Functional Annotation Final Results

Team 1 Functional Annotation

Wenyi Qiu, Tianze Song, Saurabh Gulati, Ryan Place, Dongjo Ban, Qinwei Zhuang, Kunal Agarwal, Frank Ambrosio

Background/Review

- Antibiotic Resistance
  - Antibiotic Resistance
  - Colistin Resistance
- Next Generation Sequencing
  - Bioinformatics
  - Scalability
- Functional Annotation
  - Multi-Tool Approach
  - Gene Clustering

Antibiotic Resistance

High levels of antibiotic resistance found worldwide, new data shows

News release

29 JANUARY 2018 I BANGKOK - WHO's first release of surveillance data on antibiotic resistance reveals high levels of resistance to a number of serious bacterial infections in both high- and low-income countries.

WHO's new Global Antimicrobial Surveillance System (GLASS) reveals widespread occurrence of antibiotic resistance among 500 000 people with suspected bacterial infections across 22 countries.

http://www.who.int/mediacentre/news/releases/2018/antibiotic-resistance-found/en/

- Antibiotic resistance is a rapidly growing problem
- Strains have been shown to possess resistance to last line antibiotics such as colistin
- The power of sequencing technologies is ever increasing
- Bioinformatic analysis techniques must scale up

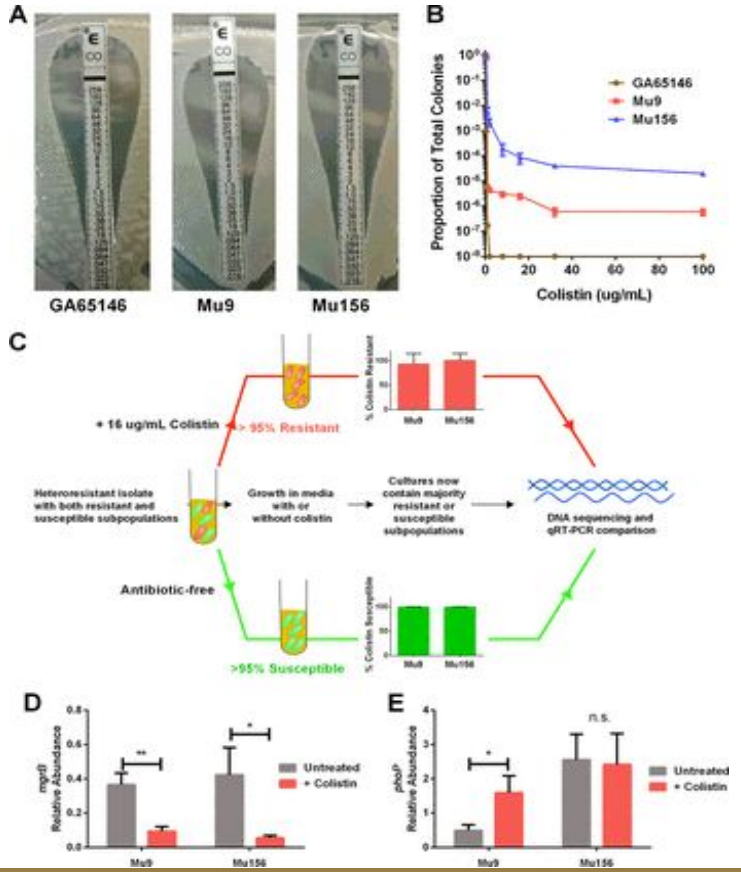
Colistin Resistance

- Significance
  - Last-Line Drug
  - Resistance Genes are Mobile
- Mechanism
  - Efflux pump
- Related Gene Family
  - MCR-1 to MCR-5 (Lipopolysaccharide modification)
  - PhoP
  - PhoQ

Heteroresistance

- Heteroresistance is an emerging concern for AMR research
- Isolates previously considered fully susceptible must now be reviewed
- Presence of a resistant subpopulation will affect treatment
- Ability to identify (culture independently) resistant subpopulation will save lives

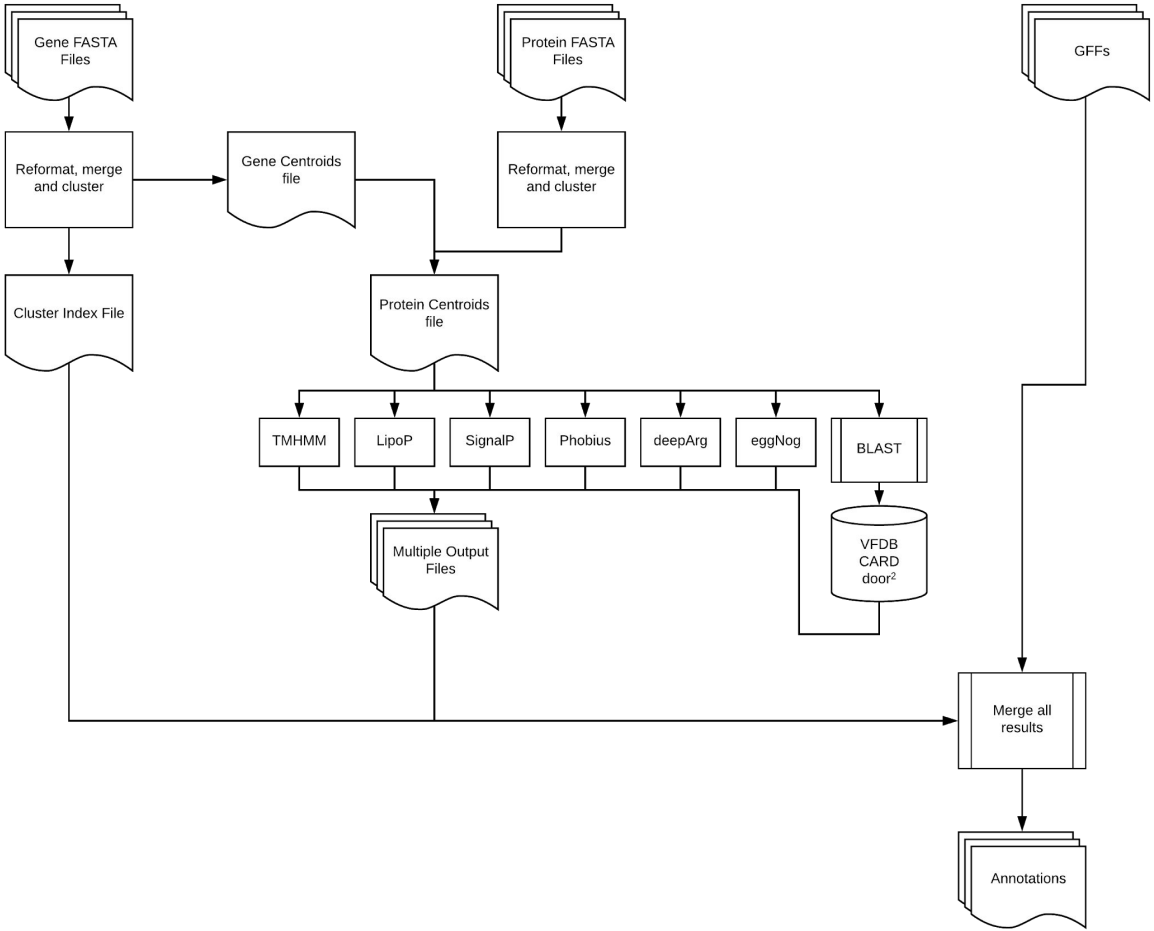
Heteroresistance and Heterosusceptibility



- Heteroresistance exists in two “states”
- Untreated Isolate
  - 95% Susceptible
  - 5% Resistant
- Antibiotic Treated Isolate
  - 5% Susceptible
  - 95% Resistent
- Relative abundance of mgrB(D) and phoP(E) determined by qPCR

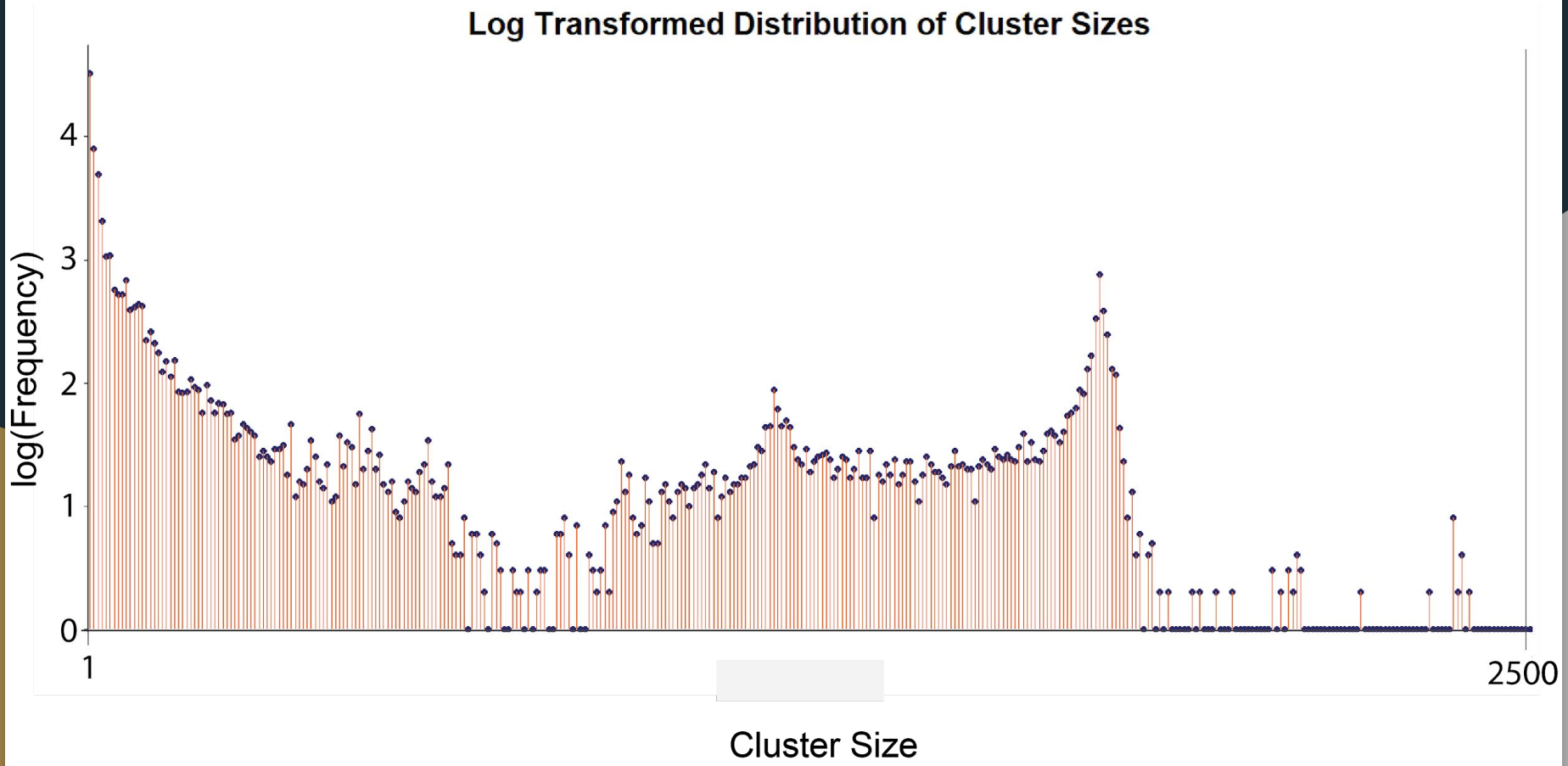
- http://mbio.asm.org/content/9/2/e02448-17.full.pdf+html

