Functional Annotation

•••

Preliminary 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
- Tools
  - Ab-initio
  - Homology based

Approach must be scalable!

Scalability

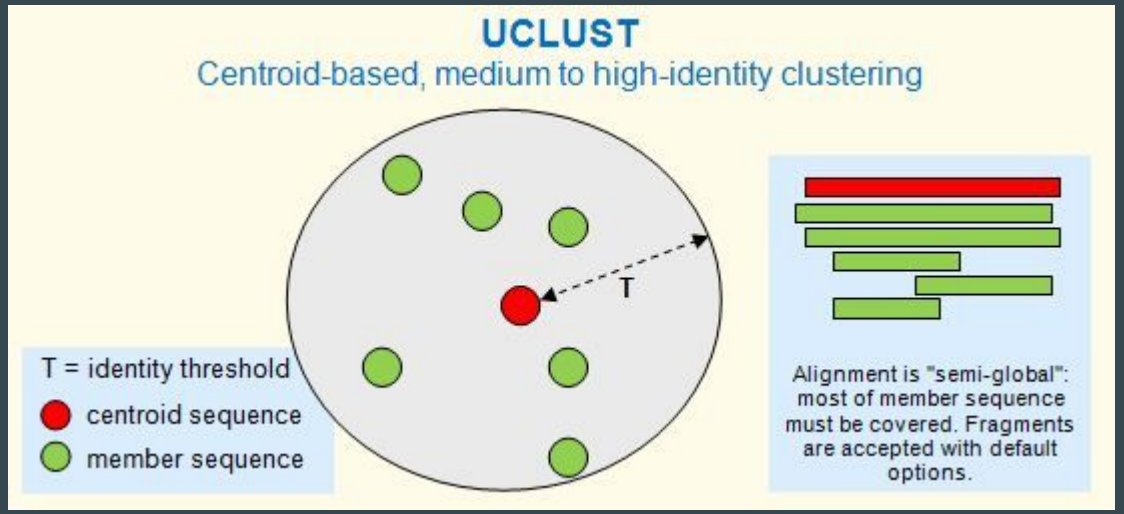
- Minimize number of searches/operations
  - Reduce query size
  - Reduce database size

- Our dataset is highly redundant

Clustering

Scalability by Clustering

Genes From Many Genomes →



↙

Clustered into Homologues

↓

Functional Annotation of Homologue Clusters →

Map Results Back to Genomes

Clustering

5 Reference Proteomes
~8Mb

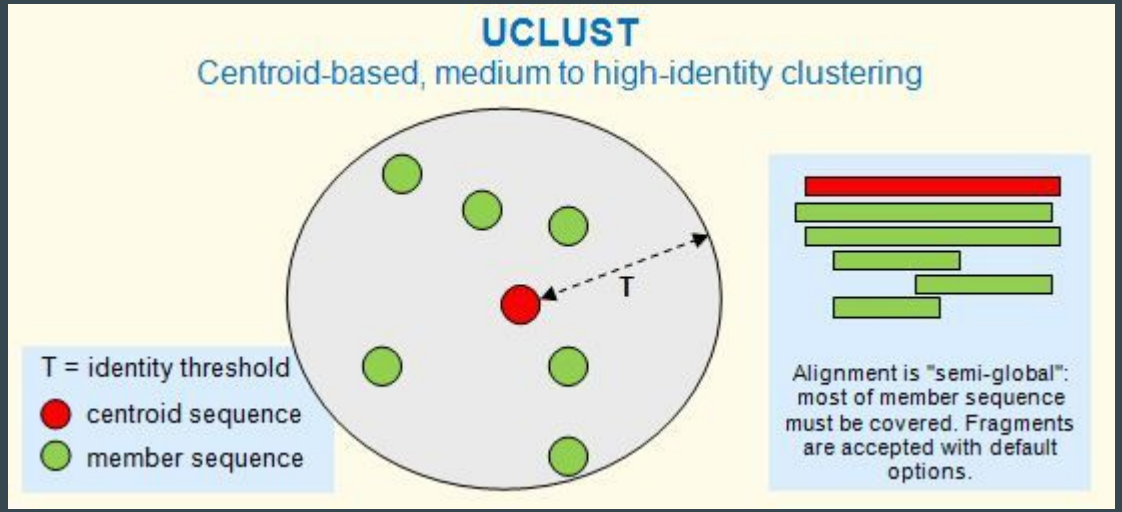
After Clustering
~2.5Mb

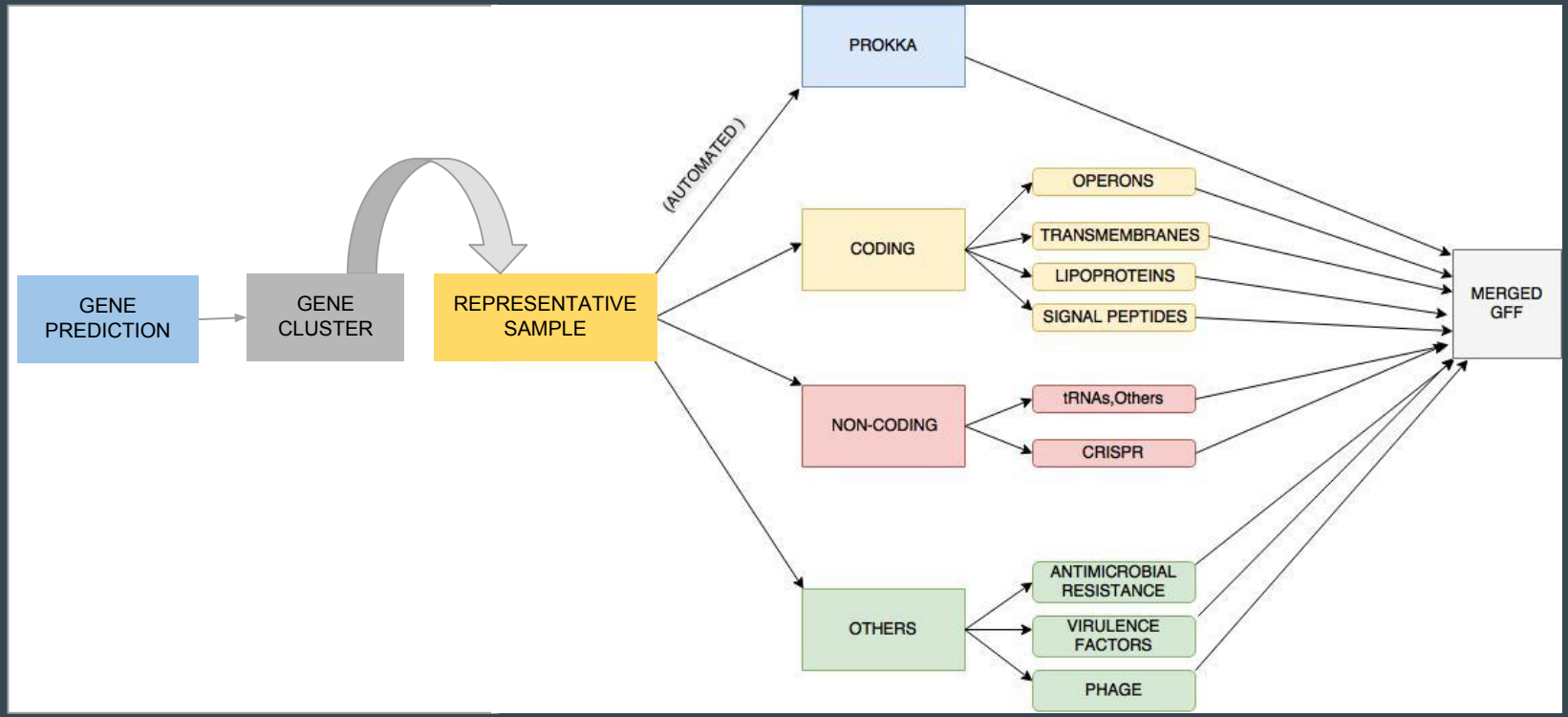
(should not change too dramatically with additional genomes)

DB	Method	Seqs./ sec.	Sens (Err) all	Sens (Err) Med.	Sens (Err) Low
Pfam	ublast	55	97 (2)	96 (1)	75 (5)
	usearch	77	92 (3)	70 (3)	22 (8)
	blastp	0.16	99 (2)	96 (3)	98 (3)
Rfam	ublast	684	99.4 (0.6)	–	96 (4)
	usearch	671	99.3 (0.7)	–	93 (7)
	blastn	2.3	99.4 (0.6)	–	97 (3)
	mega-blast.	8.8	73 (0.2)	–	35 (0)

Clustering

Homology vs. ab-initio




GENE PREDICTION
GENE CLUSTER
REPRESENTATIVE SAMPLE
(AUTOMATED)
PROKKA
CODING
NON-CODING
OTHERS
OPERONS
TRANSMEMBRANES
LIPOPROTEINS
SIGNAL PEPTIDES
tRNAs,Others
CRISPR
ANTIMICROBIAL RESISTANCE
VIRULENCE FACTORS
PHAGE
MERGED GFF

PROKKA

Command: prokka --force --outdir </path/to/output/dir> --kingdom Bacteria --genus Klebsiella --gram neg --prefix <prefix name> --rfam --rnammer <query.fasta>

Summary output:
organism: *Klebsiella* species strain
contigs: 87
bases: 5,570,351
tmRNA: 1 (Aragorn also found 1 tmRNA)
CDS: 5,282
misc_RNA: 129
tRNA: 77
sig_peptide: 500

Time: ~ 16 mins.

