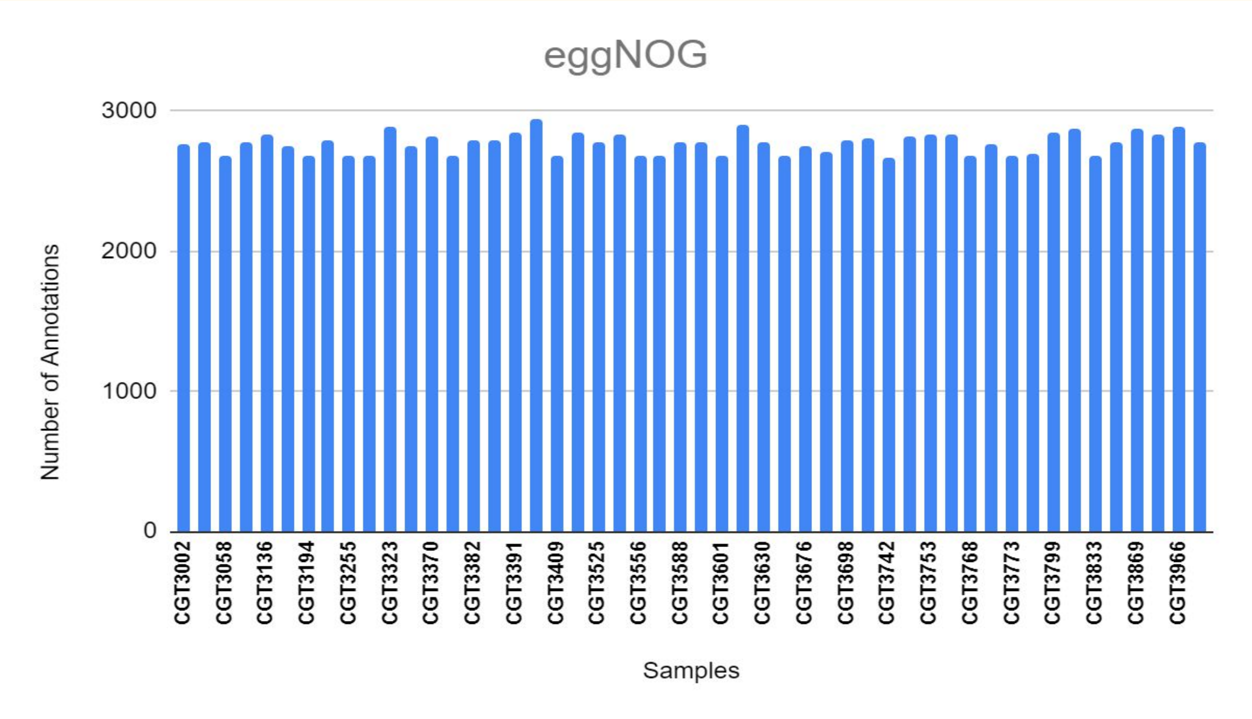


eggNOG Results

- After mapping the results of eggNog we determined what genes were left unannotated by eggNOG

- We mapped these unannotated genes to the output of InterProScan.

