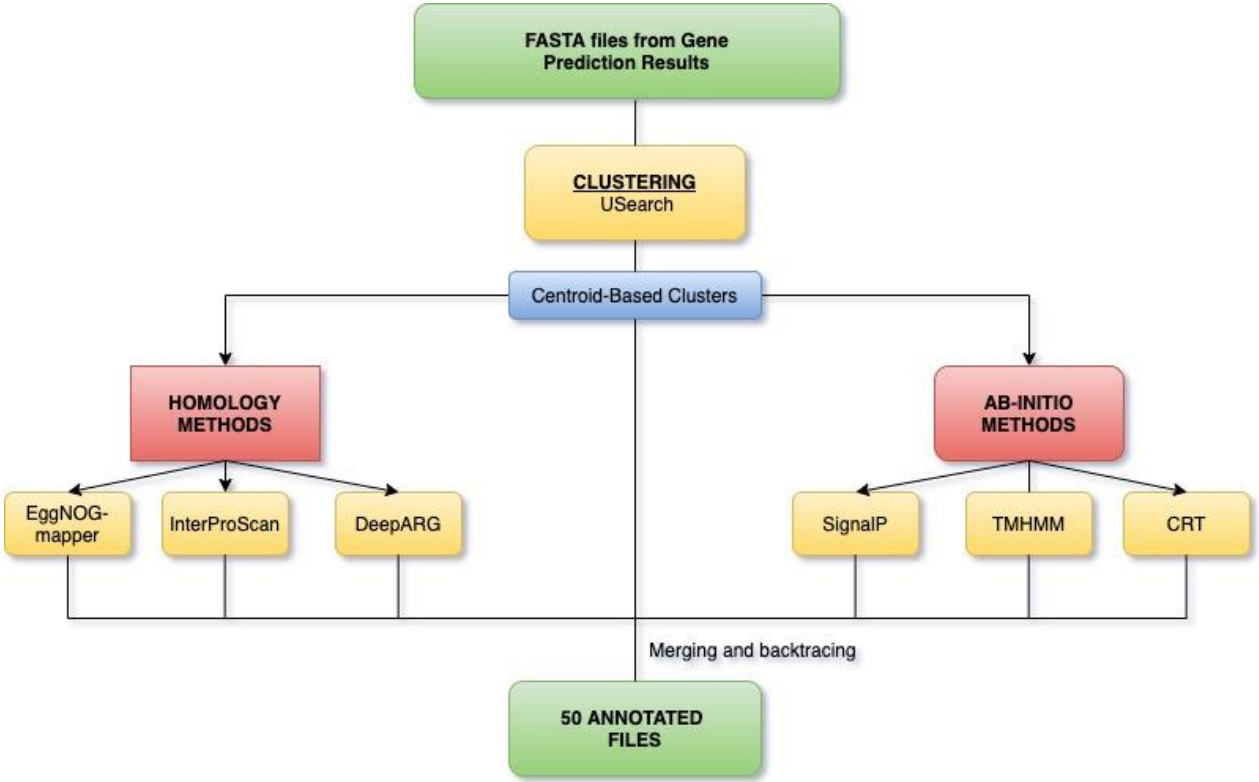
TEAM 1
FUNCTIONAL ANNOTATION RESULTS

Kenji Gerhardt, Manasa Vegesna, Shuheng Gan,
Hyeonjeong Cheon, Maria Ahmad

FINAL PIPELINE



CLUSTERING

- We utilized USearch to cluster our amino acid sequences
  - Identity threshold of 70% similarity
  - cluster_fast command
- 70% identity in amino acid sequences is a very conservative threshold
  - NCBI guidelines indicate that 40% AA identity is very likely to result in shared function, although outliers exist
  - 70% formed a good tradeoff between sureness of clusters and reduction of sequences to annotate
- The command expects uniquely named sequences in a single FASTA format file
  - We renamed the sequences from the gene prediction group to achieve this result, as each file contained the same ordered prodigal outputs
    - e.g. prodigal_sequence_1, prodigal_sequence_2…
- The files were concatenated
  - Since all of the organisms were e. coli, most genes should be shared by at least their immediate relatives.
  - No sense in annotating shared genes multiple times
- Once concatenated, there were 231894 protein sequences
- Clustering reduced this to 7361 cluster representatives - a 97% reduction in the sequences we had to annotate

AB-INITIO APPROACH

Ab-Initio Tools predict and annotate different regions of the prokaryotic genome using:

- Sequence composition
- Likelihoods within the gene models
- Gene content
- Signal detection

We tested out various tools for determining the following features of the prokaryotic genome:

- Signal Peptides
- Transmembrane Proteins
- CRISPR Sites

SIGNAL PEPTIDE PREDICTION

Ab-Initio tools take advantage of the signal peptide structure, which contains positively charged N-region, followed by a hydrophobic H-region and a neutral but polar C-region, to predict their presence in the given protein sequences.

