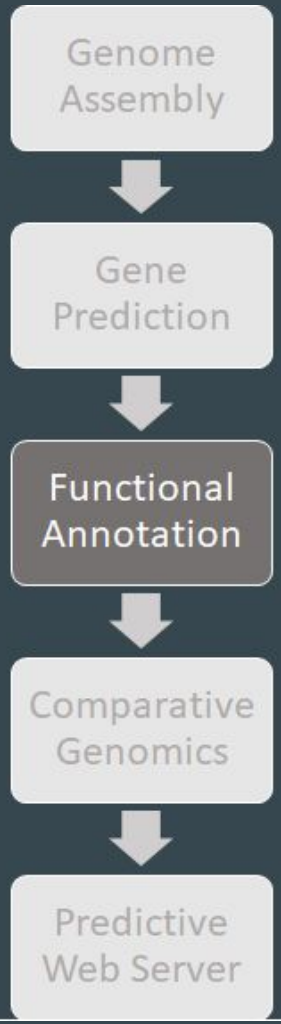
Functional Annotation

Background & Strategy

•••

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


Genome Assembly
Gene Prediction
Functional Annotation
Comparative Genomics
Predictive Web Server


Genome Assembly
Gene Prediction
Functional Annotation
Comparative Genomics
Predictive Web Server

Outline

Introduction

Approach

Tools

Preliminary Pipeline

Introduction

- What is functional annotation?
  - Assigning function to genes

- Information on
  - Biological function
  - Biochemical function
  - Regulation
  - Expression

- Objective
  - To functionally annotate 262 *Klebsiella* genomes

- Tools
  - Ab-initio
  - Homology based

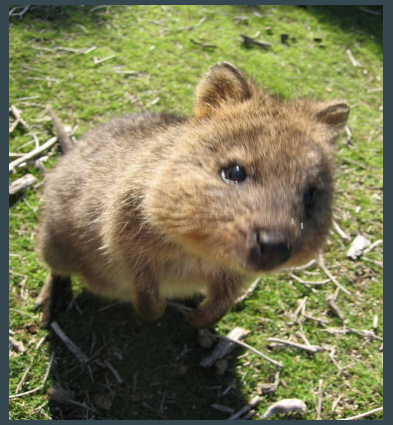
Approach

- Input
  - Predicted genes FASTA/GFF
- Processing
  - General
  - Protein Coding
  - Non-coding
  - Other
- Output
  - Merged GFF

GENERAL TOOL

...

PROKKA

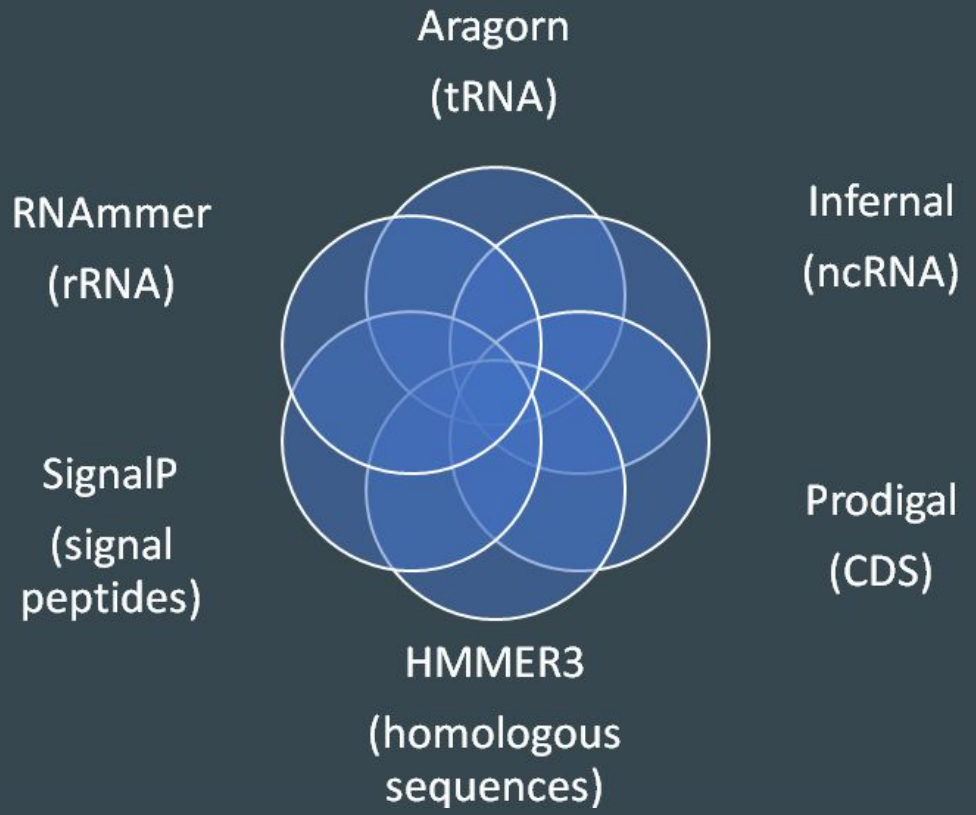


- Designed for Prokaryotic Genomes
- Input
  - Preassembled genomic DNA sequences in FASTA
  - Sequence file is the only mandatory parameter
  - Finished sequences without gaps are the ideal input, but it is expected that the typical input will be a set of scaffold sequences produced by de novo assembly software
- Pros
  - Annotate a genome in ~10 mins
  - Supports multithreading
- Cons
  - Has many dependencies
  - Annotates only some categories of genes

PROKKA

Dependencies:

- BioPerl
- BLAST+
- HMMER
- Aragorn
- Prodigal
- tbl2asn
- GNU Parallel
- Infernal



PROKKA OUTPUT

Suffix	Description of file contents
.fna	FASTA file of original input contigs (nucleotide)
.faa	FASTA file of translated coding genes (protein)
.ffn	FASTA file of all genomic features (nucleotide)
.fsa	Contig sequences for submission (nucleotide)
.tbl	Feature table for submission
.sqn	Sequin editable file for submission
.gbk	Genbank file containing sequences and annotations
.gff	GFF v3 file containing sequences and annotations
.log	Log file of Prokka processing output
.txt	Annotation summary statistics

SPECIFIC TOOLS

...