PROKKA

```
>Feature scaffold1_size427644
48      284     CDS
                        EC_number       1.-.-.-
                        gene    yghA
                        inference       ab initio prediction:Prodigal:2.6
                        inference       similar to AA sequence:UniProtKB:P0AG84
                        locus_tag       EJJAIGGM_00001
                        product putative oxidoreductase YghA
481     726     CDS
                        inference       ab initio prediction:Prodigal:2.6
                        locus_tag       EJJAIGGM_00002
                        product hypothetical protein
1687    794     CDS
                        gene    dmlR_1
                        inference       ab initio prediction:Prodigal:2.6
                        inference       similar to AA sequence:UniProtKB:P76250
                        locus_tag       EJJAIGGM_00003
                        product HTH-type transcriptional regulator DmlR
1801    2652    CDS
                        inference       ab initio prediction:Prodigal:2.6
                        locus_tag       EJJAIGGM_00004
                        product hypothetical protein
1801    1860    sig_peptide
                        inference       ab initio prediction:SignalP:4.1
                        note    predicted cleavage at residue 20
                        product putative signal peptide
2830    2970    CDS
                        gene    sra
                        inference       ab initio prediction:Prodigal:2.6
                        inference       similar to AA sequence:UniProtKB:P68191
                        locus_tag       EJJAIGGM_00005
                        product Stationary-phase-induced ribosome-associated protein
3152    4849    CDS
                        EC_number       1.1.1.38
                        gene    maeA
                        inference       ab initio prediction:Prodigal:2.6
                        inference       similar to AA sequence:UniProtKB:P26616
                        locus_tag       EJJAIGGM_00006
                        product NAD-dependent malic enzyme
5030    5737    CDS
                        inference       ab initio prediction:Prodigal:2.6
                        locus_tag       EJJAIGGM_00007
                        product hypothetical protein
6134    7849    CDS
                        gene    yheS_1
                        inference       ab initio prediction:Prodigal:2.6
                        inference       similar to AA sequence:UniProtKB:P63389
                        locus_tag       EJJAIGGM_00008
                        product putative ABC transporter ATP-binding protein YheS
```

locus_tag	ftype	gene	EC_number	product
EJJAIGGM_00001	CDS	yghA	1.-.-.-	putative oxidoreductase YghA
EJJAIGGM_00002	CDS			hypothetical protein
EJJAIGGM_00003	CDS	dmlR_1		HTH-type transcriptional regulator DmlR
EJJAIGGM_00004	CDS			hypothetical protein
	sig_peptide			putative signal peptide
EJJAIGGM_00005	CDS	sra		Stationary-phase-induced ribosome-associated protein
EJJAIGGM_00006	CDS	maeA	1.1.1.38	NAD-dependent malic enzyme
EJJAIGGM_00007	CDS			hypothetical protein
EJJAIGGM_00008	CDS	yheS_1		putative ABC transporter ATP-binding protein YheS
EJJAIGGM_00009	CDS			hypothetical protein
EJJAIGGM_00010	CDS	ppa_1	3.6.1.1	Inorganic pyrophosphatase
	sig_peptide			putative signal peptide
EJJAIGGM_00011	CDS	adhP	1.1.1.1	Alcohol dehydrogenase, propanol-preferring
EJJAIGGM_00012	CDS	gsiB_1		Glutathione-binding protein GsiB
	sig_peptide			putative signal peptide
EJJAIGGM_00013	CDS	gsiC_1		Glutathione transport system permease protein GsiC
EJJAIGGM_00014	CDS	ddpC_1		putative D,D-dipeptide transport system permease protein DdpC
EJJAIGGM_00015	CDS	gsiA_1	3.6.3.-	Glutathione import ATP-binding protein GsiA
EJJAIGGM_00016	CDS	luxA	1.14.14.3	Alkanal monooxygenase alpha chain
EJJAIGGM_00017	CDS			hypothetical protein
EJJAIGGM_00018	CDS	ydhP_1		Inner membrane transport protein YdhP
EJJAIGGM_00019	CDS			hypothetical protein
EJJAIGGM_00020	CDS			hypothetical protein
EJJAIGGM_00021	CDS	fdnI		Formate dehydrogenase, nitrate-inducible, cytochrome b556(Fdn) subunit
EJJAIGGM_00022	CDS	fdnH		Formate dehydrogenase, nitrate-inducible, iron-sulfur subunit
EJJAIGGM_00023	CDS	fdnG_1	1.1.5.6	Formate dehydrogenase, nitrate-inducible, major subunit
EJJAIGGM_00024	misc_RNA			SECIS_3

PROKKA output in the .gff format (left) and .tsv (right).

SPECIFIC TOOLS

...

Specific Tools (Based on features to be annotated)

- Protein-coding regions
  - Signaling peptides
  - Transmembrane regions
  - Lipoproteins
  - 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)

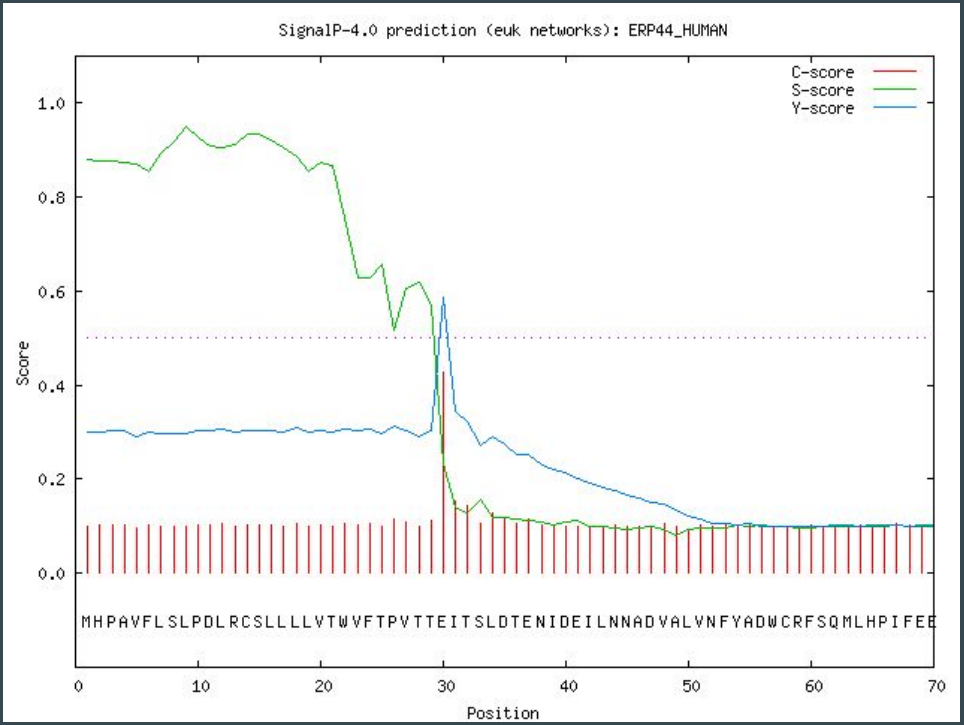
- Protein-coding regions
  - Signaling peptides
  - Transmembrane regions
  - Lipoproteins
  - Operons
  - Pathways

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

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

Signal Peptides Prediction - SignalP 4.1

- Command to run
  - ./signalp -t gram- [-f short/long/all/summary] input_file.faa > output_file.out



Signal Peptide Prediction - SignalP 4.1

Sample Output File(Short/Summary)

