Background & Strategy

Team 1 Functional Annotation

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

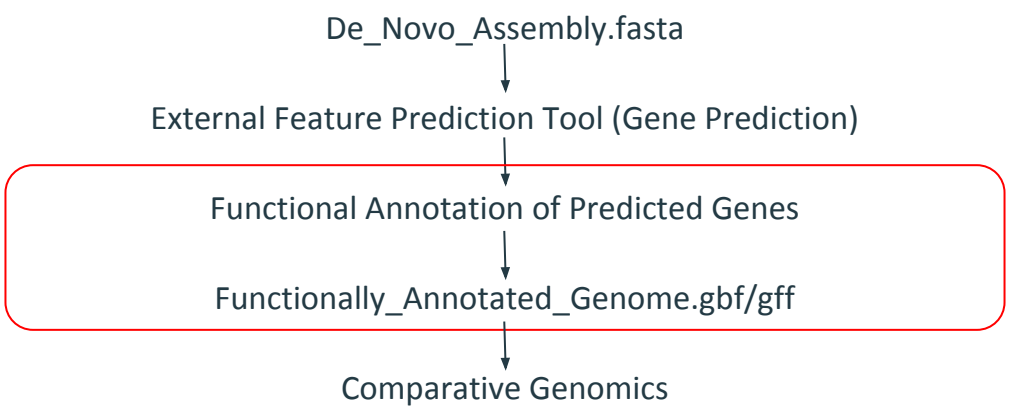
Georgia Tech

Outline

- Introduction to Functional Annotation
- Approaches used in Functional Annotation
- Tools under consideration
  - Homology based tools (Prokka, InterProScan, BLASTDOOR2, Plier-CR)
  - Ab-initio based tools (LipoP, SignalP, TMHMM, DeepARG, Patric)
  - Automated pipelines (blast2go, RAST)
- Preliminary pipeline
- References

Introduction to Functional Annotation

What is our job?



Functional Annotation

- Essential feature of the NGS analysis pipeline
- Provides the Comparative Genomics team with required data
- Uses predicted genomic feature locations from the Gene Prediction group
- Different types of genomic features are identified with different tools
    - Homology based - databases of known genes
        - Blast - compare sequence data to databases of known genomic feature functions
        - Interproscan - compares again a high quality combined redundancy-free database
    - Ab Initio - sequences with characteristic attributes of a particular subset of genomic features
        - LipoP - predicts lipoproteins
        - SignalP - predicts signal peptides
        - TMHMM - predicts transmembrane helices in proteins

Homology

Prokka

Prokaryotic Genome Functional Annotation Tool

- Developed by Victorian Bioinformatics Consortium
- Significantly faster than other tools available at the time of release
- Currently Prokka is an industry standard

Relative runtimes according to the literature

"There are various online annotation servers (Stewart *et al.* , 2009). The NCBI provides a Prokaryotic Genomes Automatic Annotation Pipeline service via email, with a turn-around time measured in days. RAST is a web server for annotating bacterial and archaeal genomes that provides annotation results in under a day (Aziz *et al.* , 2008), and xBASE2 does similar in a few hours (Chaudhuri *et al.* , 2008). These classes of tools are valuable, but they are not useful where throughput or privacy is critical.
Here we present Prokka, a command line software tool that can be installed on any Unix system. Prokka coordinates a suite of existing software tools to achieve a rich and reliable annotation of genomic bacterial sequences. Where possible, it will exploit multiple processing cores, and a typical bacterial genome can be annotated in ~10 min on a quad core desktop computer. It is well suited to iterative models of sequence analysis and integration into genomic software pipelines."

–Torsten Seemann; Prokka: rapid prokaryotic genome annotation, *Bioinformatics*, Volume 30, Issue 14, 15 July 2014, Pages