Specific Tools (Based on features to be annotated)

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

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

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

Specific Tools (Based on features to be annotated)

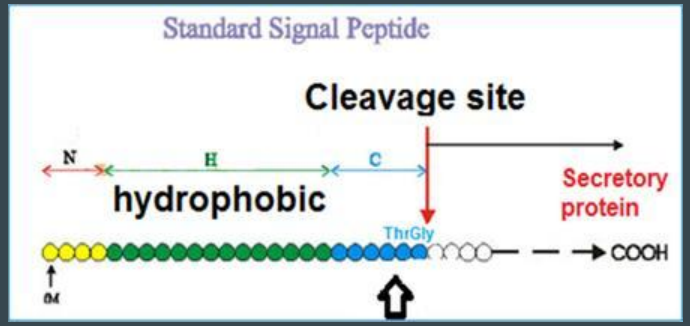
- Protein-coding regions
  - Signaling peptides
  - Transmembrane regions
  - Lipoproteins
  - Operons
  - Enzymes

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

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

Signal Peptides

- N-terminal sequences mediating translocation of proteins across the cytoplasmic membrane or the extracellular space
- Typically has 3 regions
  - a positively charged *N-region* followed by
  - a hydrophobic *H-region* and
  - a neutral but polar *C-region*.
- Relative proportion of secreted and cytoplasmic proteins – useful parameter for genome based reconstructions of organism's behaviour and metabolism



https://www.sigmaaldrich.com/content/dam/sigma-aldrich/articles/biology/chozn-tech-articles/structure-signal-peptide.jpg

	Eukaryotes	Prokaryotes	
		Gram-negative	Gram-positive
Total length (average)	22.6 aa	25.1 aa	32.0 aa
n-regions	only slightly Arg-rich	Lys+Arg-rich	
h-regions	short, very hydrophobic	slightly longer, less hydrophobic	very long, less hydrophobic
c-regions	short, no pattern	short, Ser+Ala-rich	longer, Pro+Thr-rich
-3,-1 positions	small and neutral residues	almost exclusively Ala	
+1 to +5 region	no pattern	rich in Ala, Asp/Glu, and Ser/Thr	

Comparison of different Signal Peptides Prediction Methods

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

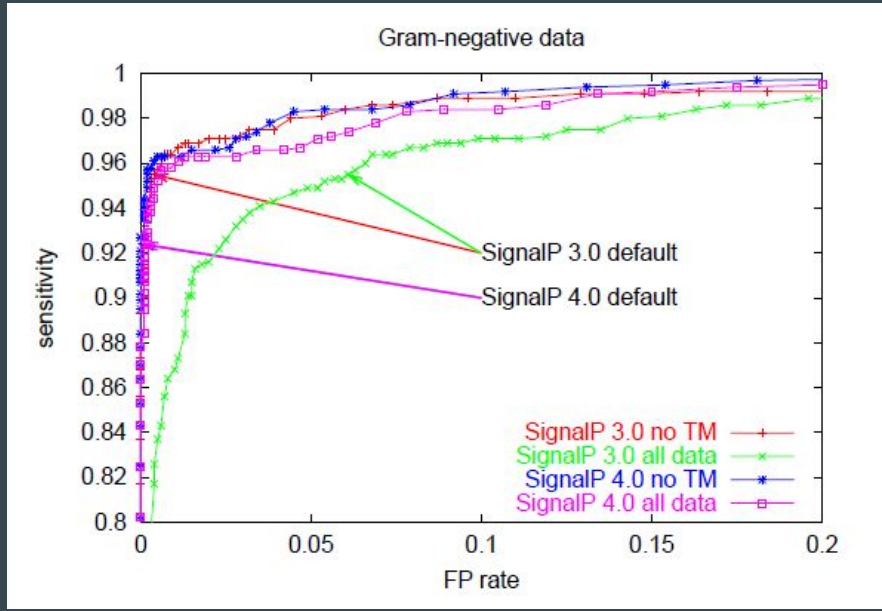
Petersen, Thomas Nordahl, et al. "SignalP 4.0: discriminating signal peptides from transmembrane regions." Nature methods 8.10 (2011): 785.

Prediction of Signal Peptides

- SignalP
- Phobius
- Philius

SignalP 4.0

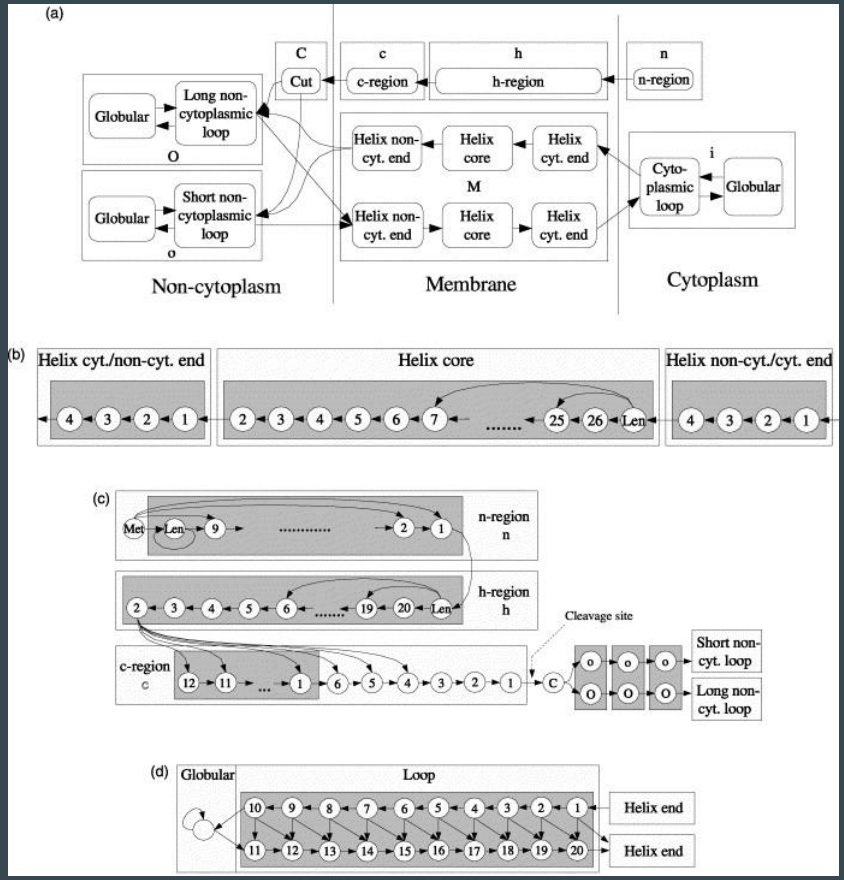
- Based on two kinds of negative data:
  - First: cytoplasmic and, for the eukaryotes, nuclear proteins
  - Second: Comprise sequences not containing signal peptides but containing transmembrane regions within the first 70 residues of the sequence.
- A purely neural network–based method



Petersen, Thomas Nordahl, et al. "SignalP 4.0: discriminating signal peptides from transmembrane regions." Nature methods 8.10 (2011): 785.