- Sample Results
  - Input Reference Genome: 5075 sequences
  - #Predicted Signal Peptides: 489
  - Time taken for execution: ~4 minutes

```
# SignalP-4.1 gram- predictions
# name                Cmax  pos  Ymax  pos  Smax  pos  Smean   D     ?  Dmaxcut    Networks-used
NC_017540.1_1         0.107  37  0.113  16  0.151   4  0.126   0.119 N  0.570      SignalP-noTM
NC_017540.1_2         0.108  23  0.131  31  0.242  27  0.137   0.134 N  0.570      SignalP-noTM
NC_017540.1_3         0.131  44  0.117  44  0.117  42  0.092   0.105 N  0.570      SignalP-noTM
NC_017540.1_4         0.123  35  0.123  35  0.133  26  0.107   0.115 N  0.570      SignalP-noTM
NC_017540.1_5         0.116  20  0.124  20  0.189  13  0.122   0.123 N  0.570      SignalP-noTM
NC_017540.1_6         0.142  48  0.139  11  0.263   3  0.196   0.166 N  0.570      SignalP-noTM
NC_017540.1_7         0.106  30  0.111  11  0.200   2  0.122   0.116 N  0.570      SignalP-noTM
NC_017540.1_8         0.130  24  0.185  11  0.455   1  0.349   0.246 N  0.510      SignalP-TM
NC_017540.1_9         0.114  21  0.138  11  0.279   1  0.178   0.157 N  0.570      SignalP-noTM
NC_017540.1_10        0.140  48  0.143  11  0.318   2  0.202   0.171 N  0.570      SignalP-noTM
NC_017540.1_11        0.156  24  0.153  24  0.241   3  0.167   0.160 N  0.570      SignalP-noTM
NC_017540.1_12        0.151  26  0.164  26  0.232  20  0.161   0.163 N  0.510      SignalP-TM
NC_017540.1_13        0.223  28  0.157  28  0.164  48  0.108   0.134 N  0.570      SignalP-noTM
NC_017540.1_14        0.201  41  0.194  16  0.483  13  0.359   0.272 N  0.570      SignalP-noTM
NC_017540.1_15        0.171  25  0.176  25  0.277   1  0.200   0.187 N  0.570      SignalP-noTM
NC_017540.1_16        0.135  30  0.126  30  0.191  29  0.095   0.111 N  0.570      SignalP-noTM
NC_017540.1_17        0.108  25  0.109  13  0.146   3  0.121   0.114 N  0.570      SignalP-noTM
NC_017540.1_18        0.129  38  0.117  16  0.165  12  0.133   0.124 N  0.570      SignalP-noTM
NC_017540.1_19        0.182  21  0.247  21  0.432  20  0.317   0.273 N  0.510      SignalP-TM
NC_017540.1_20        0.155  53  0.137  53  0.227  14  0.113   0.128 N  0.510      SignalP-TM
NC_017540.1_21        0.103  33  0.101  70  0.151   2  0.085   0.093 N  0.570      SignalP-noTM
NC_017540.1_22        0.148  27  0.136  27  0.151  26  0.117   0.127 N  0.570      SignalP-noTM
NC_017540.1_23        0.110  18  0.095  68  0.094  59  0.076   0.086 N  0.570      SignalP-noTM
NC_017540.1_24        0.114  19  0.134  19  0.193   5  0.156   0.144 N  0.570      SignalP-noTM
NC_017540.1_25        0.210  29  0.163  29  0.388   1  0.153   0.159 N  0.510      SignalP-TM
NC_017540.1_26        0.182  35  0.195  11  0.457   2  0.399   0.271 N  0.510      SignalP-TM
NC_017540.1_27        0.209  37  0.239  37  0.449  28  0.230   0.234 N  0.570      SignalP-noTM
NC_017540.1_28        0.109  44  0.147  16  0.283   8  0.209   0.169 N  0.510      SignalP-TM
NC_017540.1_29        0.128  37  0.152  11  0.305   3  0.247   0.187 N  0.510      SignalP-TM
NC_017540.1_30        0.187  65  0.110  15  0.166   9  0.121   0.114 N  0.510      SignalP-TM
NC_017540.1_31        0.712  27  0.771  27  0.959  11  0.877   0.821 Y  0.570      SignalP-noTM
NC_017540.1_32        0.109  15  0.105  59  0.140   1  0.097   0.101 N  0.570      SignalP-noTM
NC_017540.1_33        0.769  26  0.793  26  0.975  11  0.882   0.835 Y  0.570      SignalP-noTM
NC_017540.1_34        0.540  27  0.704  27  0.990  14  0.945   0.817 Y  0.570      SignalP-noTM
NC_017540.1_35        0.120  27  0.228  17  0.631  25  0.461   0.337 N  0.570      SignalP-noTM
```

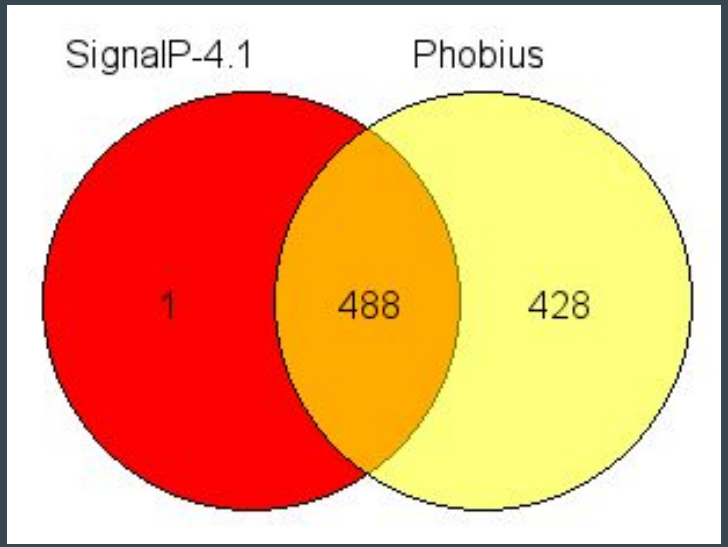
Signal Peptide Prediction - Phobius

- Input Format: FASTA file
- Sample Results
  - Input Reference Genome: 5075 sequences
  - #Predicted Signal Peptides: 916
  - Time taken for execution: ~13 minutes

Phobius prediction

SEQENCE ID	TM	SP	PREDICTION
NC_017540.1_1	0	0	i
NC_017540.1_2	0	0	i
NC_017540.1_3	0	0	i
NC_017540.1_4	0	0	i
NC_017540.1_5	0	0	i
NC_017540.1_6	0	0	i
NC_017540.1_7	0	0	i
NC_017540.1_8	8	Y	n2-13c17/18o45-64i76-98o104-121i133-151o171-194i206-226o238-262i274-295o
NC_017540.1_9	0	0	i
NC_017540.1_10	0	0	i
NC_017540.1_11	4	0	o122-146i186-206o226-247i254-273o
NC_017540.1_12	6	0	o6-24i74-92o98-120i141-162o177-197i209-239o
NC_017540.1_13	0	0	i
NC_017540.1_14	0	0	i
NC_017540.1_15	0	0	i
NC_017540.1_16	0	0	i
NC_017540.1_17	0	0	i
NC_017540.1_18	0	0	i
NC_017540.1_19	7	Y	n8-16c20/21o30-51i63-86o92-112i119-141o153-175i187-206o212-236i
NC_017540.1_20	0	0	i
NC_017540.1_21	0	0	i
NC_017540.1_22	0	0	i
NC_017540.1_23	0	0	i
NC_017540.1_24	0	0	i
NC_017540.1_25	3	0	o6-28i40-61o73-102i
NC_017540.1_26	6	0	i12-36o56-79i127-155o161-176i183-201o207-224i
NC_017540.1_27	1	0	i310-331o
NC_017540.1_28	4	0	i12-35o55-74i86-103o109-128i
NC_017540.1_29	6	0	i12-39o84-110i122-143o163-181i202-227o258-281i
NC_017540.1_30	8	0	i15-34o40-58i67-91o277-303i315-340o371-392i413-436o484-506i
NC_017540.1_31	0	Y	n9-17c26/27o

Signal Peptide Prediction - SignalP 4.1 vs Phobius



iii: *Gram-negative bacterial sequences*					
	All Sequences			Only TM	No TM
Method	SP corr.	CS sens. (%)	CS prec. (%)	FP-rate (%)	SP corr.
SignalP 4.0	**0.848**	65.4	70.8	1.5	0.882
SignalP-TM	0.815	61.5	**75.3**	**1.1**	0.839
SignalP-noTM	0.497	71.2	26.1	35.8	**0.948**
SignalP 3.0 NN	0.542	74.0	30.8	28.5	0.925
SignalP 3.0 HMM	0.477	**76.9**	26.1	39.2	0.931
PrediSi	0.479	75.0	27.2	35.6	0.901
SPEPlip	0.429	70.2	21.4	45.1	0.891
Signal-CF*	0.288	73.1	13.8	78.1	0.698
Signal-3L*	0.287	73.1	13.5	81.1	0.714
SignalBlast SP1	0.530	39.4	14.6	25.4	0.767
SignalBlast SP2	0.252	18.3	3.2	72.8	0.543
SignalBlast SP3	0.642	34.6	22.8	11.5	0.836
SignalBlast SP4	0.387	39.4	9.4	46.1	0.635
Phobius	0.586	73.1	33.6	23.3	0.920
Philius	0.639	**76.9**	26.1	15.7	0.872
MEMSAT3	0.084	0.0	0.0	17.8	0.312
MEMSAT-SVM	0.497	1.0	0.6	16.4	0.780
SPOCTOPUS	0.510	33.7	18.6	20.5	0.848

Table E. Benchmarking of signal peptide and cleavage site predictions on the comparison dataset for all three organism groups. 'SP corr.' denotes signal peptide correlation, while 'CS sens.' denotes cleavage site sensitivity (the percentage of actual cleavage sites that are predicted correctly) and 'CS prec.' denotes cleavage site precision (the percentage of predicted cleavage sites that are correct). 'FP-rate' (false positive rate) is the percentage of transmembrane sequences that are incorrectly predicted as a signal peptide. 'No TM' denotes the test where there were no transmembrane sequences in the data, *i.e.* only the first negative set was used. Note that performance for SignalP 4.0 and SignalP-TM are identical for Gram-positive bacterial sequences, since SignalP 4.0 does not use the combination scheme for this organism group. The methods indicated with a star (*) can only make predictions for sequences longer than 50 aa. For those methods the evaluation sets were reduced by 4, 2, and 22 sequences for the Eukaryote, Gram-positive and Gram-negative sets, respectively.

LipoP

Prediction of lipoproteins in Gram-negative bacteria.

