Functional Annotation

•••

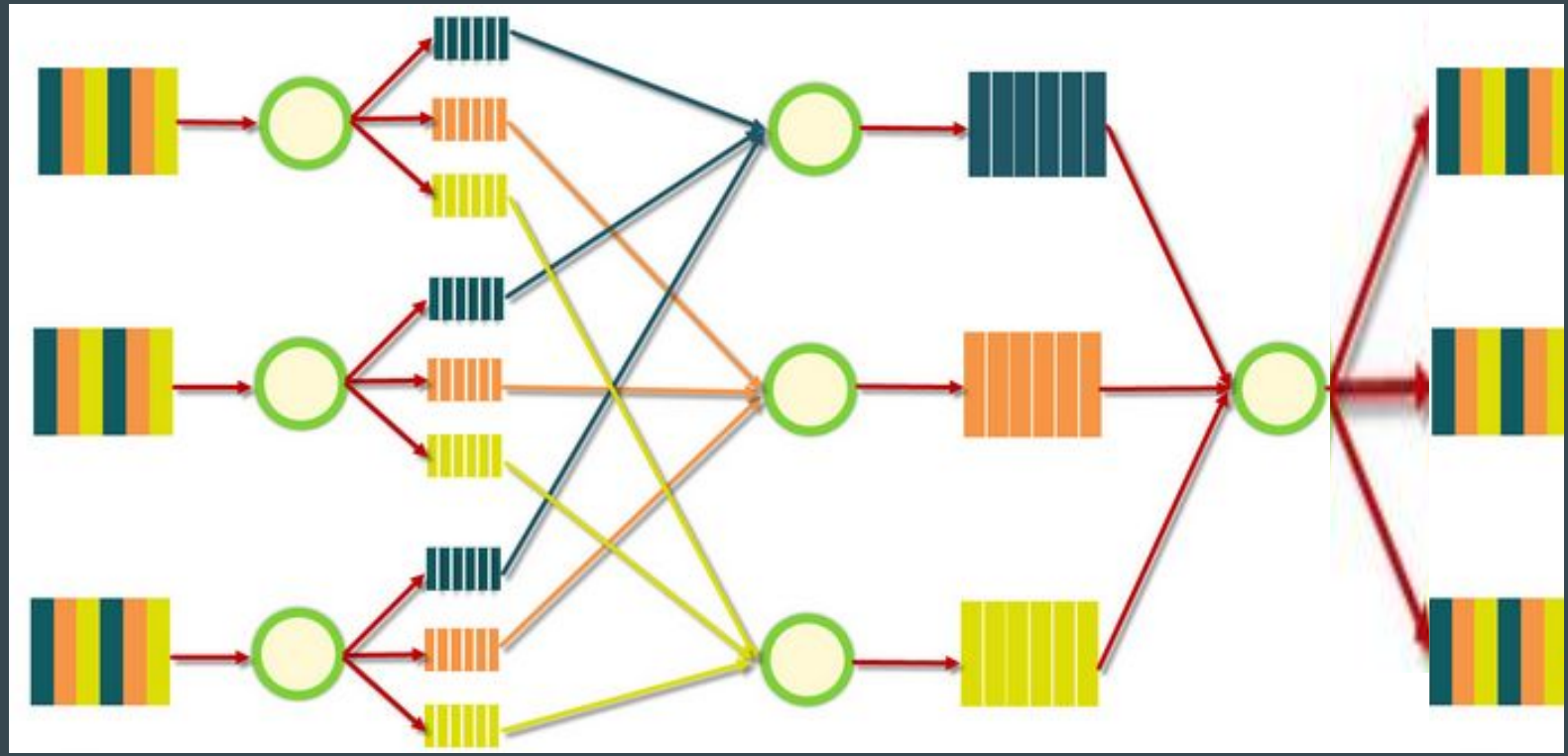
Final Results

Team 2:
Siu Lung Ng, Prachiti Prabhu, Brian Merritt, Rong Jin, Xinyu Wang, Jacob Boswell, Jiani Long, Pooja Khurana, Yinquan Lu, Shrey Mathur

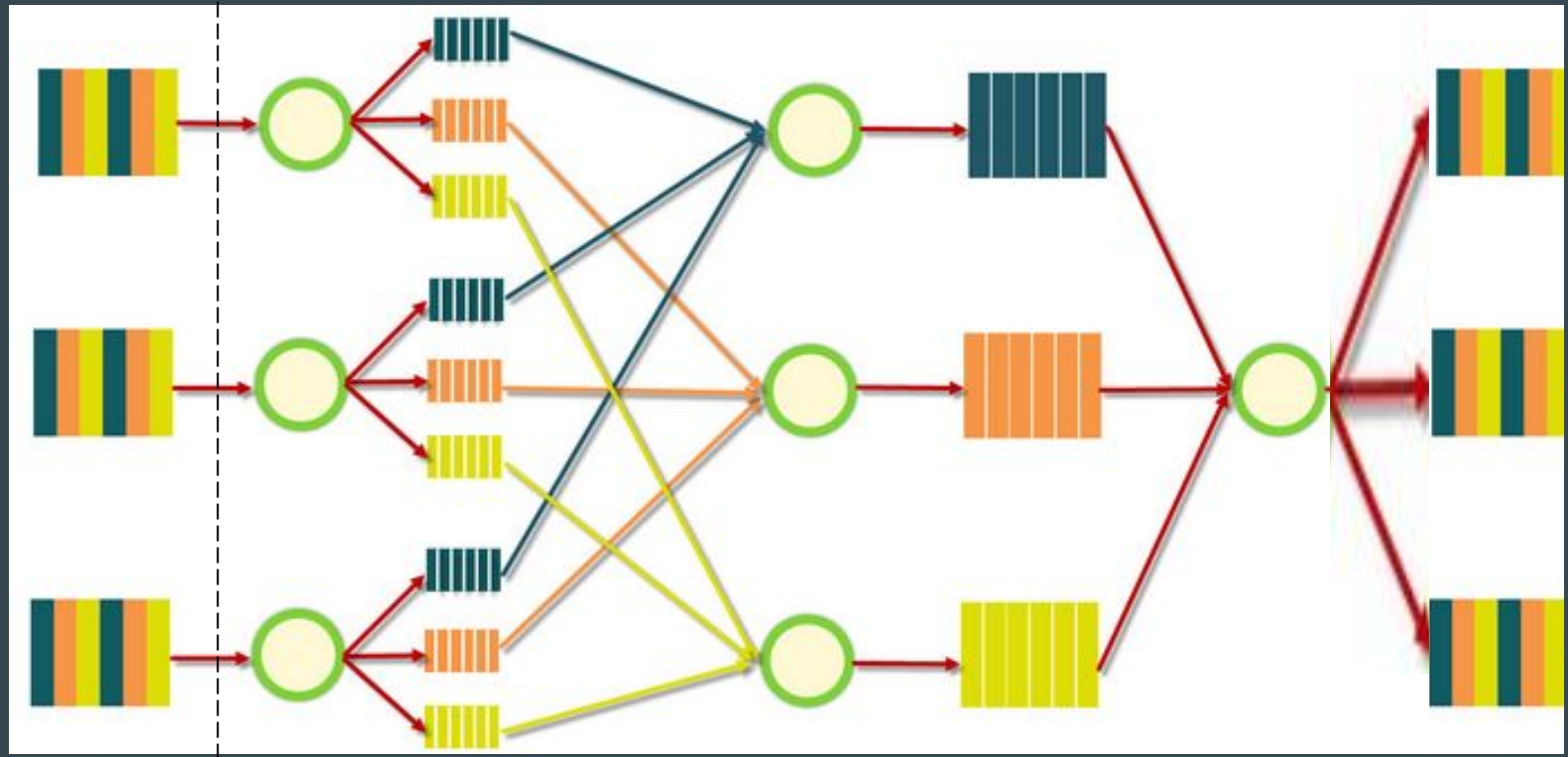
Introduction

- Objective
  - To functionally annotate 258 *Klebsiella* genomes
- Approach
  - SHOULD BE SCALABLE!!!
  - Reduce query size
- Deliverable
  - Wrapper script to run all the tools
  - One .gff file per genome