Phobius

- A combined transmembrane protein topology and signal peptide predictor.
- Based on a hidden Markov model (HMM)
- Models the different sequence regions of a signal peptide and the different regions of a transmembrane protein in a series of interconnected states.
- Pros:
  - Superior over SignalP in prediction of TM helices
  - More accurate on mixed TM/SP proteins than the best TM-only and SP-only predictors
- For SP-only proteins, it is more conservative than SignalP



The layout of the HMM model. (a) Overview of the model. (b) The TM helix submodel. (c) The signal peptide submodel. (d) The cytoplasmic and short non-cytoplasmic loop submodels.

Phobius

Table 1. Accuracy of transmembrane topology predictions by Phobius, TMHMM 2.0, HMMTOP 2.1, a combination of SignalP-HMM and TMHMM, and a combination of SignalP-HMM and HMMTOP, measured for different data sets

Proteins contain	Both-TM-and-SP		TM-only		SP-only	Neither-TM-nor-SP
Test set sequences	**New**	**All**	**New**	**All**	**All**	**All**
Phobius	94.1%	91.1%	53.9%	63.6%	96.1%	98.2%
TMHMM 2.0	70.6%	71.1%	44.5%	65.2%	73.9%	98.7%
HMMTOP 2.1	52.9%	51.1%	50.8%	66.8%	37.2%	85.0%
TMHMM-SignalP	88.2%	86.7%	39.5%	58.7%	98.0%	99.2%
HMMTOP-SignalP	88.2%	82.2%	45.0%	59.1%	89.6%	86.0%

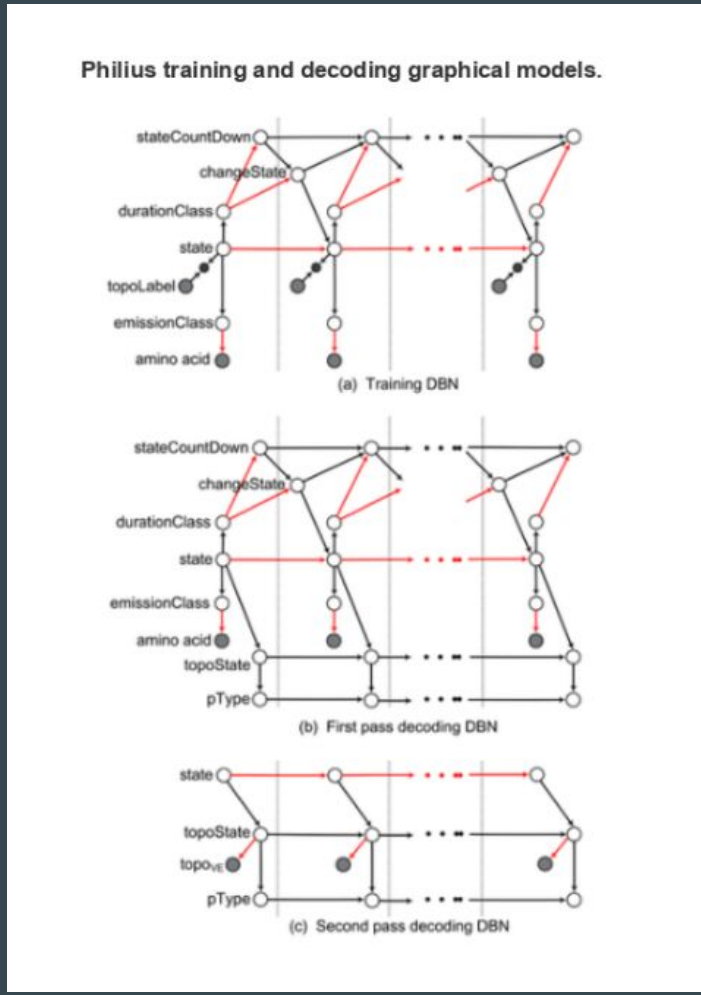
Table 2. Errors in signal peptide predictions made by Phobius and SignalP V2.0.2b

Proteins contain	Both-TM-and-SP	TM-only	SP-only		Neither-TM-nor-SP
Error type	**False negatives**	**False positives**	**False negatives**		**False positives**
Test set sequences	**All**	**All**	**New**	**All**	**All**
Phobius	4.4%	7.7%	2.4%	3.5%	2.3%
SignalP-NN	2.2%	42.9%	2.2%	2.3%	4.8%
SignalP-HMM	0.0%	19.0%	0.6%	1.4%	4.0%

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.

Philius

- Combined transmembrane topology and signal peptide predictor
- Based on dynamic Bayesian networks (DBN)
- Inspired by Phobius and tackles the problem of discriminating among four basic types of proteins:
  - globular (G),
  - globular with a signal peptide (SP+G),
  - transmembrane (TM), and
  - transmembrane with a signal peptide (SP+TM).
- Also predicts the location of the signal peptide cleavage site and the complete topology for membrane proteins.
- Gives a set of confidence measures at three different levels:
  - at the level of protein type,
  - at the level of the individual topology segment (e.g., inside, membrane, outside),
  - at the level of the full topology



Philius

Dataset	Count	Acc	Prec	Sens	Spec	cc	C-Site
Eukaryotes	1192	0.97	0.97	0.97	0.97	0.94	72.4%
Gram−	370	0.97	0.91	0.98	0.97	0.92	87.8%
Gram+	167	0.97	0.81	0.96	0.97	0.86	62.3%
All	1729	0.97	0.98	0.97	0.97	0.94	74.7%

The negative set contained 1087 globular proteins.
doi:10.1371/journal.pcbi.1000213.t005

		Accuracy		Precision		Sensitivity		Specificity		Matthews C	
Protein Type	Data %	Phobius	Philius	Phobius	Philius	Phobius	Philius	Phobius	Philius	Phobius	Philius
TM, SP+TM	11%	0.98	0.98	0.79	0.87	0.91	0.92	0.98	0.98	0.83	0.88
SP+G	48%	0.96	0.95	0.97	0.95	0.94	0.95	0.97	0.96	0.92	0.91
G	41%	0.97	0.97	0.97	0.97	0.96	0.95	0.98	0.98	0.94	0.93

doi:10.1371/journal.pcbi.1000213.t001

Transmembrane Proteins

- High rate of false discovery, due to homology with other hydrophobic proteins such as signal peptides

Phobius

- ML trained tool for differentiation of transmembrane proteins from others
- False discovery of signal peptides reduced from 26% to 4% (over previous tools)
- False discovery of transmembrane proteins reduced from 19% to 8%