Based on hidden Markov model

Command: perl LipoP -short <Input FASTA>

Example Results:

```
# scaffold1|size403193 SpI score=17.1976 margin=10.92408 cleavage=27-28
# scaffold2|size340902 SpI score=15.5525 margin=10.21524 cleavage=29-30
# scaffold3|size331074 SpI score=14.7302 margin=13.02795 cleavage=20-21
# scaffold4|size324596 SpI score=4.27174 margin=1.47737 cleavage=28-29
# scaffold5|size295481 SpI score=15.5545 margin=15.755413 cleavage=18-19
# scaffold6|size272962 SpI score=10.8871 margin=9.66931 cleavage=36-37
# scaffold7|size271798 SpI score=17.5344 margin=14.94773 cleavage=25-26
# scaffold8|size261158 SpI score=14.3406 margin=13.936765 cleavage=19-20
# scaffold9|size237569 SpI score=7.65902 margin=4.87739 cleavage=32-33
# scaffold10|size219630 SpI score=16.9006 margin=15.69714 cleavage=24-25
# scaffold11|size204495 SpI score=9.4103 margin=9.186328 cleavage=22-23
# scaffold12|size176635 SpI score=9.32899 margin=5.26222 cleavage=33-34
# scaffold13|size161491 SpI score=12.4825 margin=7.37331 cleavage=27-28
# scaffold14|size149520 SpI score=15.2614 margin=13.89573 cleavage=27-28
# scaffold15|size262683 SpI score=9.81478 margin=9.504438 cleavage=20-21
# scaffold16|size135172 SpI score=6.9 margin=7.100913 cleavage=18-19
# scaffold17|size129360 SpI score=16.0197 margin=16.220613 cleavage=20-21
# scaffold18|size108795 SpI score=16.2098 margin=14.03159 cleavage=21-22
```

```
# NC_016845.1 SpI score=17.1009 margin=11.05799 cleavage=31-32
# NC_016838.1 SpI score=9.14976 margin=9.350673 cleavage=19-20
# NC_016846.1 SpI score=15.1474 margin=15.348313 cleavage=19-20
# NC_016839.1 SpI score=16.6144 margin=6.4663 cleavage=23-24
# NC_016840.1 SpI score=9.06892 margin=7.76956 cleavage=26-27
# NC_016847.1 SpI score=24.2514 margin=14.85952 cleavage=28-29
# NC_016841.1 SpI score=13.4394 margin=8.35687 cleavage=25-26
```

LipoP

Running time: run command “time perl LipoP -short <Input FASTA>”

```
# scaffold107|size478 SpI sco
# scaffold108|size426 SpI sco
# scaffold109|size415 SpI sco
Running for 1 seconds
real    0m0.529s
user    0m0.606s
sys     0m0.034s
```

```
# NC_016847.1 SpI score=
# NC_016841.1 SpI score=
Running for 0 seconds
real    0m0.267s
user    0m0.258s
sys     0m0.040s
```

File Size:
5.6 MB (assembled) ; 5.5 MB (reference)
Contains 5,745,742 bases (assembled), 5,682,322 bases (reference)

LipoP

LipoP is also used for Transmembrane regions.

Input Reference Genome: 5541 sequences

Predicted Transmembrane : 867

Much faster than other tools, like Phobius(~15min)

```
# NC_011283.1_12 TMH score=6.80174 margin=7.002653
# NC_011283.1_14 TMH score=12.1204 margin=12.321313
# NC_011283.1_16 TMH score=1.43192 margin=1.632833
# NC_011283.1_17 TMH score=5.86663 margin=6.067543
# NC_011283.1_20 TMH score=10.2266 margin=10.427513
# NC_011283.1_30 TMH score=12.1182 margin=12.319113
# NC_011283.1_32 TMH score=0.66513 margin=0.866043
# NC_011283.1_36 TMH score=10.7816 margin=10.982513
# NC_011283.1_42 TMH score=8.20291 margin=8.403823
# NC_011283.1_50 TMH score=7.75446 margin=7.955373
# NC_011283.1_53 TMH score=9.67056 margin=9.871473
# NC_011283.1_55 TMH score=8.52701 margin=4.98544
# NC_011283.1_59 TMH score=4.70729 margin=4.6789318
# NC_011283.1_66 TMH score=4.63746 margin=4.838373
# NC_011283.1_75 TMH score=2.19019 margin=2.391103
# NC_011283.1_79 TMH score=7.87516 margin=8.076073
# NC_011283.1_80 TMH score=14.699 margin=7.29236
# NC_011283.1_84 TMH score=9.59383 margin=9.794743
# NC_011283.1_88 TMH score=16.9619 margin=17.162813
# NC_011283.1_91 TMH score=5.65457 margin=5.855483
# NC_011283.1_101 TMH score=7.95103 margin=3.53982
# NC_011283.1_126 TMH score=4.43705 margin=4.637963
# NC_011283.1_145 TMH score=10.0948 margin=10.295713
# NC_011283.1_150 TMH score=11.9934 margin=12.194313
# NC_011283.1_151 TMH score=12.6783 margin=12.879213
# NC_011283.1_152 TMH score=-0.115433 margin=0.08548
# NC_011283.1_160 TMH score=14.6227 margin=14.823613
# NC_011283.1_170 TMH score=4.24254 margin=4.443453
# NC_011283.1_175 TMH score=6.46243 margin=6.663343
# NC_011283.1_176 TMH score=7.221 margin=7.421913
```

DOOR2

Download operon tables for *Klebsiella pneumoniae*

Eutil fasta files based on GI numbers in the operon table

Makeblastdb and and blastp queries

Filter and Match hits back to the operon table

DOOR2

- Download operon tables

Total 6321 Operons, 31991 Genes

OperonID	GI	Synonym	Start	End	Strand	Length	COG	Product
1263312	386032584	N2242_00(	5238	7022	+	594	-	bifunctional isocitrate dehydrogenase kinase/phosphatase protein
1263312	386032583	N2242_00(	3831	5135	+	434	-	isocitrate lyase
1263312	386032582	N2242_00(	2150	3751	+	533	-	malate synthase
1263313	386032595	N2242_00(	19223	20170	-	315	-	DeoR family transcriptional regulator
1263313	386032594	N2242_00(	18349	19152	-	267	-	sorbitol-6-phosphate 2-dehydrogenase
1263313	386032593	N2242_00(	17932	18339	-	135	-	PTS system mannose/fructose/sorbose family transporter subunit IIA
1263313	386032592	N2242_00(	17438	17932	-	164	-	PTS system mannose/fructose/sorbose family transporter subunit IIB
1263313	386032591	N2242_00(	16572	17372	-	266	-	PTS system mannose-specific transporter subunit IIC
1263313	386032590	N2242_00(	15736	16560	-	274	-	PTS system mannose/fructose/sorbose family transporter subunit IID
1263313	386032589	N2242_00(	14431	15663	-	410	-	hypothetical protein
1263313	386032588	N2242_00(	13629	14438	-	269	-	shikimate 5-dehydrogenase

DOOR2

- Fetch fasta files based on GI numbers

wget -nv -O operon.table "https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=nuccore&id=386032584&rettype=fasta&retmode=text"

- Makeblastdb

makeblastdb -in operon.table -dbtype prot

DOOR2

- Blastp, filter and match hits back to table

blastp -db operon.table -query SRR3982229.gff.fasta -num_threads 4 -evalue 1e-10 -outfmt "6 stitle qseqid sseqid qcovs pident evalue" > SRR3982229

(~7 minutes per sample)

stitle	qseqid	sseqid	qcovs	pident	evalue
YP_002238633.1 YgbK domain protein [Klebsiella pneumoniae 342]	scaffold84|size540_1	YP_002238633.1	100	100	2.12E-19
YP_002238613.1 YgbK domain protein [Klebsiella pneumoniae 342]	scaffold84|size540_1	YP_002238613.1	100	100	2.12E-19
YP_001335265.1 hypothetical protein KPN_01604 [Klebsiella pneumoniae subsp. pneumoniae M	scaffold84|size540_1	YP_001335265.1	100	100	2.30E-19
YP_005226819.1 ygbK domain protein [Klebsiella pneumoniae subsp. pneumoniae HS11286]	scaffold84|size540_1	YP_005226819.1	100	100	2.32E-19
YP_006636427.1 hypothetical protein A79E_2632 [Klebsiella pneumoniae subsp. pneumoniae 10	scaffold84|size540_1	YP_006636427.1	100	100	2.34E-19
YP_005954636.1 ygbK domain protein [Klebsiella pneumoniae KCTC 2242]	scaffold84|size540_1	YP_005954636.1	100	100	2.34E-19
YP_005957005.1 IS150 putative transposase [Klebsiella pneumoniae KCTC 2242]	scaffold83|size624_1	YP_005957005.1	99	99.507	2.15E-152
YP_005955660.1 IS150 putative transposase [Klebsiella pneumoniae KCTC 2242]	scaffold83|size624_1	YP_005955660.1	99	99.507	2.15E-152

InterProScan 5 & eggNOG-mapper commands

./interproscan.sh -goterms -iprlookup -pa -o [output.gff] -i [input_File.fasta] -f GFF3

