Functional Annotation

Background & Strategy

Team 1: Jiyeong Choi
Asmita Kishor Lagwankar
Chloe Elizabeth Pryor
Hannah Snyder
Likitha Venkatesh
Jiahong Zang

What is functional annotation?

Describe the biochemical and biological function of proteins

Types of Functional Annotation

Ab-Initio

- *Ab initio* - “from the beginning”

- No external evidence is available to identify a gene
- Mathematical models
- Does not need experimental data

Homology Based

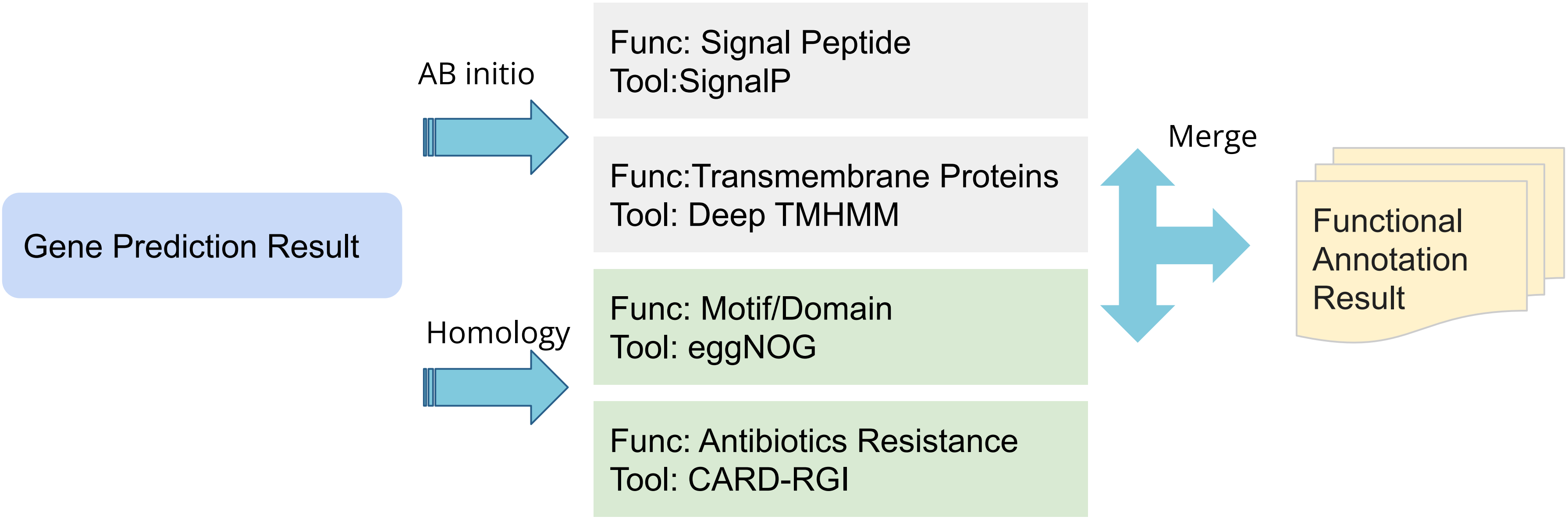
- Evidence Based Annotation

- Rely on comparison between sequences
- Uses information about known structure of related proteins to predict unknown

Gene functional Annotation Strategy

We are going to choose 4 different aspects of functions to predict. (Signal Peptide, Transmembrane Proteins, Motif/Domain and Antibiotics Resistance).

Use 2 homology based methods to predict 2 functions and use 2 ab initio methods to predict other 2 functions.