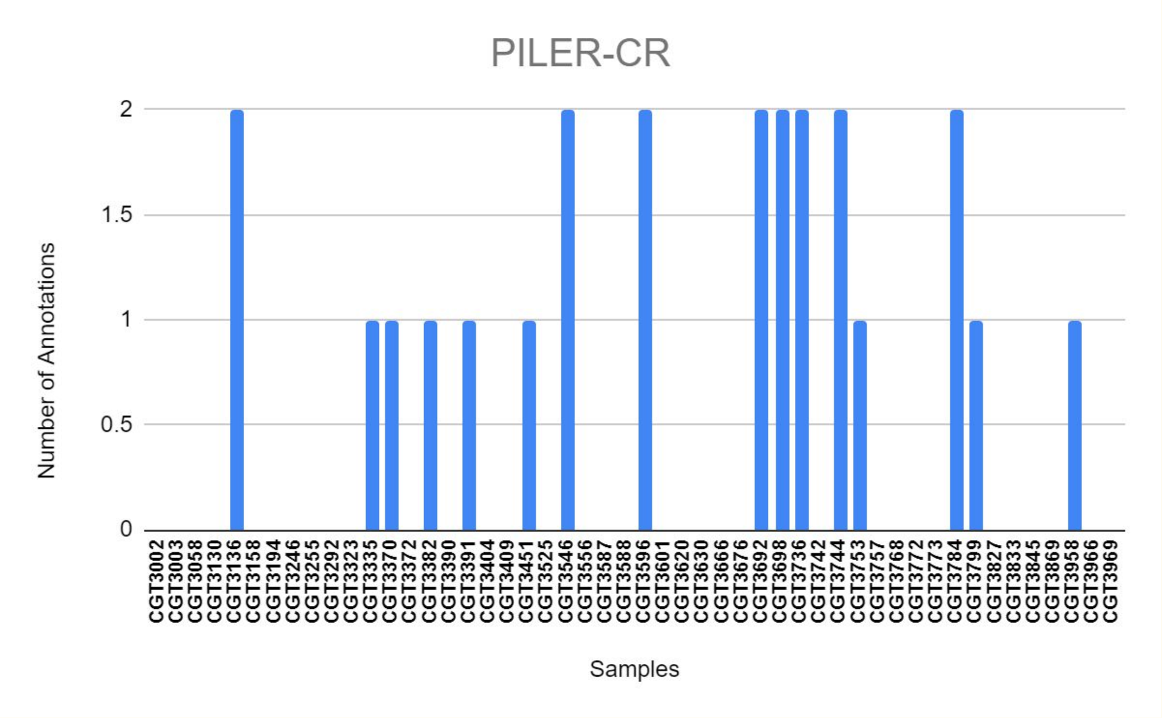
- From this we were able to gain more coverage



PILER-CR Results

- Identifies CRISPR repeats which play a major role in bacteria’s antiviral defense system
- The results shown are from non-coding regions of genomes
- Only one genome had a CRISPR array in coding region of genome
- Command used-

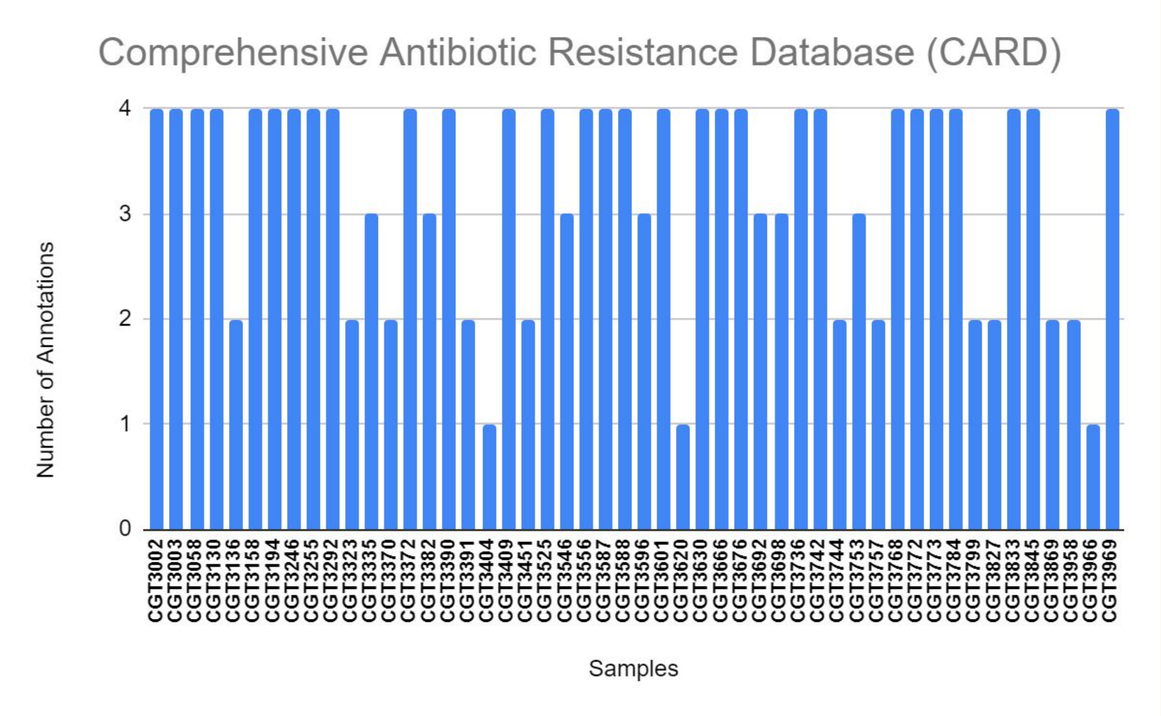
```
./pilercr -in <input_file> -out <output_file>
```