LipoP

- Prediction of lipoproteins in Gram-negative bacteria.
- Based on hidden Markov model (HMM), able to predict 96.8% of the lipoproteins(SPaseII-cleaved proteins) correctly with only 0.3% false positives in a set of SPaseI-cleaved, cytoplasmic, and transmembrane proteins.
- Able to identify 92.9% of the lipoproteins included in a Gram-positive test set
- Classifies sequences into **SpI** (signal peptide), **SpII (lipoprotein signal peptide), TMH** (n-terminal transmembrane helix), **CYT (everything else)**
- Offer a Linux version that can be incorporated into our own site (offer free academic download)
    - Command: LipoP -short [Input.fasta] > [Output.gff] ("-short" summarizes the best prediction)

LipoP

Input:
FASTA (can take in both genome sequences and amino acid sequences)

```
>5H2A_CRIGR you can have comments after the ID
MEILCEDNTSLSSIPNSLMQVDGDSGLYRNDFNSRDANSSDASNWTIDGENRTNLSFEGYLPPTCLSILHL
QEKNWSALLTAVVIILTIAGNILVIMAVSLEKKLQNATNYFLMSLAIADMLLGFLVMPVSMLTILYGYRWP
LPSKLCAVWIYLDVLFSTASIMHLCAISLDRYVAIQNPIHHSRFNSRTKAFLKIIAVWTISVGVSMPIPVF
GLQDDSKVFKQGSCLLADDNFVLIGSFVAFFIPLTIMVITYFLTIKSLQKEATLCVSDLSTRAKLASFSFL
PQSSLSSEKLFQRSIHREPGSYTGRRTMQSISNEQKACKVLGIVFFLFVVMWCPFFITNIMAVICKESCNE
HVIGALLNVFVWIGYLSSAVNPLVYTLFNKTYRSAFSRYIQCQYKENRKPLQLILVNTIPALAYKSSQLQA
GQNKDSKEDAEPTDNDCSMVTLGKQQSEETCTDNINTVNEKVSCV
```

Output:
GFF format

The output format is essentially in GFF format. The default (long) output format looks like this:

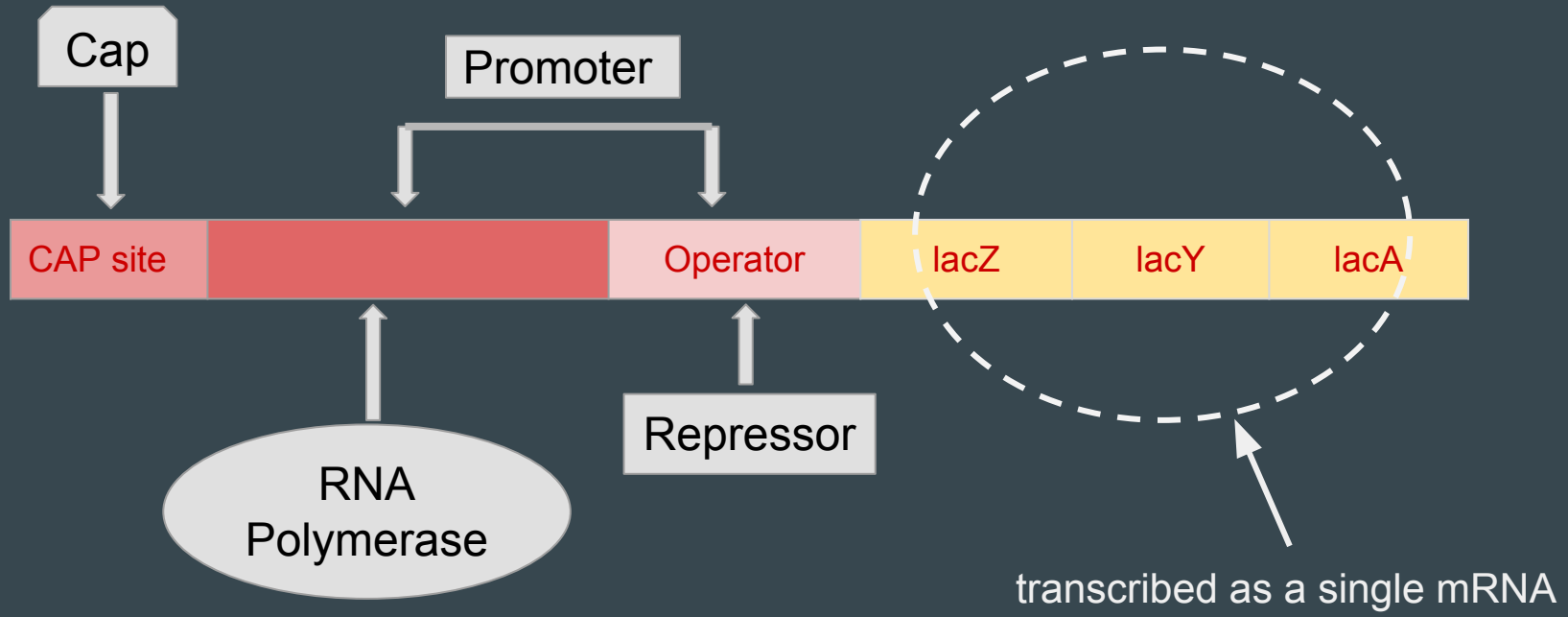
```
# ANIA_NEIGO SpII score=29.6052 margin=11.2327 cleavage=18-19 Pos+2=G
# Cut-off=-3
ANIA_NEIGO      LipoP1.0:Best   SpII    1       1       29.6052
ANIA_NEIGO      LipoP1.0:Margin SpII    1       1       11.2327
ANIA_NEIGO      LipoP1.0:Class  SpI     1       1       18.3725
ANIA_NEIGO      LipoP1.0:Class  CYT     1       1       -0.200913
ANIA_NEIGO      LipoP1.0:Signal CleavII 18      19      29.6052 # FALAA|CGGEQ Pos+2=G
ANIA_NEIGO      LipoP1.0:Signal CleavI  24      25      18.0333 # GGEQA|AQAPA
ANIA_NEIGO      LipoP1.0:Signal CleavI  20      21      15.9259 # LAACG|GEQAA
ANIA_NEIGO      LipoP1.0:Signal CleavI  26      27      12.0794 # EQAAQ|APAET
ANIA_NEIGO      LipoP1.0:Signal CleavI  25      26      11.4077 # GEQAA|QAPAE
ANIA_NEIGO      LipoP1.0:Signal CleavI  27      28      9.40252 # QAAQA|PAETP
```

Operon

- Clusters of coregulated genes
- Physically close in the genome
- All turned on or off together
- Co-transcribed as a single mRNA