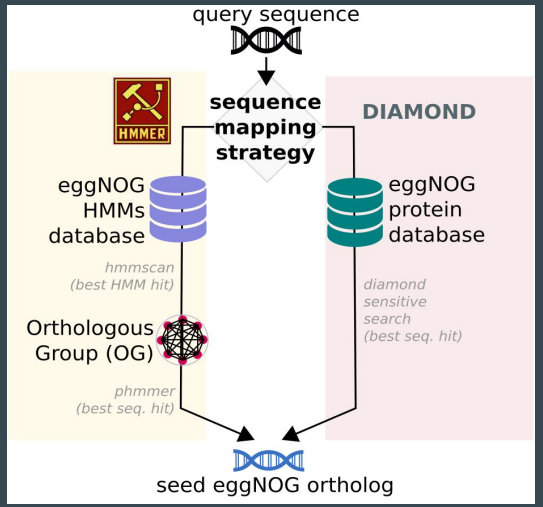
- Can also output TSV, XML, HTML, & JSON
- Specify GO & pathway with -goterms & -pa, respectively

python emapper.py -m hmm -i [input_file.fasta] --output [output_file] -d [database] --usemem

- Precomputed HMM models
- --usemem loads database into memory.
  - Entire bact database is ~32GB.
  - For our server, you will need to read entirely from disk or limit database by taxon

python emapper.py -m diamond -i [input_file.fasta] --output [output_file]

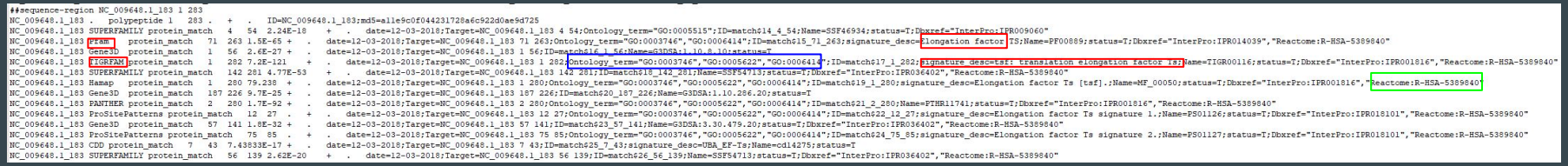
- Tailored for large sequence counts
- Does not scale linearly (slower for small amount of sequences)
- Efficient BLAST alternative



https://doi.org/10.1093/molbev/msx148

Example Outputs

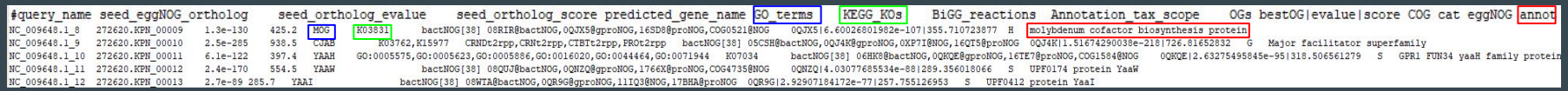
InterProScan 5



```
##sequence-region NC_009648.1_183 1 283
NC_009648.1_183 .  polypeptide 1   283 .   +   .   ID=NC_009648.1_183;md5=a11e9c0f044231728a6c922d0ae9d725
NC_009648.1_183 SUPERFAMILY protein_match   4   54  2.24E-18    +   .   date=12-03-2018;Target=NC_009648.1_183 4 54;Ontology_term="GO:0005515";ID=match$14_4_54;Name=SSF46934;status=T;Dbxref="InterPro:IPR009060"
NC_009648.1_183 Pfam   protein_match   71  263 1.5E-65 +   .   date=12-03-2018;Target=NC_009648.1_183 71 263;Ontology_term="GO:0003746","GO:0006414";ID=match$15_71_263;signature_desc=Elongation factor TS;Name=PF00889;status=T;Dbxref="InterPro:IPR014039","Reactome:R-HSA-5389840"
NC_009648.1_183 Gene3D  protein_match   1   56  2.6E-27 +   .   date=12-03-2018;Target=NC_009648.1_183 1 56;ID=match$16_1_56;Name=G3DSA:1.10.8.10;status=T
NC_009648.1_183 TIGRFAM protein_match   1   282 7.2E-121    +   .   date=12-03-2018;Target=NC_009648.1_183 1 282;Ontology_term="GO:0003746","GO:0005622","GO:0006414";ID=match$17_1_282;signature_desc=tsf: translation elongation factor Ts;Name=TIGR00116;status=T;Dbxref="InterPro:IPR001816","Reactome:R-HSA-5389840"
NC_009648.1_183 SUPERFAMILY protein_match   142 281 4.77E-53 +   .   date=12-03-2018;Target=NC_009648.1_183 142 281;ID=match$18_142_281;Name=SSF54713;status=T;Dbxref="InterPro:IPR036402","Reactome:R-HSA-5389840"
NC_009648.1_183 Hamap   protein_match   1   280 75.238  +   .   date=12-03-2018;Target=NC_009648.1_183 1 280;Ontology_term="GO:0003746","GO:0005622","GO:0006414";ID=match$19_1_280;signature_desc=Elongation factor Ts [tsf].;Name=MF_00050;status=T;Dbxref="InterPro:IPR001816","Reactome:R-HSA-5389840"
NC_009648.1_183 Gene3D  protein_match   187 226 9.7E-25 +   .   date=12-03-2018;Target=NC_009648.1_183 187 226;ID=match$20_187_226;Name=G3DSA:1.10.286.20;status=T
NC_009648.1_183 PANTHER protein_match   2   280 1.7E-92 +   .   date=12-03-2018;Target=NC_009648.1_183 2 280;Ontology_term="GO:0003746","GO:0005622","GO:0006414";ID=match$21_2_280;Name=PTHR11741;status=T;Dbxref="InterPro:IPR001816","Reactome:R-HSA-5389840"
NC_009648.1_183 ProSitePatterns protein_match   12  27  .   +   .   date=12-03-2018;Target=NC_009648.1_183 12 27;Ontology_term="GO:0003746","GO:0005622","GO:0006414";ID=match$22_12_27;signature_desc=Elongation factor Ts signature 1.;Name=PS01126;status=T;Dbxref="InterPro:IPR018101","Reactome:R-HSA-5389840"
NC_009648.1_183 Gene3D  protein_match   57  141 1.8E-32 +   .   date=12-03-2018;Target=NC_009648.1_183 57 141;ID=match$23_57_141;Name=G3DSA:3.30.479.20;status=T;Dbxref="InterPro:IPR036402","Reactome:R-HSA-5389840"
NC_009648.1_183 ProSitePatterns protein_match   75  85  .   +   .   date=12-03-2018;Target=NC_009648.1_183 75 85;Ontology_term="GO:0003746","GO:0005622","GO:0006414";ID=match$24_75_85;signature_desc=Elongation factor Ts signature 2.;Name=PS01127;status=T;Dbxref="InterPro:IPR018101","Reactome:R-HSA-5389840"
NC_009648.1_183 CDD protein_match   7   43  7.43833E-17 +   .   date=12-03-2018;Target=NC_009648.1_183 7 43;ID=match$25_7_43;signature_desc=UBA_EF-Ts;Name=cd14275;status=T
NC_009648.1_183 SUPERFAMILY protein_match   56  139 2.62E-20    +   .   date=12-03-2018;Target=NC_009648.1_183 56 139;ID=match$26_56_139;Name=SSF54713;status=T;Dbxref="InterPro:IPR036402","Reactome:R-HSA-5389840"
```

Databases of Note: SMART, TIGRFAM, Pfam, SUPERFAMILY, PANTHER, CATH-Gene3D
Annotations can be redundant, increasing computation time

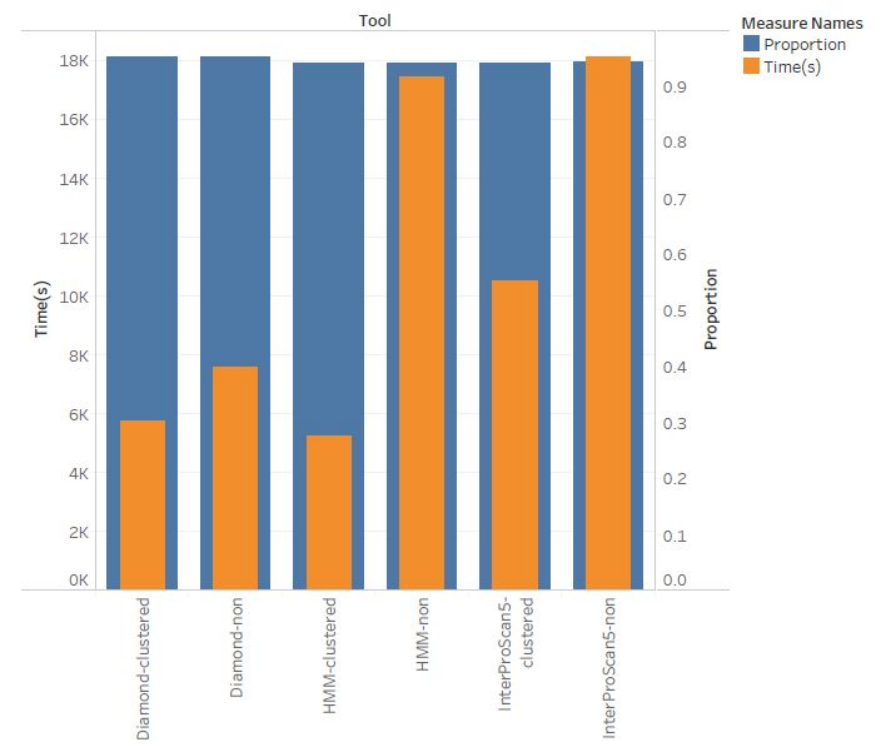
eggNOG-mapper



```
#query_name seed_eggNOG_ortholog    seed_ortholog_evalue    seed_ortholog_score predicted_gene_name GO_terms    KEGG_KOs    BiGG_reactions  Annotation_tax_scope    OGs bestOG|evalue|score COG cat eggNOG annot
NC_009648.1_8   272620.KPN_00009    1.3e-130    425.2   MOG     KO3831      bactNOG[38] 08RIR@bactNOG,0QJX5@proNOG,16SD8@proNOG,COG0521@NOG     0QJX5|6.60026801982e-107|355.710723877  H   molybdenum cofactor biosynthesis protein
NC_009648.1_9   272620.KPN_00010    2.5e-285    938.5   CJAB    KO3762,K15977   CRNDt2rpp,CRNt2rpp,CTBTt2rpp,PROt2rpp    bactNOG[38] 05CSH@bactNOG,0QJ4K@proNOG,0XP7I@NOG,16QT5@proNOG   0QJ4K|1.51674290038e-218|726.81652832   G   Major facilitator superfamily
NC_009648.1_10  272620.KPN_00011    6.1e-122    397.4   YAAH    GO:0005575,GO:0005623,GO:0005886,GO:0016020,GO:0044464,GO:0071944  KO7034      bactNOG[38] 06HK8@bactNOG,0QKQE@proNOG,16TE7@proNOG,COG1584@NOG     0QKQE|2.63275495845e-95|318.506561279   S   GPR1 FUN34 yaaH family protein
NC_009648.1_11  272620.KPN_00012    2.4e-170    554.5   YAAW        bactNOG[38] 08QUJ@bactNOG,0QNZQ@proNOG,176EX@proNOG,COG4735@NOG     0QNZQ|4.03077685534e-88|289.356018066   S   UPF0174 protein YaaW
NC_009648.1_12  272620.KPN_00013    2.7e-89 285.7   YAAI        bactNOG[38] 08WTA@bactNOG,0QR9G@proNOG,11IQ3@NOG,17BHA@proNOG   0QR9G|2.92907184172e-77|257.755126953   S   UPF0412 protein YaaI
```