CARD-RGI Results

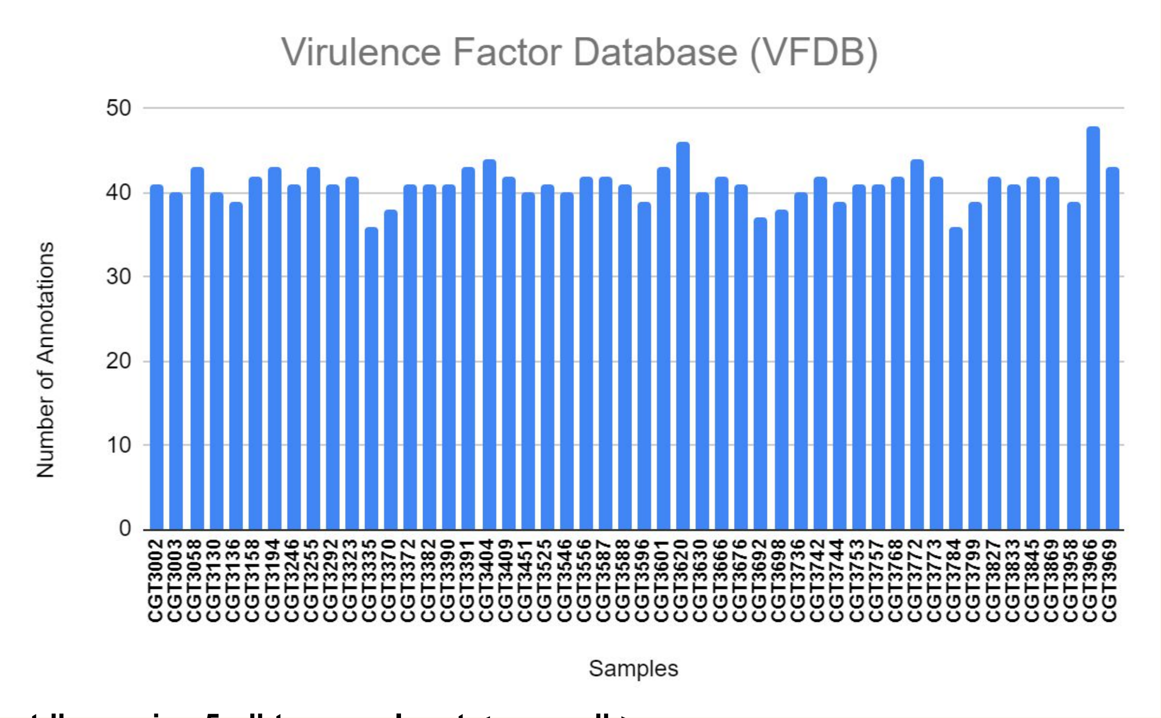
- CARD is a rigorously curated collection of characterized, peer-reviewed ARG which is monthly updated
- The results shown are from coding regions of genomes
- No antibiotic resistance genes were present in the non-coding region
- Command used-

```
rgi -i <input_file> -o <output_file>
```



VFDB Results

- VFDB is an integrated and comprehensive online resource for virulence factors of bacterial pathogens (recently updated in 2019)
- The results shown are from coding regions of genomes
- No virulence genes were present in the non-coding region
- Commands used-