(accurate prediction of operons can improve the functional annotation of genes within operons)

Lac Operon


Cap
Promoter
CAP site
Operator
lacZ
lacY
lacA
RNA Polymerase
Repressor
transcribed as a single mRNA

Operon (tool)

DOOR^2 **D**atabase of pr**O**karyotic **O**pe**R**ons

- Comprehensive, the largest public operon databases available
- Web based, need to manually download data

DOOR^2 itself uses data-mining classifier algorithm to predict operons

Workflow with DOOR^2

Download All Operon tables for *Klebsiella pneumoniae*

Extract fasta files for all GI numbers in the table (NCBI Eutil)

Blast query fasta (from last group) against reference DB

Match the results back to original Operon table

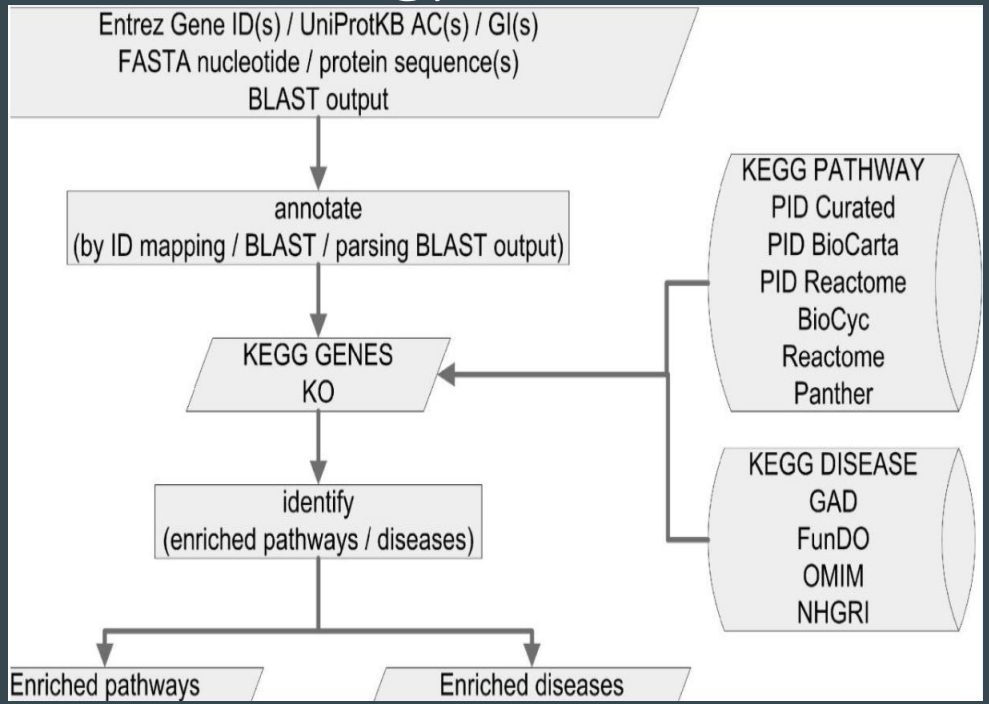
Total 6321 Operons, 32413 Genes

Species	NCs	Genes	Operons
Klebsiella pneumoniae 342	NC_011282(P) NC_011281(P) NC_011283(C)	5768	1114
K. pneumoniae KCTC 2242	NC_017540(C) NC_017541(P)	5264	1004
K. pneumoniae NTUH-K2044	NC_006625(P) NC_012731(C)	5262	1048
K. pneumoniae subsp. pneumoniae 1084	NC_018522(C)	5067	982
K. pneumoniae subsp. pneumoniae HS11286	NC_016847(P) NC_016838(P) NC_016846(P) NC_016841(P) NC_016845(C) NC_016839(P) NC_016840(P)	5867	1140
K. pneumoniae subsp. pneumoniae MGH 78578	NC_009648(C) NC_009649(P) NC_009653(P) NC_009652(P) NC_009650(P) NC_009651(P)	5185	1033

Pathway Annotation & Gene Ontology Tools

Kobas 2.0

- Tailored for disease-related pathways
- Maps genes to the KEGG Genes and KEGG Orthology databases
  - Stored in their SQL relational database
- BLAST against database
  - E-value cutoff (default) of <10e-05 and rank <6



Kobas 2.0 Pipeline

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3125809/

Pathway Annotation Tools & Gene Ontology

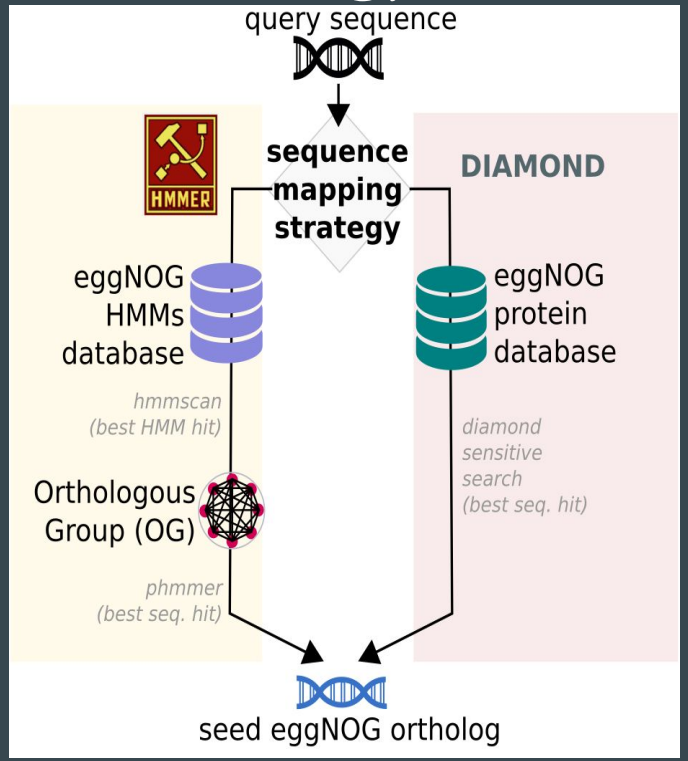
- InterProScan
  - Commonly used Homology based annotation tool
  - Predicts domains & important sites
    - Active sites, binding sites, etc
  - Utilizes protein entries from multiple databases
    - Each database has their own area of focus
  - Last updated 1st March, 2018
    - Aims for 8 week/update to database
  - Outputs in variety of formats
    - XML, JSON, GFF3, TSV
  - Optional parameters for Gene Ontology and Pathways
    - --goterms & --iprlookup/-pa