Approach

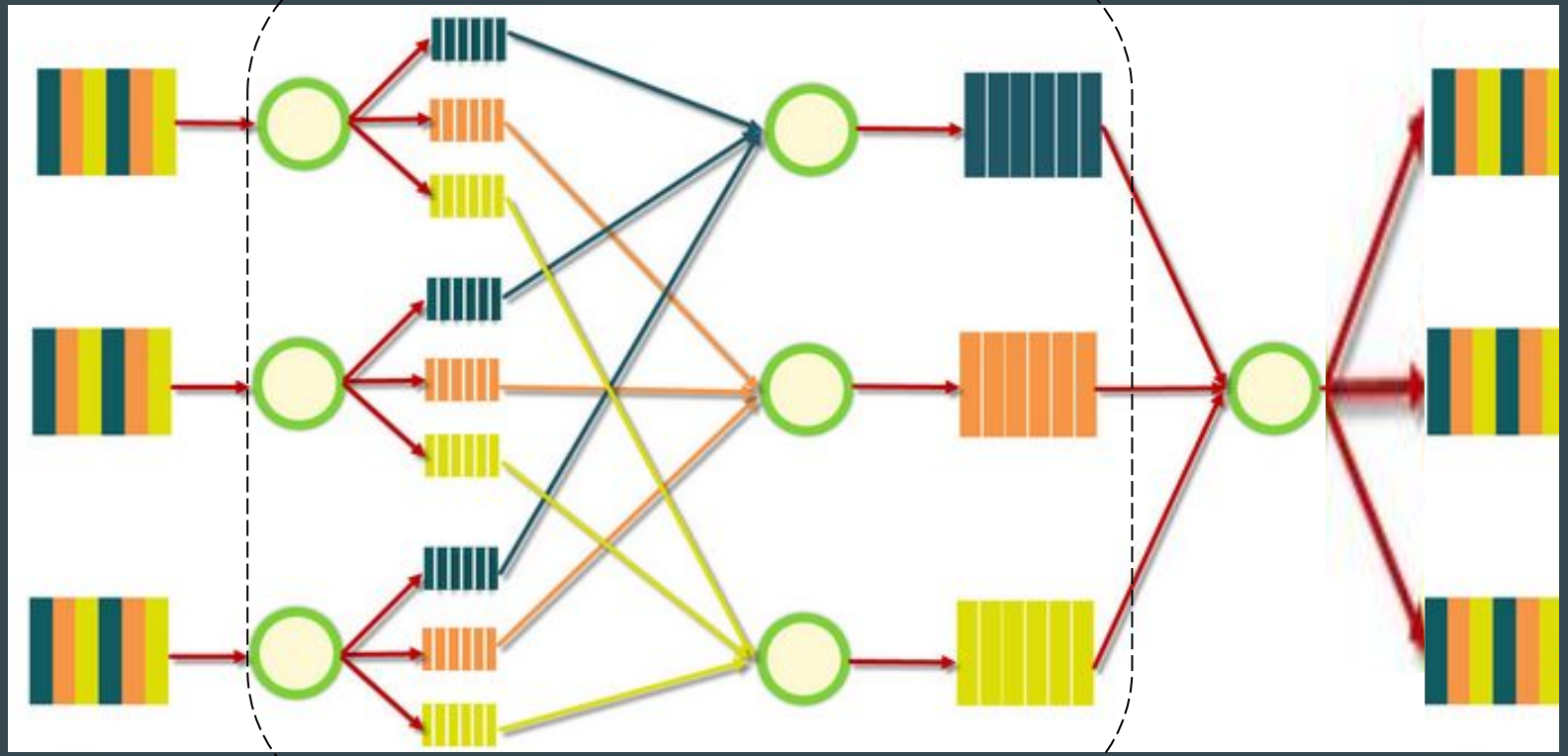


FASTA files

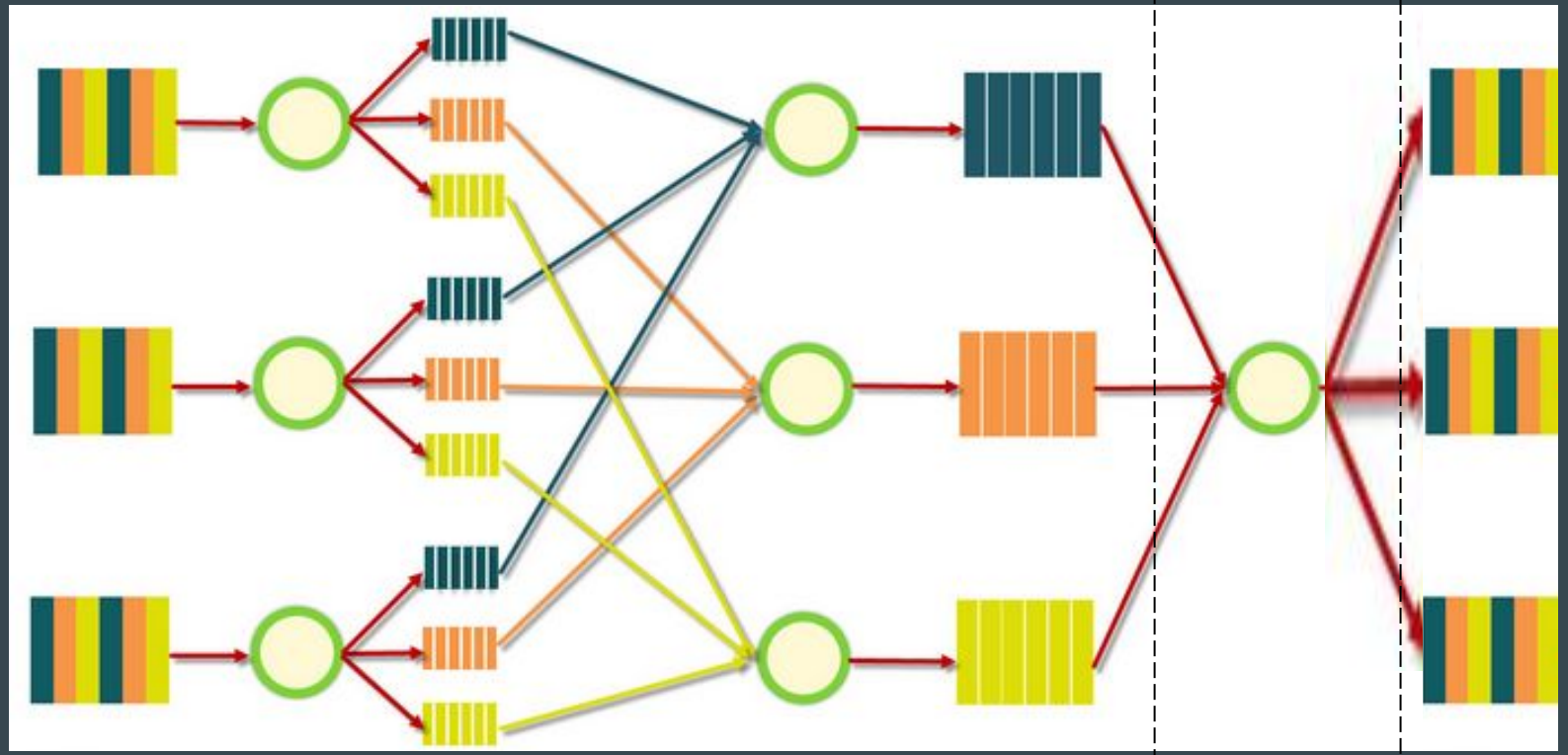


Input files (258)