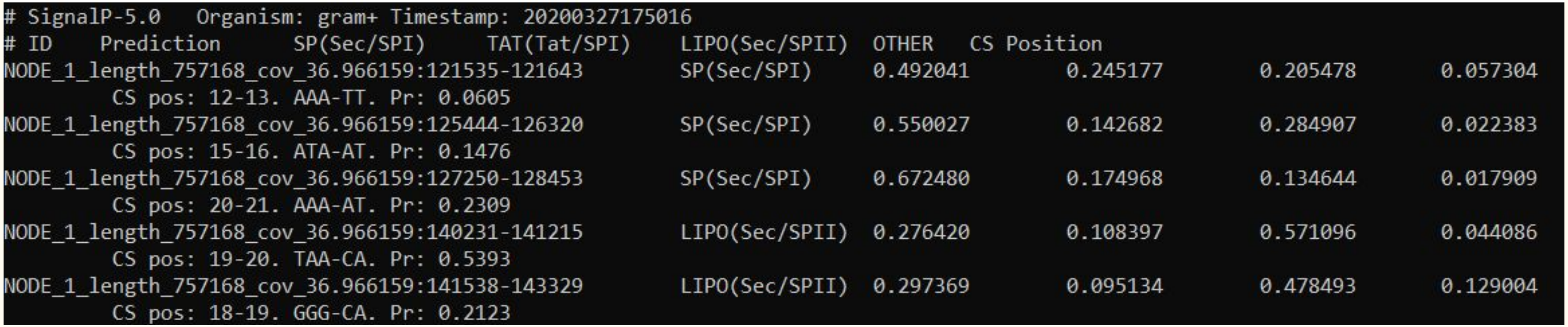
```
makeblastdb -in <input_db> -parse_seqids -blastdb_version 5 -dbtype nucl -out <name_db>
blastn -db <name_db> -query <input_file> -out <output_file> -max_hsps 1 -max_target_seqs 1 -num_threads 4 -evalue 1e-5
```

SignalP

SignalP 5.0 is a deep neural network-based method combined with conditional random field classification and optimized transfer learning for improved SP prediction. [4]

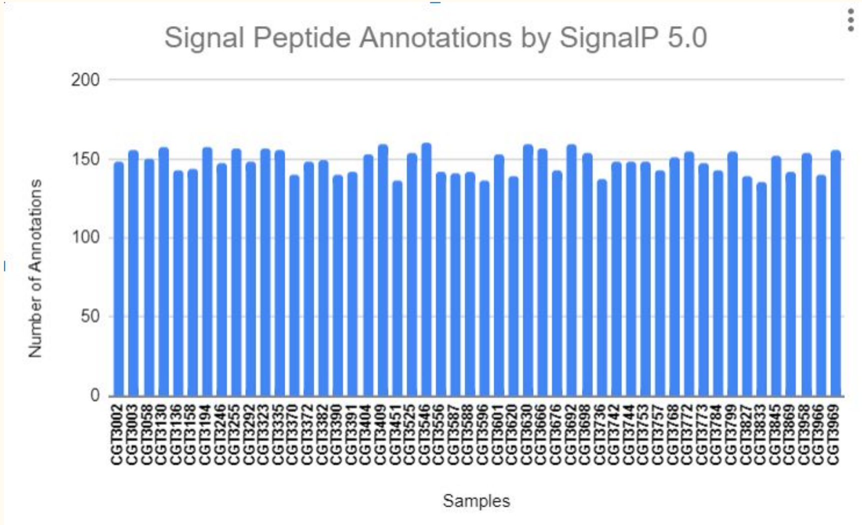
Characterizes between signaling peptides and lipoproteins