Gene Prediction Result

AB initio

Func: Signal Peptide
Tool:SignalP

Func:Transmembrane Proteins
Tool: Deep TMHMM

Homology

Func: Motif/Domain
Tool: eggNOG

Func: Antibiotics Resistance
Tool: CARD-RGI

Merge

Functional Annotation Result

Ab - Initio Tools

Signal Peptides

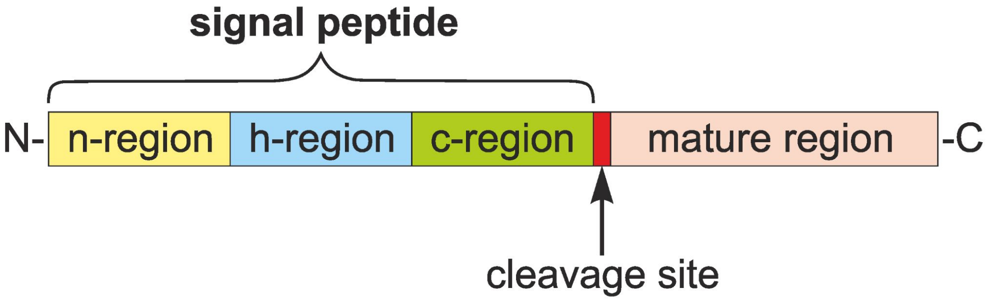
- Signal peptides (SPs) are short amino acid sequences that control protein secretion and translocation in all living organisms.

- They play an important role in targeting proteins for secretion or for transfer to specific organelles for further processing

- SP prediction tools enable identification of proteins that follow the general secretory or twin-arginine translocation (Tat) pathway and predict the position in the sequence where a signal peptidase (SPase) cleaves the SP.

Tools:

- SignalP
- Phobius
- LipoP
- TargetP

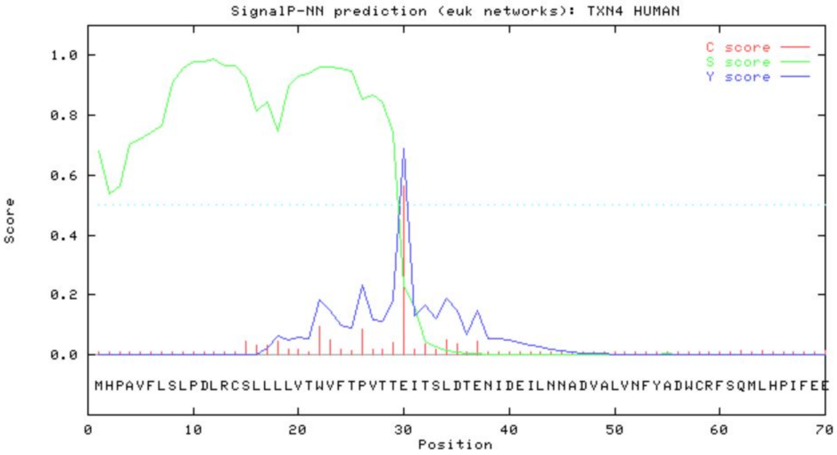


SignalP

- It uses a deep neural network-based method for improved SP prediction
- Predicts the presence and location of signal peptide cleavage sites in amino acid sequences from different organisms
- Input: Amino acid sequence in fasta format
- SignalP truncates every sequence to 70 amino acids before prediction

- Output:

1. Raw cleavage site score(C-score): This is the output from the cleavage site networks which are trained to distinguish Sp cleavage sites from everything else
2. Signal peptide score (S-score): It is to distinguish positions within SPs from positions in the mature part of the proteins and from the proteins without Sps.
3. Combined Cleavage site score(Y-score): it is a combination of the C-score and the slope of the S-score, resulting in a better cleavage site prediction than the raw C-score alone.