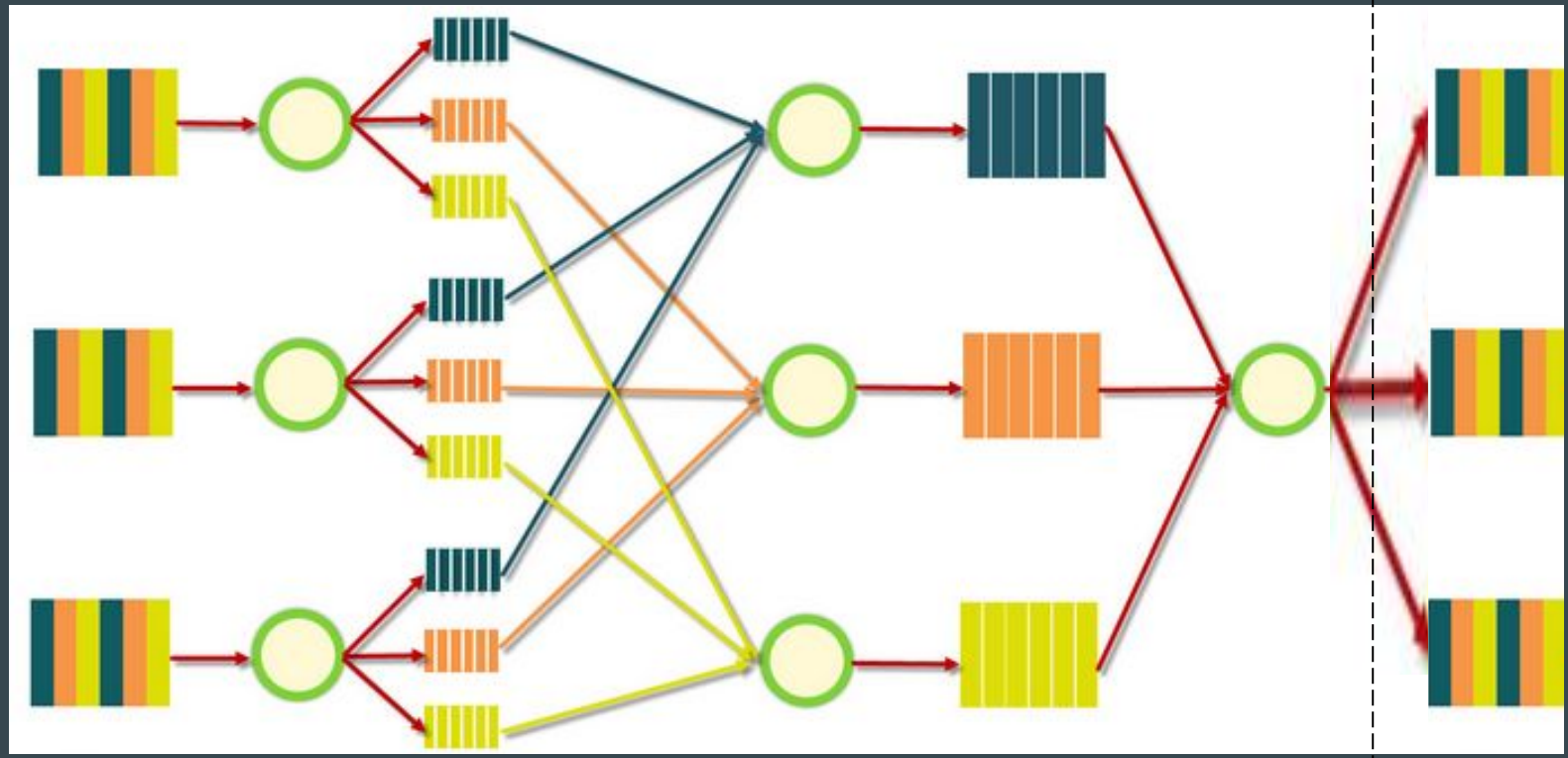
Cluster genes based on similarity



Analyze Representatives for all the Clusters

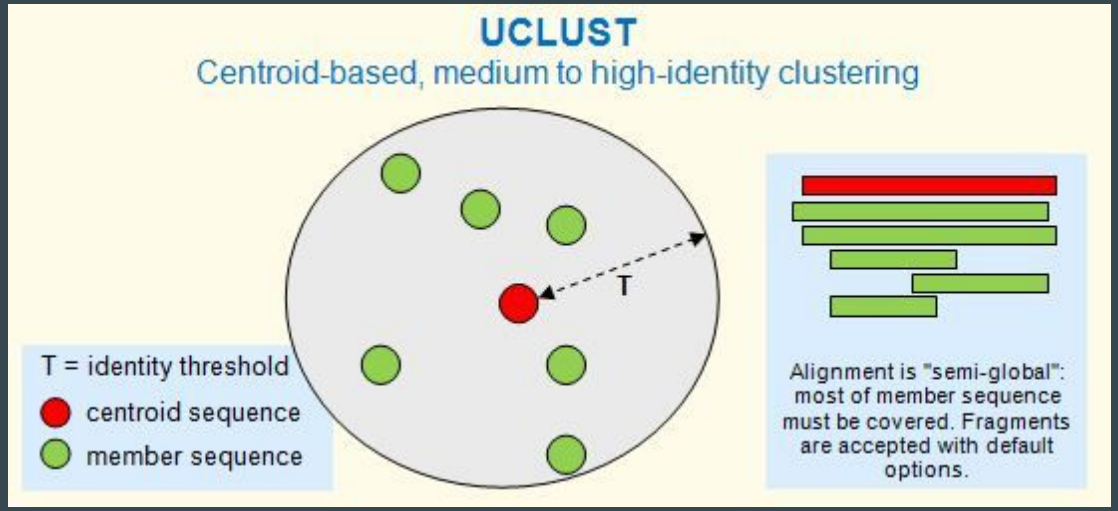


Map to original FASTA file



Output GFF files (258)

Recall: Clustering Algorithm

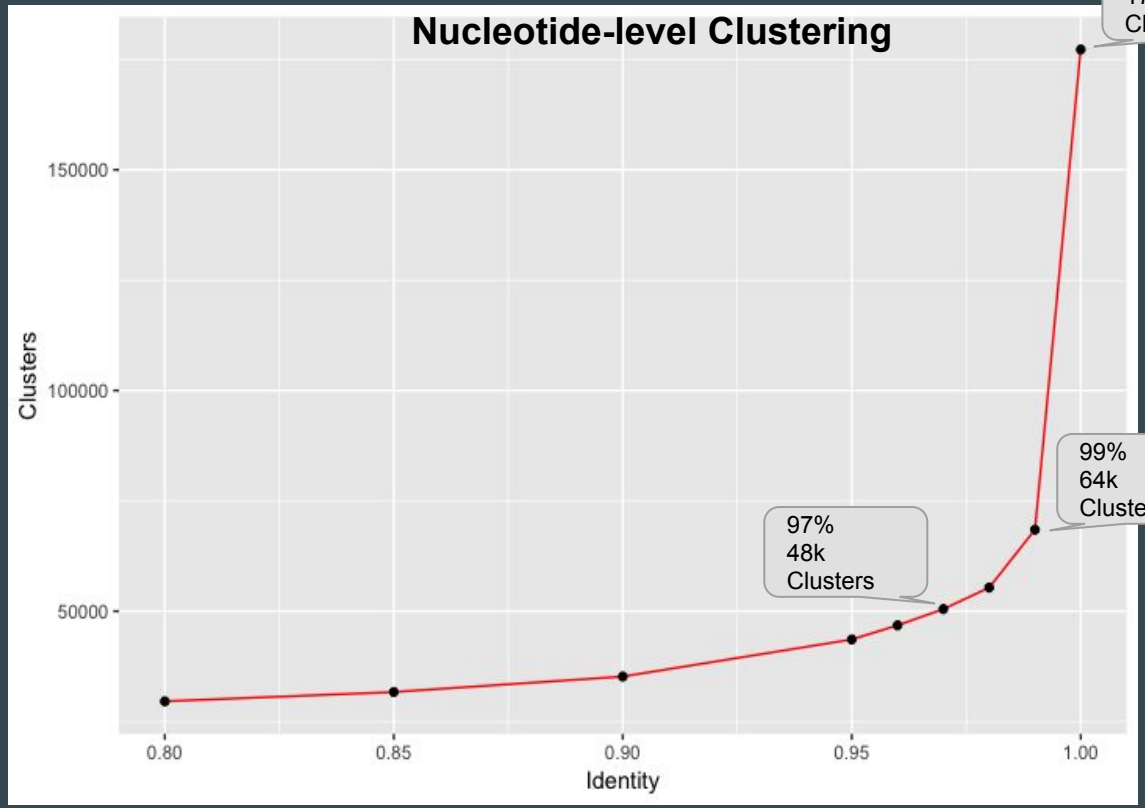


Annotate cluster centroid

Map annotation to member sequences

S	5	9651	*	.	*	*	*	SRR5666411_scaffold34|size34192:94-9745 *	
H	5	9651	100.0	.	0	9651	=	SRR5666525_scaffold34|size28538:93-9744	SRR5666411_scaffold34|size34192:94-9745
H	5	9621	99.6	+	0	0	30I42MI17MD9561M	SRR5666398_scaffold25|size74516:46101-55722	SRR5666411_scaffold34|size34192:94-9745
H	5	9621	99.6	+	0	0	30I42MI17MD9561M	SRR5666447_scaffold36|size28758:343-9964	SRR5666411_scaffold34|size34192:94-9745
H	5	9621	99.6	+	0	0	30I42MI17MD9561M	SRR5666455_scaffold27|size53228:24813-34434	SRR5666411_scaffold34|size34192:94-9745
H	5	9621	99.6	+	0	0	30I42MI17MD9561M	SRR5666530_scaffold36|size29764:377-9998	SRR5666411_scaffold34|size34192:94-9745
H	5	9621	99.6	+	0	0	30I42MI17MD9561M	SRR5666546_scaffold23|size94830:66415-76036	SRR5666411_scaffold34|size34192:94-9745
H	5	9621	99.6	+	0	0	30I42MI17MD9561M	SRR5666613_scaffold22|size94880:66465-76086	SRR5666411_scaffold34|size34192:94-9745
H	5	9621	99.6	+	0	0	30I42MI17MD9561M	SRR5666616_scaffold27|size51164:22749-32370	SRR5666411_scaffold34|size34192:94-9745
S	6	9492	*	.	*	*	*	SRR5666521_scaffold4|size330908:130432-139924 *	
S	7	9492	*	.	*	*	*	SRR5666587_scaffold2|size639562:422808-432300 *	
S	8	7863	*	.	*	*	*	SRR4065638_scaffold3|size362160:317474-325337 *	
H	8	7863	100.0	.	0	7863	=	SRR4065640_scaffold12|size203304:158618-166481	SRR4065638_scaffold3|size362160:317474-325337
H	8	7863	100.0	.	0	7863	=	SRR5666480_scaffold12|size203304:158618-166481	SRR4065638_scaffold3|size362160:317474-325337
S	9	7305	*	.	*	*	*	SRR5666527_scaffold8|size295575:288268-295573 *	
S	10	6378	*	.	*	*	*	SRR4065679_scaffold20|size66143:2-6380 *	
S	11	6321	*	.	*	*	*	SRR4065638_scaffold4|size346292:23959-30280 *	