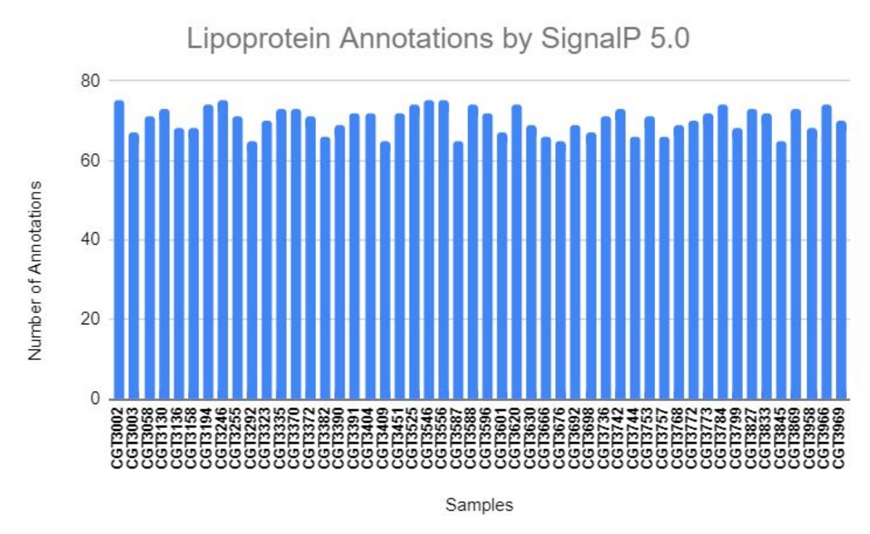
Outputs probability of predictions and position of protein in the sequence



```
# SignalP-5.0   Organism: gram+ Timestamp: 20200327175016
# ID     Prediction      SP(Sec/SPI)     TAT(Tat/SPI)     LIPO(Sec/SPII)  OTHER   CS Position
NODE_1_length_757168_cov_36.966159:121535-121643        SP(Sec/SPI)     0.492041        0.245177        0.205478        0.057304
        CS pos: 12-13. AAA-TT. Pr: 0.0605
NODE_1_length_757168_cov_36.966159:125444-126320        SP(Sec/SPI)     0.550027        0.142682        0.284907        0.022383
        CS pos: 15-16. ATA-AT. Pr: 0.1476
NODE_1_length_757168_cov_36.966159:127250-128453        SP(Sec/SPI)     0.672480        0.174968        0.134644        0.017909
        CS pos: 20-21. AAA-AT. Pr: 0.2309
NODE_1_length_757168_cov_36.966159:140231-141215        LIPO(Sec/SPII)  0.276420        0.108397        0.571096        0.044086
        CS pos: 19-20. TAA-CA. Pr: 0.5393
NODE_1_length_757168_cov_36.966159:141538-143329        LIPO(Sec/SPII)  0.297369        0.095134        0.478493        0.129004
        CS pos: 18-19. GGG-CA. Pr: 0.2123
```