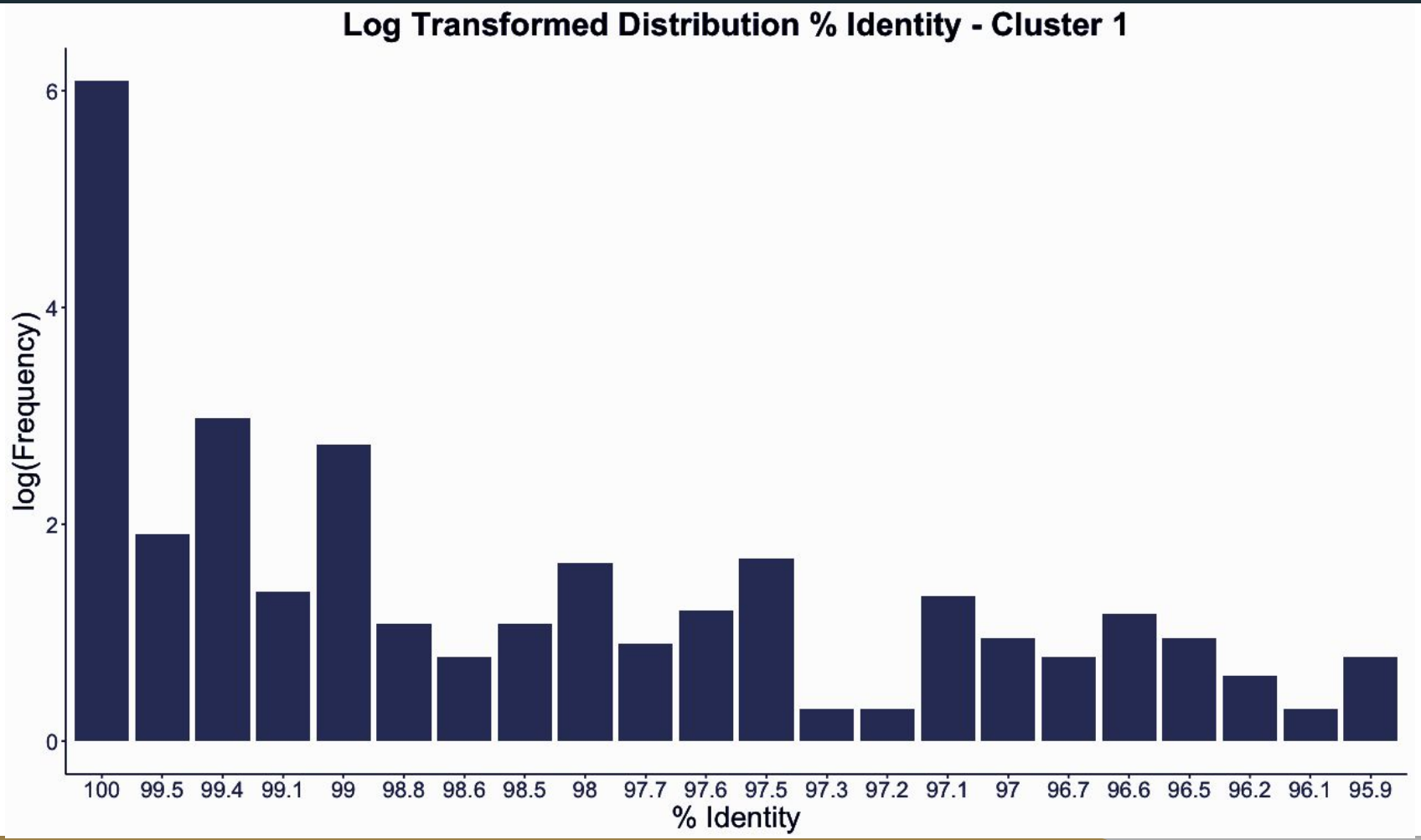
Pipeline

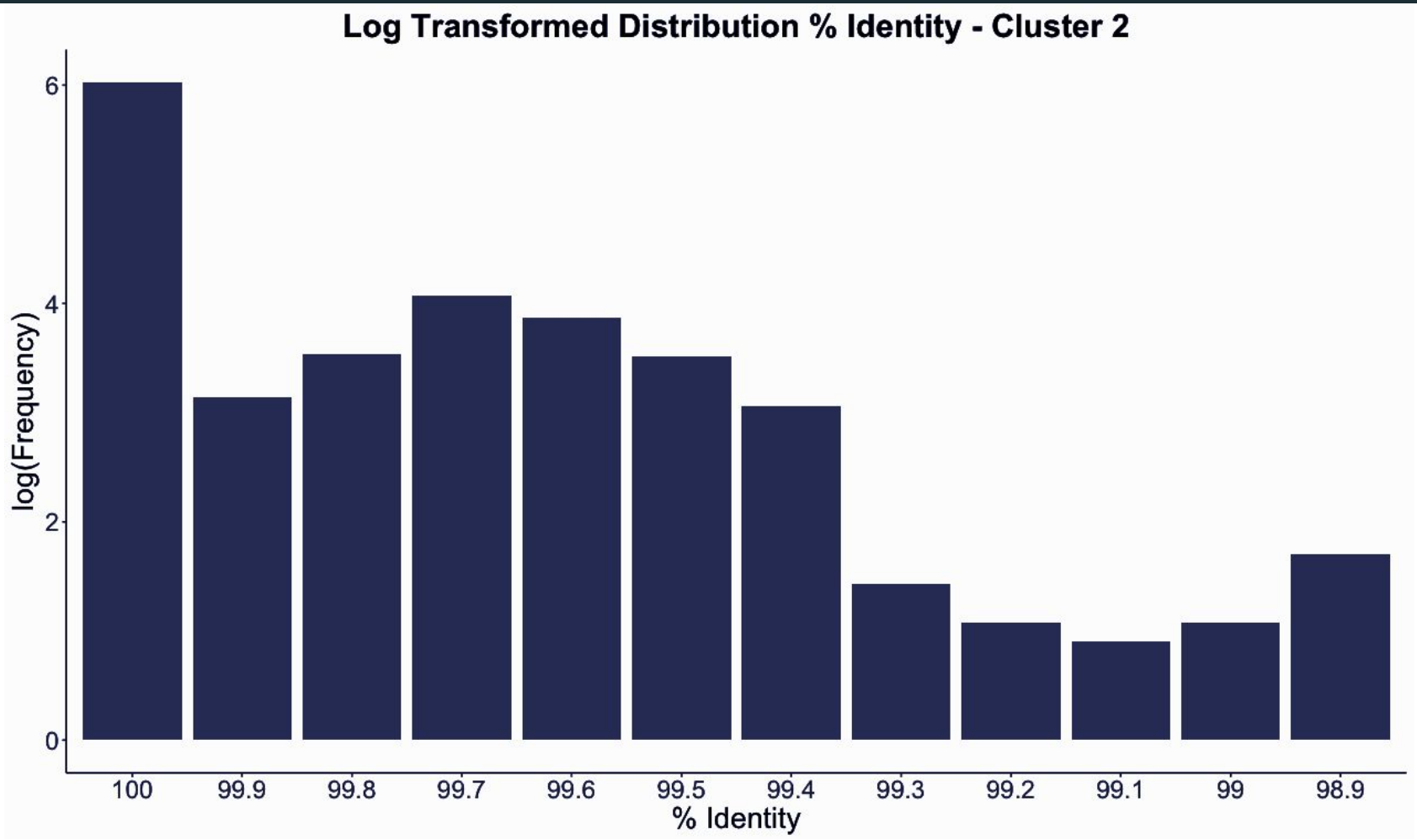

Gene FASTA Files
Protein FASTA Files
GFFs
Reformat, merge and cluster
Gene Centroids file
Reformat, merge and cluster
Cluster Index File
Protein Centroids file
TMHMM
LipoP
SignalP
Phobius
deepArg
eggNog
BLAST
Multiple Output Files
VFDB
CARD
door2
Merge all results
Annotations