Transmembrane Proteins

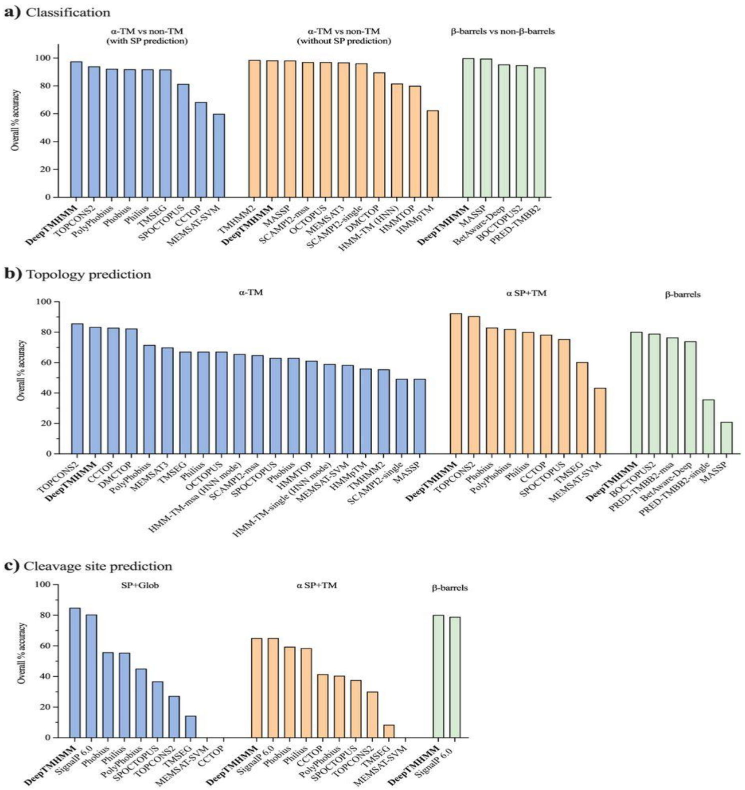
Transmembrane helices are essential components of transmembrane proteins.

TM proteins are membrane-bound receptors and channels with particular pharmacological significance (therapeutic or vaccine target)

Numerous transmembrane proteins serve as passageways that allow the passage of particular substances across the membrane.

Tools:
Deep TMHMM(Hidden Markov Model)
TOPCONS(consensus based approach)
HMMTOP(hidden Markov Model)
MemSAT-SVM(Support Vector Machine)

Deep TMHMM



- a deep learning protein language model-based algorithm
- encoder consists of three components: a pre-trained language model (ESM-1b), a bi-directional LSTM and a dense layer with drop-out
- takes a protein sequence as input and outputs the corresponding per-residue sequence of labels

Homology Based Tools

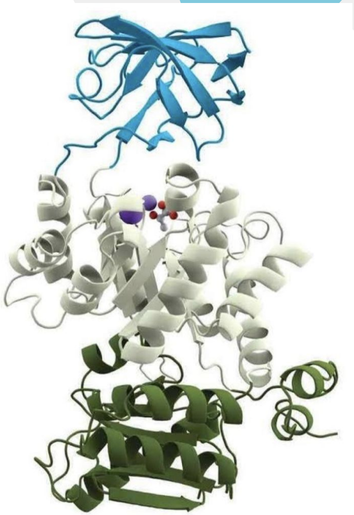
Domain / Motif

Domain

Structural, functional, or evolutionary units of protein

Relatively short (20-100 residues)

Proteins can comprise a single domain or a combination of domains



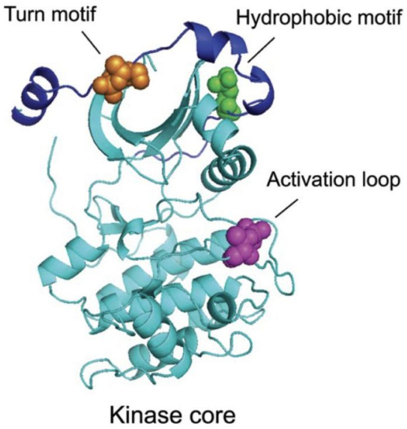
Motif

A region of protein that has a specific structures

Very short (5-20 residues)