Tools we tested: SignalP, LipoP, TatP ---> **Selected Tool: SignalP 5.0**

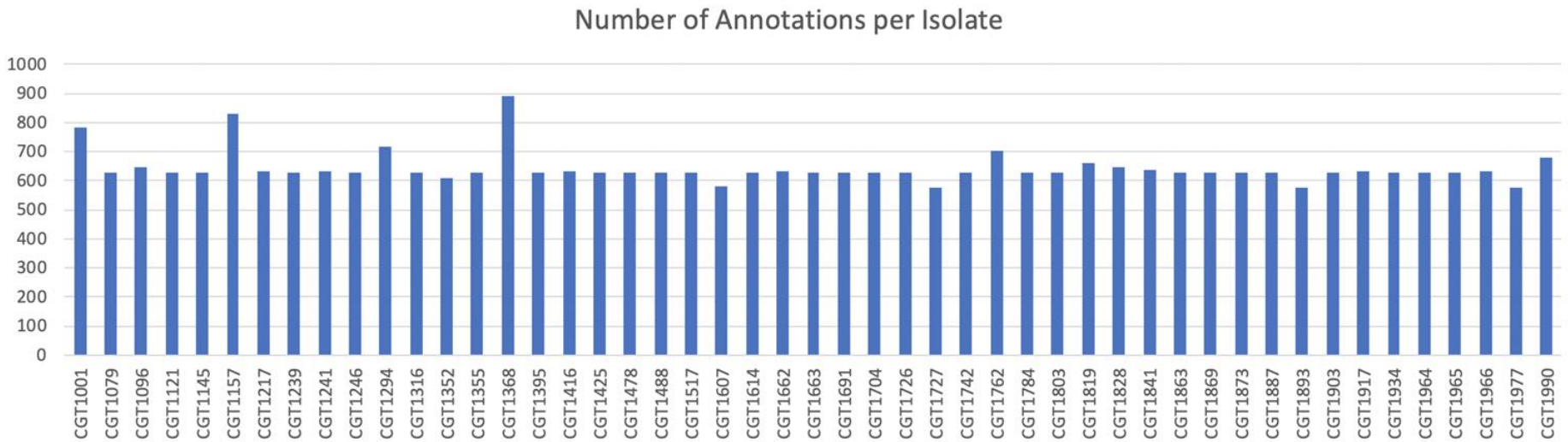
- Has known to perform well on gram-negative bacterial proteins
- Based on deep convolutional and recurrent neural networks
- Predicts all three types of signal peptides: Sec signal peptide, Lipoprotein and Tat signal peptide
- Relatively fast and provides relevant information for us

SignalP 5.0 Output

```
##gff-version 3
CGT1001_trim_assembled_protein.faa;Prodigal_3929        SignalP-5.0     signal_peptide  1       33      0.999763        .       .       Note=TAT
CGT1990_trim_assembled_protein.faa;Prodigal_4807        SignalP-5.0     signal_peptide  1       27      0.427436        .       .       .
CGT1990_trim_assembled_protein.faa;Prodigal_1607        SignalP-5.0     signal_peptide  1       23      0.998455        .       .       .
CGT1352_trim_assembled_protein.faa;Prodigal_4667        SignalP-5.0     signal_peptide  1       23      0.898289        .       .       .
CGT1368_trim_assembled_protein.faa;Prodigal_1019        SignalP-5.0     signal_peptide  1       21      0.995477        .       .       .
CGT1001_trim_assembled_protein.faa;Prodigal_1568        SignalP-5.0     signal_peptide  1       32      0.921151        .       .       .
CGT1368_trim_assembled_protein.faa;Prodigal_1945        SignalP-5.0     signal_peptide  1       42      0.618004        .       .       .
CGT1001_trim_assembled_protein.faa;Prodigal_3398        SignalP-5.0     signal_peptide  1       21      0.548601        .       .       .
CGT1368_trim_assembled_protein.faa;Prodigal_2183        SignalP-5.0     signal_peptide  1       45      0.990578        .       .       Note=TAT
CGT1368_trim_assembled_protein.faa;Prodigal_2335        SignalP-5.0     signal_peptide  1       28      0.975166        .       .       .
CGT1001_trim_assembled_protein.faa;Prodigal_4610        SignalP-5.0     signal_peptide  1       21      0.999160        .       .       .
CGT1001_trim_assembled_protein.faa;Prodigal_2396        SignalP-5.0     lipoprotein_signal_peptide      1       16      0.839056        .       .
CGT1294_trim_assembled_protein.faa;Prodigal_2856        SignalP-5.0     signal_peptide  1       22      0.993453        .       .       .
CGT1001_trim_assembled_protein.faa;Prodigal_1186        SignalP-5.0     signal_peptide  1       27      0.998516        .       .       Note=TAT
```

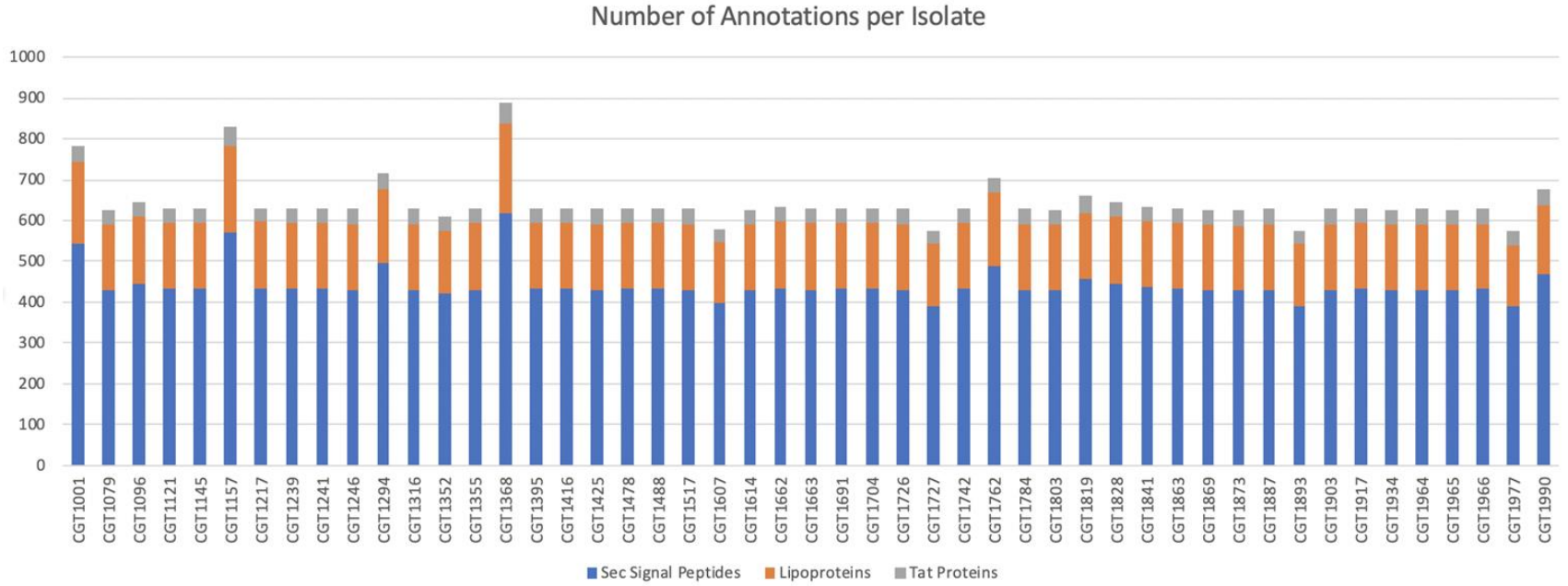
signalp -fasta <input file path> -org gram- -format -short -prefix <output_file_path> -gff3