Signature database
CATH-Gene3D
CDD
HAMAP
PANTHER
Pfam
PIRSF
PRINTS
ProDom
PROSITE patterns
PROSITE profiles
SFLD
SMART
SUPERFAMILY
TIGRFAMs

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3998142/pdf/btu031.pdf
Member databases: https://www.ebi.ac.uk/interpro/release_notes.html

Pathway Annotation Tools & Gene Ontology

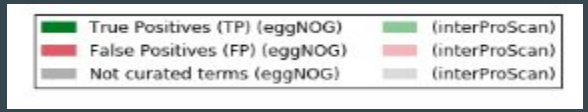
- eggNOG-mapper
  - Database of Orthologous Groups (OGs) and functional annotations
  - Extends the Clusters of Orthologous groups (COG) methodology of graph-based clustering
    - Any 2 proteins from different lineages that still belong to the same COG are orthologs
  - Considered more precise than homology searches, such as BLAST or InterProScan
    - Avoids transferring annotations for duplicate genes involved with functional divergence
  - Reportedly 2.5x faster speed than InterProScan and 15x faster than BLAST
    - DIAMOND mode as sensitive as BLAST-based methods
    - Much faster than HMM. Tailored for >1000 sequences
  - Can reduce total database size by taxonomic constraints
    - ~20GB annotation database
    - ~20GB fasta files
    - ~130GB optimized database all three domains (euk, bact, arch)



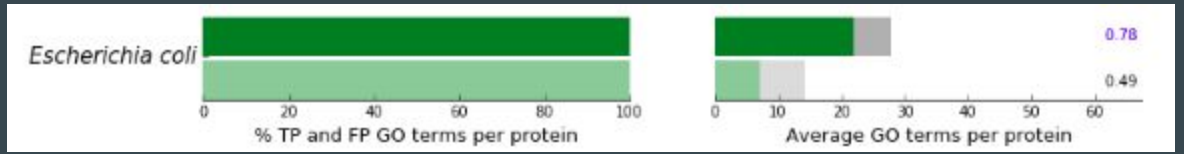
https://doi.org/10.1093/molbev/msx148

InterProScan 5 & eggNOG-mapper

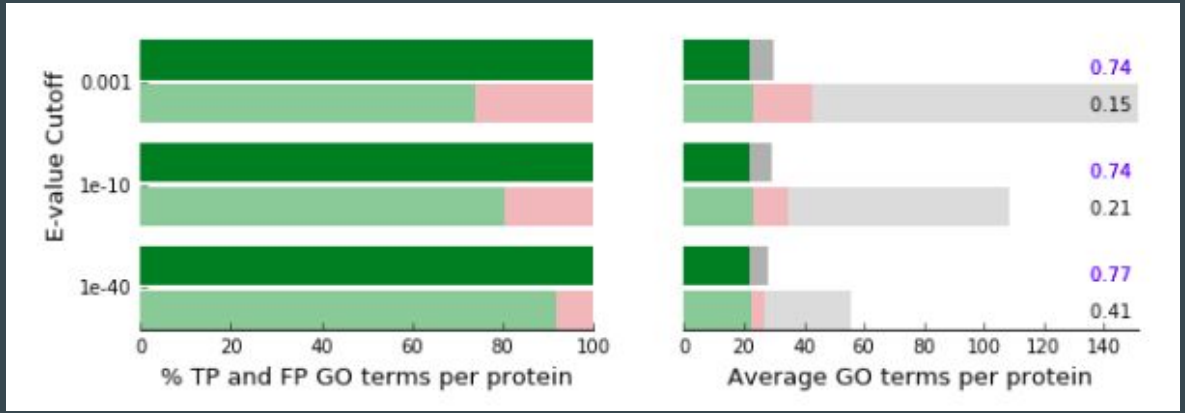
- Local Machine
  - Quad-core, 12GB RAM, SSD
  - InterProScan: ~3.2 hours
  - eggNOG-hmm mode: 31 minutes
- eggNOG Reported Times
  - 6 hours InterProScan
  - ~15 minutes eggNOG-mapper DIAMOND mode
  - Annotate 300-400 proteins with 10 cpus & under /dev/shm filesystem (memory based)



HMM mode and InterProScan



DIAMOND mode and InterProScan



https://academic.oup.com/mbe/article/34/8/2115/3782716
https://github.com/jhcepas/emapper-benchmark/blob/master/benchmark_analysis.ipynb

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

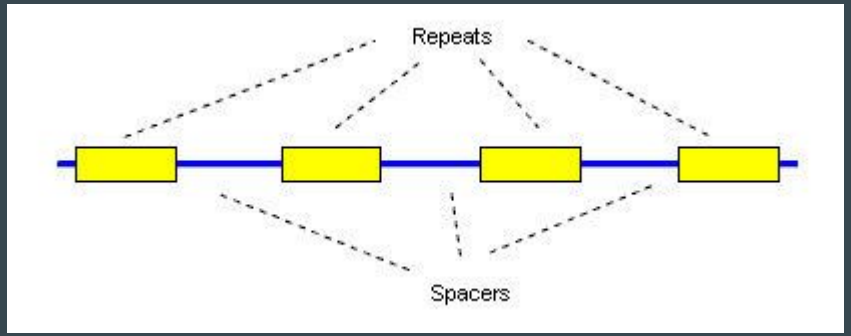
Non-coding RNA

- Rfam
  - Database of ncRNA
  - Group ncRNA into families using multiple sequence alignments and covariance models

- Infernal
  - Identifies tRNAs, rRNAs, other ncRNAs using primary and secondary structure homology searches

CRISPR : **C**lustered **R**egularly **I**nterspaced **S**hort **P**alindromic **R**epeats.

- Found in nearly 40% of all bacterial species
- Related with Bacterial immunity regulation, cell defense mechanism, DNA rearrangement, replication and regulation.
- Act as tools for evolutionary studies and strain production



-*Klebsiella* has an unusually high proportion of self-targeting spacers
-linked to programmed cell-death, gene regulation, and antibiotic resistance.

CRISPR : **C**lustered **R**egularly **I**nterspaced **S**hort **P**alindromic **R**epeats.

Tools:

- PilerCR :
  - Fast (completing a 5 Mb genome in around 5 seconds)
  - High sensitivity (may approach 100% with default parameters)
  - Output: Fasta file

- CRT :
  - Fast and memory efficient (nearly 6 million bases in about 3 seconds)
  - High Recall rate and quality
  - Only supports FASTA file format.
  - Lower precision compared to PilerCR.

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

Annotation of Antibiotic Resistance Genes

Which antibiotics are our samples resistant to?

What are the molecular mechanisms of resistance?

Which classes of proteins are conferring resistance?