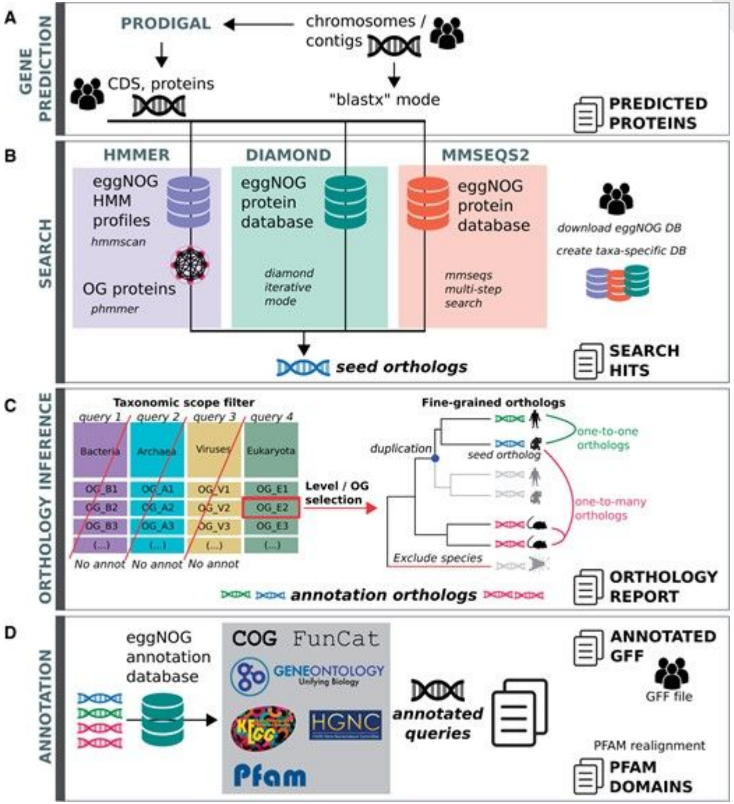
The most conserved region in a domain, candidates for functionally important sites