Databases of Note: COG, KEGG, GO, Pfam/SMART

Gene Ontology & Pathway Annotation



Tool	Time(s)	Annotations	Genes	Proportion
Diamond-non	7615	25519	26807	0.951952848
HMM-non	17433	25232	26807	0.941246689
InterProScan5-non	18120	25309	26807	0.944119073
Diamond-clustered	5744	25516	26807	0.951840937
HMM-clustered	5245	25231	26807	0.941209386
InterProScan5-clustered	10516	25231	26807	0.941209386

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

ncRNAs

- Tools used in prediction group will predict and annotate at the same time
  - tRNAscan, Aragorn, and Rfam

CRISPR

- **Pilercr1.06**
  - Current version: 1.06 (Jan 20, 2007)
  - Command Line: ./pilercr -in <input_file.fasta> -out <output_file>
  - Running time: <2s/genom
  - Identify the characteristic signature of CRISPR repeats; repeats and spacers are within the expected ranges of length and sequence conservation (parameters of the algorithm and can be changed by the user)

- **CRT**
  - Current version: 1.1 (March 14, 2007)
  - Command Line: java -cp CRT1.2-CLI.jar crt [options] inputFile [outputFile]
  - Running time: <2s/genom
  - Screens for exact k-mer/k-nucleotide repeats in a genome, and concatenates the neighbouring repeats into candidate CRISPRs

CRISPR

Klebsiella pneumoniae Genomes	CRT (Blast against to CRISPRdb)	Pilercr1.06 (Blast against to CRISPRdb)	NCBI Annotation
CP007727.1	0/0	0/0	0
NC_011283.1	0/0	0/0	0
NC_016845.1	0/0	0/0	0
NZ_CP008827.1	0/0	0/0	0
NC_009648.1	0/1	0/0	0
NC_017540.1	0/0	0/1	0
NC_012731.1	2/2	2/2	2
NC_018522.1	2/2	2/2	2

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

Resistance Gene Identifier (CARD)

Command:

rgi -t protein -i clustered.fa

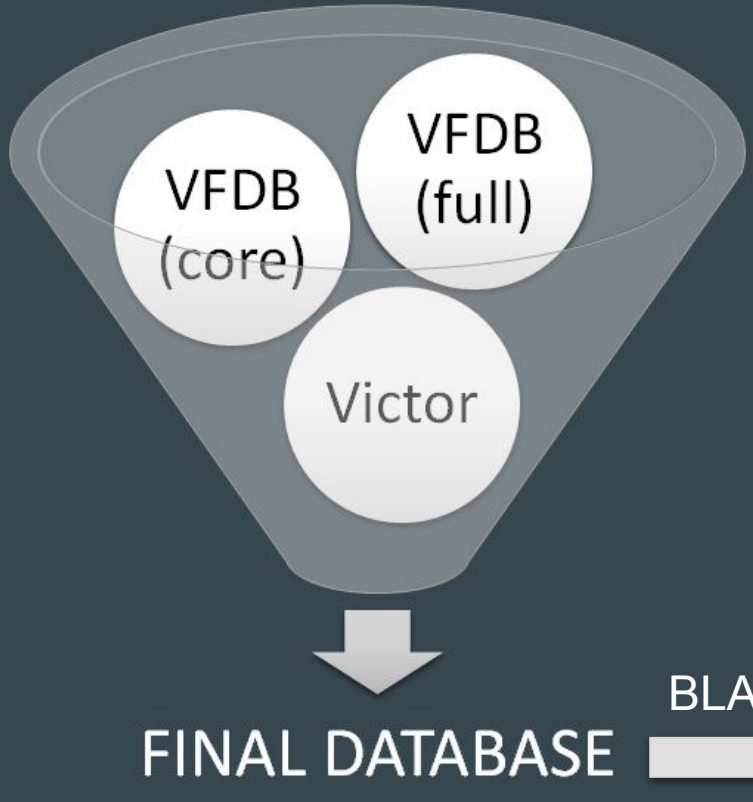
For 5 genomes: ~30 minutes

For clustered equivalent: ~10 minutes

(should not change much with additional genomes)

Result accuracy not yet evaluated

Virulence Factors



Fetch the fasta files from VFDB and VICTOR using “wget” command

Build a FINAL “Combined Database” using “makeblastdb”

BLAST the query sequence against the combined database using “blastn”

Time taken to generate results: ~19 secs. for 5 test files

Output in the form of table containing details:
query name, query start, query stop, percent identity, evalue, name of virulence factor

Test Run with Reference Assembly

```
NC_016845.1     3448881 3458372 99.937  0.0     VFG034118(gi:218705473) (irp1) High-molecular-weight nonribosomal peptide/polyketide synthetase 1 [Yersiniabactin siderophore (CVF458)] [Escherichia coli O17:K52:H
18 str. UMN026]
NC_016845.1     3448881 3458372 99.831  0.0     VFG000362(gb|NP_405471) (irp1) yersiniabactin biosynthetic protein Irp1 [Yersiniabactin (VF0136)] [Yersinia pestis CO92]
NC_016846.1     9489    10258   88.586  0.0     VFG012785(gi:3776756) (ospC4) - [Mxi-Spa TTSS effectors controlled by VirB (CVF463)] [Shigella boydii Sb227]
~
~
~
```