Clustering Justification



Protein-level Clustering	
Identity	# Clusters
100%	85k
97%	31k

Summary

Tools	Methods	Predict Features	Protein/DNA	Input files	Total Time
PHASTER	Ab-initio Homology	Prophage	DNA	Assembled Genomes	> 2 days (vary)
PilerCR	Ab-initio	CRISPR	DNA	Assembled Genomes	~10 mins
VFDB	Homology	Virulence Factors	DNA	Clustered Genes (97%)	~2.5 mins
LipoP	Ab-initio Homology	Lipoproteins, Transmembrane	DNA	Clustered Genes (100%)	~15 mins
SignalP	Neural Network	Signal peptide	Protein	Clustered Proteins (100%)	~55 mins
RGI	Ab-initio Homology	Antibiotic Resistance	Protein	Clustered Proteins (100%)	~15 mins
DOOR	Homology	Operon	Protein	Clustered Proteins (97%)	~20 mins
EggNOG	Homology	Pathway Gene Ontology	Protein	Clustered Proteins (97%)	~3 hours

Specific Tools (Based on features to be annotated)

- Protein-coding regions
  - Signaling peptides
  - Transmembrane regions
  - Operons
  - Pathways

- Non-coding RNA
  - rRNA, tRNA and sRNA
  - CRISPR

- Others:
  - Antibiotic resistance
  - Virulence factors
  - Prophage genes

Specific Tools (Based on features to be annotated)

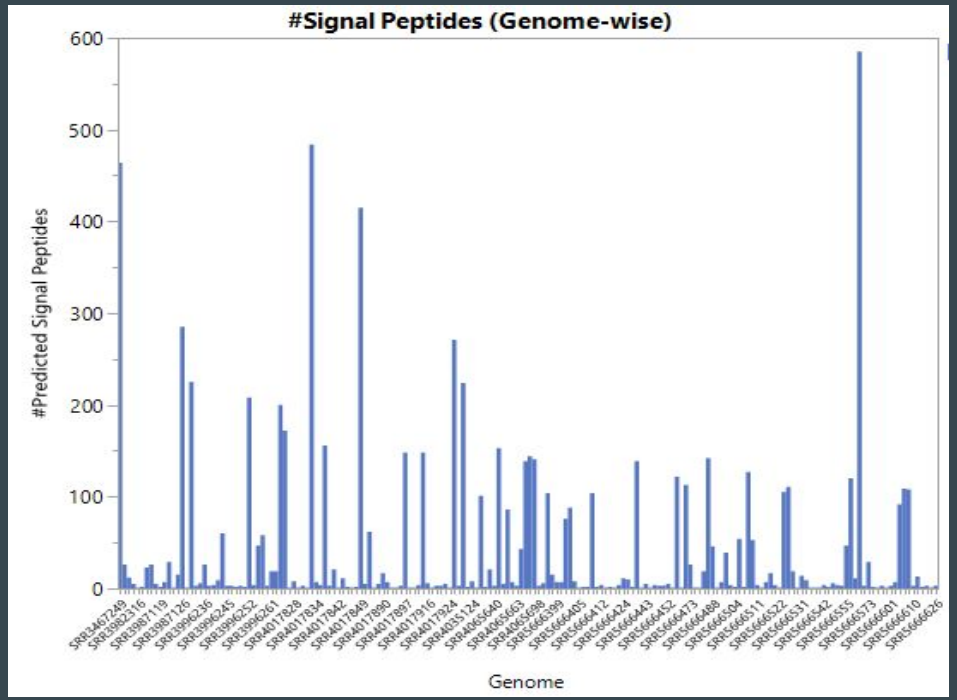
- Protein-coding regions
  - Signaling peptides
  - Transmembrane regions
  - Operons
  - Pathways

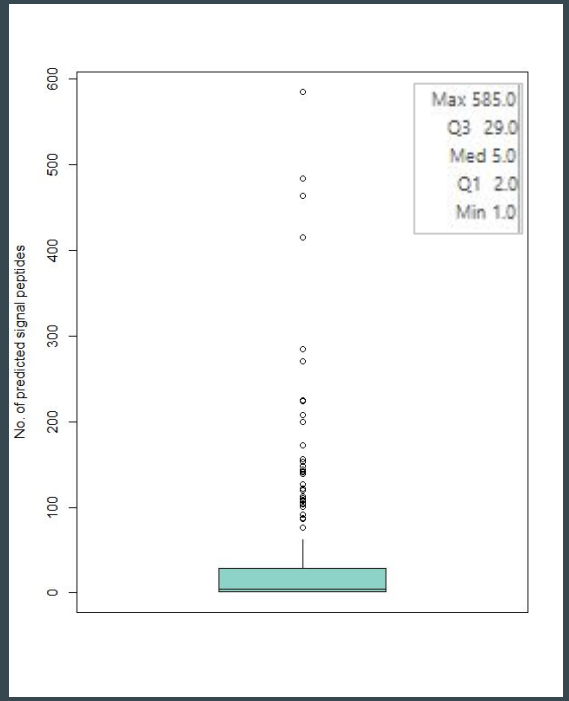
- Non-coding RNA
  - rRNA, tRNA and sRNA
  - CRISPR

- Others:
  - Antibiotic resistance
  - Pathway
  - Prophage genes
  - Virulence factors

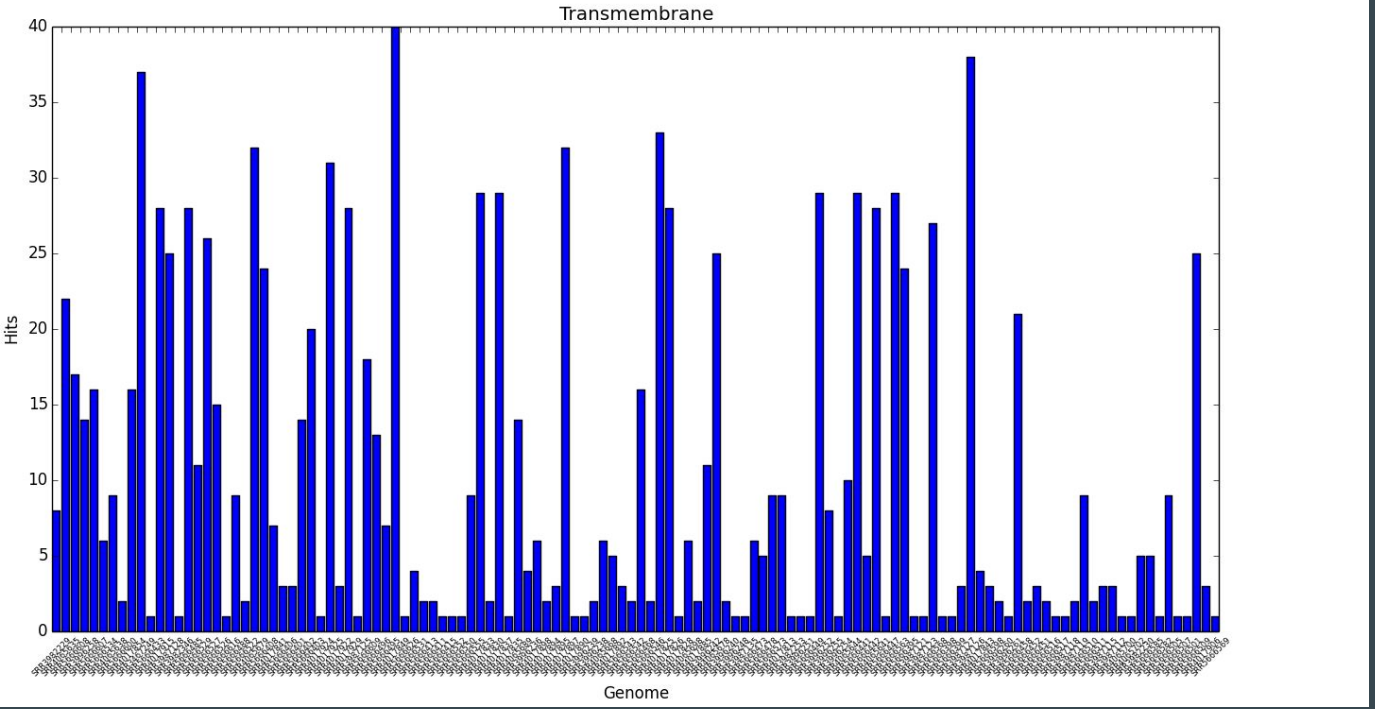
Signal Peptide Prediction - SignalP-4.1

Clustered Proteins (100% Identity)