Used as a base of protein classification



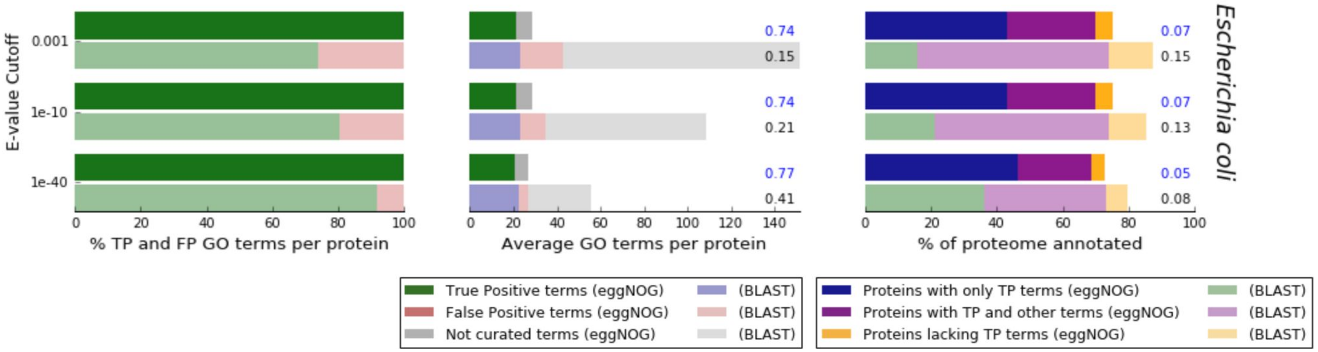
eggNOG-mapper

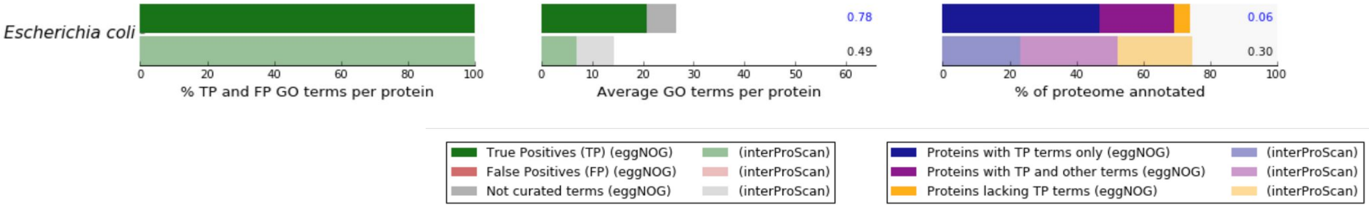
- A tool for fast functional annotation of novel sequences
- It uses precomputed orthologous groups (OGs) and phylogenies from the eggNOG database
- STEP1 Sequence Mapping
  - Search for significant matches in the precomputed collection
  - Hmmer, DIAMOND, MMseqs2
  - The best matching sequence is stored as seed ortholog
- STEP2 Orthology Inference
  - Retrieve fine-grained orthologs from pre-analyzed eggNOG phylogenetic trees
- STEP3 Annotation
  - All functional descriptors available for the retrieved orthologs are transferred to the corresponding query proteins
  - Provides PFAM and SMART protein domain predictions.
    - With very little impact on computational cost



eggNOG-mapper

- Runs ~15 x faster than BLAST and at least 2.5 x faster than InterProScan
- A higher precision than traditional homology searches





Antibiotics Resistance

Background:

Antibiotics and antifungals save lives, but their use can contribute to the development of resistant germs.

Antibiotics resistance is accelerated when the presence of antibiotics and antifungals pressure bacteria and fungi to adapt.

Basic mechanism:

random mutation of bacterial DNA generates a wide variety of genetic changes.

Through mutation and selection, bacteria can develop defense mechanisms against antibiotics.

For example, some bacteria have developed biochemical “pumps” that can remove an antibiotic before it reaches its target, while others have evolved to produce enzymes to inactivate the antibiotics

The antibiotics resistant germs survive and multiply. These surviving germs have resistance traits in their DNA that can spread to other germs.

The Comprehensive Antibiotic Resistance Database(CARD) & Resistance Gene Identifier (RGI)

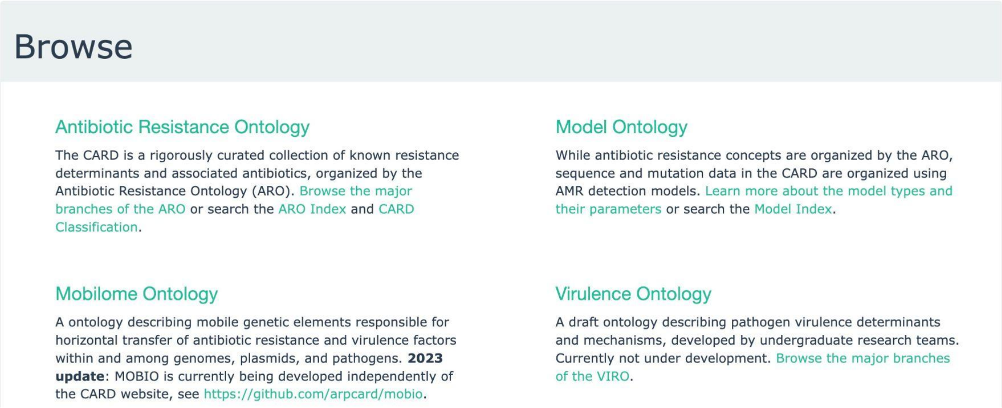
> The Comprehensive Antibiotic Resistance Database (CARD) is a biological database that collects and organizes reference information on antimicrobial resistance genes, proteins and phenotypes.

> RGI: The application uses reference data from the CARD; RGI analyses can be performed via the CARD website RGI portal, via use of a Galaxy wrapper for the Galaxy platform, or alternatively install RGI from Conda or run RGI from Docker

CARD

Use or Download Copyright & Disclaimer

Help Us Curate: github.com/arpcard/amr_curation



Browse

Antibiotic Resistance Ontology

The CARD is a rigorously curated collection of known resistance determinants and associated antibiotics, organized by the Antibiotic Resistance Ontology (ARO). Browse the major branches of the ARO or search the ARO Index and CARD Classification.