Relative runtimes according to the literature

Time required to functionally annotate a single genome on a quad core computer

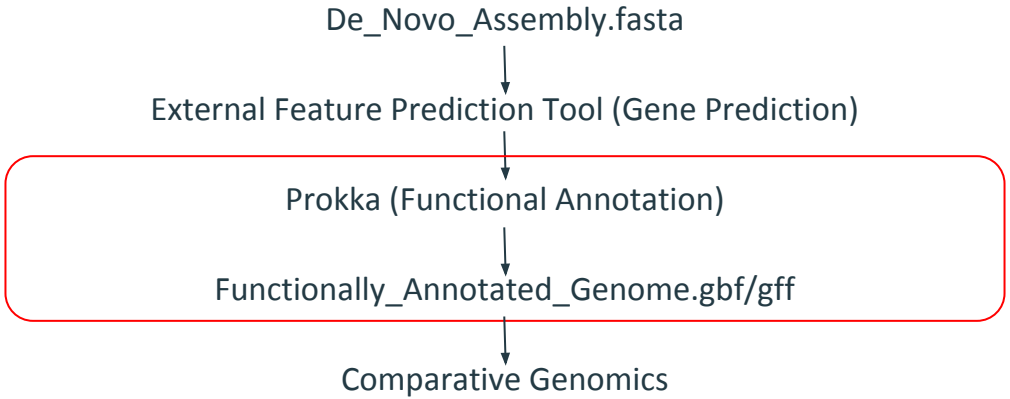
PGAAP - Days

RAST - Hours

PROKKA - Minutes

Prokka

What does it do?



Prokka

How does it annotate genes?

- Prodigal identifies the coordinates of potential genes and passes this information to Prokka
- Prokka then uses a hierarchical series of sequence/database comparisons
  - Small - Most Trustworthy
    - CARD Database - continuously curated database of AMR genes
  - Medium - Domain Specific
    - UniProtKB/Swiss-Prot - manually annotated, non-redundant
  - Large - Protein Families
    - RefSeq
    - Pfam
    - TIGRFAMs

RAST

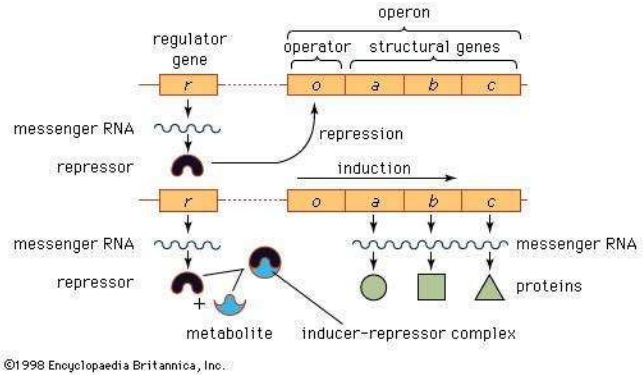
- Rapid Annotation using Subsystem Technology
- Predominantly offered as a web service
- Offered on command line as RASTtk (RAST Tool Kit)
- Upon initial consideration this tools does not appear to scale well
- Required for the use of PATRIC (ML Algorithm)

Interproscan

- Interproscan is a large tool that predicts proteins and functional domains, and summarized their output.
- This is done by 15 applications bundled within the software.
    - (CDD, Coils, Gene3D, HAMAP, MobiDB-lite, PANTHER, Pfam, PIRSF, PRINTS, ProDom, PROSITE, SFLD, SMART, SUPERFAMILY, TIGRFAMs)
    - 4 more applications can also be bundled within (TMHMM, Phobius, and SignalP)
- Computationally intensive
- Allows user to specify which applications to use

DOOR2

- Operon database with ~1.3 million operons for ~2 thousand prokaryotic genomes
  - Operon: a unit of genomic feature with linked function
- Has no command line
  - Use database with BLAST



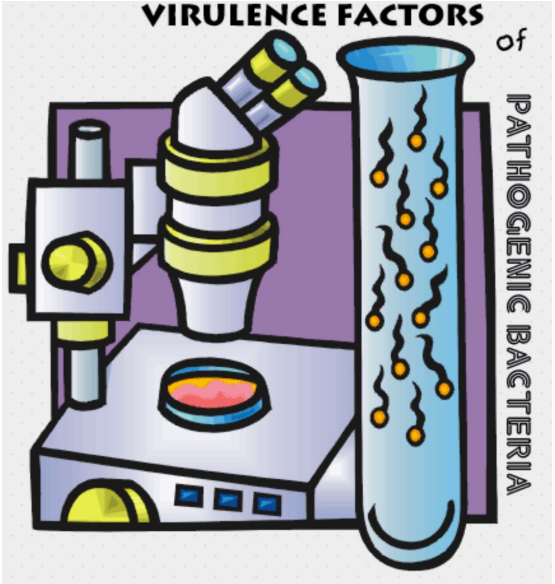
VFDB

- an integrated and comprehensive online resource for curating information about virulence factors of bacterial pathogens