Test Run with Team 2 Genome Assembly Results

```
scaffold1|size427644    376880  380623  86.035  0.0     gi|16445223:1817022-1820765 Escherichia coli O157:H7 str. EDL933 chromosome, complete genome
scaffold1|size427644    38469   42205   76.679  0.0     gi|16445223:1817022-1820765 Escherichia coli O157:H7 str. EDL933 chromosome, complete genome
scaffold1|size427644    413868  414959  84.826  0.0     gi|117622295:1270445-1271536 Escherichia coli APEC O1, complete genome
scaffold2|size413645    230423  233047  88.343  0.0     gi|16763390:2373710-2376346 Salmonella enterica subsp. enterica serovar Typhimurium str. LT2 chromosome, complete genome
scaffold2|size413645    191402  194128  85.327  0.0     gi|205351346:2416815-2419541 Salmonella enterica subsp. enterica serovar Gallinarum str. 287/91 chromosome, complete genome
scaffold3|size383805    26216   28702   99.839  0.0     VFG043625(gi:238896388) (mrkC) fimbrial biogenesis outer membrane usher protein mrkC precursor [type 3 fimbriae (biofilm formation) (AI076)] [Klebsiella pn
eumoniae subsp. pneumoniae NTUH-K2044]
scaffold3|size383805    38935   41547   94.266  0.0     VFG043616(gi:206577896) (fimD) outer membrane usher protein fimD [type 1 fimbriae (AI075)] [Klebsiella pneumoniae 342]
scaffold4|size340664    241819  244038  76.854  0.0     gi|16445223:4066968-4069262 Escherichia coli O157:H7 str. EDL933 chromosome, complete genome
scaffold4|size340664    177376  178515  84.307  0.0     gi|218698419:2234381-2235505 Escherichia coli IAI39, complete genome
scaffold5|size331356    270048  272149  84.670  0.0     gi|16763390:661273-663378 Salmonella enterica subsp. enterica serovar Typhimurium str. LT2 chromosome, complete genome
scaffold5|size331356    270044  272148  84.408  0.0     gi|29140543:2333603-2335708 Salmonella enterica subsp. enterica serovar Typhi str. Ty2 chromosome, complete genome
scaffold6|size318597    296     2868    85.753  0.0     gi|16763390:2802115-2804688 Salmonella enterica subsp. enterica serovar Typhimurium str. LT2 chromosome, complete genome
scaffold6|size318597    192178  194414  83.564  0.0     gi|207855516:2991566-2993800 Salmonella enterica subsp. enterica serovar Enteritidis str. P125109 chromosome, complete genome
scaffold7|size271828    107919  111135  88.630  0.0     gi|49175990:30816-34037 Escherichia coli str. K-12 substr. MG1655 chromosome, complete genome
scaffold7|size271828    238769  240918  86.704  0.0     gi|29140543:4734841-4736991 Salmonella enterica subsp. enterica serovar Typhi str. Ty2 chromosome, complete genome
scaffold8|size248034    166065  168185  88.177  0.0     gi|207855516:3815173-3817284 Salmonella enterica subsp. enterica serovar Enteritidis str. P125109 chromosome, complete genome
scaffold8|size248034    21833   23374   93.839  0.0     gi|16763390:4076433-4077974 Salmonella enterica subsp. enterica serovar Typhimurium str. LT2 chromosome, complete genome
scaffold9|size203741    201502  203106  84.773  0.0     gi|170679574:3932412-3934019 Escherichia coli SMS-3-5, complete genome
scaffold9|size203741    201502  203106  84.639  0.0     gi|215485161:3954556-3956163 Escherichia coli O127:H6 str. E2348/69, complete genome
scaffold10|size193900   27568   29861   80.061  0.0     gi|16445223:4066968-4069262 Escherichia coli O157:H7 str. EDL933 chromosome, complete genome
scaffold10|size193900   62221   63906   82.454  0.0     gi|16763390:1915221-1916906 Salmonella enterica subsp. enterica serovar Typhimurium str. LT2 chromosome, complete genome
scaffold11|size182822   163696  166221  87.732  0.0     VFG034483(gi:222155097) (ecpC) hypothetical protein [E. coli common pilus (ECP) (CVF625)] [Escherichia coli O83:H1 str. LF82]
scaffold11|size182822   163696  166221  87.732  0.0     VFG018216(gi:91209353) (ecpC) hypothetical protein YagX precursor [E. coli common pilus (ECP) (CVF625)] [Escherichia coli UTI89]
scaffold12|size135437   32999   35415   83.169  0.0     gi|74310614:4633502-4635943 Shigella sonnei Ss046, complete genome
scaffold12|size135437   32999   35415   83.004  0.0     gi|110804074:4479325-4481766 Shigella flexneri 5 str. 8401, complete genome
scaffold14|size108818   31009   33681   78.439  0.0     gi|16445223:2247962-2250637 Escherichia coli O157:H7 str. EDL933 chromosome, complete genome
scaffold14|size108818   35805   38400   77.605  0.0     gi|123440403:2409235-2411850 Yersinia enterocolitica subsp. enterocolitica 8081, complete genome
scaffold15|size106624   35231   36166   98.825  0.0     VFG044322(gi:238894719) (iroE) hypothetical protein [Salmochelin (IA015)] [Klebsiella pneumoniae subsp. pneumoniae NTUH-K2044]
```

Phaster

- Run time: ~10mins/genome
- 3 Output files
- Disadvantages:
  - No options
  - Their server only takes one file at a time
- What I did:
  - Wrote script to check the status every 5 mins
  - Download and unzip the result when it's completed

Output File 1: Summary.txt

- Intact: >90
- Questionable: 70 - 90
- Incomplete: < 70

```
Criteria for scoring prophage regions (as intact, questionable, or incomplete):
Method 1:
1. If the number of certain phage organism in this table is more than or equal to 100% of the total number of CDS of the region,
   the region is marked with total score 150. If less than 100%, method 2 and 3 will be used.

Method 2:
1. If the number of certain phage organism in this table  is more than 50% of the total number of CDS of the region, that phage
   organism is considered as the major potential phage for that region; the percentage of the total number
   of that phage organism in this table in the  total number of proteins of the region is calculated and
   then multipled by 100; the percentage of the length of that phage organism in this table in the length
   of the region is calculated and then multipled by 50 (phage head's encapsulation capability is considered).

Method 3:
1. If any of the specific phage-related keywords (such as 'capsid', 'head', 'integrase', 'plate', 'tail', 'fiber',
   'coat', 'transposase', 'portal', 'terminase', 'protease' or 'lysin') are present, the score will be increased
   by 10 for each keyword found.
2. If the size of the region is greater than 30 Kb, the score will be increased by 10.
3. If there are at least 40 proteins in the region, the score will be increased by 10.
4. If all of the phage-related proteins and hypothetical proteins constitute more than 70% of
   the total number of proteins in the region, the score will be increased by 10.

Compared the total score of method 2 with the total score of method 3, the bigger one is chosen as the total score of the region.
If the region's total score is less than 70, it is marked as incomplete; if between 70 to 90, it is marked as questionable; if greater than 90, it is marked as intact.



Totally 4 intact prophage regions have been identified.
```

Output File 1: Summary.txt

- Region
- Completeness
- Specific Keyword
- Region Position
- Most Common Phage Name

- Preliminary Results
  - *Klebsiella pneumoniae* Reference Genome
    - 8 Prophage Regions
      - 4 Regions are intact (Score >=100)
  - SRR3467249 - 1 out of 5 Regions is intact
  - SRR3982229 - 1 out of 6 Regions is intact
  - SRR3982230 - 1 out of 7 Regions is intact
  - SRR3982253 - 3 out of 6 Regions are intact
  - SRR3982316 - 1 out of 5 Regions is intact

REGION	REGION_LENGTH	COMPLETENESS(score)	SPECIFIC_KEYWORD	REGION_POSITION	TRNA_NUM	TOTAL_PROTEIN_NUM	PHAGE_HIT_PROTEIN_NUM	HYPOTHETICAL_PROTEIN_NUM	PHAGE+HYPO_PROTEIN_PERCENTAGE	BACTERIAL_PROTEIN_NUM	ATT_SITE_SHOWUP	PHAGE_SPECIES_NUM	MOST_COMMON_PHAGE_NAME(hit_genes_count)	FIRST_MOST_COMMON_PHAGE_NUM	FIRST_MOST_COMMON_PHAGE_PERCENTAGE	GC_PERCENTAGE
1	13.5Kb	questionable(75)	integrase,transposase,head	581768-595336	0	15	11	4	100%	0	yes	5	PHAGE_Entero_P4_NC_001609(7),PHAGE_Salmon_SJ46_NC_031129(1),PHAGE_Entero_N15_NC_001901(1),PHAGE_Erwini_ENT90_NC_019932(1),PHAGE_Serrat_Eta_NC_021563(1),PHAGE_Ralsto_RSY1_NC_025115(1),PHAGE_Salmon_SP_004_NC_021774(1),PHAGE_Rueger_DSS3_P1_NC_025428(1),PHAGE_Shewan_1/41_NC_025458(1),PHAGE_Stenot_S1_NC_011589(1),PHAGE_Entero_HK022_NC_002166(1),PHAGE_Entero_WPhi_NC_005056(1),PHAGE_Entero_186_NC_001317(1),PHAGE_Yersin_PY54_NC_005069(1),PHAGE_Salmon_Fels_1_NC_010391(1),PHAGE_Entero_fiAA91_ss_NC_022750(1),PHAGE_Xantho_vB_XveM_DIBBI_NC_017981(1),PHAGE_Burkho_KL3_NC_015266(1)	7	46.66%	50.39%