Model Ontology

While antibiotic resistance concepts are organized by the ARO, sequence and mutation data in the CARD are organized using AMR detection models. Learn more about the model types and their parameters or search the Model Index.

Mobilome Ontology

A ontology describing mobile genetic elements responsible for horizontal transfer of antibiotic resistance and virulence factors within and among genomes, plasmids, and pathogens. **2023 update**: MOBIO is currently being developed independently of the CARD website, see https://github.com/arpcard/mobio.

Virulence Ontology

A draft ontology describing pathogen virulence determinants and mechanisms, developed by undergraduate research teams. Currently not under development. Browse the major branches of the VIRO.

Resistance Gene Identifier (RGI)

Input data:

If **DNA FASTA** sequences are submitted, RGI first predicts complete open reading frames (ORFs) using Prodigal (ignoring those less than 30 bp) and analyzes the predicted protein sequences.

If **protein FASTA** sequences are submitted, RGI skips ORF prediction and uses the protein sequences directly.

Model:

Protein Homolog Models (PHM) detect protein sequences based on their similarity to a curated reference sequence, using curated BLASTP bit-score cut-offs.

Protein Variant Models (PVM) perform a similar search as Protein Homolog Models (PHM), but secondarily screen query sequences for curated sets of mutations to differentiate them from antibiotic susceptible wild-type alleles.

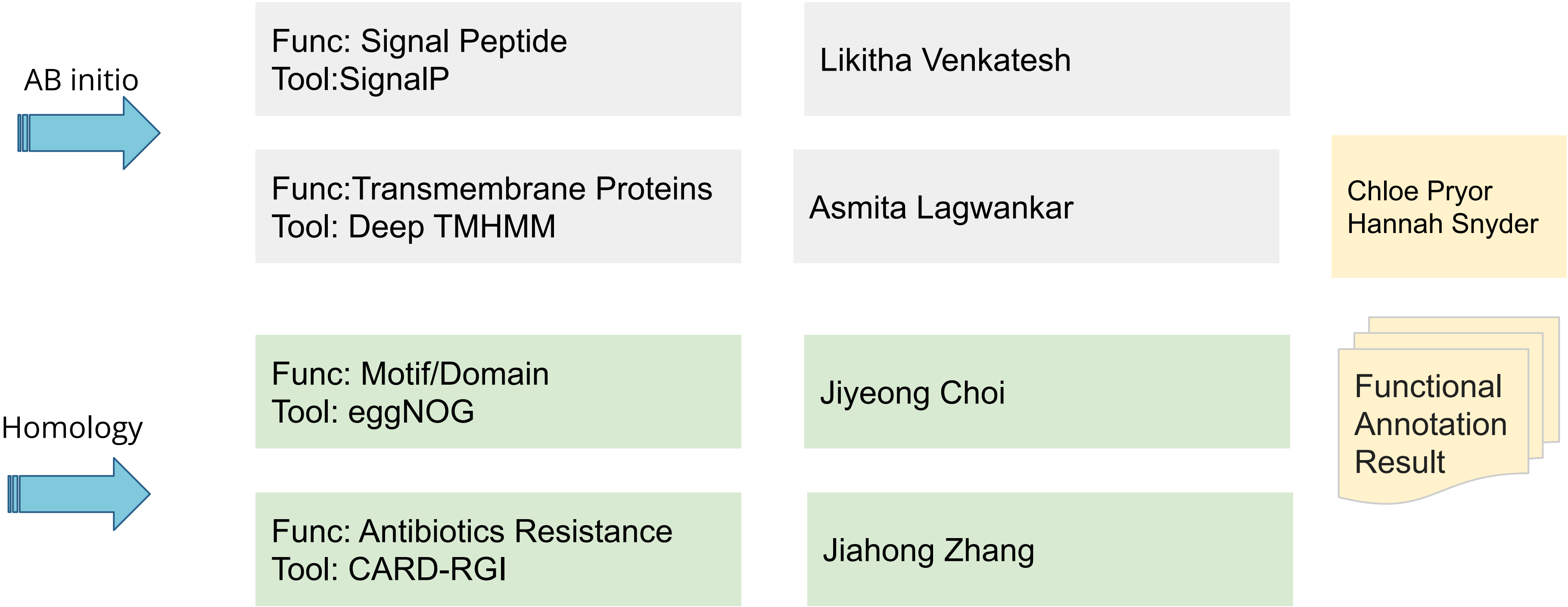
Protein Overexpression Models (POM) are similar to Protein Variant Models (PVM) in mapped resistance variants. POMs are restricted to regulatory proteins and report both wild-type sequences and/or sequences with mutations leading to overexpression of efflux complexes.