LipoP

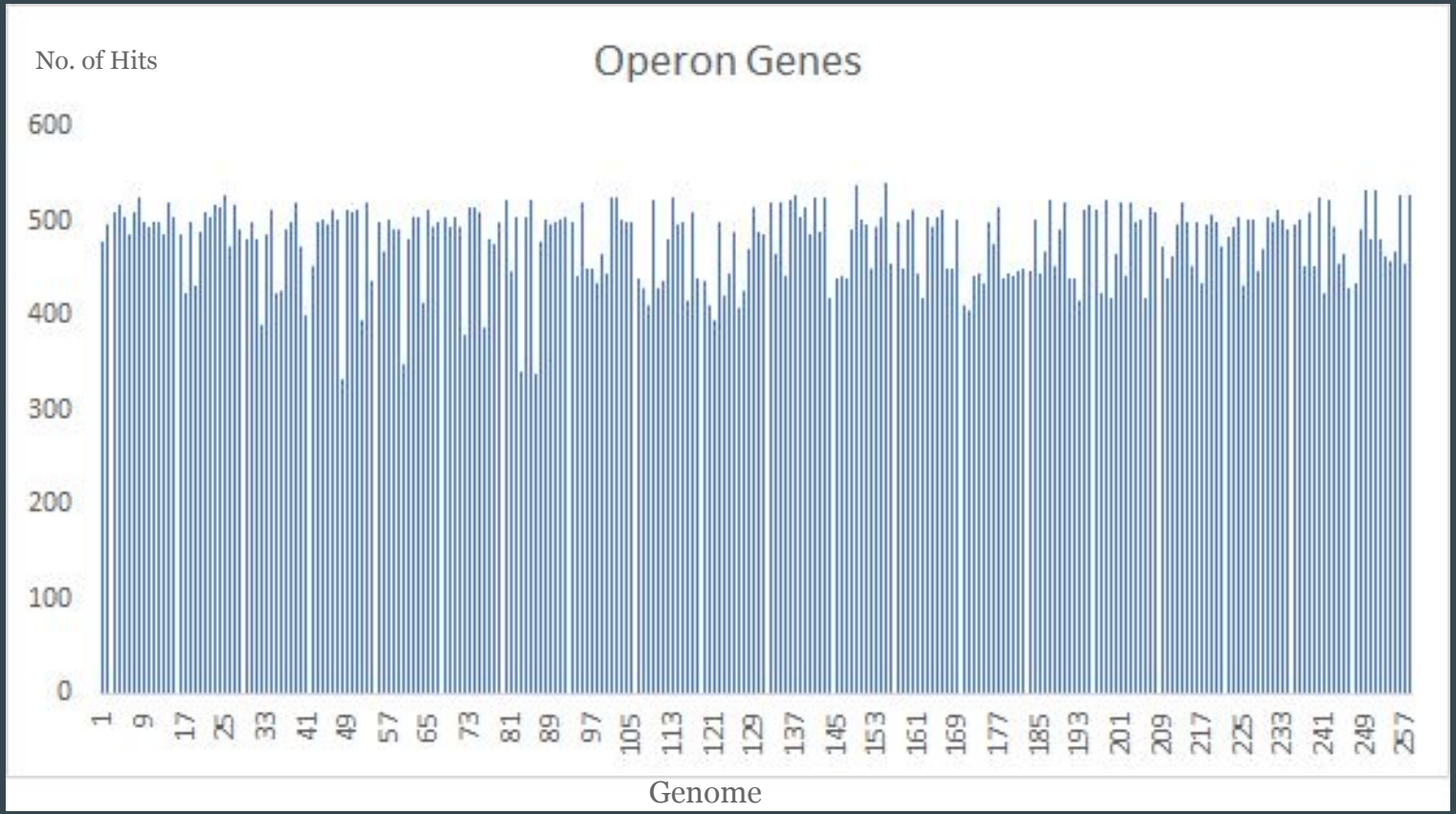


DOOR - operons

SRR5666568_scaffold24|size99086_46:51520-52476	6	318	YP_002241132.1	0.0	1	313	1	operon ID = 471676;COG0181H;porphobilinogen deaminase
SRR5666568_scaffold24|size99086_49:54429-55625	14	397	YP_002241135.1	0.0	14	397	1	operon ID = 471676;COG3071H;protoheme IX biogenesis protein
SRR5666568_scaffold24|size99086_53:60464-61543	1	355	YP_002241140.1	0.0	1	354	1	operon ID = 471677;-;4-alpha-L-fucosyltransferase
SRR5666568_scaffold24|size99086_61:68988-70004	1	337	YP_002241148.1	0.0	11	347	1	operon ID = 471677;COG3765M;lipopolysaccharide biosynthesis protein WzzE

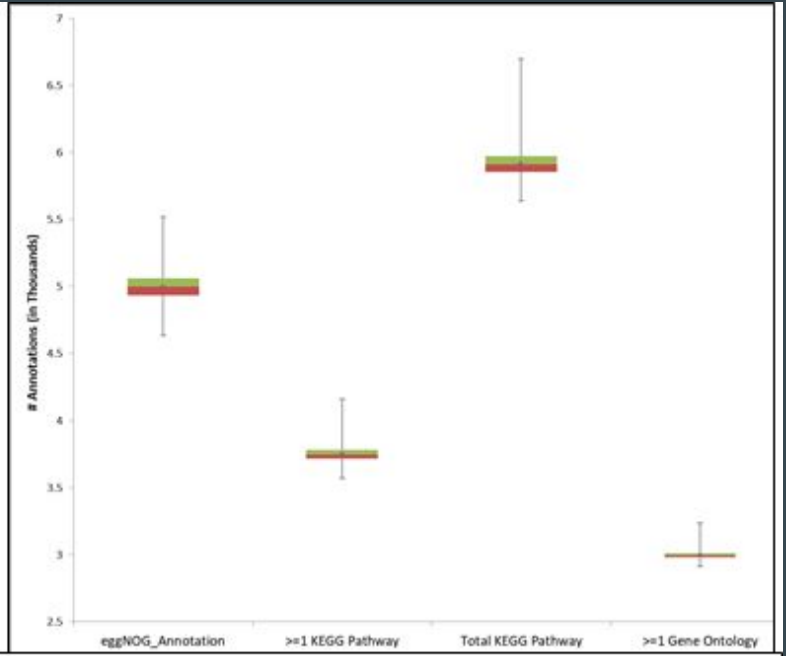
SRR5666568_scaffold42|size36644_19:17982-18923	5	313	YP_002238285.1	0.0	2	310	1	operon ID = 471138;COG0834ET;ABC transporter substrate-binding protein
SRR5666568_scaffold42|size36644_20:18939-20255	1	437	YP_002238284.1	0.0	1	437	1	operon ID = 471138;COG2141C;Ytnj protein
SRR5666568_scaffold42|size36644_21:20286-21407	1	371	YP_002238283.1	0.0	1	371	1	operon ID = 471138;COG1473R;amidohydrolase family protein

SRR5666568_scaffold14|size159094_133:139347-140579	2	403	YP_002238440.1	0.0	12	413	1	operon ID = 471163;COG0624E;allantoate amidohydrolase
SRR5666568_scaffold14|size159094_134:140606-141847	1	411	YP_002238439.1	0.0	1	411	1	operon ID = 471163;COG0075E;aminotransferase class V
SRR5666568_scaffold14|size159094_140:145918-147501	3	527	YP_002238433.1	0.0	4	528	1	operon ID = 471162;COG0405E;gamma-glutamyltransferase
SRR5666568_scaffold14|size159094_144:149642-150574	6	310	YP_002238403.1	0.0	6	310	1	operon ID = 471158;COG0726G;urate catabolism protein


No. of Hits
Operon Genes
600
500
400
300
200
100
0
1
9
17
25
33
41
49
57
65
73
81
89
97
105
113
121
129
137
145
153
161
169
177
185
193
201
209
217
225
233
241
249
257
Genome