SignalP Final Result



Average Number of Annotations per isolate: **642.52 signal peptides**

SignalP - Prediction of Different Signal Peptides



Sec Signal Peptides: **441.23** , Lipoproteins: **164.27** , Tat Proteins: **37.02**

TRANSMEMBRANE PROTEIN PREDICTION

Transmembrane proteins contain crucial components for cell-cell signaling, mediate the transport of ions and solutes across the membrane. Transmembrane helices are a basic type of transmembrane proteins

Tools we tested: TMHMM, HMMTop, Phobius ---> **Tool Selected: TMHMM**

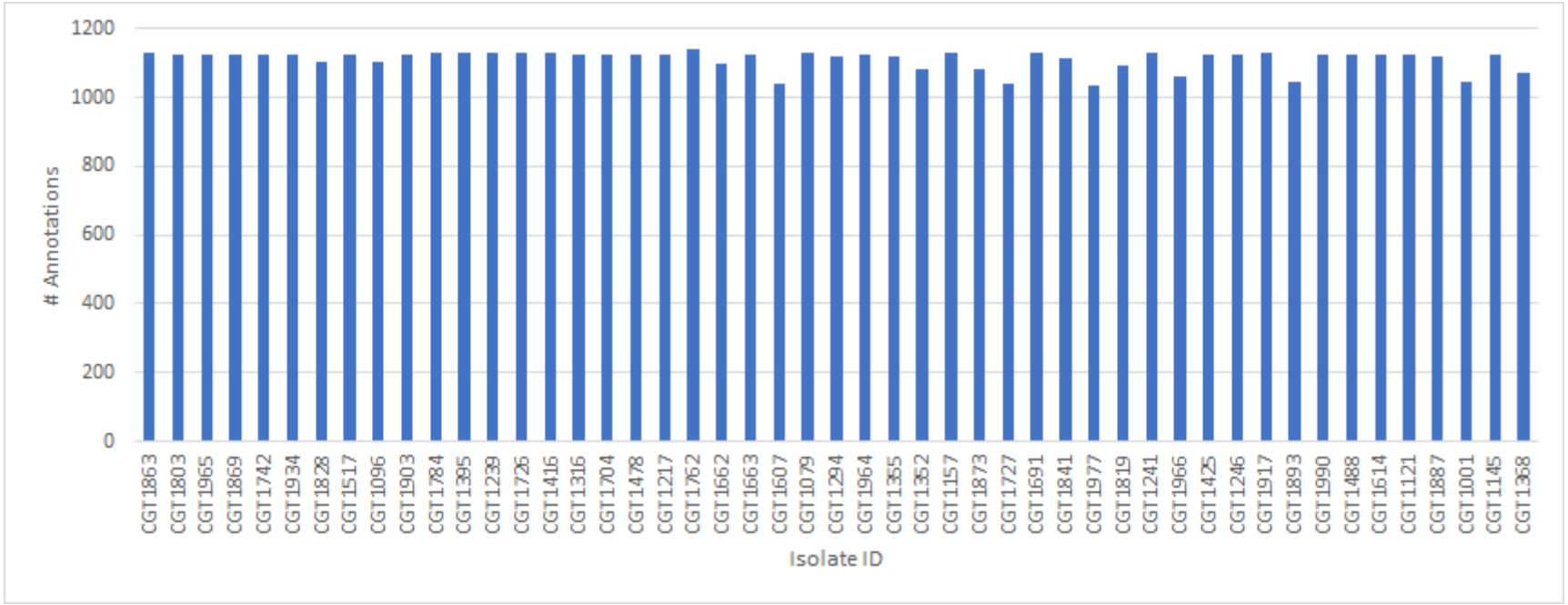
- Transmembrane Topology Prediction tool
- Based on Hidden Markov Models
- Easy to use and fast
- Memory-Efficient