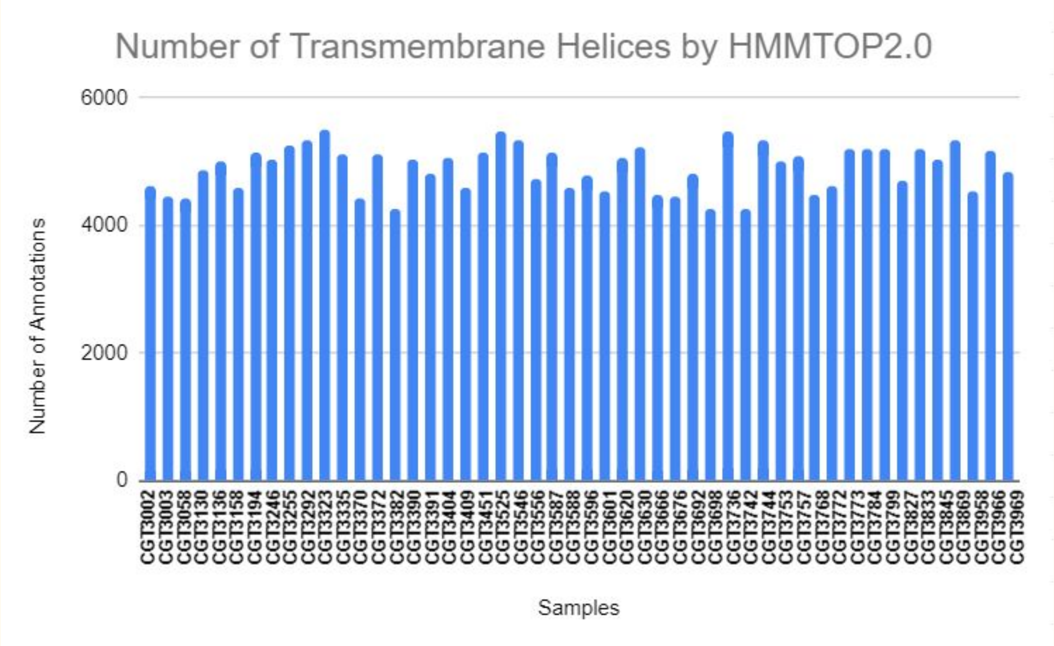
SignalP Results

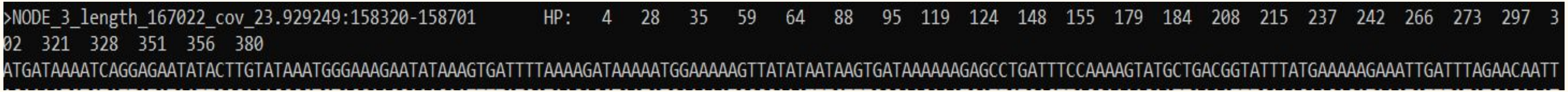




HMMTOP Results

The HMMTOP transmembrane topology prediction server predicts transmembrane proteins, transmembrane helices, and their start and end positions in the sequence.

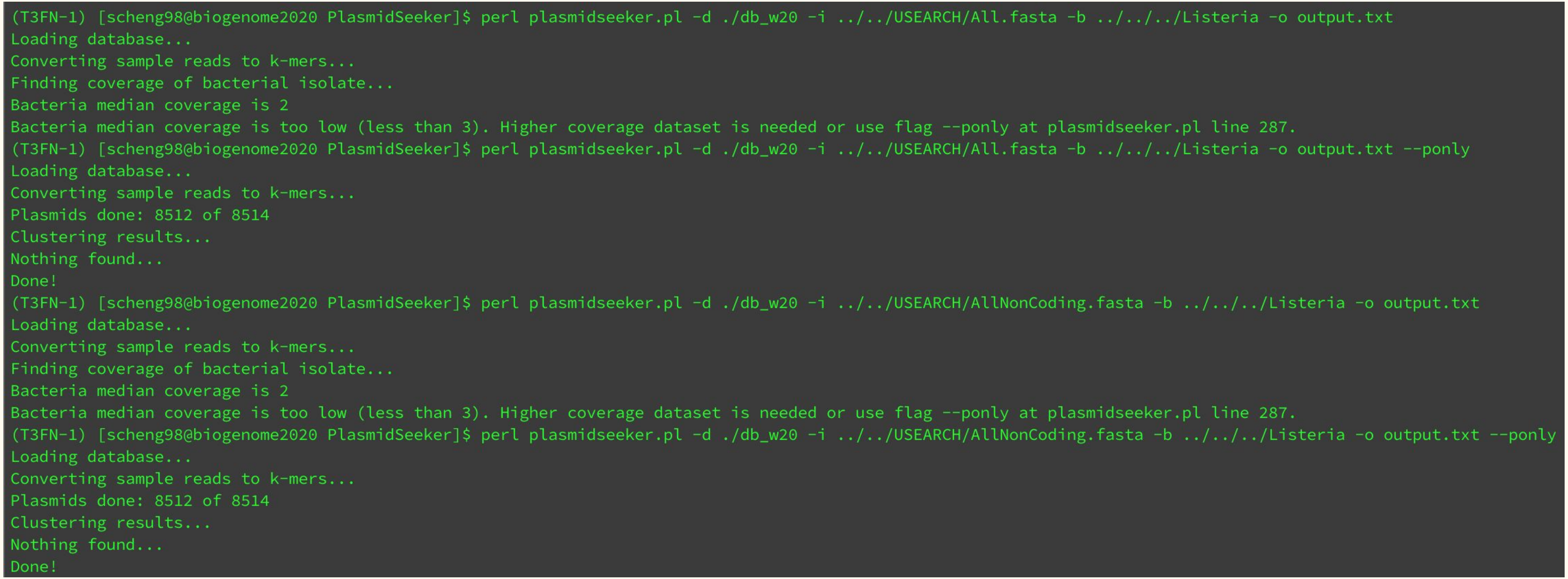




```
>NODE_3_length_167022_cov_23.929249:158320-158701       HP:   4   28   35   59   64   88   95  119  124  148  155  179  184  208  215  237  242  266  273  297  3
02  321  328  351  356  380
ATGATAAAATCAGGAGAATATACTTGTATAAATGGGAAAGAATATAAAGTGATTTTAAAAGATAAAAATGGAAAAAGTTATATAATAAGTGATAAAAAAGAGCCTGATTTCCAAAAGTATGCTGACGGTATTTATGAAAAAGAAATTGATTTAGAACAATT
```