Summary statistics

Type of data	Number
Bacterial pathogens (by genus)	74
of which with full information	30
Virulence factors	1796
of which experimentally verified	532
Bacteria strains involved	926
of which with complete genome	504
VF-related genes (non-redundant) (Set B)*	30178
of which with curation (Set A)*	2599
Related literatures	2508

*uncharacterized genes in PAIs were excluded



VFDB

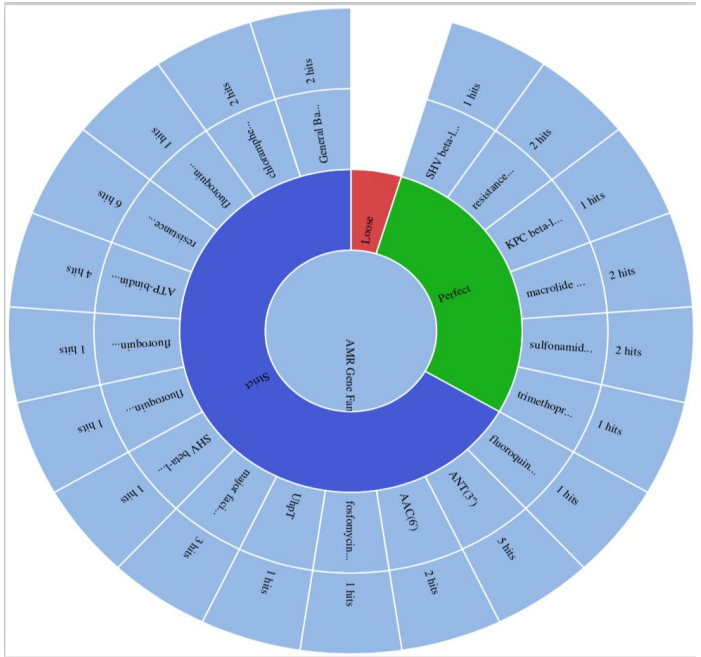
- VFs collected based on PubMed papers, GenBank files.
- Could be searched by text, blast or function category keywords.
- The DB could be download in fasta format and the size (compressed) is really small. (DNA sequences,9.01MB, proteins sequences 4.86MB)
- Easy to blast in command line

CARD

The Comprehensive Antibiotic Resistance Database

- What it do:
  - Describe antimicrobial molecules and their targets, resistance mechanisms, genes and mutations, and their relationships
  - Predict antibiotic resistance genes from sequence data

- Why choose it:
  - Rigorously Curated
  - Updated Frequently (monthly)



Uniprot Database

- Composed of four categories:
  - Uniprot Knowledgebase (UniProtKB)
    - Functional information of proteins
      - Swiss-Prot (manually annotated and reviewed)
      - TrEMBL (automatically annotated and not reviewed)
  - Uniprot Reference Clusters
    - Clustered sets of sequences from UniProtKB and UniParc
  - UniParc
    - Non-redundant database for protein sequences
- Also contains a proteomes database
  - Has 29 reference *Klebsiella* proteomes

Databases

- Uniprot
- Interpro
- CARD
- DeepARG-DB
- ARDB
- VFDB
- RefSeq
- DOOR2

CRISPR

- **C**lustered **R**egularly **I**nterspaced **S**hort **P**alindromic **S**equences
- They are parts of genomes containing repetitive short (20-30 bp) sequences separated by unique spacer sequences.
- These repeats are found in 40% of all bacteria species.
- Have several biological functions such as-
    - Host cell defense
    - DNA rearrangement
    - Replication and Regulation
- These sequences can be used as a tool for evolutionary study and strain typing.