TMHMM Output

CGT1001_trim_assembled_protein.faa;Prodigal_4116	TMHMM2.0	outside	1	124
CGT1001_trim_assembled_protein.faa;Prodigal_2519	TMHMM2.0	outside	1	80
CGT1990_trim_assembled_protein.faa;Prodigal_4299	TMHMM2.0	outside	1	68
CGT1001_trim_assembled_protein.faa;Prodigal_4031	TMHMM2.0	outside	1	163
CGT1662_trim_assembled_protein.faa;Prodigal_446	TMHMM2.0	outside	1	455
CGT1001_trim_assembled_protein.faa;Prodigal_1388	TMHMM2.0	outside	1	9
CGT1001_trim_assembled_protein.faa;Prodigal_1388	TMHMM2.0	TMhelix	10	32
CGT1001_trim_assembled_protein.faa;Prodigal_1388	TMHMM2.0	inside	33	35
CGT1368_trim_assembled_protein.faa;Prodigal_3747	TMHMM2.0	inside	1	8
CGT1368_trim_assembled_protein.faa;Prodigal_3747	TMHMM2.0	TMhelix	9	31
CGT1368_trim_assembled_protein.faa;Prodigal_3747	TMHMM2.0	outside	32	336
CGT1368_trim_assembled_protein.faa;Prodigal_3747	TMHMM2.0	TMhelix	337	359
CGT1368_trim_assembled_protein.faa;Prodigal_3747	TMHMM2.0	inside	360	365
CGT1368_trim_assembled_protein.faa;Prodigal_3747	TMHMM2.0	TMhelix	366	388
CGT1368_trim_assembled_protein.faa;Prodigal_3747	TMHMM2.0	outside	389	391

cat <input file path> | tmhmm > <output file path>

TMHMM Helices Annotation Count per Isolate



Average count: **1110.388 helices annotations**

CRISPR

CRISPR is a family of DNA sequences found in prokaryotic organisms. These sequences are derived from the DNA fragments of viruses which previously infected the organism. They can be used in the **immune response** of the cell against future infections, by detecting and destroying DNA from similar viruses.

Cas9 is the enzyme which uses the CRISPR sequences to recognize and cleave strands of DNA complementary to the CRISPR site.

CRISPR-Cas9 complex can be used to **edit the genes** within an organism.

CRISPR PREDICTION

PilerCR
- Fast
- Easy to download & implement

  pilercr -in <input_file> -out <output_file> -minrepeat <N> -minspacer <N> -minrepeatratio <N>

CRISPR Recognition Tool (CRT)
- Fast
- Requires more dependencies than PilerCR
- **CRT predicts more genes but PilerCR has higher precision, therefore we went with PilerCR as to reduce potential false positives**

  java -cp CRT1.2-CLI.jar crt [options] inputFile outputFile

Neither PilerCR nor CRT predicted CRISPRs in the *E. coli* genome. The following were the three conditions used for PilerCR:

Minimum repeat: 16, minimum spacer: 8, minimum repeat ratio: 0.9
Minimum repeat: 14, minimum spacer: 4, minimum repeat ratio: 0.9
Minimum repeat: 6, minimum spacer: 3, minimum repeat ratio: 0.8

None of the parameter cases found CRISPR repeats.
CRISPR are found in between 40 - 50% of sequenced bacterial genomes

HOMOLOGY APPROACH AND DATABASES

- Homologous genes that have recently diverged usually share function
    - By finding homologous genes, we're looking to transfer annotation on known genes to our predicted genes.
- When we search a gene against a database, the search is looking for homology between our gene sequences and those in the database to determine what our genes' function will be
- Need specific and quality databases which limit search size
- Want to especially look for **antibiotic resistance genes (ARGs)** which will be most useful to the comparative genomics group

HOMOLOGY APPROACH AND DATABASES

- EggNOG-mapper
  - gammaproteobacteria-specific database
  - Command: `python emapper.py -i <input_file> --output <output_file> -m diamond -d bact -o <output directory>`
- Interproscan
  - Multiple databases which include db for protein motifs, domains, families, conserved domains, protein chemical capabilities
  - Command: `interproscan.sh -i <input_file_name> -dp -d <output_directory> -appl <databases_you_choose> -f <output_format> -t <sequence_type>`
- DeepARG
  - Includes the CARD and ARDB databases for antibiotic resistance genes (ARGs)
  - Command: `python ./deepARG.py --align --genes --type prot --input <gene-like_sequences_fasta_file> --out <output_file_name>`