Output File 2: Detail.txt

```
>NC_016845.1 Klebsiella pneumoniae subsp. pneumoniae HS11286 chromosome, complete genome

CDS_POSITION                    BLAST_HIT                                                                                    EVALUE          prophage_PRO_SEQ
-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

#### region 1 ####
581768..581781                  attL                                                                                         N/A             GAGTCCGGCCTTCG
581952..583214                  PHAGE_Entero_P4_NC_001609: integrase; PP_00519; phage(gi9627511)                              0.0
MKLNARQVETAKPKDKTYKMADGGGLYLEVSAKGSKYWRMKYRRPSDKKEDRLAFGVWPTVTLAQARTKRDEAKKLLVQGIDPKAEQKEAQAENSGAYTFETIAREWHASNKRWSEDHRSRVLRYLELYIFPHIGSSDIRQLKTSHLLAPIKKVDASGKHDVAQRLQQRVTAIMRYAVQNDYIDSNPASDMAG
ALSTTKARHYPALPSSRFPEFLARLAAYRGRVMTRIAVELSLLTFVRSSELRFARWDEFDFAKSLWRIPAKREEIKGVRYSYRGMKMKEEHIVPLSRQAIVLLEQLKQISGDKELLFPGDHDATKVMSENTVNGALRAMGYDTKTEVCGHGFRTMARGALGESGLWSDDAIERQLSHSERNNVRAAYIHTSEH
LDERRLMVQWWADYLKSNEGKIVTPYEFAKIRKS
complement(583272..584282)      hypothetical; PP_00520                                                                       N/A
MIYITSVSKETSMTKYEVKMKKNIRVIPKNIHSKLRRLGNTVVAGTSIAFTENQLKNGMLEHLGIYFDNDVIKYEASIIPDPLQGKYSFKNVFGEEVIRKDLPKETHYTEIESPNWGDSSNGTHTVRLPHDKYPRDIIPPKLIAIEINHKKSSDSYFIFNFRATRILEKNSNKFDDELLFDLNLLQENLGKCG
VENADKPISTYADTLIVSWDIFPPGSKEEILARIFKGKNITDDKKAVAENRYEFFMSLEPKKIVTGNSTFSNYIGAMLEDDLVVFENIEYGNAIYILYDDWDEISKLSRIDLLSGRAGSNFDRIIHSGNWKEEVRKKVATGRL
complement(584770..585849)      hypothetical; PP_00521                                                                       N/A
MNIGHLNFFKVNKCGLYKVNDNNTYGLELSETFDLIQDWVGTKSLALTIPWDPKEKPNRSKCYCKDIYKDENTGDFLIMLWKSDTDSTGSLLGASEDGEIGSSSVVKYTNSYRGKKVIWGRPCFYWVIPELETIVSIKFDHSICDSELFQDYVHSSITNRVKHSKRVKNKTEKGYIRLSNTDDDDLYKYMYRF
DMKLRSLETTHTELGKLIPKITHIVRRETIIINPNDARADWLKTFSTLVPFVSGKKNTRTRQIEIKAEAKPSLNEVKEIIEKYSSEDREKRLWDNVGFATDKGITWVDKYRMRDILNVPSEDYSTYSAAYIYEQISHKRKEFISPILKESKALQSQSRINKASGED
586099..587022                  PHAGE_Entero_HK022_NC_002166: IS903 transposase; PP_00522; phage(gi9634149)                   0.0
VAKQKFKITNWPTYNKALINRGSITFWLDDEAIQAWYESATPSSRGRPQRYSDLAITTVLVIKRVFRLTLRAAQGFIDSIFSLMNVPLRCPDYSCVSRRAKSVNISFKTPTRGEIAHLVIDSTGLKVFGEGEWKVKKHGQERRRIWRKLHLAVDSKTHEIICADLSLNNVTDSEAFPGLIRQTHRKIRSAAAD
GAYDTRLCHDELRRKKISALIPPRKGAGYWPGEYADRNRAVANQRMTGSNARWKWTTDYNRRSIAETAMYRVKQLFGGSLTLRDYDGQVAEAMALVRALNKMTKAGMPESVRIA
complement(587656..588222)      PHAGE_Entero_P4_NC_001609: amber mutation-suppressing protein; PP_00523; phage(gi9627520)    2.48e-75
MTTLTLQQAFEACQKNETAWLNRKAELAAAEQEYQEQVLAGDDRIPAIMQELRDIIDVKKWEINQAAGRYIRSHEAVQRISIRNRLNDFMQAHGTELAATLAPELMGLSQQPALLTGHALDRSAHYLREALSVWLSTGEEINYAAEDSDILTAIGFRPDAASRVDNQEKYTPAQSLIYARRRTELASK
complement(588240..588485)      PHAGE_Entero_186_NC_001317: B protein; PP_00524; phage(gi9634083)                             8.88e-18
MMHCPFCKKSAHARTSRYLSENVKQRYHQCTNIECSATFRTTEAIDEVIRPPAEKAPPVAEPVTPPAPRNVQGCYSSPYRH
complement(588482..589219)      PHAGE_Entero_P4_NC_001609: head size determination protein sid; PP_00525; phage(gi9627518)    1.59e-95
MTDTTFIPDYLKPALERLAEARAAHLEQARLMEDTLTAITRAEEQKAELEQDNGSDTRTWRAAFRAGGAMLTDELKSGHIERVARRELAQECHNLTEVLAFERDQLKATCNSTARAFRQAHHAVLSKYAEEELNRALNDTLGPLVRAMVLKAEVMANPLANTTGHQGYTEPEKEVMHQVVTFLTGKVSAFSVT
PADEPVLSLTGFPAVALAHMDHDAASTPGERKVWQEKIRQREADLKARGLLP
```

Output File 3: phage_regions.fna

```
>1	581768-595336
GAGTCCGGCCTTCGGCACCATTAGTACTTCCAAGACCATCCGAGAAAGTCCAATTATCCCTTAAAAATCA
ATGCTTGCAGCGATTTTTACGTCCTGAGTCGTCCGAGGTTGTCCGTTGAAATCCGGATGTCATTGGGGGC
ATAATTGGGGGCATCTTAACTTCGATTAGAAATGTGCCCCCAAAATGAAGCTCAACGCCAGGCAGGTAGA
GACCGCAAAGCCCAAAGACAAAACCTACAAAATGGCTGATGGCGGCGGTTTGTATCTCGAAGTTTCGGCT
AAGGGGTCTAAATACTGGCGCATGAAATACAGACGCCCCTCTGACAAAAAAGAGGATCGCCTTGCTTTTG
GTGTTTGGCCAACTGTGACGCTTGCTCAGGCAAGAACCAAGCGCGACGAAGCTAAAAAGCTGTTAGTACA
GGGCATTGACCCAAAAGCCGAACAGAAAGAAGCTCAGGCCGAGAATTCGGGGGCATATACTTTCGAAACA
ATTGCTCGCGAATGGCATGCCAGTAACAAGCGATGGAGTGAAGACCATCGATCGCGCGTTCTTCGCTATC
TTGAGCTTTATATCTTCCCTCATATCGGTTCGTCCGACATTCGCCAGCTCAAAACCAGCCACCTGTTAGC
CCCGATTAAAAAAGTTGATGCCAGTGGCAAACATGACGTCGCGCAGCGTCTTCAGCAGCGTGTAACGGCC
ATTATGCGCTATGCCGTTCAGAACGATTACATAGACTCAAACCCGGCCAGTGATATGGCTGGTGCTTTAT
CAACAACCAAAGCAAGACACTATCCAGCTTTACCCTCTAGCCGTTTCCCTGAATTTCTTGCTCGTCTGGC
GGCATATCGTGGTCGTGTAATGACACGGATCGCCGTAGAGCTTTCATTGCTCACTTTTGTACGTTCCAGT
GAACTCCGTTTTGCACGTTGGGATGAATTCGATTTTGCTAAATCTCTTTGGCGTATACCTGCAAAGCGTG
AAGAAATTAAAGGCGTACGTTACTCCTACCGCGGCATGAAGATGAAAGAGGAACATATAGTTCCGCTTAG
TCGACAAGCTATAGTGTTGTTAGAGCAGCTTAAGCAAATCAGTGGTGATAAAGAGCTACTTTTTCCGGGA
GATCACGACGCAACGAAGGTCATGAGTGAAAACACAGTAAACGGTGCGTTGCGTGCTATGGGCTATGATA
CGAAAACAGAGGTCTGTGGGCATGGATTCAGGACGATGGCGCGCGGCGCGCTGGGGGAATCGGGTTTATG
GAGTGATGATGCTATAGAACGCCAACTGAGTCATTCAGAGCGTAATAATGTACGTGCAGCATATATTCAC
ACTTCTGAACATTTAGATGAACGCCGTTTAATGGTGCAATGGTGGGCTGATTATCTCAAATCAAATGAAG
GCAAAATAGTCACACCTTATGAATTTGCCAAAATAAGAAAAAGCTGATAGCAGAAATAAACGCCTCTACA
AAAAGAAGGCGTTTATTTATATTAAAATATTGAATCATAATCTTCCAGTAGCAACTTTTTTTCGTACTTC
CTCTTTCCAATTACCACTGTGGATTATTCTATCAAAATTACTTCCTGCTCTTCCTGATAACAAATCAATT
CGAGATAGTTTGCTGATTTCATCCCAGTCGTCATAGAGAATATATATCGCATTGCCATATTCTATGTTCT
CAAAAACAACCAGATCGTCCTCTAGCATAGCTCCAATGTAATTACTGAACGTACTATTTCCAGTAACAAT
TTTTTTAGGTTCAAGGCTCATGAAGAACTCATATCTATTTTCAGCAACGGCTTTCTTATCGTCAGTTATG
TTTTTTCCTTTGAAAATTCTTGCTAAAATTTCTTCTTTGCTACCTGGCGGAAATATGTCCCATGAGACTA
TAAGTGTATCAGCATAAGTAGATATAGGTTTATCAGCATTCTCAACGCCACATTTACCGAGATTTTCTTG
AAGAAGATTTAGATCAAATAGCAACTCATCATCAAACTTATTTGAATTTTTTTCTAAAATTCTAGTTGCT
CTGAAATTAAATATAAAATAGCTGTCCGAGGACTTCTTGTGATTAATTTCAATTGCTATAAGTTTTGGTG
GGATTATGTCTCTGGGGTATTTATCGTGAGGTAATCTAACAGTATGTGTACCGTTAGAGCTATCCCCCCA
GTTAGGTGACTCAATTTCGGTATAGTGAGTTTCTTTTGGGAGGTCTTTTCTGATGACTTCCTCGCCGAAA
ACATTTTTAAATGAATATTTACCTTGCAATGGATCAGGAATGATCGATGCCTCATATTTAATAACATCAT
TATCAAAATATATTCCGAGATGCTCTAACATTCCATTTTTTAATTGATTCTCTGTAAATGCGATCGAAGT
GCCAGCAACGACTGTATTGCCAAGCCTTCTTAACTTGCTATGAATGTTTTTAGGTATAACACGTATATTC
```

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?