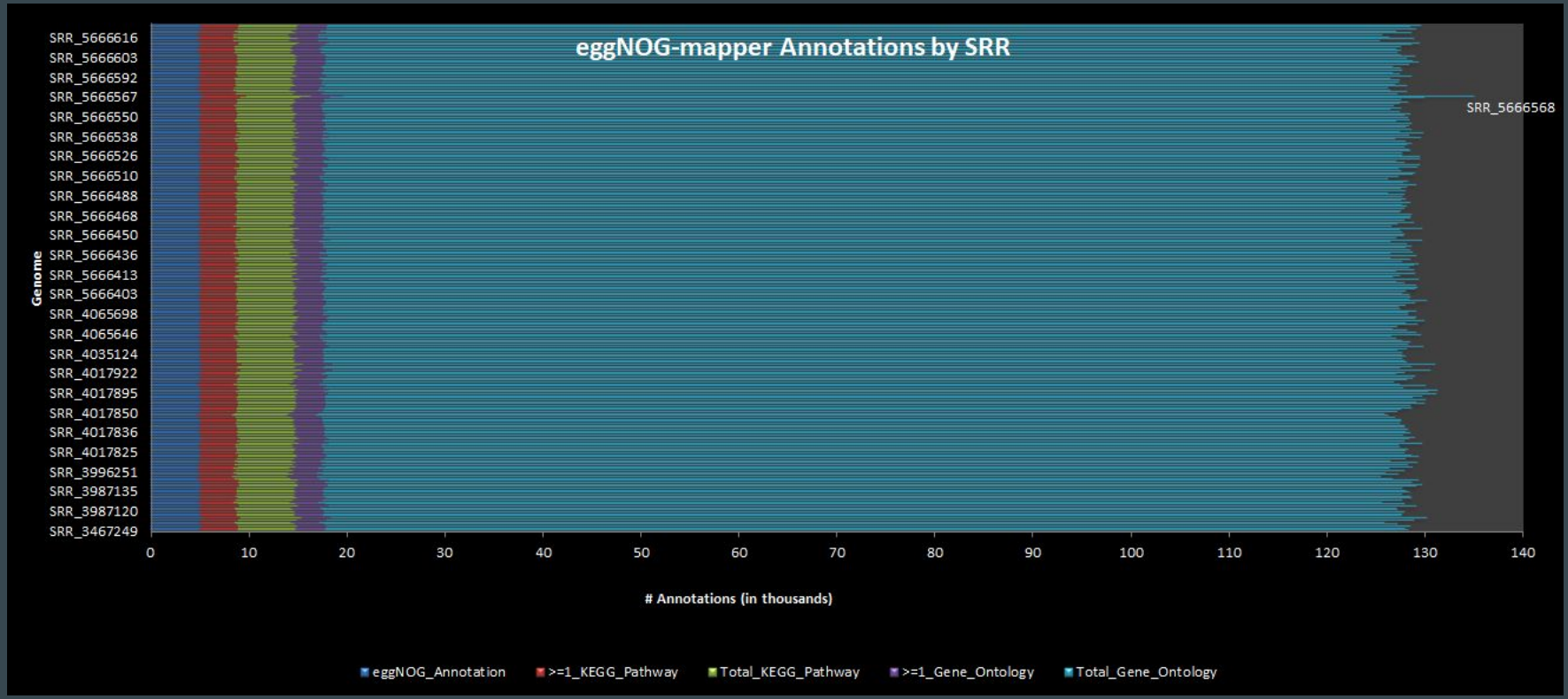
eggNOG-mapper

Annotated Sequences	Protein Sequences	Proportion
1288878	1376279	0.936494708



	eggNOG_Annotation	>=1 KEGG Pathway	Total KEGG Pathway	>=1 Gene Ontology
Mean	4995.651163	3748.391473	5917.48062	2993.72093
Standard Deviation: SD	111.9776347	65.56416233	108.3383589	31.8561887
Maximum Value: Max	5518	4161	6697	3236
3rd Quartile - Q3	5060.5	3782.75	5970.75	3012
Median	4998	3745	5909	2993.5
1st Quartile - Q1	4934	3715	5854	2978
Minimum Value: Min	4632	3571	5636	2912

eggNOG-mapper



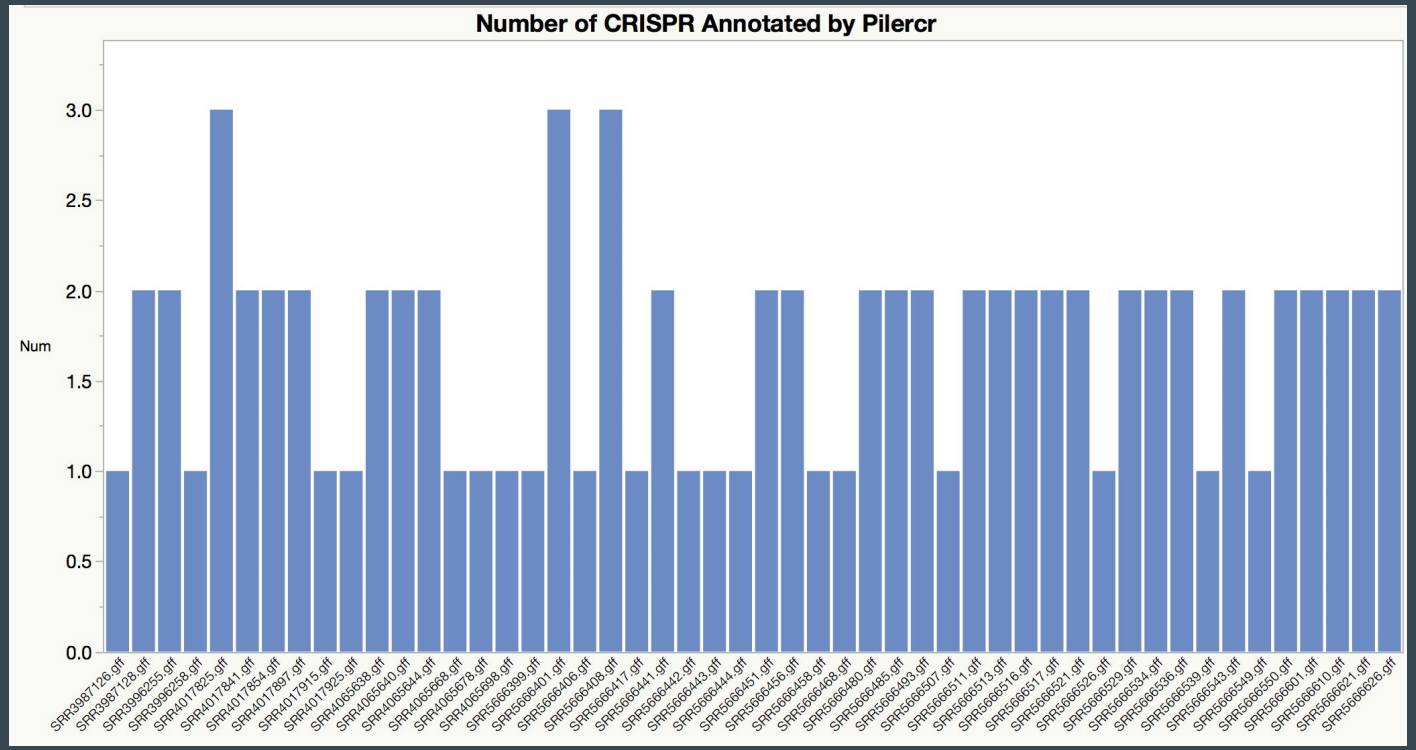
Specific Tools (Based on features to be annotated)

- Protein-coding regions
  - Transmembrane regions
  - Enzymes
  - Signaling peptides
  - Operons
  - Lipoproteins
- Others:
  - Antibiotic resistance
  - Pathway
  - Prophage genes
  - Virulence factors

- Non-coding RNA
  - rRNA, tRNA and sRNA
  - CRISPR

PILER-CR

Total 84 hits among 50 genomes



Specific Tools (Based on features to be annotated)

- Protein-coding regions
  - Transmembrane regions
  - Enzymes
  - Signaling peptides
  - Operons
  - Lipoproteins

- Non-coding RNA
  - rRNA, tRNA and sRNA
  - CRISPR

- Others:
  - Antibiotic resistance
  - Virulence factors
  - Prophage genes

Prophage: Phaster (PHAge Search Tool Enhanced Release)

GFF Example:

scaffold4|size340664 phaster PHAGE_Salmon_RE_2010_NC_019488:_gene_D_protein 306622 307638 0.0 . . region2;PHAGE_Salmon_RE_2010_NC_019488: gene D protein; PP_01463; phage(gi418489726);intact(150)

scaffold4|size340664 phaster PHAGE_Salmon_RE_2010_NC_019488:_tail_protein 307719 308198 3.18e-90 . . region2;PHAGE_Salmon_RE_2010_NC_019488: tail protein; PP_01464; phage(gi418489725);intact(150)

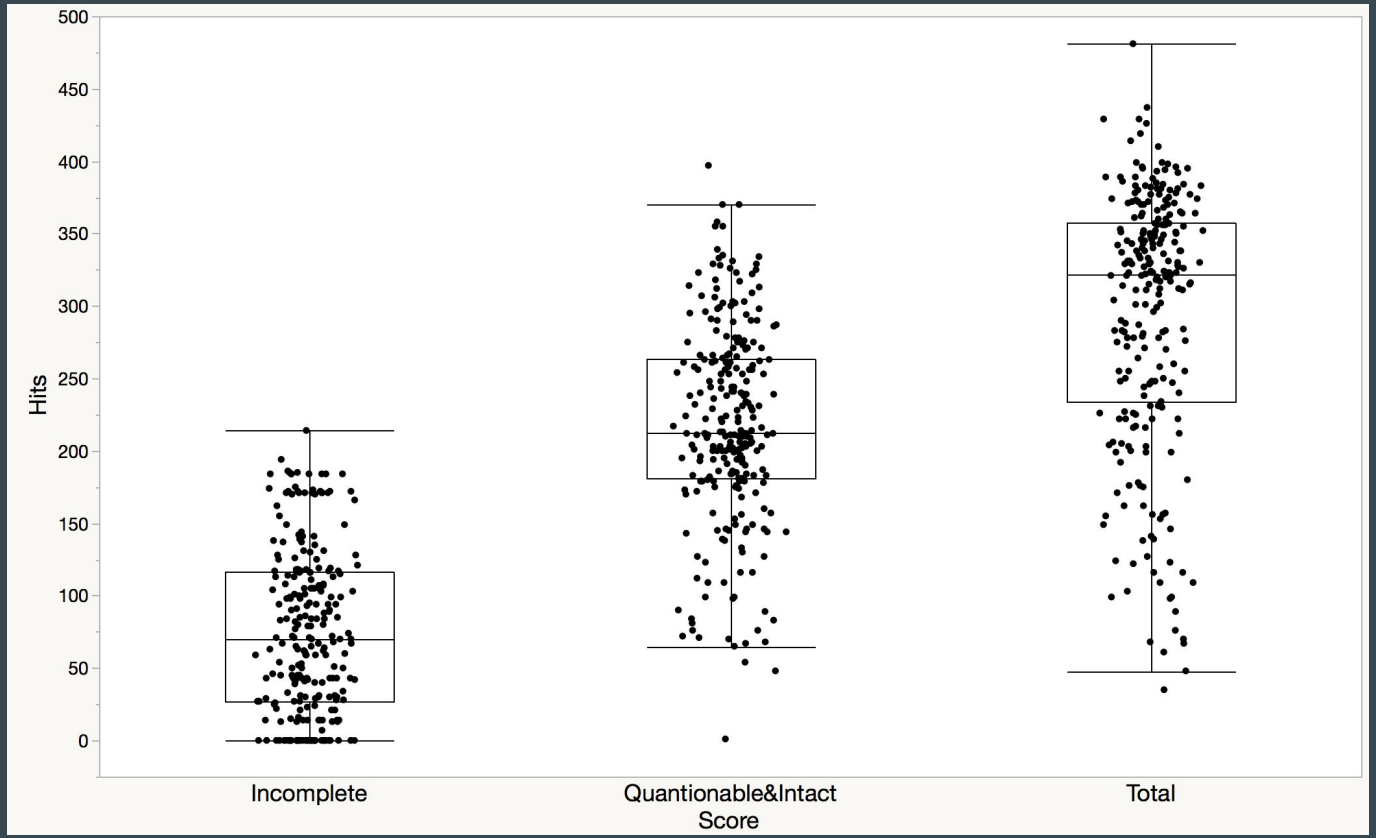
scaffold4|size340664 phaster PHAGE_Salmon_RE_2010_NC_019488:_tail_tape_measure_protein 308201 310828 4.03e-88 . . region2;PHAGE_Salmon_RE_2010_NC_019488: tail tape measure protein; PP_01465; phage(gi418489724);intact(150)