PlasmidSeeker

Command Line: perl plasmidseeker.pl -d ./db_w20 -i <input.fasta> -b <closest species> -o <output>



```
(T3FN-1) [scheng98@biogenome2020 PlasmidSeeker]$ perl plasmidseeker.pl -d ./db_w20 -i ../../USEARCH/All.fasta -b ../../../Listeria -o output.txt
Loading database...
Converting sample reads to k-mers...
Finding coverage of bacterial isolate...
Bacteria median coverage is 2
Bacteria median coverage is too low (less than 3). Higher coverage dataset is needed or use flag --ponly at plasmidseeker.pl line 287.
(T3FN-1) [scheng98@biogenome2020 PlasmidSeeker]$ perl plasmidseeker.pl -d ./db_w20 -i ../../USEARCH/All.fasta -b ../../../Listeria -o output.txt --ponly
Loading database...
Converting sample reads to k-mers...
Plasmids done: 8512 of 8514
Clustering results...
Nothing found...
Done!
(T3FN-1) [scheng98@biogenome2020 PlasmidSeeker]$ perl plasmidseeker.pl -d ./db_w20 -i ../../USEARCH/AllNonCoding.fasta -b ../../../Listeria -o output.txt
Loading database...
Converting sample reads to k-mers...
Finding coverage of bacterial isolate...
Bacteria median coverage is 2
Bacteria median coverage is too low (less than 3). Higher coverage dataset is needed or use flag --ponly at plasmidseeker.pl line 287.
(T3FN-1) [scheng98@biogenome2020 PlasmidSeeker]$ perl plasmidseeker.pl -d ./db_w20 -i ../../USEARCH/AllNonCoding.fasta -b ../../../Listeria -o output.txt --ponly
Loading database...
Converting sample reads to k-mers...
Plasmids done: 8512 of 8514
Clustering results...
Nothing found...
Done!
```

Phaster

Required complete genome sequence to run online

IslandViewer

Required gene bank format file or embl format file to search in the database

References

1. Huerta-Cepas, Jaime, et al. "Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper." Molecular biology and evolution 34.8 (2017): 2115-2122.
2. Edgar, Robert C. "Search and clustering orders of magnitude faster than BLAST." Bioinformatics 26.19 (2010): 2460-2461.
3. Buchfink, Benjamin, Chao Xie, and Daniel H. Huson. "Fast and sensitive protein alignment using DIAMOND." Nature methods 12.1 (2015): 59.
4. Armenteros, José Juan Almagro, et al. “SignalP 5.0 Improves Signal Peptide Predictions Using Deep Neural Networks.” Nature News, Nature Publishing Group, 18 Feb. 2019, www.nature.com/articles/s41587-019-0036-z.
5. Simon. “HMMTOP Transmembrane Topology Prediction Server.” OUP Academic, Oxford University Press, 1 Sept. 2001, academic.oup.com/bioinformatics/article/17/9/849/206573.
6. Barrangou R. The roles of CRISPR-Cas systems in adaptive immunity and beyond. Curr Opin Immunol. 2015;32:36–41. doi:10.1016/j.coi.2014.12.008
7. Zhang, Jiayu, et al. "The CRISPR-Cas9 system: a promising tool for discovering potential approaches to overcome drug resistance in cancer." RSC advances 8.58 (2018): 33464-33472.
8. Edgar, Robert C. "PILER-CR: fast and accurate identification of CRISPR repeats." BMC bioinformatics 8.1 (2007): 18.
9. Bland, Charles, et al. "CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats." BMC bioinformatics 8.1 (2007): 209.
10. Arango-Argoty, Gustavo, et al. "DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data." Microbiome 6.1 (2018): 1-15.
11. Alcock, Brian P., et al. "CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database." Nucleic acids research 48.D1 (2020): D517-D525.