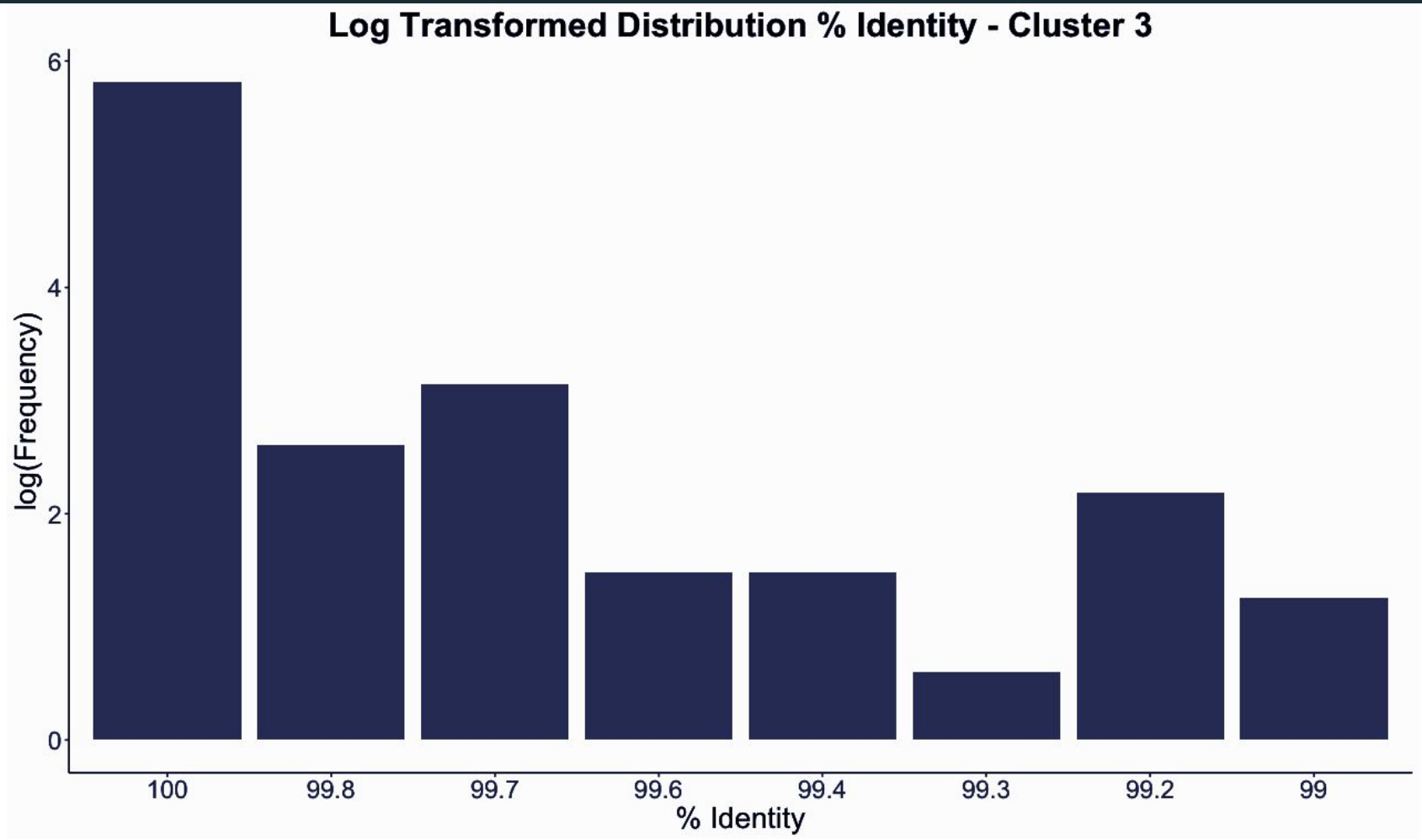
Clustering

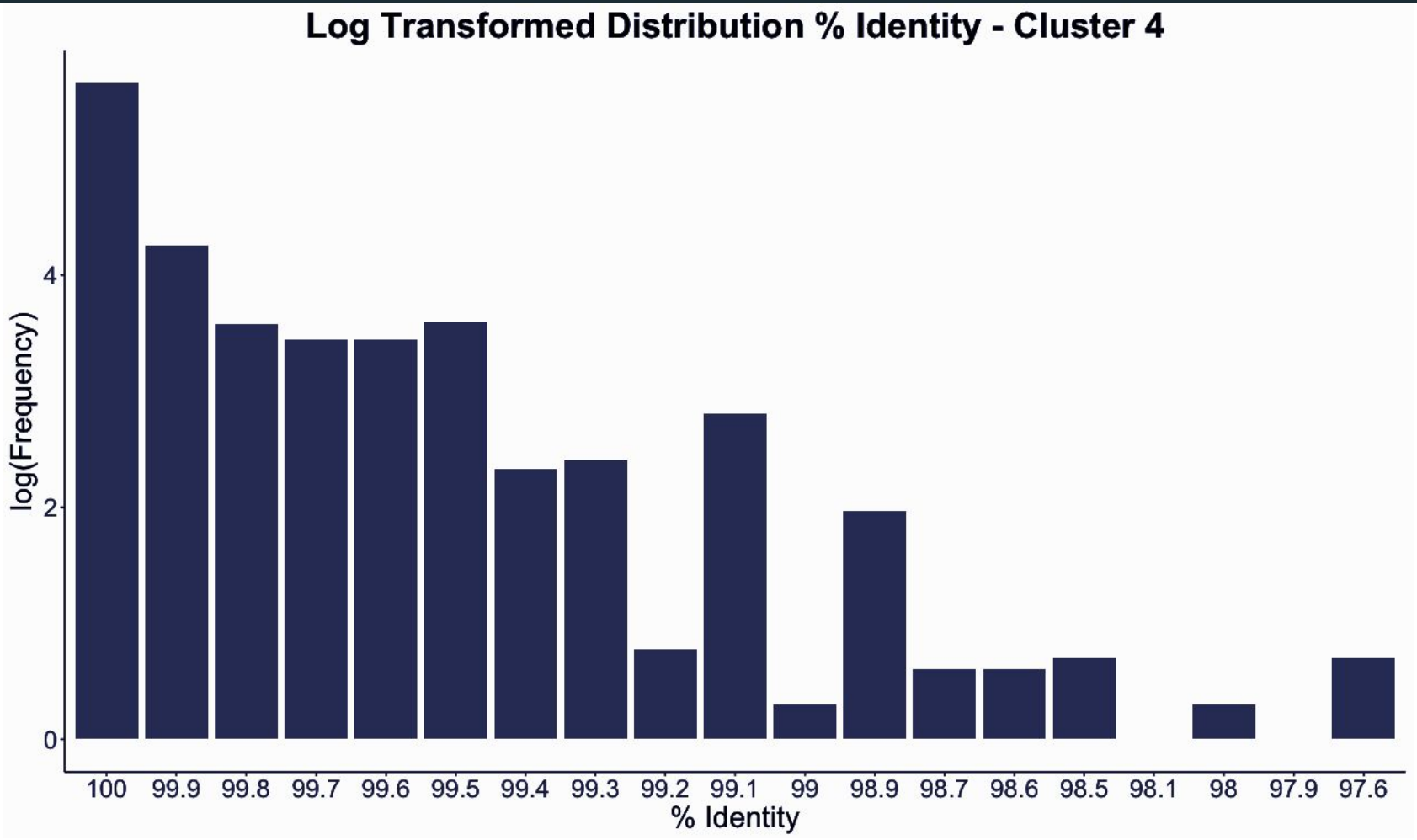
- Tool used: uCLUST (uSEARCH)
- Identity threshold: 0.99
- # of clusters: 63,127
- # of singletons: 32,792
- Max size of clusters: 2,458