PILER-CR

- Tool designed to identify and classify CRISPR repeats.
- Algorithm:-
  - Local alignment of genome with itself, construction of self-similarity plot (dot plot).
  - Identification of contiguous sets of bases (called a pile) , each of which is covered by at least 1 local alignment.
  - Construction of graph with piles as nodes.
  - Draft array identification, each pile is tested as the possible first member of a putative CRISPR array.
  - Array refinement, applies heuristics to improve the final inferred array from previous step.
  - Merge adjacent arrays with similar consensus sequences and similar spacer lengths.
  - Cluster consensus sequences into similar groups and carry out multiple alignment.
  - Report generation

PILER-CR

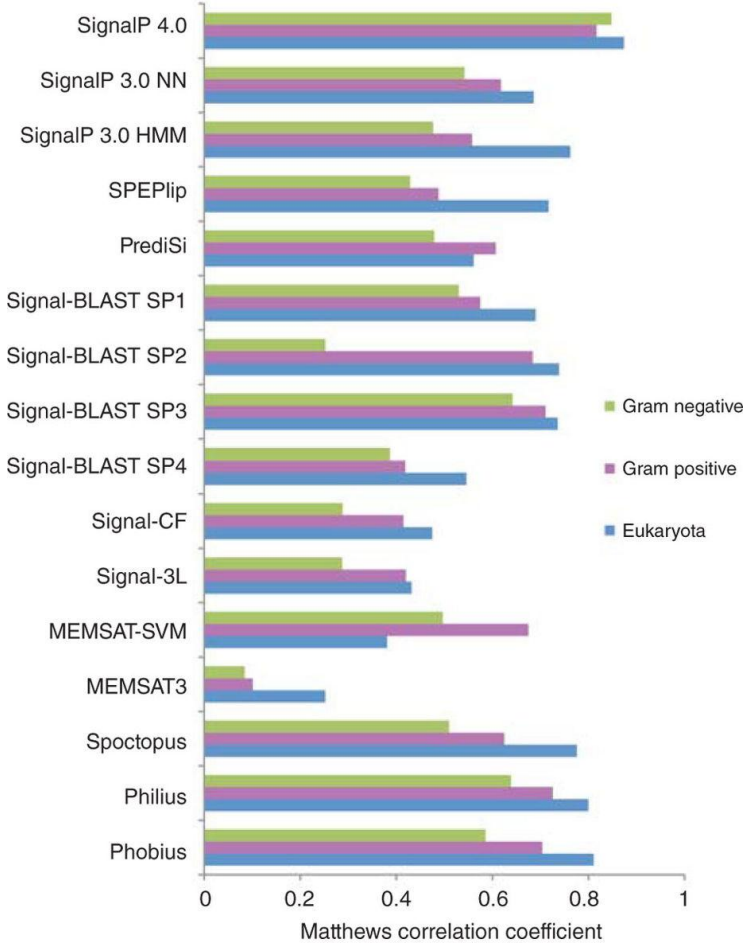
- This tool works really fast, completes a 5Mb genome in around 5 seconds on a current desktop computer.
- When tested for 346 prokaryotic genomes it completed in 15 minutes on a 2GHz desktop computer.
- Has very high sensitivity (>90%).

Ab Initio

SignalP

- Predict signal peptides from amino acid sequences
    - Input: FASTA
    - Output: GFF
- Utilizes a neural network approach
    - Discriminate between signal peptides and transmembrane regions
    - Two networks
        i. SignalP-TM : transmembrane sequence used in training
        ii. SignalP-noTM: trained without transmembrane sequence
    - Selection scheme
        i. If a transmembrane helix is > 4 residues… use SignalP-TM prediction
        ii. Else… use SignalP-noTM prediction

SignalP


SignalP 4.0
SignalP 3.0 NN
SignalP 3.0 HMM
SPEPlip
PrediSi
Signal-BLAST SP1
Signal-BLAST SP2
Signal-BLAST SP3
Signal-BLAST SP4
Signal-CF
Signal-3L
MEMSAT-SVM
MEMSAT3
Spoctopus
Philius
Phobius
Gram negative
Gram positive
Eukaryota
0
0.2
0.4
0.6
0.8
1
Matthews correlation coefficient

SignalP

- Version 4.0
  - Not as sensitive as v3.0...
  - Version 4.1 offers cutoff options which partially addresses this issue (increases FP rate)

iii: *Gram-negative bacterial sequences*

Method	All Sequences			Only TM	No TM
	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

LipoP

- Tool to predict lipoprotein producing genes in bacteria
- Based on Hidden Markov Model
- Trained on sequences from Gram-negative bacteria samples from SwissProt
- Classifies genes into 4 categories based on proteins they produce
    - Signal peptide (Signal peptidase I-cleaved proteins)
    - Lipoprotein signal peptide (Signal peptidase II-cleaved proteins)
    - N-terminal membrane helices
    - Cytoplasmic
- Advantage:
    - 94.6% of lipoproteins in test set were correctly classified

Juncker, A. S., Willenbrock, H., Von Heijne, G., Brunak, S., Nielsen, H., & Krogh, A. (2003). Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Science*, *12*(8), 1652-1662.