Ribosomal RNA (rRNA) Gene Variant Models (RVM) are similar to Protein Variant Models (PVM).RVMs include a rRNA reference sequence (often from antibiotic susceptible wild-type alleles), a curated bit-score cut-off, and mapped resistance variants.

How to Decide on What Tools to Use

- Performance
  - Memory, resources, speed, consistency
  - Documentation, reproducibility, ease of use
- Quality of annotations
  - Completeness, accuracy, documented usage
  - Alignment scores, bit scores, percent identity, domains detected
- ORForise annotation comparisons
  - Work by gene prediction team
  - https://github.com/NickJD/ORForise

Work Delegation



AB initio	Func: Signal Peptide Tool:SignalP	Likitha Venkatesh	
	Func:Transmembrane Proteins Tool: Deep TMHMM	Asmita Lagwankar	Chloe Pryor Hannah Snyder
Homology	Func: Motif/Domain Tool: eggNOG	Jiyeong Choi	Functional Annotation Result
	Func: Antibiotics Resistance Tool: CARD-RGI	Jiahong Zhang	

References

McArthur, A. G., Waglechner, N., Nizam, F., Yan, A., Azad, M. A., Baylay, A. J., ... & Wright, G. D. (2013). The comprehensive antibiotic resistance database. *Antimicrobial agents and chemotherapy*, *57*(7), 3348-3357.

DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks | bioRxiv

Alcock, B. P., Huynh, W., Chalil, R., Smith, K. W., Raphenya, A. R., Wlodarski, M. A., ... & McArthur, A. G. (2023). CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Research*, *51*(D1), D690-D699.

Loewenstein, Y., Raimondo, D., Redfern, O.C. *et al.* Protein function annotation by homology-based inference. *Genome Biol* 10, 207 (2009). https://doi.org/10.1186/gb-2009-10-2-207

Sinha, S., Lynn, A.M. & Desai, D.K. Implementation of homology based and non-homology based computational methods for the identification and annotation of orphan enzymes: using *Mycobacterium tuberculosis* H37Rv as a case study. *BMC Bioinformatics* 21, 466 (2020). https://doi.org/10.1186/s12859-020-03794-x

U.S. National Library of Medicine. (n.d.). *NCBI Prokaryotic Genome Annotation Pipeline*. National Center for Biotechnology Information. Retrieved from https://www.ncbi.nlm.nih.gov/genome/annotation_prok/

Yandell, M., Ence, D. A beginner's guide to eukaryotic genome annotation. *Nat Rev Genet* 13, 329–342 (2012). https://doi.org/10.1038/nrg3174

Aroul-Selvam, R., Hubbard, T., & Sasidharan, R. (2004). Domain insertions in protein structures. *Journal of molecular biology*, *338*(4), 633–641. https://doi.org/10.1016/j.jmb.2004.03.039

Carlos P Cantalapiedra, Ana Hernández-Plaza, Ivica Letunic, Peer Bork, Jaime Huerta-Cepas, eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale, *Molecular Biology and Evolution*, Volume 38, Issue 12, December 2021, Pages 5825–5829, https://doi.org/10.1093/molbev/msab293

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., von Mering, C., & Bork, P. (2017). Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Molecular biology and evolution*, *34*(8), 2115–2122. https://doi.org/10.1093/molbev/msx148

Questions?