scaffold16|size101770 phaster hypothetical 86927 87262 N/A . . region5;hypothetical; PP_03803;questionable(70)

scaffold2|size413645 phaster PHAGE_Entero_IME11_NC_019423:_hypothetical_protein 2497 3018 1.38e-40 . . region1;PHAGE_Entero_IME11_NC_019423: hypothetical protein; PP_00414; phage(gi414090513);incomplete(30)

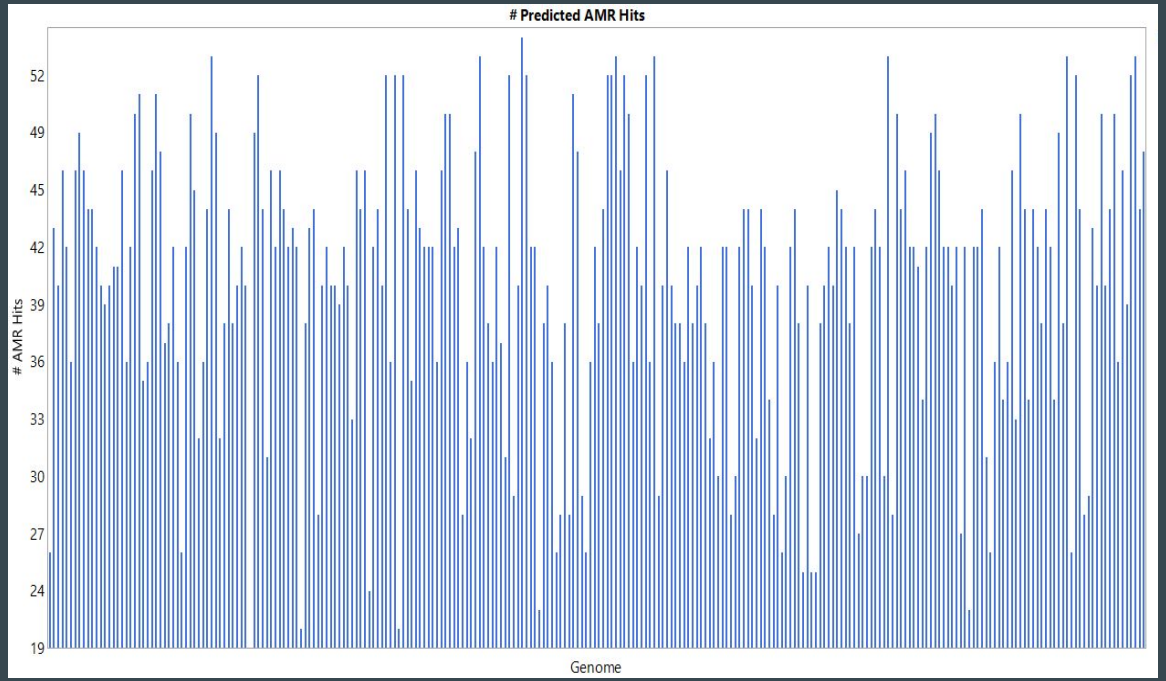
Prophage: Phaster (PHAge Search Tool Enhanced Release)

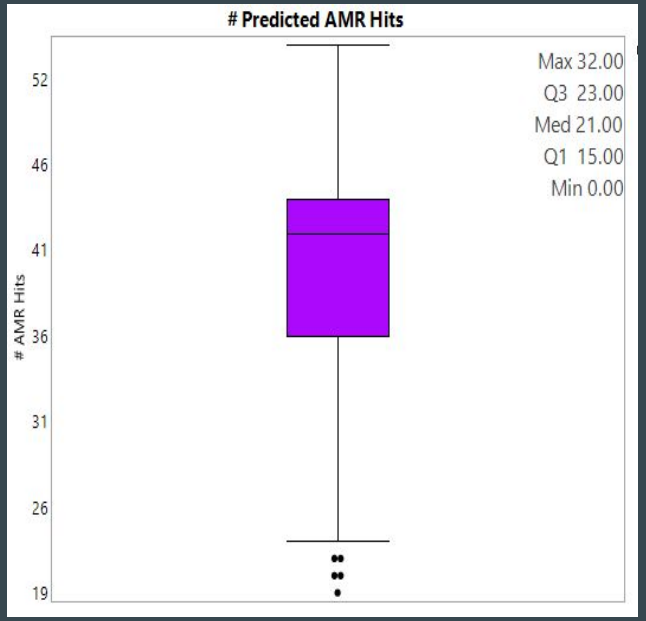
Number of phage gene hits in each genome