TMHMM

- Tool used to detect transmembrane proteins.
- Helical membrane proteins follow a *“grammar”* in which cytoplasmic and non-cytoplasmic loops alternate.
- Based on Hidden Markov Model
- The HMM has several sub-models corresponding to a different region of a membrane protein such as:- globular, helix core, helix cap etc.
- The sub-models have multiple states to model the lengths of the various regions.
- Has very high (>99%) sensitivity and specificity in the test data set.

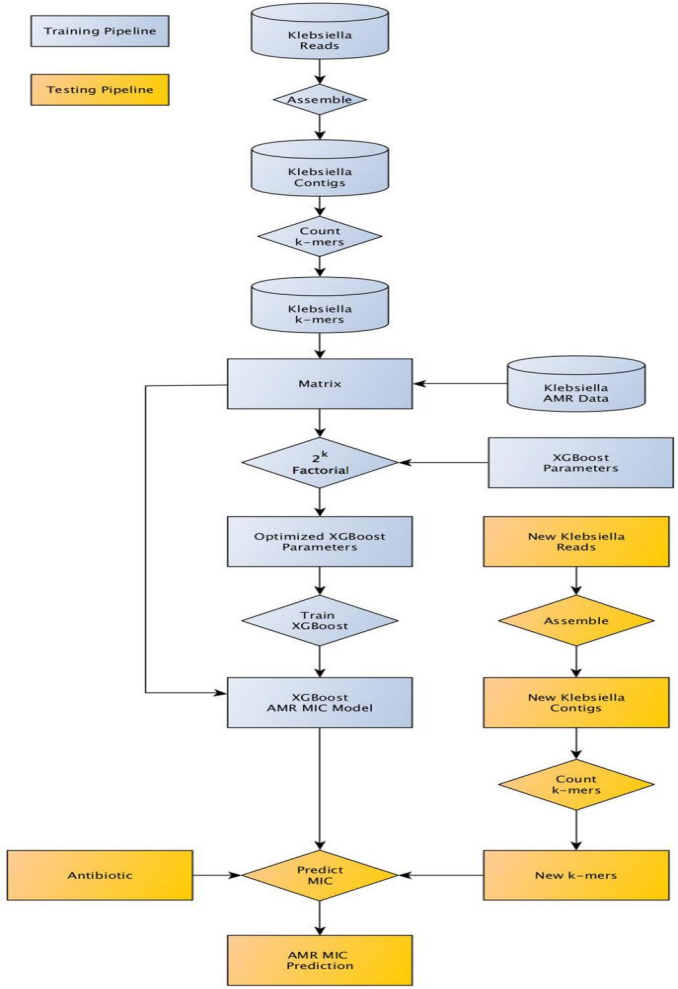
J Mol Biol. 2001 Jan 19;305(3):567-80. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. Krogh A1, Larsson B, von Heijne G, Sonnhammer EL.

PATRIC3

- Trains a model for predicting minimum inhibitory concentrations (MICs)
- Uses RAST as the annotation tool
- Requires a large set of genomes with associated phenotypic data
- Uses Adaboost or XGBoost to train the model

PATRIC3

- PATRIC is a suite of tools including RASTtk
- Machine learning algorithm determines associations between genomic feature and phenotypic antibiotic resistance
- A model has been trained for Klebsiella
- This model does not predict colistin resistance



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5765115/

DeepARG

- Predict antibiotic resistance genes (ARGs) from metagenomic data or assembly genome.

Input	Output
FASTA file	blast tab delimited alignment file
BLAST tabular file	annotated reads

- Uses a neural network approach
  - Developed a new ARG database name DeepARG-DB: collection of CARD, ARDB and UniProt
    - 30 antibiotic categories, 2149 groups, and 14,933 references sequences*.
  - Developed two models:
    - DeepARG-SS: constructed for short read sequences (NGS data)
    - DeepARG-LS: constructed for full gene length sequences
  - Training model is time-consuming, but has already been done and released.
  - Prediction time will be in order of minutes**.