Discovery of known AMR genes (homology)

Prediction of new AMR genes (*ab initio)*

The Old and New Standards

ARDB

-DB of 27,000 known genes

-Homology search tool

-Last updated June 2009

IMPORTANT: An up to date resource for antibiotic resistance is the CARD database at http://arpcard.mcmaster.ca. All data in ARDB are also found in CARD, but CARD is actively being maintained.

CARD

-DB of 35,000 genomes

-Homology search (Diamond)

-Protein variant search

-Protein overexpression search

-Command-line tool for parallelization

-Updated 2018

CARD - RGI toolkit

Annotation Classifications

Perfect - 100% matches to curated reference sequences or annotated variants known to confer AMR

Strict - Close matches to known sequences. Includes secondary screen for key mutations likely to confer AMR.

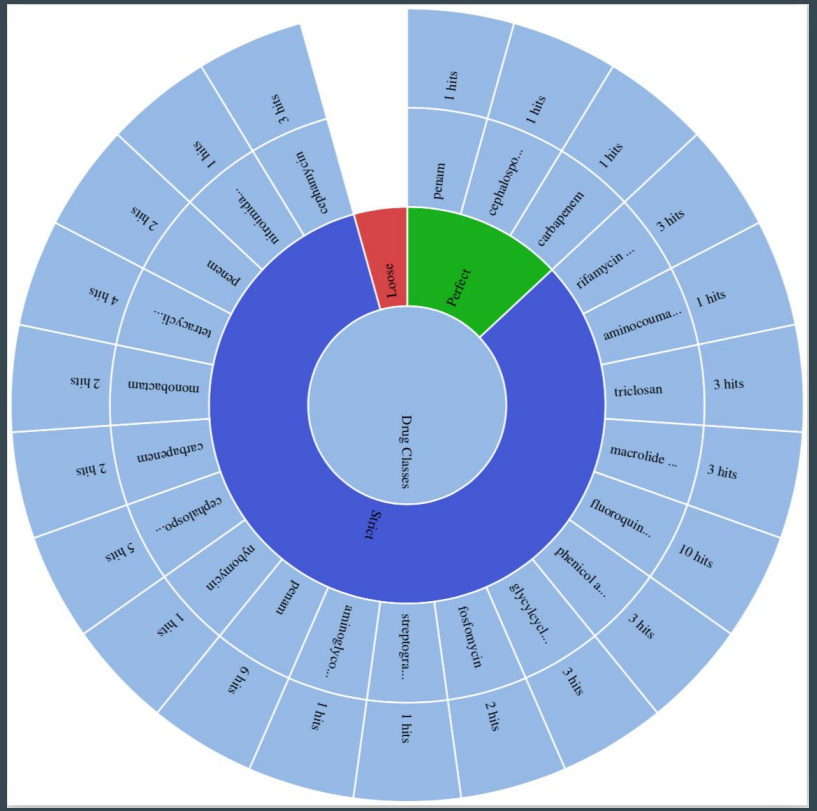
Loose - Detects distant homologues of known AMR genes and partial hits to key motifs. Discovers new targets for experimental phenotyping

Results from Reference *K. pneumoniae*

~ 30 seconds

- 1 Perfect match
- 16 Strict matches
- 401 Loose matches

Browsable displays organized by drug class, gene family, or resistance mechanism.



Other Tools

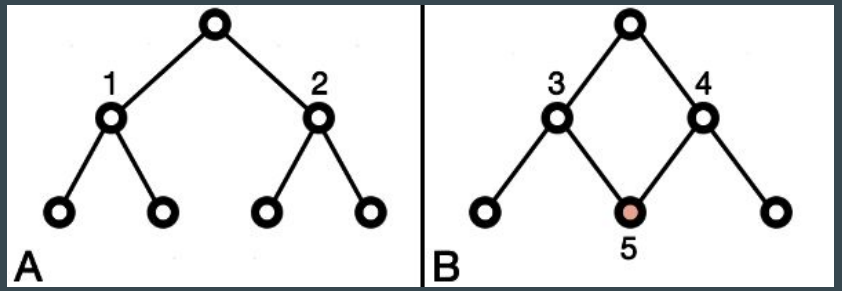
MEGARes

-DB of 4000 AMR genes

-Designed for big data

-Fixes artificial count inflations inherent to CARD data structure

-Updated 2017



FARME DB

-DB composed exclusively of short-read metagenomic data

-Captures unannotated sequences involved in AMR

-Unique annotations not found in other databases

-Updated 2017

Virulence Factors

Tools Available:

Rapid Virulence Annotation (RVA) -web

Virulence Searcher -web

Victors -no *Klebsiella*

Pathogen Host Interactions (PHI-base) -requires translation to protein

Virulence Factors Database (VFDB)

Virulence Factors

Tools Available:

Rapid Virulence Annotation (RVA)

Virulence Searcher

Victors

Pathogen Host Interactions (PHI-base)

Virulence Factors Database (VFDB)

Virulence Factors

Virulence Factors DataBase (VFDB)

Comprehensive database of known and predicted genes

Updated every four years (last update: 2016)

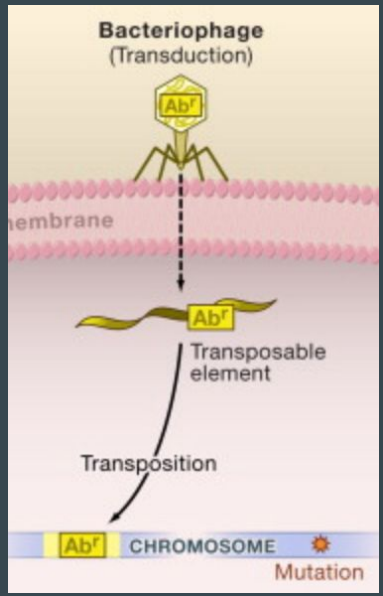
Results contain the name of the virulence factors and its associated function