HOMOLOGY RESULT - eggNOG-Mapper

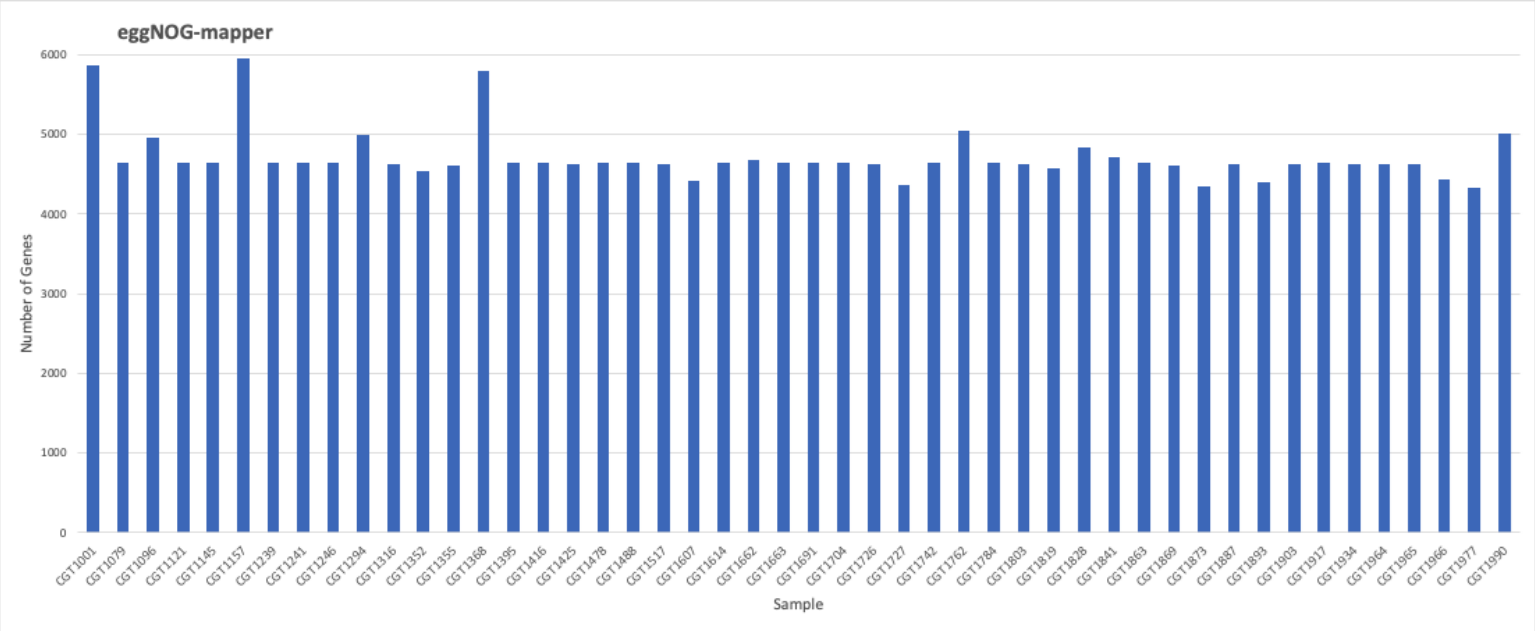
key	seed_eggNOG_ortholo	seed_ortholog_	seed_ortholog_	best_tax_level	Preferred _name	GOs	KEGG_ko	KEGG_Pathway	BRITE	COG.Functional .cat.	eggNOG.free	sample	seq_in_samp
CGT1001_t	316407.8568	2.00E-113	415.2	Escherichia	eutQ		ko:K04019,kc	ko00564,ko011(	ko00000,ko0	E	ethanolamin	CGT1001_tri	Prodigal_3
CGT1001_t	469008.B21_02306	7.50E-214	749.6	Escherichia	eutG	GO:0003674,G(	ko:K04022	ko00010,ko011(	ko00000,ko0	C	Ethanolamin	CGT1001_tri	Prodigal_10
CGT1001_t	198214.SF2473	2.80E-154	551.2	Gammaproteo bacteria	pdxK	GO:0003674,G(	ko:K00868	ko00750,ko011(	ko00000,ko0	H	Pyridoxal kin	CGT1001_tri	Prodigal_35
CGT1001_t	1028307.EAE_00365	1.90E-37	161.4	Enterobacter	ptsH	GO:0003674,G(	ko:K02768,kc	ko00051,ko005(	ko00000,ko0	G	PTS HPr com	CGT1001_tri	Prodigal_38
CGT1001_t	155864.EDL933_3558	1.40E-223	781.9	Escherichia	yfeO	GO:0005575,G(	ko:K03281		ko00000	P	ion-transport	CGT1001_tri	Prodigal_57
CGT1001_t	198214.SF2431	3.00E-181	641	Gammaproteo bacteria	cscR		ko:K02529		ko00000,ko0	K	Transcriptior	CGT1001_tri	Prodigal_82
CGT1001_t	198214.SF2430	5.20E-294	1016.1	Gammaproteo bacteria	cscA	GO:0005575,G(	ko:K01193	ko00052,ko005(	ko00000,ko0	G	invertase	CGT1001_tri	Prodigal_83
CGT1001_t	155864.EDL933_2561	2.80E-246	857.4	Escherichia						T	PhoQ Sensor	CGT1001_tri	Prodigal_97
CGT1001_t	316407.1743	6.60E-187	659.8	Escherichia	add	GO:0003674,G(	ko:K01488	ko00230,ko011(	ko00000,ko0	F	Belongs to th	CGT1001_tri	Prodigal_111
CGT1001_t	155864.EDL933_2587	6.30E-120	436.8	Escherichia	rnfE	GO:0003674,G(	ko:K02560,kc	ko00540,ko011(	ko00000,ko0	C	Part of a me	CGT1001_tri	Prodigal_120
CGT1001_t	155864.EDL933_2599	5.00E-37	159.8	Escherichia	ydhI					S	Protein of un	CGT1001_tri	Prodigal_132
CGT1001_t	316407.8568	8.40E-40	169.1	Escherichia	ydhL		ko:K06938		ko00000	S	Protein of un	CGT1001_tri	Prodigal_137
CGT1001_t	316407.8568	7.20E-107	393.7	Escherichia	ydhO	GO:0000270,G(	ko:K01183,kc	ko00520,ko011(	ko00000,ko0	M	A murein DD	CGT1001_tri	Prodigal_144