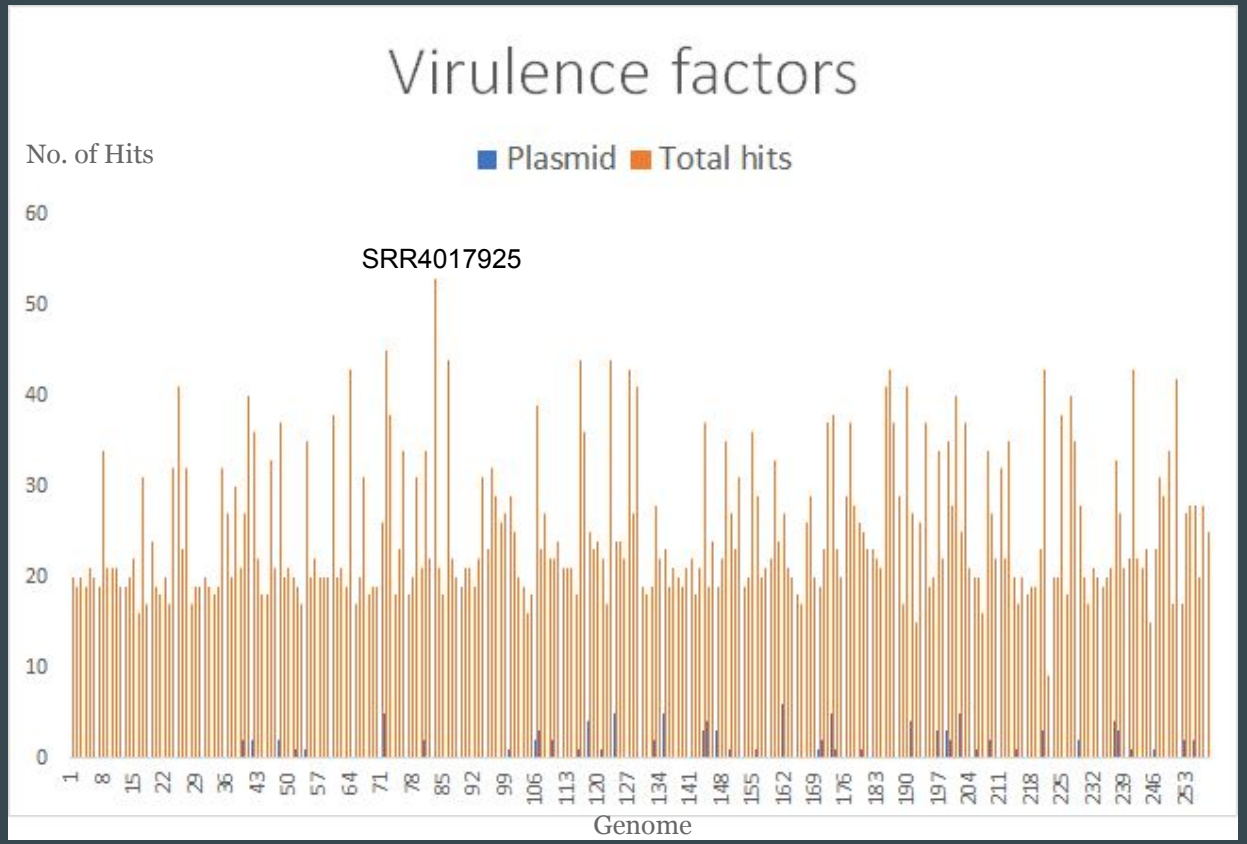
Antibiotic Resistance: RGI

ab-initio and homology
proteins clustered at 100% identity

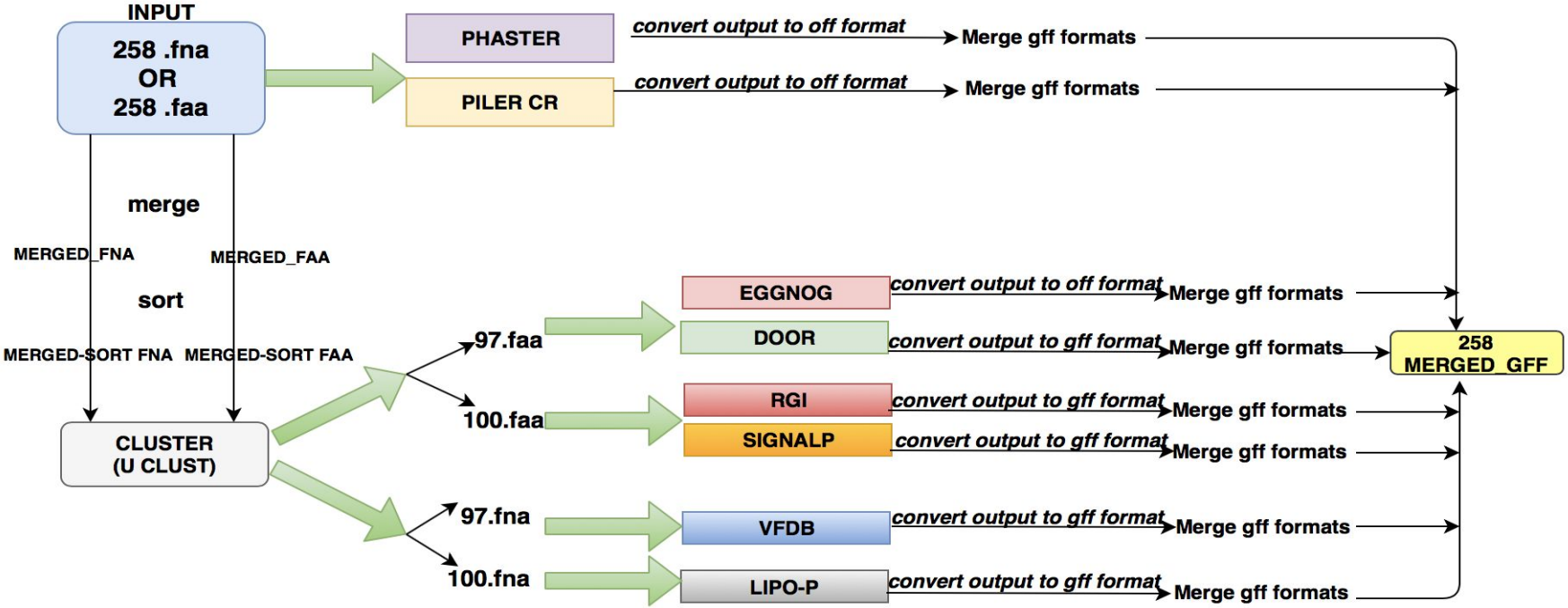




Virulence Factors: VFDB



Final Annotation Pipeline (./finalPipeline.sh)



merged_GFF

```
scaffold50	phaster	PHAGE_Entero_lato_NC_001422:_minor_spike_protein	3907	4893	0.0	.	.	region11;PHAGE_Entero_lato_NC_001422:_mino
scaffold50	phaster	PHAGE_Entero_lato_NC_001422:_major_spike_protein	4902	5429	4.98e-127	.	.	region11;PHAGE_Entero_lato_NC_0014
scaffold50	DOOR	operon	4903	5074	2.52e-107	.	.	Name=operon ID = 2388555;COG0456R;acetyltransferase GNAT family
scaffold50	phaster	PHAGE_Entero_lato_NC_001422:_capsid_protein	5540	6823	0.0	.	.	region11;PHAGE_Entero_lato_NC_001422:_capsid_prote
scaffold50	phaster	PHAGE_Entero_lato_NC_001422:_external_scaffolding_protein	6976	7434	3.80e-109	.	.	region11;PHAGE_Entero_lato
scaffold50	phaster	PHAGE_Entero_lato_NC_001422:_DNA_replication_initiation_protein_gpA	7688	9229	0.0	.	.	region11;PHAGE_Entero_lato
scaffold50	phaster	PHAGE_Entero_lato_NC_001422:_minor_spike_protein	9293	10174	0.0	.	.	region11;PHAGE_Entero_lato_NC_001422:_mino
scaffold51	DOOR	operon	5015	5352	0.0	.	.	Name=operon ID = 2387414;COG1609K;DNA-binding transcriptional regulator CytR
scaffold51	rgi	AMR_gene	7202	8083	0.0	.	0	KPC-3 : antibiotic inactivation enzyme; determinant of beta-lactam resistance
scaffold54	Infernal-1.1.2/Rfam	RNAI	6317	6419	1.8e-15	+	.	Target_Name:RNAI;Model:cm;Target/RF_ACC:RF00106;Clan_Name:-;GC:0.50
scaffold55	DOOR	operon	4501	4837	0.0	.	.	Name=operon ID = 328176;COG1609K;putative periplasmic binding protein/LacI transcriptional
scaffold56	DOOR	operon	1860	1982	1.67e-63	.	.	Name=operon ID = 470727;-;hypothetical protein
scaffold57	Infernal-1.1.2/Rfam	group-II-D1D4-7	5739	5887	1e-25	+	.	Target_Name:group-II-D1D4-7;Model:cm;Target/RF_ACC:RF02012;Clan_Na
scaffold58	Infernal-1.1.2/Rfam	5S_rRNA	9	124	3.9e-15	+	.	Target_Name:5S_rRNA;Model:cm;Target/RF_ACC:RF00001;Clan_Name:CL00113;GC:0.
scaffold58	Infernal-1.1.2/Rfam	tRNA	3934	4006	1.4e-15	+	.	Target_Name:tRNA;Model:cm;Target/RF_ACC:RF00005;Clan_Name:CL00001;GC:0.51
scaffold58	Infernal-1.1.2/Rfam	tRNA	4020	4101	2.1e-12	+	.	Target_Name:tRNA;Model:cm;Target/RF_ACC:RF00005;Clan_Name:CL00001;GC:0.59
```

References

- Seemann, Torsten. “PROKKA: Rapid Prokaryotic Genome Annotation” *Bioinformatics* 30.14 (2014):2068-2069
- Petersen, Thomas Nordahl, et al. "SignalP 4.0: discriminating signal peptides from transmembrane regions." *Nature methods* 8.10 (2011): 785.
- Käll, Lukas, Anders Krogh, and Erik LL Sonnhammer. "A combined transmembrane topology and signal peptide prediction method." *Journal of molecular biology* 338.5 (2004): 1027-1036.
- Juncker, Agnieszka S. et al. “Prediction of Lipoprotein Signal Peptides in Gram-Negative Bacteria.” *Protein Science : A Publication of the Protein Society* 12.8 (2003): 1652–1662.
- Mao, Xizeng et al. “DOOR 2.0: Presenting Operons and Their Functions through Dynamic and Integrated Views.” *Nucleic Acids Research* 42.Database issue (2014): D654–D659.
- Jones, Philip et al. “InterProScan 5: Genome-Scale Protein Function Classification.” *Bioinformatics* 30.9 (2014): 1236–1240.

References

- Jensen, Lars Juhl et al. “eggNOG: Automated Construction and Annotation of Orthologous Groups of Genes.” *Nucleic Acids Research* 36.Database issue (2008): D250–D254.
- Nawrocki, Eric P., Diana L. Kolbe, and Sean R. Eddy. “Infernal 1.0: Inference of RNA Alignments.” *Bioinformatics* 25.10 (2009): 1335–1337.
- Edgar, Robert C. “PILER-CR: Fast and Accurate Identification of CRISPR Repeats.” *BMC Bioinformatics* 8 (2007): 18.
- Jia, Baofeng et al. “CARD 2017: Expansion and Model-Centric Curation of the Comprehensive Antibiotic Resistance Database.” *Nucleic Acids Research* 45.Database issue (2017): D566–D573.
- Chen, Lihong et al. “VFDB: A Reference Database for Bacterial Virulence Factors.” *Nucleic Acids Research* 33.Database Issue (2005): D325–D328.
- Arndt, David et al. “PHASTER: A Better, Faster Version of the PHAST Phage Search Tool.” *Nucleic Acids Research* 44.Web Server issue (2016): W16–W21.

Questions?