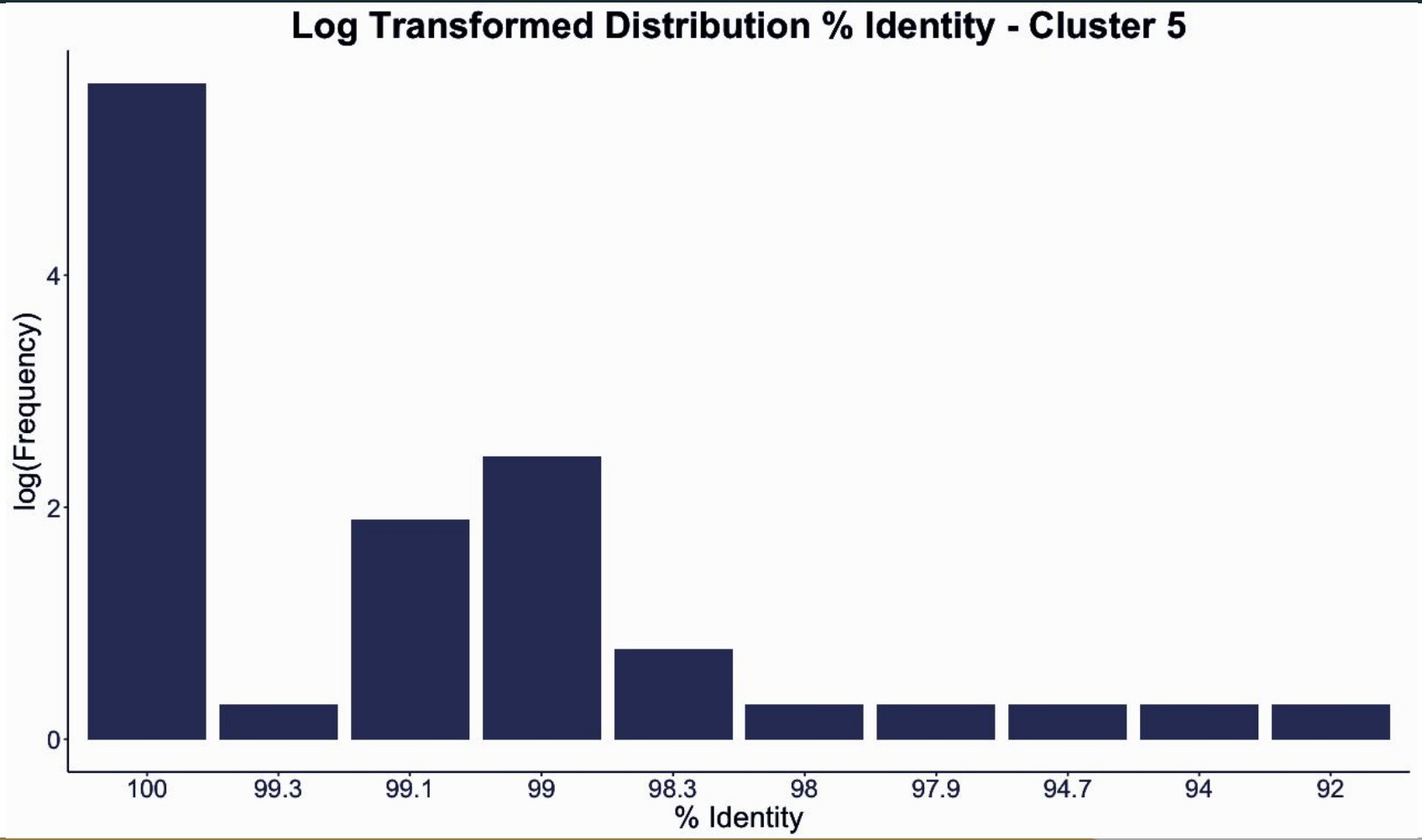

Log Transformed Distribution of Cluster Sizes
log(Frequency)
4
3
2
1
0
1
2500
Cluster Size


Log Transformed Distribution % Identity - Cluster 1
log(Frequency)
0
2
4
6
100
99.5
99.4
99.1
99
98.8
98.6
98.5
98
97.7
97.6
97.5
97.3
97.2
97.1
97
96.7
96.6
96.5
96.2
96.1
95.9
% Identity


Log Transformed Distribution % Identity - Cluster 2
log(Frequency)
6
4
2
0
100
99.9
99.8
99.7
99.6
99.5
99.4
99.3
99.2
99.1
99
98.9
% Identity


Log Transformed Distribution % Identity - Cluster 3
log(Frequency)
6
4
2
0
100
99.8
99.7
99.6
99.4
99.3
99.2
99
% Identity


Log Transformed Distribution % Identity - Cluster 4
log(Frequency)
4
2
0
100
99.9
99.8
99.7
99.6
99.5
99.4
99.3
99.2
99.1
99
98.9
98.7
98.6
98.5
98.1
98
97.9
97.6
% Identity


Log Transformed Distribution % Identity - Cluster 5
log(Frequency)
0
2
4
100
99.3
99.1
99
98.3
98
97.9
94.7
94
92
% Identity

Annotation

- Eggnog
- SignalP
- LipoP
- VFDB
- CARD
- DeepARG
- DOOR^2

Homology Based Functional Annotation

Uses databases of genomic features with known function

- Accuracy is dependent on database quality
  - Garbage in garbage out
- Databases for AMR genomic features are added to on a regular basis
- CARD and VFDB are examples of databases of homologous genomic features

Interproscan

Command:
time /projects/data/team1_functionalAnnotation/interproscan-5.28-67.0/interproscan.sh -applPfamA, CDD, HAMAP, PROSITEPATTERNS, PROSITEPROFILES, SFLD, SMART, SUPERFAMILY, TIGRFAM, Phobius, TMHMM, SignalP_GRAM_NEGATIVE -i <input.faa> -f gff3

Number of annotations for clustered centroids: 49831 (total sequences for analysis: 63127)

Running time: ~48h (could have been way faster if using cluster mode instead of standalone mode)

```
real    2895m30.857s
user    6041m13.085s
sys     377m0.426s
```

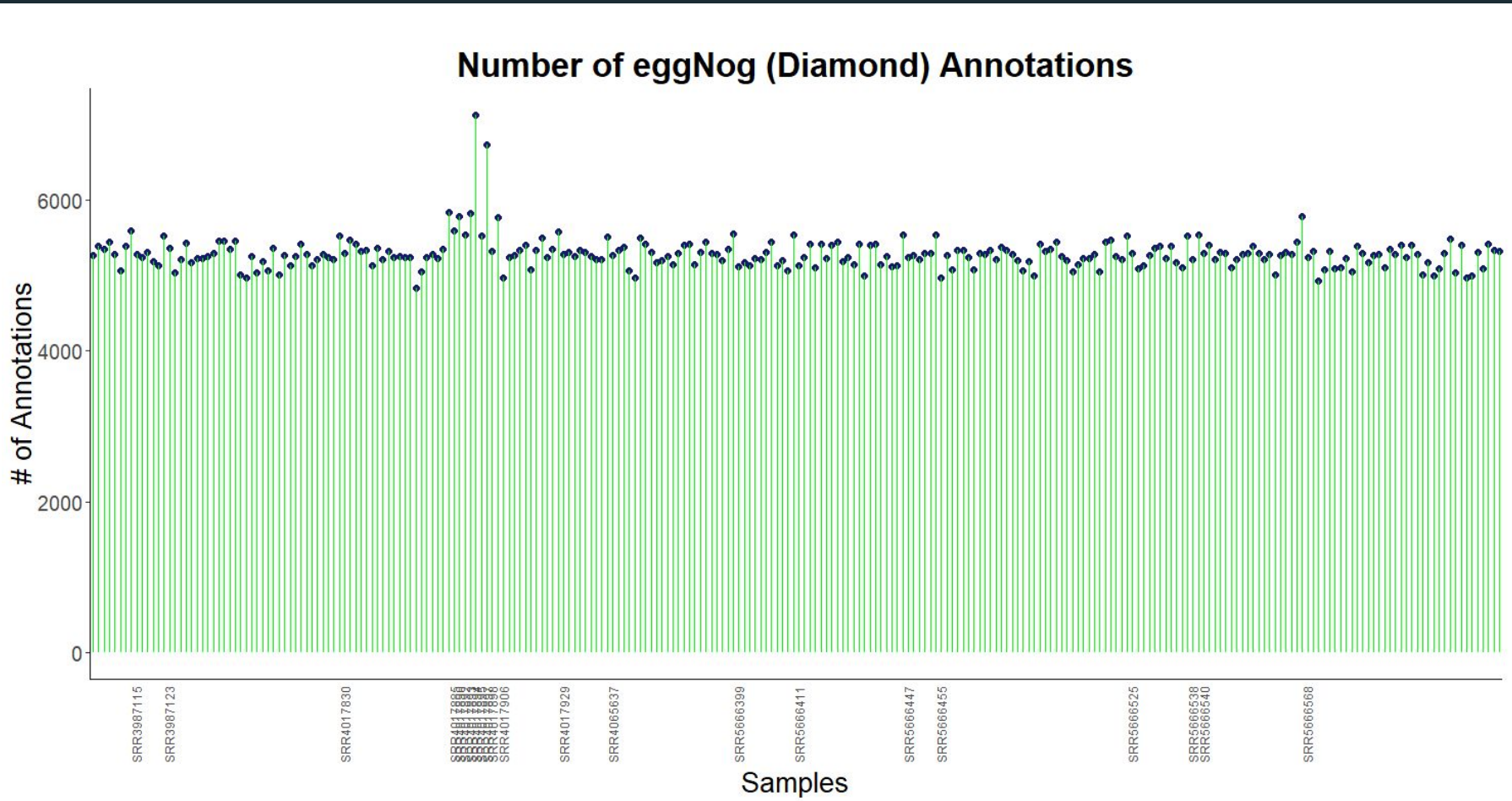
eggNog

Command:

time python emapper.py -i <input.faa> --output test1_faa_maNOG -m diamond --translate --usemem --cpu 4

Number of annotations for clustered centroids: 49353 (total sequences for analysis: 63127)

Running time: ~2h


Number of eggNog (Diamond) Annotations
# of Annotations
6000
4000
2000
0
SRR3987115
SRR3987123
SRR4017830
SRR4017929
SRR4065637
SRR5666399
SRR5666411
SRR5666447
SRR5666455
SRR5666525
SRR5666538
SRR5666540
SRR5666568
Samples

eggNog vs. InterProScan

- eggNog able to perform just as well in terms of proteome coverage and precision
- 2.5 times faster than InterProScan
- Annotation sources kept up-to-date

Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper

Jaime Huerta-Cepas,[†,1] Kristoffer Forslund,[†,1] Luis Pedro Coelho,[1] Damian Szklarczyk,[2,3] Lars Juhl Jensen,[4] Christian von Mering,[2,3] and Peer Bork*[,1,5,6,7]

CARD

Input: Protein sequence collection of all samples (mergedProtein.fasta)

Command used:

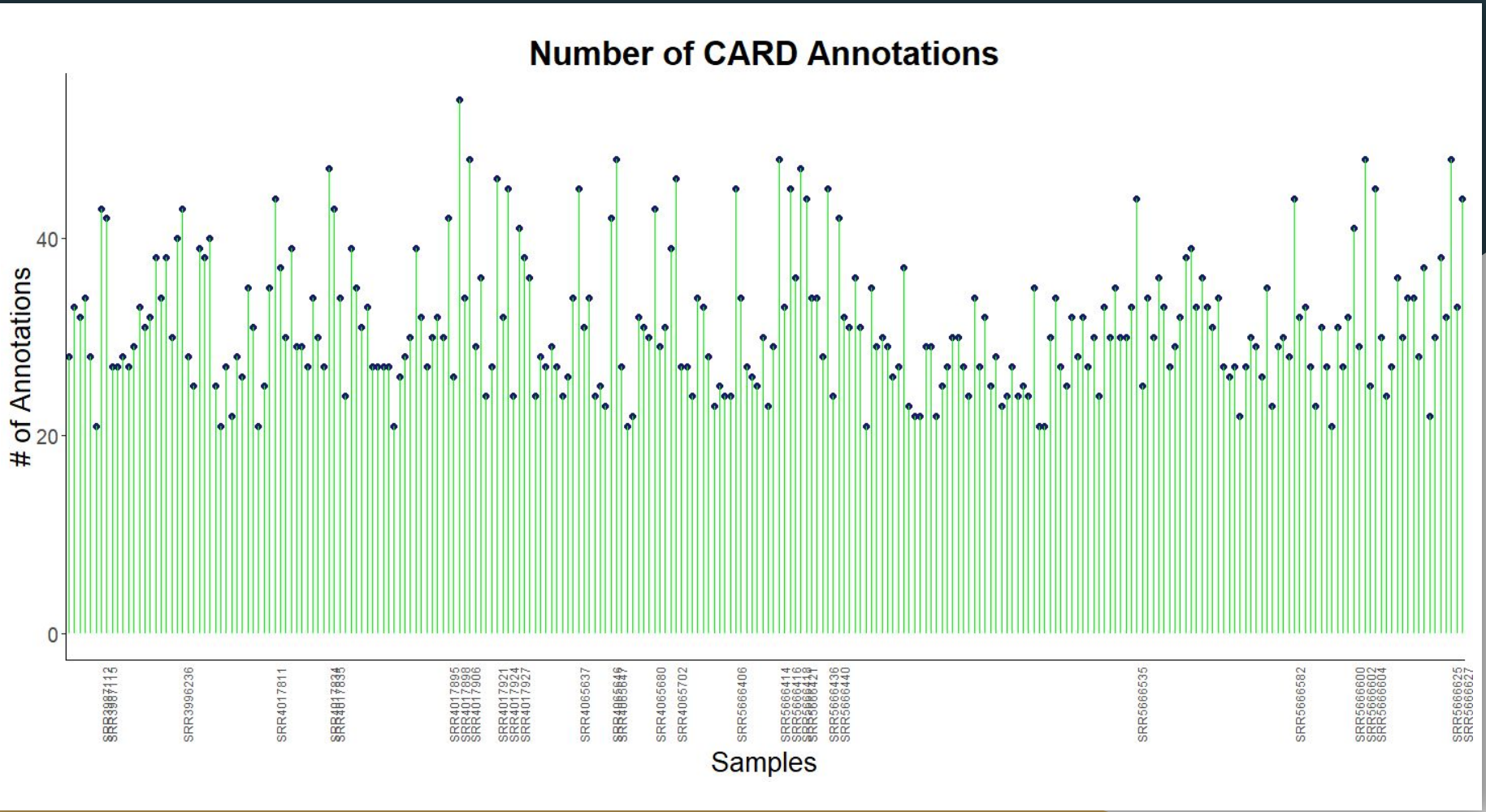
```
time blastp -query ../final_fna_faa/mergedProtein.fasta -db protein_fasta_protein_homolog_model.fasta -outfmt "6 qseqid qstart qend qlen length qcovs pident evalue stitle" -max_hsps 1 -max_target_seqs 1 -num_threads 5 > CARD_larger_top1_nucl.out

awk '{if($6>=90 && $7>=90) print $0}' CARD_larger_top1_nucl.out >card_90_90.out
```

Size of database: 2239

Number of queries return >90% coverage and >90% identity: 8051 (# of queries: 1,501,569)

Running time: 248 mins


Number of CARD Annotations
# of Annotations
0
20
40
Samples
SRR3996236
SRR4017811
SRR4017895
SRR4017898
SRR4017906
SRR4017921
SRR4017924
SRR4017927
SRR4065637
SRR4065680
SRR4065702
SRR5666406
SRR5666414
SRR5666416
SRR5666436
SRR5666440
SRR5666535
SRR5666582
SRR5666600
SRR5666602
SRR5666604
SRR5666625
SRR5666627

VFDB

Input: Nucleotide sequence collection of all samples (mergedNucleotide.fasta)

Command used:

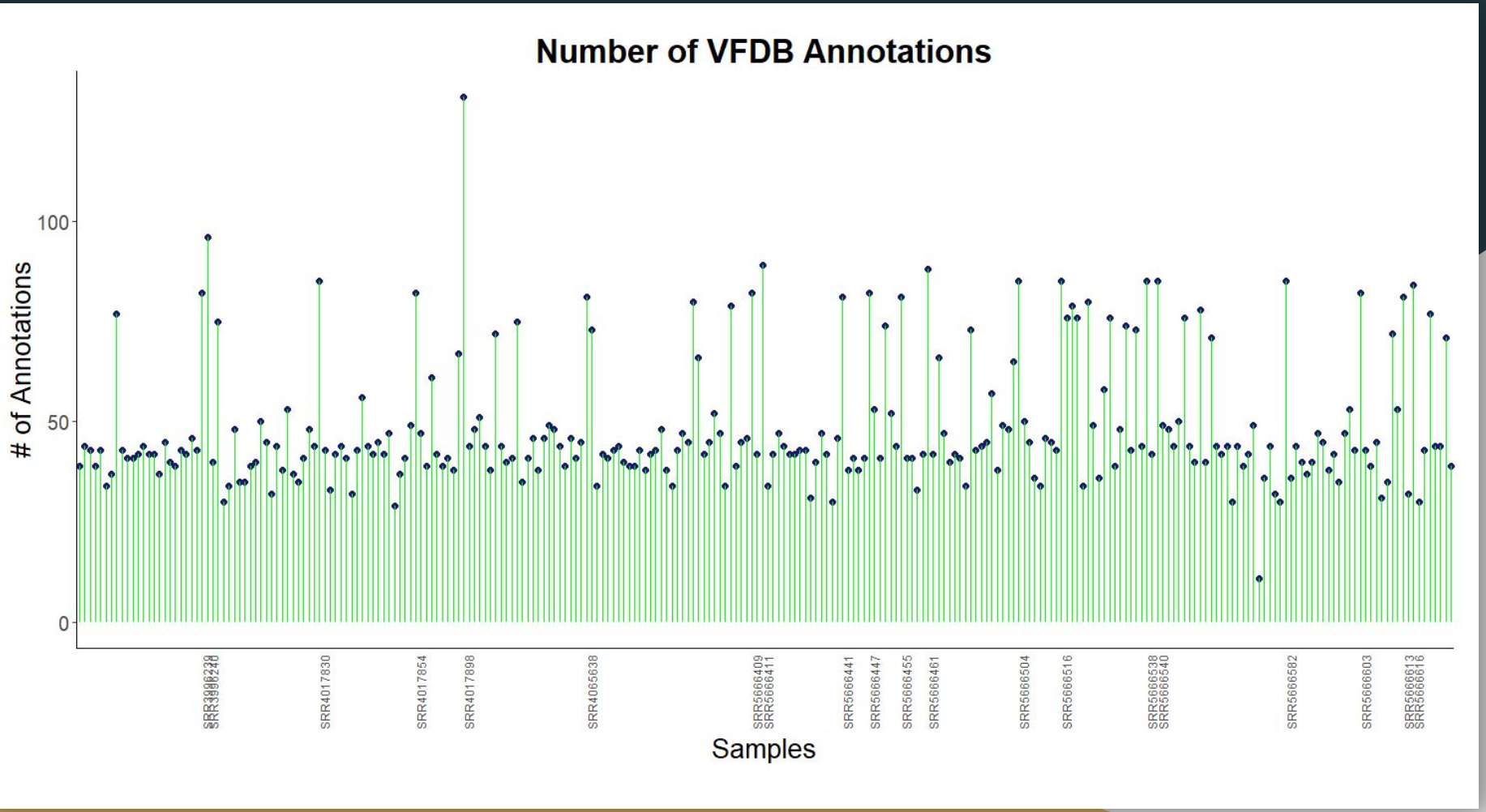
time blastn -query <input.fna> -db ./VFDB_setB_nt.fas -outfmt "6 qseqid qstart qend qlen length qcovs pident evalue stitle" -dust no -max_hsps 1 -max_target_seqs 1 > VFDB_setB_top1_nucl.out

awk '{if($6>=90 && $7>=90) print $0}' CARD_larger_top1_nucl.out >card_90_90.out

Size of database: 30318

Number of queries return >90% coverage and >90% identity: 12,517(# of queries: 1,501,569)

Running time: ~7 mins


Number of VFDB Annotations
# of Annotations
100
50
0
SRR4017830
SRR4017854
SRR4017898
SRR4065638
SRR5666409
SRR5666411
SRR5666441
SRR5666447
SRR5666455
SRR5666461
SRR5666504
SRR5666516
SRR5666538
SRR5666540
SRR5666582
SRR5666603
Samples

DOOR^2

Input: Protein sequence collection of clustered centroids (centroidsProtein.fasta)

Command used:

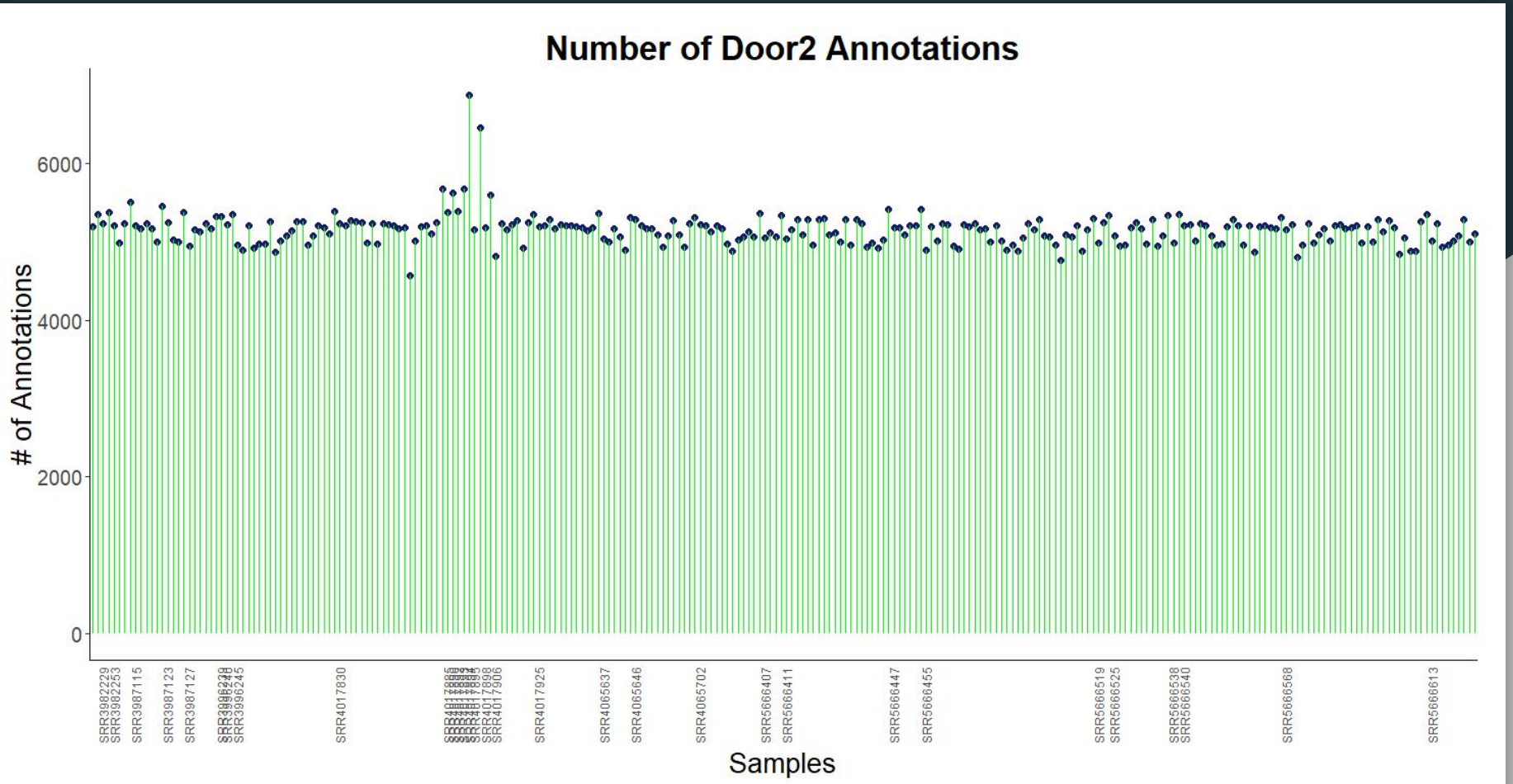
blastp -db kop_final.table -query ../centroidsProtein.fasta -out $out -max_target_seqs 1 -max_hsps 1 -num_threads 6 -outfmt “6 qseqid sseqid qstart qend evalue pident qcovs” > operon_intermediate.txt

Created a perl script that uses other files to link the blast output sseqid to its GID located in the the DOOR^2 operon tables and adds the operon id, cog, and product to the blast results

Size of database: 97073

Number of queries return >=90% coverage and >=90% identity: 38469(# of queries: 63127)

Running time: ~270 minutes


Number of Door2 Annotations
# of Annotations
6000
4000
2000
0
SRR3982229
SRR3982253
SRR3987115
SRR3987123
SRR3987127
SRR3996245
SRR4017830
SRR4017925
SRR4065637
SRR4065646
SRR4065702
SRR5666407
SRR5666411
SRR5666447
SRR5666455
SRR5666519
SRR5666525
SRR5666538
SRR5666540
SRR5666568
SRR5666613
Samples

DeepARG

Output files: (% identity range: 30% to 100%; e-value < 1e-10)

1) ARG: probability >= 0.8; best-hit: found in DeepARG-DB (from CARD, ARDB, and UNIPROT)
2) Potential ARG: a) probability >= 0.8; best-hit: “undefined” (one of the 30 categories defined by the deepARG model).
   b) probability < 0.8

Command used:

time python ../deeparg-ss/deepARG.py --align --type nucl --genes --input mergedNucleotide.fasta --out ./raw_fasta_gene/larged_merged_nucl.out

Running time: 44 mins

DeepARG

Output files: (% identity range: 30% to 100%; e-value < 1e-10)

1) ARG: probability >= 0.8; best-hit: found in DeepARG-DB (from CARD, ARDB, and UNIPROT)
2) Potential ARG: a) probability >= 0.8; best-hit: undefined (one of the 30 categories defined by the deepARG model.
b) probability < 0.8

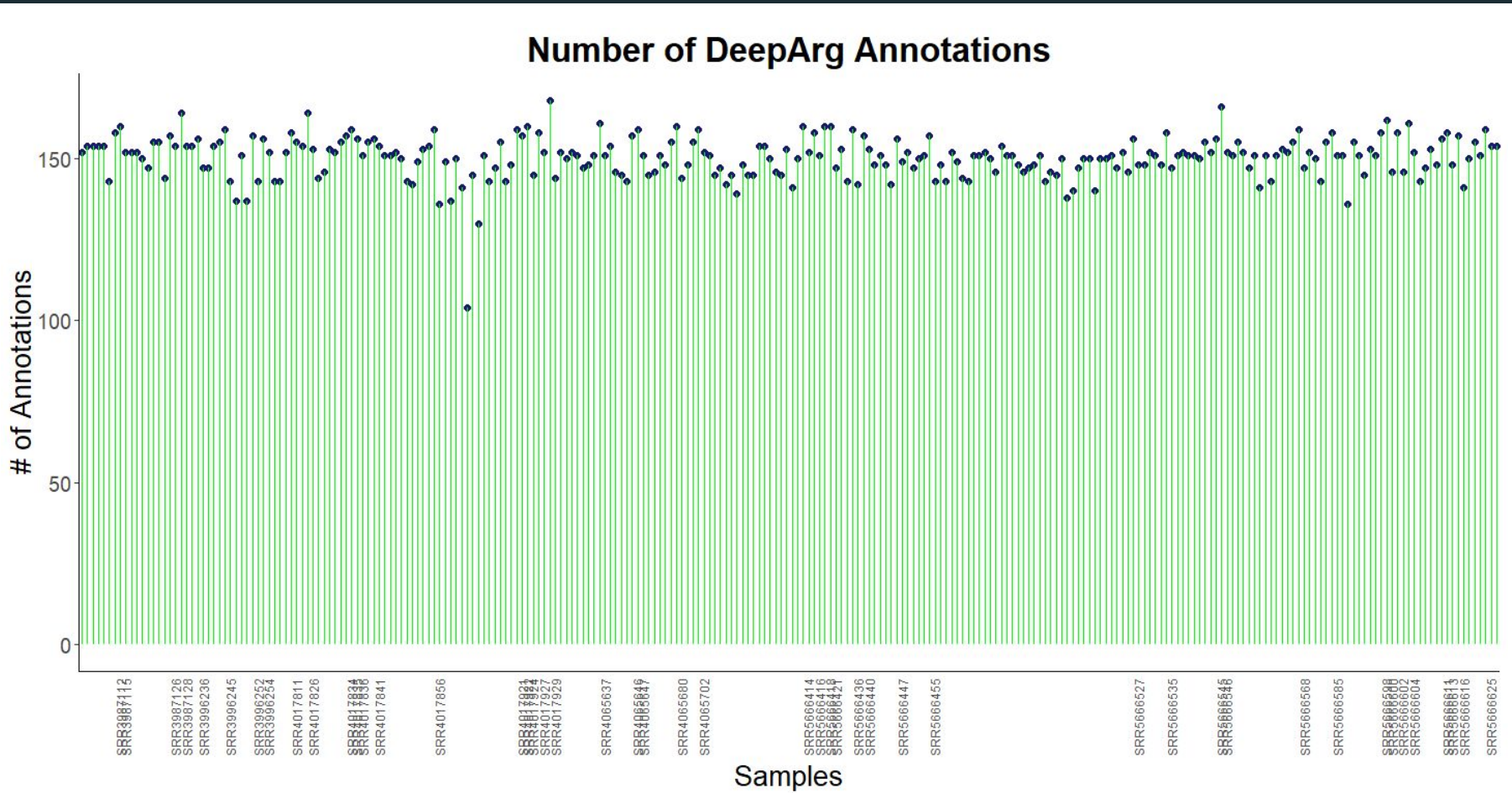
% identity distribution	ARG	Potential ARG
[30, 50)	30070 (77.5%)	43960 (94.7%)
[50, 70)	4827 (12.4%)	861 (1.9%)
[70, 90)	3201 (8.3%)	1077 (2.3%)
[90, 100]	701 (1.8%)	538 (1.1%)
Total	38799 (100%)	46436 (100%)

	ARG	Potential ARG
Total ARGs	38799 (100%)	46436 (100%)
Unique ARGs	261	67
Average number	~149	~693