No command line tool

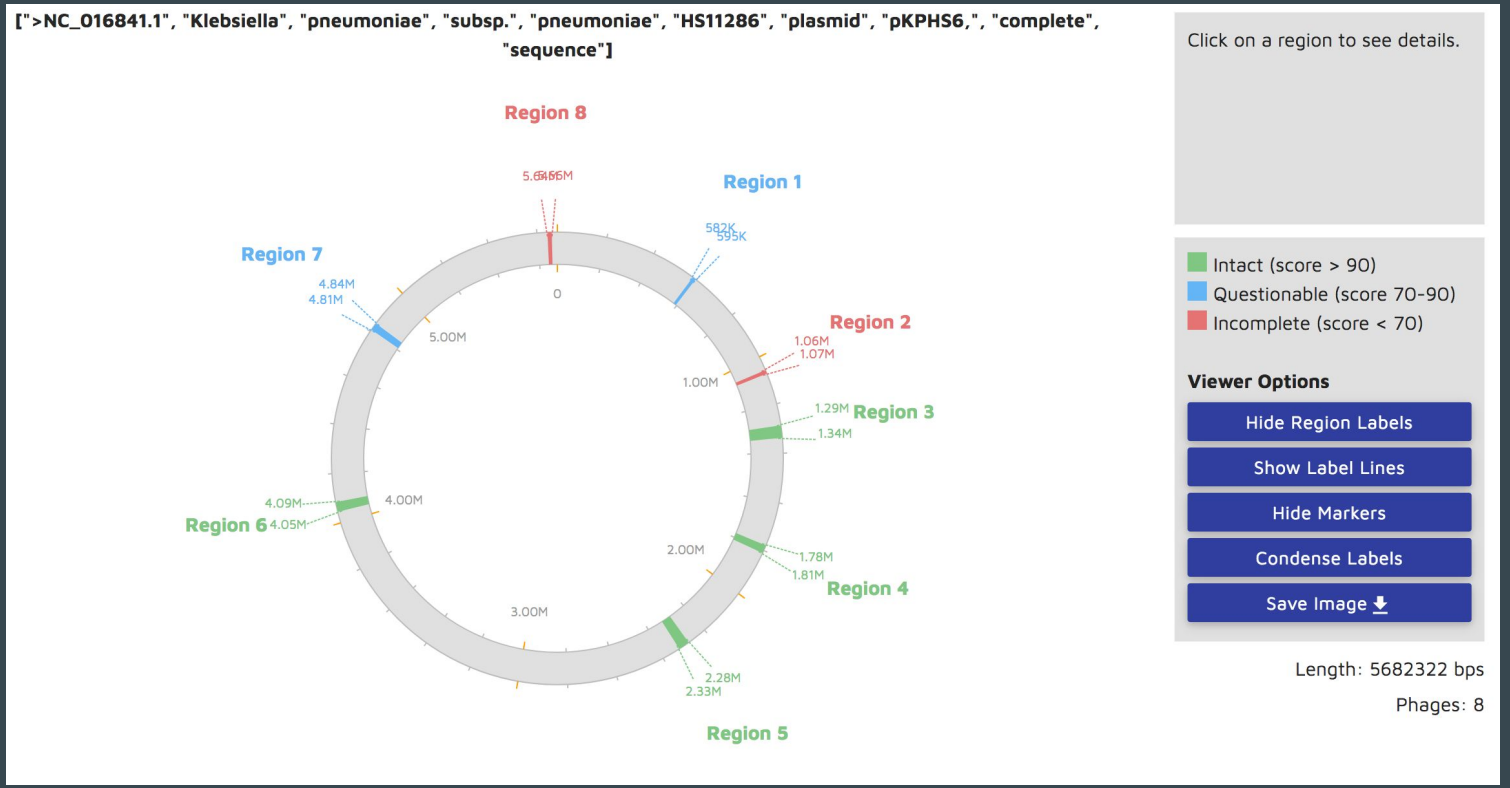
Can only be run using BLAST+

Prophage Genes

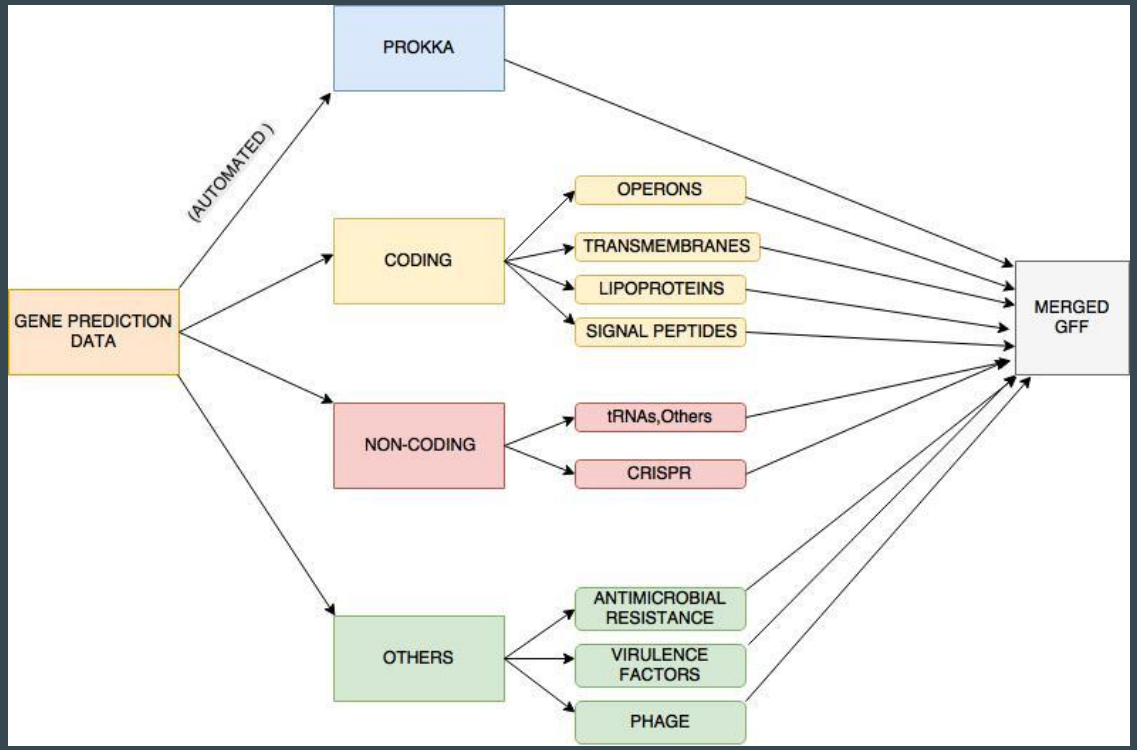
- PHAge Search Tool Enhanced Release (PHASTER)
  - Prophage
    - Bacteriophage DNA that has been integrated into bacterial genome/plasmid
  - PHASTER
    - Annotation of prophage sequences in bacteria
    - Search for similarity against the database
    - 3 mins from raw sequences or 1.5 mins from a pre-annotated GenBank file
    - URL API for large samples
      - Using wget to UPLOAD files and GET results



Prophage Preliminary Results



Proposed Pipeline



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.
- Xie, Chen et al. “KOBAS 2.0: A Web Server for Annotation and Identification of Enriched Pathways and Diseases.” *Nucleic Acids Research* 39.Web Server issue (2011): W316–W322.
- 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?