Arango-Argoty, G., Garner, E., Pruden, A., Heath, L. S., Vikesland, P., & Zhang, L. (2018). DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome*, *6*(1), 23.

DeepARG

- Comparison between DeepARG and the typical best hit approach*:
  - Both can reach high precision (low false positive rates)
  - DeepARG is able to yield consistently lower false negative rates
    - Best hit approach has high identity cutoffs (80% or 90%). DeepARG does not have a strict identity cutoff.
      - In reality, ARGs can have a significant e-value but a low identity (such as 30%) when comparing to known ARGs in databases.
    - DeepARG is capable of predicting novel ARGs of known categories** in DeepARG-DB, while best hit approach usually performs poorly in finding novel ARGs.
    - DeepARG performs better than best hit approach in PseudoARG test***.

Arango-Argoty, G., Garner, E., Pruden, A., Heath, L. S., Vikesland, P., & Zhang, L. (2018). DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome*, *6*(1), 23.

Limitations of DeepARG

Like all neural network approaches, the prediction accuracy and specificity of DeepARG strongly relies on the quality of ARGs in the database.

The limitations of DeepARG includes:

- Cannot predict antibiotic resistance that arises from SNPs.
- Can only predict whether a gene or read belongs to one of the 30 categories that are considered by the model.
- Predicting power relies on the quality of the training database
    - Annotation errors will adversely affect the prediction of the models
- Downstream validation is still needed*.

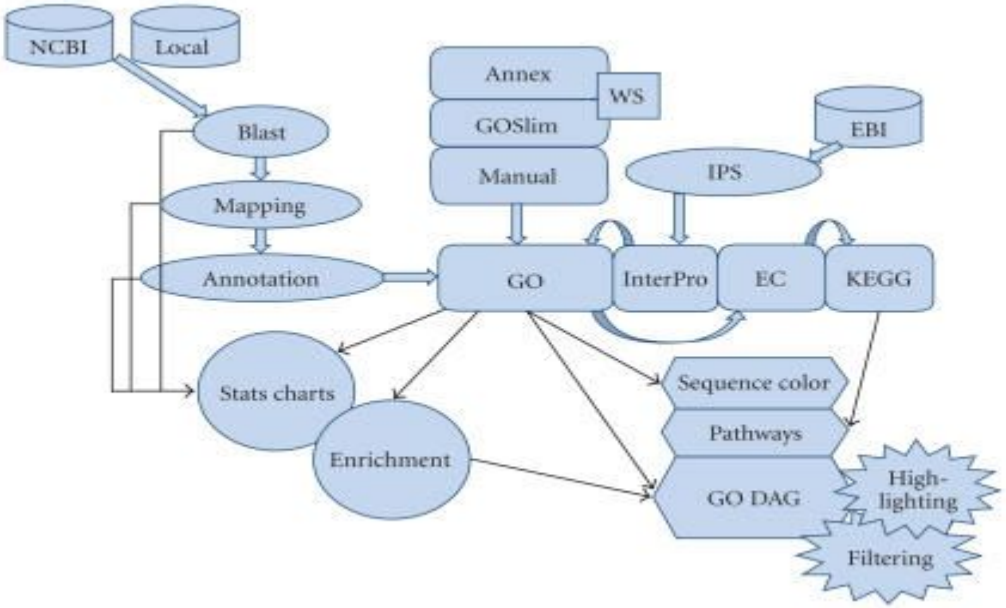
Arango-Argoty, G., Garner, E., Pruden, A., Heath, L. S., Vikesland, P., & Zhang, L. (2018). DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome*, *6*(1), 23.