DeepARG

Interesting results: “fosfomycin” is one of the thirty categories generated by deepARG (already trained).

```
qzhuang8@ubuntu:~/7210/func_annotation/final_fna_faa$ grep 'colistin' ./raw_fasta_gene/larged_merged_nucl.out.mapping.potential.ARG |wc -l
0
qzhuang8@ubuntu:~/7210/func_annotation/final_fna_faa$ grep 'colistin' ./raw_fasta_gene/larged_merged_nucl.out.mapping.potential.ARG |wc -l
0
qzhuang8@ubuntu:~/7210/func_annotation/final_fna_faa$ grep 'fosfomycin' ./raw_fasta_gene/larged_merged_nucl.out.mapping.ARG |wc -l
514
qzhuang8@ubuntu:~/7210/func_annotation/final_fna_faa$ grep 'fosfomycin' ./raw_fasta_gene/larged_merged_nucl.out.mapping.potential.ARG |wc -l
0
```


Number of DeepArg Annotations
# of Annotations
150
100
50
0
SRR3987126
SRR3987128
SRR3996236
SRR3996245
SRR3996252
SRR3996254
SRR4017811
SRR4017826
SRR4017841
SRR4017856
SRR4017927
SRR4017929
SRR4065637
SRR4065680
SRR4065702
SRR5666436
SRR5666440
SRR5666447
SRR5666455
SRR5666527
SRR5666535
SRR5666568
SRR5666585
SRR5666616
SRR5666625
Samples

Ab Initio Functional Annotation

- Looks for intrinsic characteristics of particular gene feature types
- Signal Peptide and Transmembrane Proteins can be identified in this way
    - These regions are of particular importance to this project because of their significance to AMR

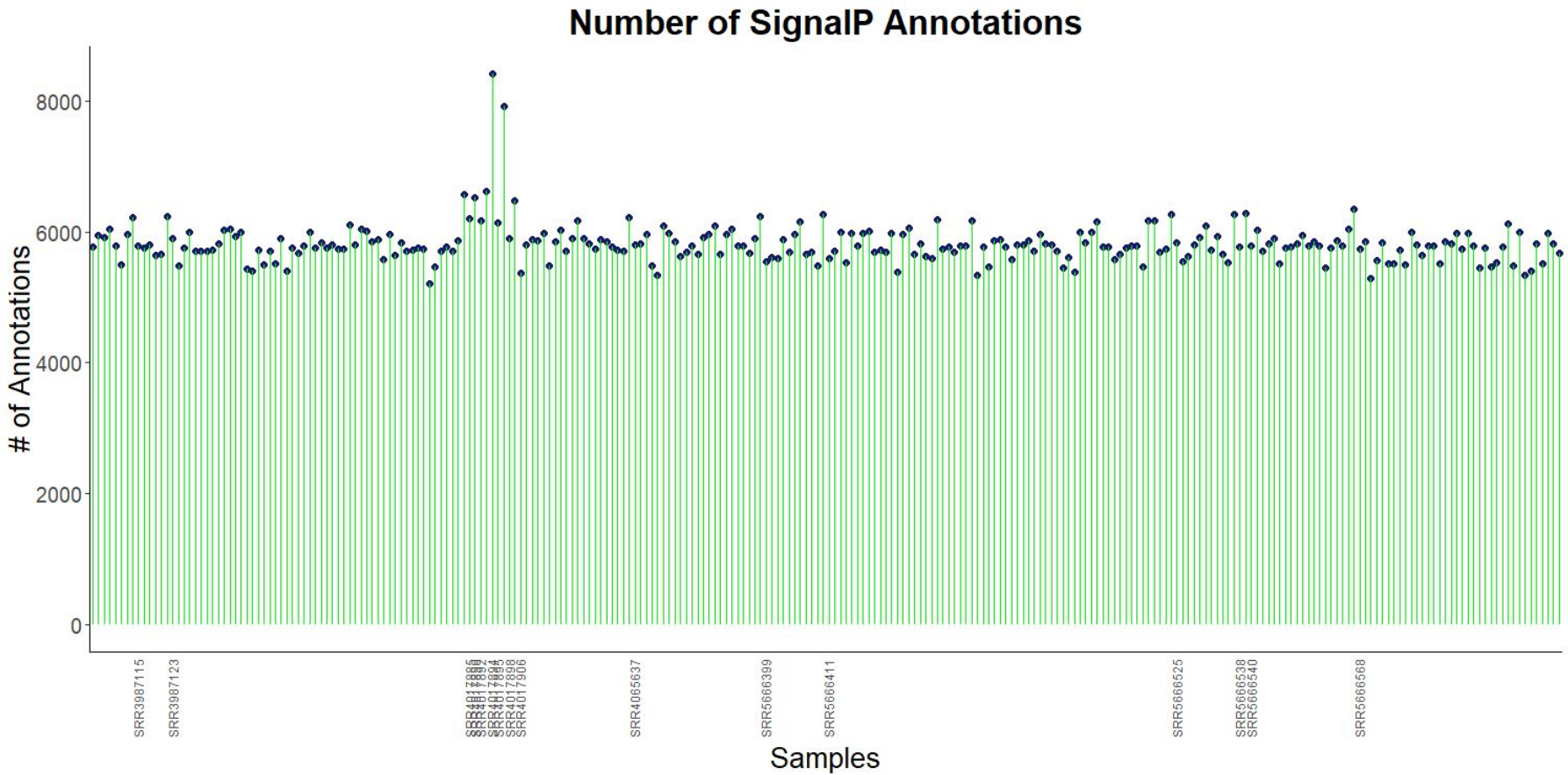
SignalP

Input: Protein sequence collection of clustered centroids (centroidsProtein.fasta)

Algorithm: combination of several trained neural networks

Command used:

signalp -t <organism_type> -f <output_format> <input_file> > <output_file>


Number of SignalP Annotations
# of Annotations
8000
6000
4000
2000
0
SRR3987115
SRR3987123
SRR4065637
SRR5666399
SRR5666411
SRR5666525
SRR5666538
SRR5666540
SRR5666568
Samples

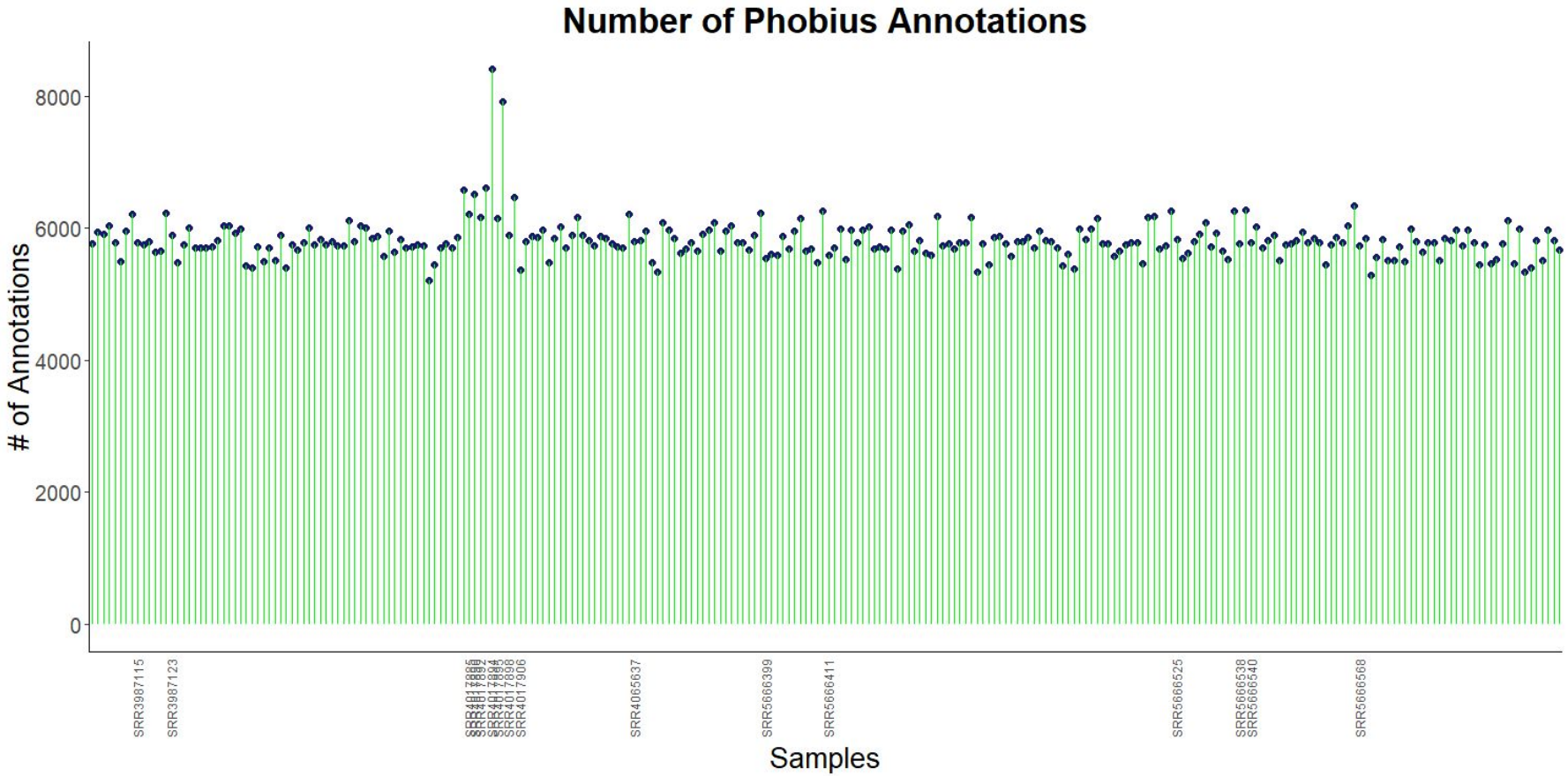
Phobius

Input: Protein sequence collection of clustered centroids (centroidsProtein.fasta)