CGT1001_t	155864.EDL933_2620	5.30E-74	283.5	Escherichia	ribE	GO:0003674,G(	ko:K00793	ko00740,ko011(	ko00000,ko0	H	Riboflavin sy	CGT1001_tri	Prodigal_152
CGT1001_t	155864.EDL933_4422	1.30E-134	485.7	Escherichia	mlaE	GO:0005575,G(	ko:K02066	ko02010,map02	ko00000,ko0	Q	Part of the A	CGT1001_tri	Prodigal_190
CGT1001_t	155864.EDL933_4418	1.40E-43	181.8	Escherichia	yrbA	GO:0003674,G(	ko:K07390		ko00000,ko0	K	Belongs to th	CGT1001_tri	Prodigal_194
CGT1001_t	1440052.EAKF1_ch275	7.10E-178	629.8	Escherichia	ispB	GO:0003674,G(	ko:K00805,kc	ko00900,ko011	ko00000,ko0	H	Polyprenyl sy	CGT1001_tri	Prodigal_197
CGT1001_t	1440052.EAKF1_ch275	4.00E-40	170.2	Escherichia	rpmA	GO:0000027,G(	ko:K02899	ko03010,map03	br01610,ko0(	J	Ribosomal L:	CGT1001_tri	Prodigal_199
CGT1001_t	316407.8568	2.40E-254	884.4	Escherichia	dacB	GO:0000003,G(	ko:K07259	ko00550,map00	ko00000,ko0	M	D-alanyl-D-a	CGT1001_tri	Prodigal_202
CGT1001_t	1440052.EAKF1_ch277	2.40E-59	234.6	Escherichia	deaD	GO:0000027,G(	ko:K05591,kc	ko03018,map03	ko00000,ko0	F	DEAD-box RN	CGT1001_tri	Prodigal_220
CGT1001_t	481805.EcolC_0537	9.50E-228	795.8	Escherichia	mtr	GO:0003333,G(	ko:K03834,ko:K03835,ko:K03(		ko00000,ko0	E	Tryptophan-s	CGT1001_tri	Prodigal_222
CGT1001_t	198214.SF3196	5.20E-47	193.4	Gammaproteo bacteria	yhbQ	GO:0003674,G(	ko:K07461		ko00000	L	endonucleas	CGT1001_tri	Prodigal_228
CGT1001_t	316407.8568	2.40E-127	461.5	Escherichia	yraI	GO:0003674,G(	ko:K07346,ko:K07353,ko:K15(		ko00000,ko0	M	Part of the y	CGT1001_tri	Prodigal_240
CGT1001_t	155864.EDL933_3688	0	1163.7	Escherichia	hscA	GO:0000166,G(	ko:K04043,kc	ko03018,ko0421	ko00000,ko0	O	Chaperone ir	CGT1001_tri	Prodigal_264
CGT1001_t	1440052.EAKF1_ch347	3.30E-55	220.7	Escherichia	iscA	GO:0003674,G(	ko:K05997,ko:K13628		ko00000,ko0	S	Is able to tra	CGT1001_tri	Prodigal_266
CGT1001_t	155864.EDL933_3694	4.60E-85	320.5	Escherichia	iscR	GO:0003674,G(	ko:K04487,kc	ko00730,ko011(	ko00000,ko0	K	Regulates th	CGT1001_tri	Prodigal_269