Automated Pipeline

BLAST2GO

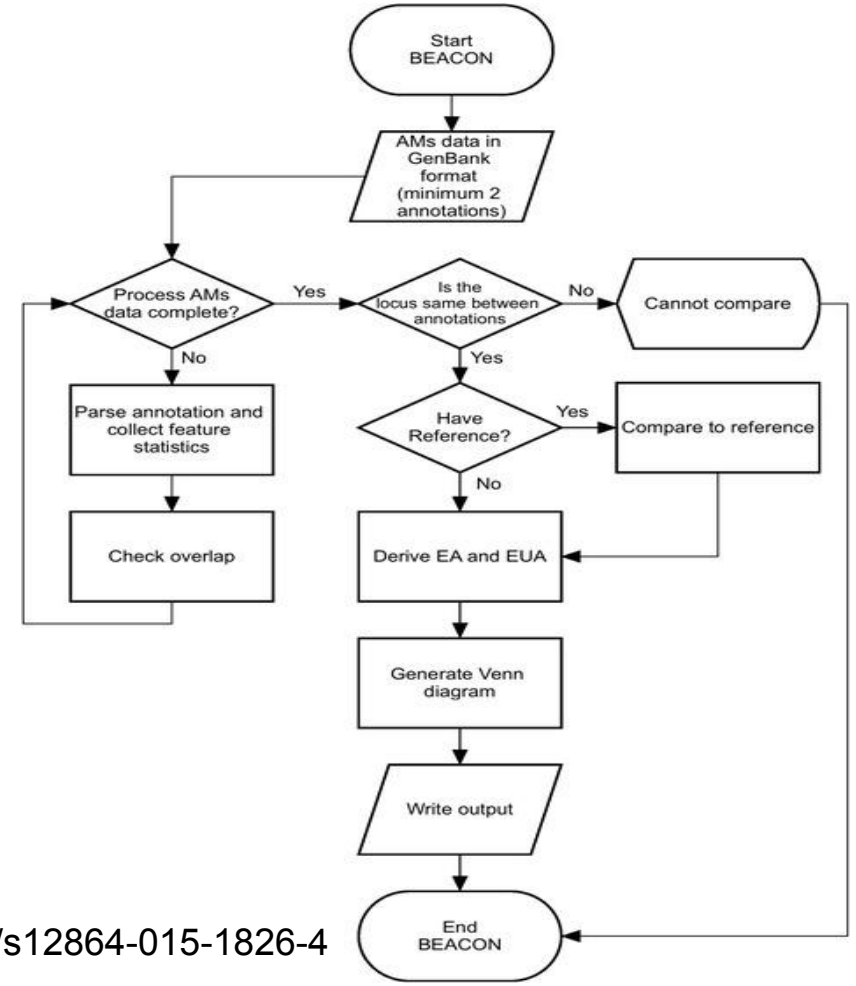
- Tool for functional annotation of novel sequences and the analysis of annotation data.
- Annotation based on homology
- Performs functional annotation in three steps:
  - Homologous sequences are identified using BLAST
  - Mapping is done to retrieve Gene Ontology terms
  - Annotation to select functions
- Supports annotation databases like InterPro, Enzyme Codes, KEGG pathways and GO.
- Command line version is not free, desktop version has limits on features.

BLAST2GO


NCBI
Local
Blast
Mapping
Annotation
Annex
WS
GOSlim
Manual
EBI
IPS
GO
InterPro
EC
KEGG
Stats charts
Enrichment
Sequence color
Pathways
GO DAG
High-lighting
Filtering