Algorithm: Hidden Markov Model

Command used:

phobius.pl -<output_format> <input_file> > <output_file>


Number of Phobius Annotations
# of Annotations
8000
6000
4000
2000
0
SRR3987115
SRR3987123
SRR4065637
SRR5666399
SRR5666411
SRR5666525
SRR5666538
SRR5666540
SRR5666568
Samples

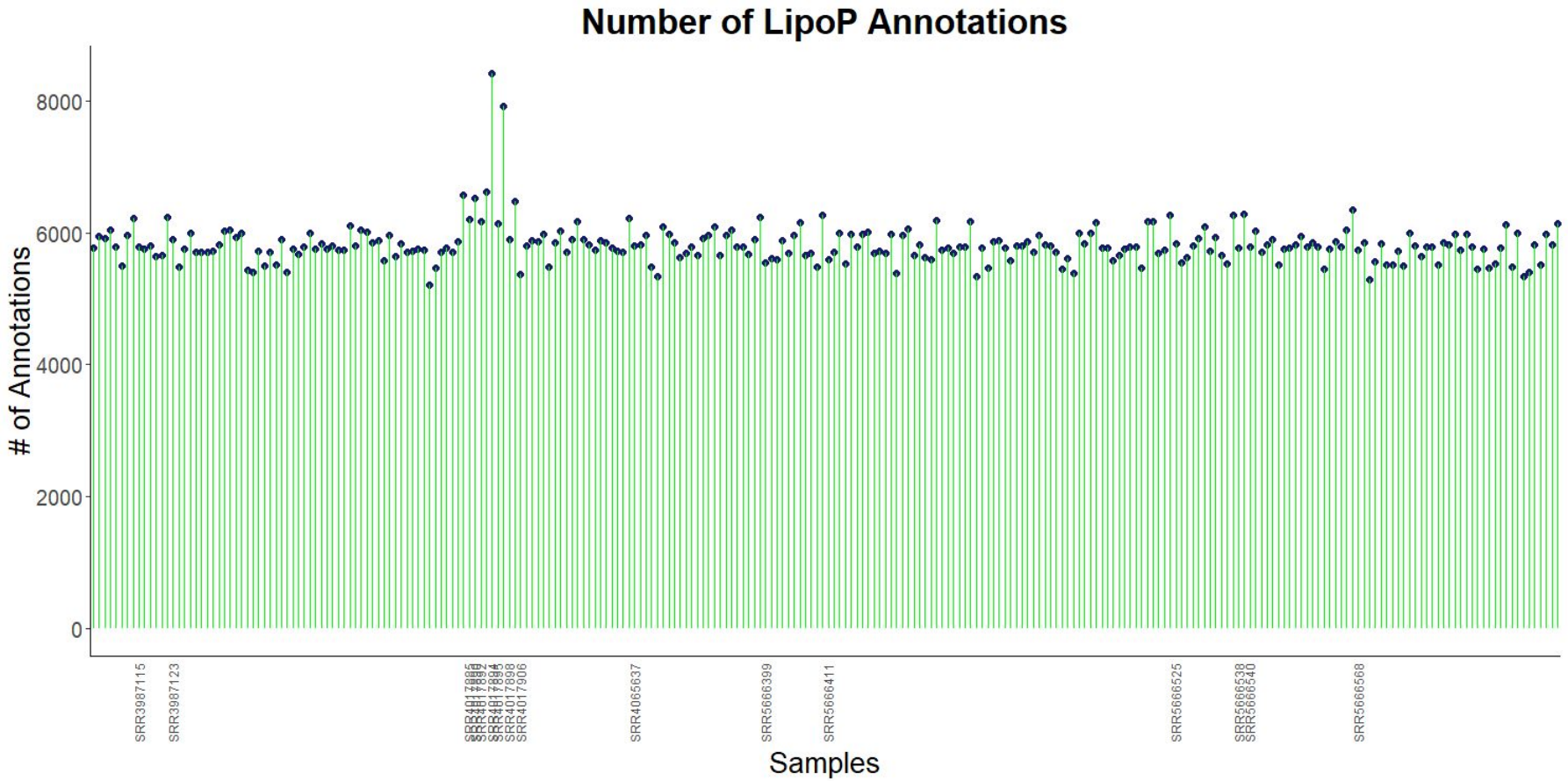
LipoP

Input: Protein sequence collection of all samples (mergedProtein.fasta)

Algorithm: Hidden Markov Model

Command used:

LipoP -<output_format> -<input_file> **>** <output_file>


Number of LipoP Annotations
# of Annotations
8000
6000
4000
2000
0
SRR3987115
SRR3987123
SRR4017898
SRR4017906
SRR4065637
SRR5666399
SRR5666411
SRR5666525
SRR5666568
Samples

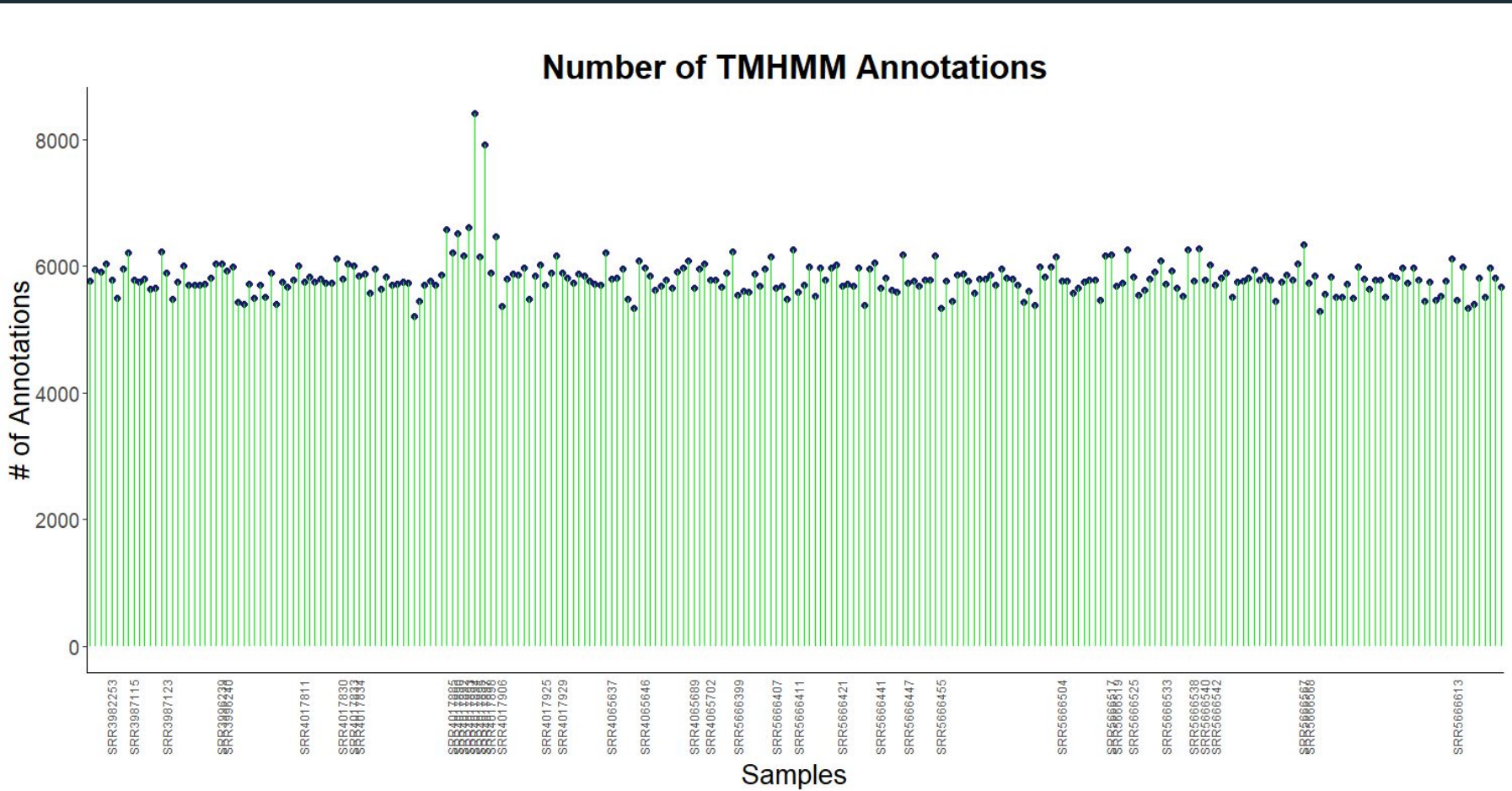
TMHMM

Input: Protein sequence collection of clustered centroids (centroidsProtein.fasta)

Algorithm: Hidden Markov Model

Command used:

tmhmm -<output_format> <input_file> **>** <output_file>


Number of TMHMM Annotations
# of Annotations
8000
6000
4000
2000
0
SRR3982253
SRR3987115
SRR3987123
SRR4017811
SRR4017830
SRR4017834
SRR4017906
SRR4017925
SRR4017929
SRR4065637
SRR4065646
SRR4065689
SRR4065702
SRR5666399
SRR5666407
SRR5666411
SRR5666421
SRR5666441
SRR5666447
SRR5666455
SRR5666504
SRR5666525
SRR5666533
SRR5666538
SRR5666540
SRR5666542
SRR5666613
Samples

Overall statistics

Tools / Database	Total # of annotations	Average # of annotations
CARD	8,051	31
DeepARG	38,799	150
Door2	1,330,879	5158
Eggnog (diamond)	1,364,546	5289
GeneMark.hmm	110,235	427
LipoP	1,502,024	5822
Phobius	1,501,560	5820
Prodigal	1,391,789	5395
SignalP	1,501,569	5820
TMHMM	1,501,577	5820
VFDB	12,517	49