Category	Clusters of Orthologous Groups of proteins (COGs)
J	translation, including ribosome structure and biogenesis
L	replication, recombination and repair
K	transcription
O	molecular chaperones and related functions
M	cell wall structure and biogenesis and outer membrane
N	secretion, motility and chemotaxis
T	signal transduction
P	inorganic ion transport and metabolism
C	energy production and conversion
G	carbohydrate metabolism and transport
E	amino acid metabolism and transport
F	nucleotide metabolism and transport
H	coenzyme metabolism
I	lipid metabolism
D	cell division and chromosome partitioning
R	general functional prediction only; S, no functional prediction

HOMOLOGY RESULT - eggNOG-Mapper

Average: 4717.13 / Maximum: 5956 / minimum: 4340

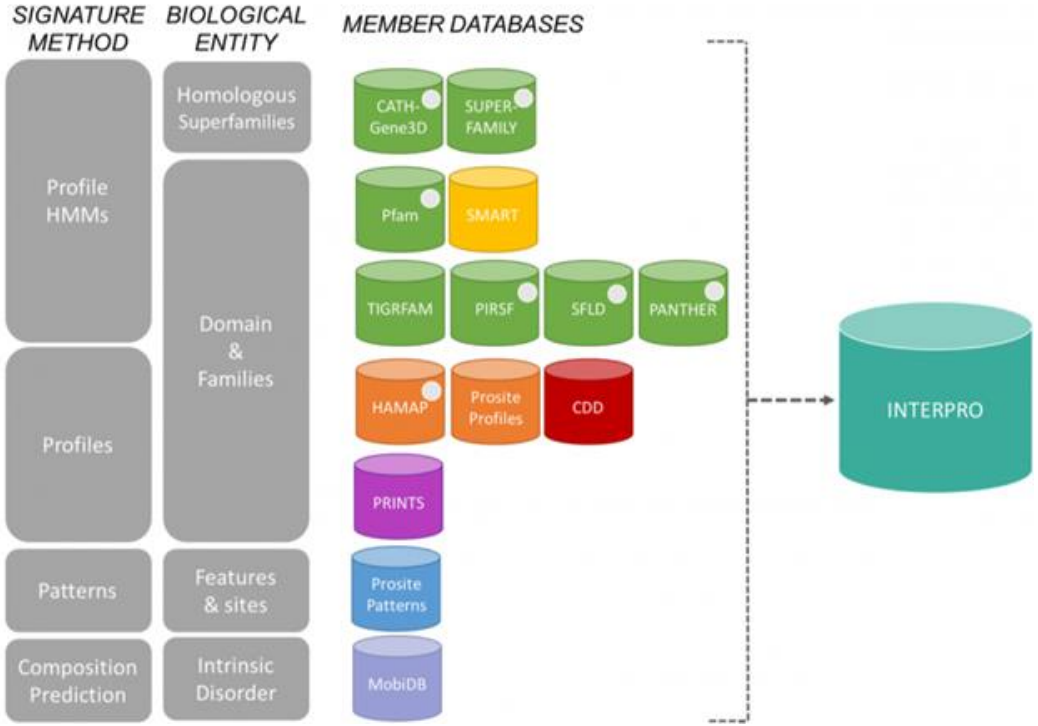
Runtime: ~6 h



HOMOLOGY RESULT - Interproscan

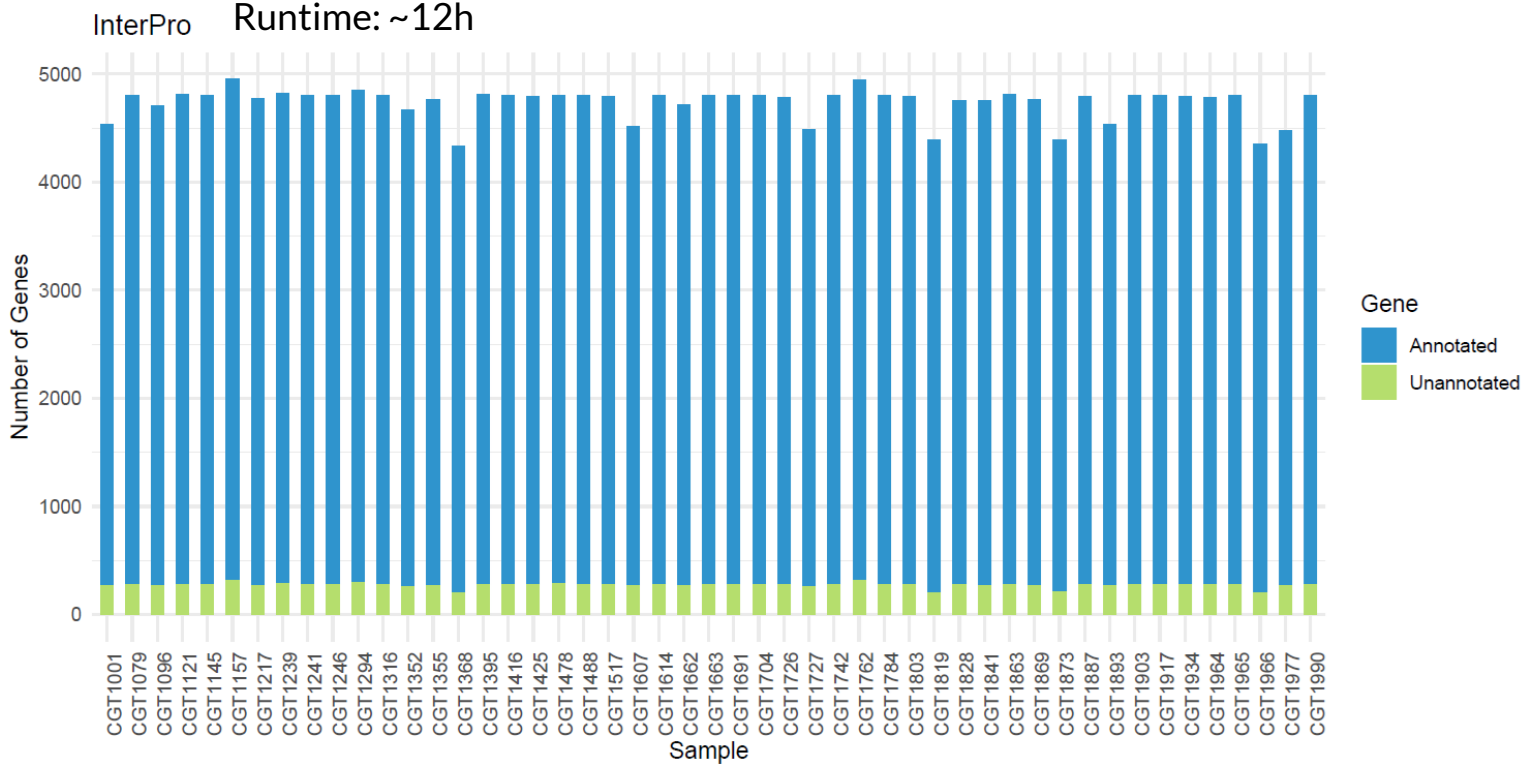
- InterProScan allows sequences to be scanned against protein signatures from 14 databases.

- Signatures are predictive models constructed from multiple sequence alignments that can be used to classify proteins.
  - patterns
  - profiles
  - fingerprints
  - hidden Markov models