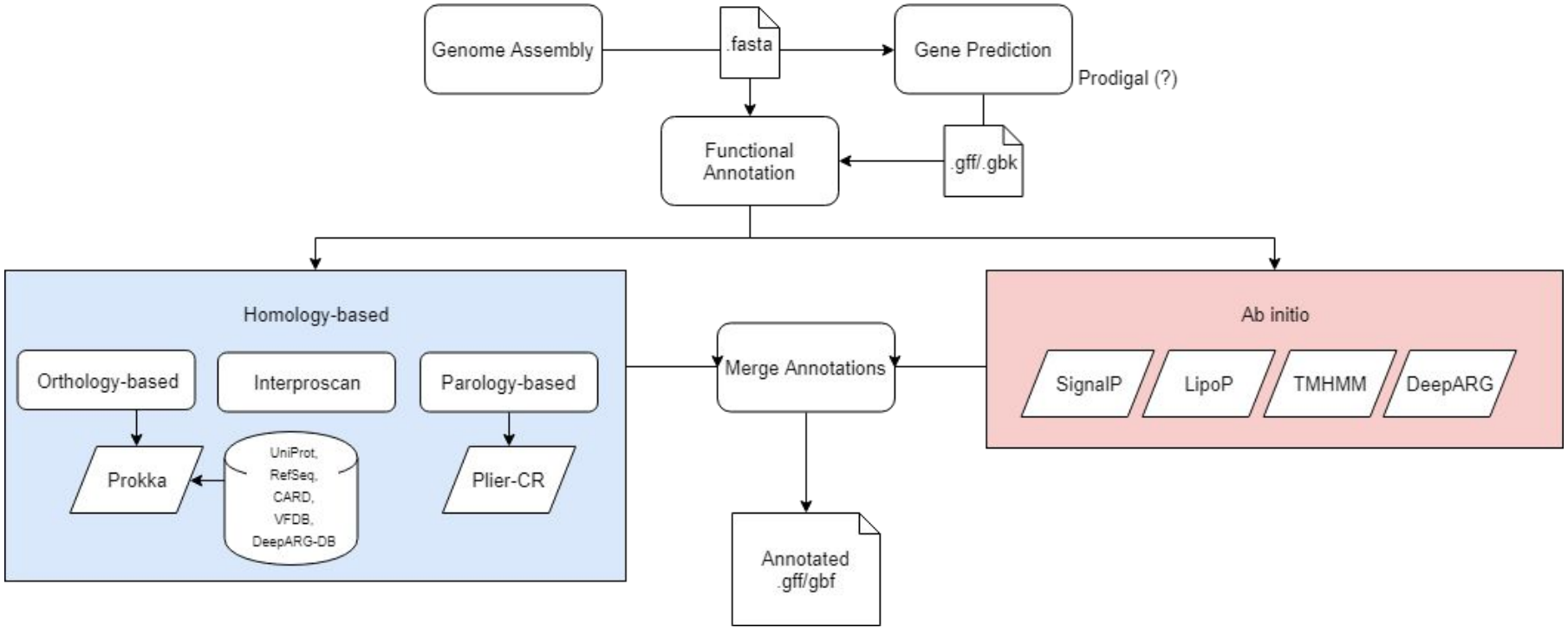
BEACON

- Compares annotations produced by various annotation methods (AMs)
- Using this comparison a consensus may be reached for loci with discrepancies between predicted genomic feature function
- Can combine annotations from various AMs to generate an Extended Union Annotation



https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-015-1826-4

Proposed Pipeline



References

Bendtsen, J. D., Nielsen, H., von Heijne, G., & Brunak, S. (2004). Improved prediction of signal peptides: SignalP 3.0. *Journal of molecular biology*, *340*(4), 783-795.

Petersen, T. N., Brunak, S., von Heijne, G., & Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature methods*, *8*(10), 785.

Nielsen, H. (2017). Predicting secretory proteins with SignalP. *Protein Function Prediction: Methods and Protocols*, 59-73.

Jia, B., Raphenya, A. R., Alcock, B., Waglechner, N., Guo, P., Tsang, K. K., … McArthur, A. G. (2017). CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Research*, *45*(Database issue), D566–D573. http://doi.org/10.1093/nar/gkw1004

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., … Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics, 30(9), 1236–1240. http://doi.org/10.1093/bioinformatics/btu031

http://csbl.bmb.uga.edu/DOOR/index.php

Arango-Argoty, G., Garner, E., Pruden, A., Heath, L. S., Vikesland, P., & Zhang, L. (2018). DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome*, *6*(1), 23.

Conesa, A., & Götz, S. (2008). Blast2GO: A comprehensive suite for functional analysis in plant genomics. *International journal of plant genomics*, *2008*.

References

Juncker, A. S., Willenbrock, H., Von Heijne, G., Brunak, S., Nielsen, H., & Krogh, A. (2003). Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Science*, *12*(8), 1652-1662.

The UniProt Consortium. (2015). UniProt: a hub for protein information. Nucleic Acids Research, 43(Database issue), D204–D212. http://doi.org/10.1093/nar/gku989

J Mol Biol. 2001 Jan 19;305(3):567-80. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. Krogh A1, Larsson B, von Heijne G, Sonnhammer EL.

Nguyen, M., Brettin, T., Long, S. W., Musser, J. M., Olsen, R. J., Olson, R., … Davis, J. J. (2018). Developing an *in silico* minimum inhibitory concentration panel test for *Klebsiella pneumoniae*. *Scientific Reports*, *8*, 421. http://doi.org/10.1038/s41598-017-18972-w