HOMOLOGY RESULT - Interproscan

Average 94% (4452) genes get annotations in each sample
Runtime: ~12h



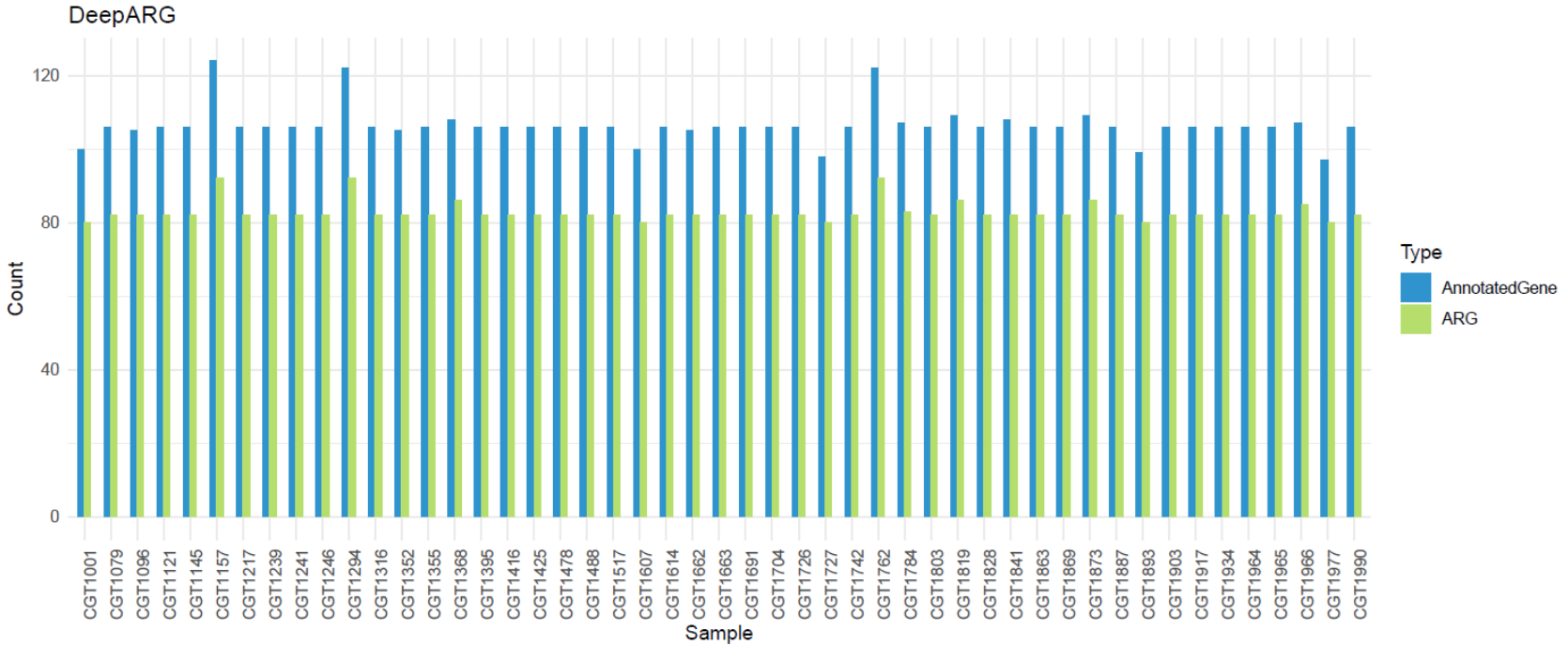
HOMOLOGY RESULT - DeepARG

- a deep learning tool that annotate antibiotic resistance genes in metagenomes.
- composed of two models for two types of input:
  - DeepARG-SS for short sequence reads from Next Generation Sequencing (NGS)
  - DeepARG-LS for long gene-like sequences from assembled samples.
- Databases: ARDB and CARD

Output:
ARG: prediction probability >= 0.8
Potential ARG: prediction probability < 0.8

HOMOLOGY RESULT - DeepARG

Average 106 genes are annotated as ARG and 83 kinds of ARG in each sample
Runtime: ~46s



Final Merging

- After all tools have run their course, the resulting outputs are propagated out to the other members of each cluster (aside from the representative).
- After which, the results are split into 1 file per sample and tool, containing the annotations of each gene in each sample as applied by each tool.
    - Contents of the annotations of each tool are quite disparate, so we left the data intact as much as possible
- We have a developed a script to pipeline the whole process
    - Still has a few tweaks to finalize it

QUESTIONS?

REFERENCES

- Clustering:
  - Joshi, T., & Xu, D. (2007). Quantitative assessment of relationship between sequence similarity and function similarity. *BMC genomics*, *8*(1), 222.
  - Zou, Q., Lin, G., Jiang, X., Liu, X., & Zeng, X. (2020). Sequence clustering in bioinformatics: an empirical study. *Briefings in bioinformatics*, *21*(1), 1-10.
  - Pearson W. R. (2013). An introduction to sequence similarity ("homology") searching. Current protocols in bioinformatics, Chapter 3, Unit3.1. https://doi.org/10.1002/0471250953.bi0301s42

REFERENCES

- Ab-Initio Methods:
  - Viklund, H., & Elofsson, A. (2004). Best α-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information. *Protein Science*, *13*(7), 1908-1917.
  - Romine, M. F. (2011). Genome-wide protein localization prediction strategies for gram negative bacteria. *BMC genomics*, *12*(S1), S1.
  - Armenteros, J. J. A., Tsirigos, K. D., Sønderby, C. K., Petersen, T. N., Winther, O., Brunak, S., ... & Nielsen, H. (2019). SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nature biotechnology*, *37*(4), 420-423.
  - Krogh, A., Larsson, B., Von Heijne, G., & Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of molecular biology*, *305*(3), 567-580.
  - Bland, C., Ramsey, T. L., Sabree, F., Lowe, M., Brown, K., Kyrpides, N. C., & Hugenholtz, P. (2007). CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC bioinformatics*, *8*, 209. https://doi.org/10.1186/1471-2105-8-209
  - Edgar, R.C. (2007) PILER-CR: fast and accurate identification of CRISPR repeats, *BMC Bioinformatics*, Jan 20;8:18.
  - Godde JS1, Bickerton A. J Mol Evol. 2006 Jun;62(6):718-29. Epub 2006 Apr 11. The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes.
  - C. Dı´ez-Villasenor, C. Almendros, J. Garcı´a-Martı´nez and F. J. M. Mojica. Diversity of CRISPR loci in Escherichia coli Departamento de Fisiologı´a, Genetica y Microbiologı´a, Facultad de Ciencias, Universidad de Alicante, E-03080, Spain.
  - Hille, F., & Charpentier, E. (2016). CRISPR-Cas: biology, mechanisms and relevance. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *371*(1707), 20150496. https://doi.org/10.1098/rstb.2015.0496
  - Barrangou, R., Doudna, J. Applications of CRISPR technologies in research and beyond. *Nat Biotechnol* 34, 933–941 (2016). https://doi.org/10.1038/nbt.3659

REFERENCES

- Homology:
  - Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068-2069.
  - Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+
  - Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature methods*, *12*(1), 59.
  - Finn, R. D., Attwood, T. K., Babbitt, P. C., Bateman, A., Bork, P., Bridge, A. J., ... & Gough, J. (2017). InterPro in 2017—beyond protein family and domain annotations. *Nucleic acids research*, *45*(D1), D190-D199.
  - Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., Von Mering, C., & Bork, P. (2017). Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Molecular biology and evolution*, *34*(8), 2115-2122.
  - Arango-Argoty, G., Garner, E., Pruden, A., Heath, L. S., Vikesland, P., & Zhang, L. (2018). DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome*, *6*(1), 1-15.