Monday Meeting presentation

21/10/2024

PanRes 2

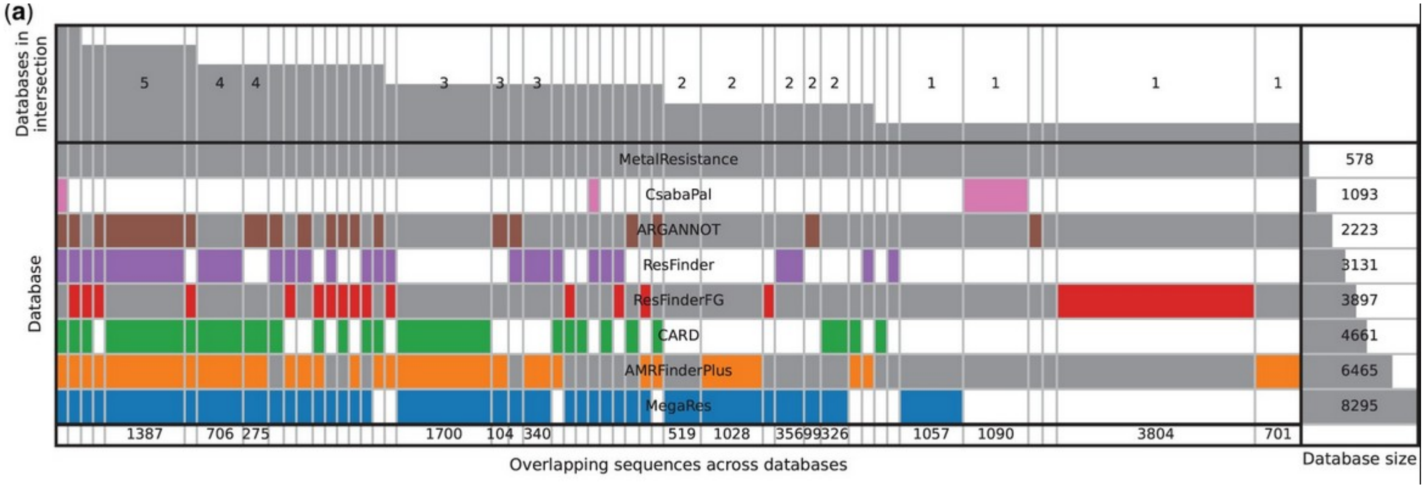
Nikiforos Pyrounakis
Research assistant

PanRes db

- Pan Resistance

- Collection of genes that encode resistance to antibiotic drugs, heavy metals and biocides:
    - ResFinder
    - ResFinderFG
    - CARD
    - MegaRes
    - AMRFinderPlus
    - ARGANNOT
    - CsabaPal*

PanRes db

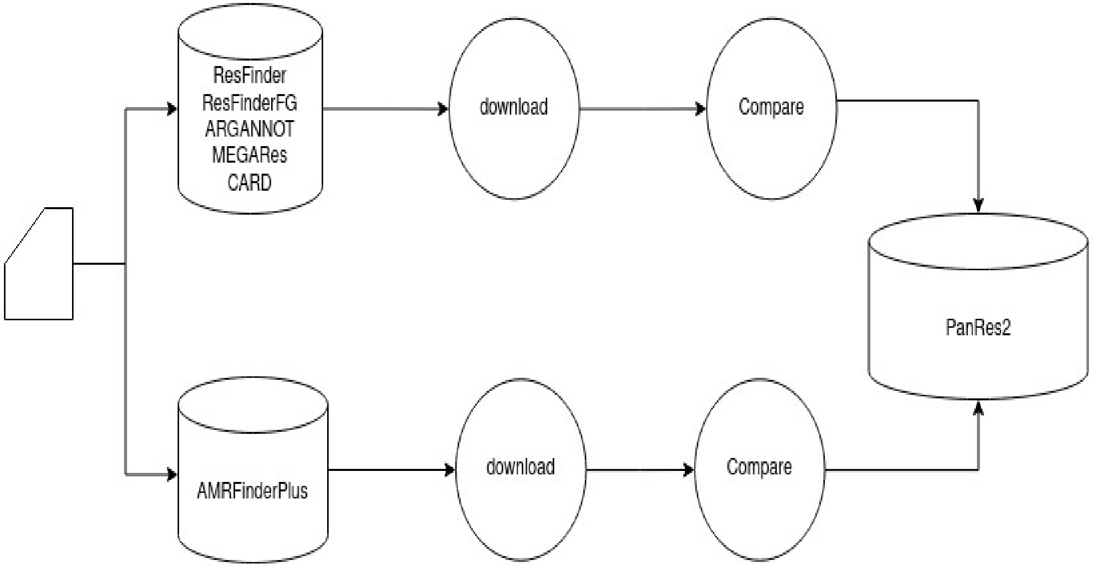


- 14078 unique genes
- Grouped based on 90% identity and 90% coverage
- Lengths range between 93 and 5972 bp

PanRes db 2

- Automatic db update
- Protein sequences for each gene
- PDB structures for each gene
- HMM profile for each high-homology cluster

Automated update



- Check databases for new genes
- Compare new db version with the old
- Update PanRes 2 with new genes

Automated update

- Databases were downloaded either using git or wget

- AMRFinderPlus special case
  - ✗ Extract sequences that had type:
    - ✗ Type: AMR – subtype:AMR
    - ✗ Type: STRESS – subtype:METAL
    - ✗ Type: STRESS – subtype:BIOCIDE

    - ✗ Use esearch and efetch to get sequences from NCBI

- All the new sequences are compared with the old ones to check for new genes

Protein sequences - Prodigal

- **Single**:
  - Used on well-assembled genomes of high quality.
  - Uses a statistical model that its being trained based on the input data.
  - Re-trains the model iteratively to fine-tune the predictions

- **Meta**:
  - Used for metagenomic data.
  - Relies on pre-built models.
  - Sacrifices some of the specificity in favor of more flexible gene detection.

Protein sequences - Validation

- Prodigal Single:
  - ➔ 14138 sequences
    - ✓ Multiple translations: 67
    - ✓ 3 Ratio: ~95%
    - ✓ Non 3 Ration: ~ 5%

- Prodigal Meta:
  - ➔ 14212 sequences
    - ✓ Multiple translation: 116
    - ✓ 3 Ratio: ~ 93%
    - ✓ Non 3 Ratio: ~ 7%

Protein sequences - Validation

Single	Meta
14138	14212
13985	13966
13460	13280
13293*	13091*

1. Remove genes that had multiple translations

2. Remove genes that had gene_length/protein_length ratio <> 3 (73% MEGAres)

3. Remove genes that were located in reverse strand with incorrect start or stop codon?

PDB Structures

- Genes with 100% identity: 505
- Genes with 99%-80% identity: 5968
- Genes with 79%-50% identity: 2269
- Genes with 49%-20% identity: 4710

RCSB PDB

PROTEIN DATA BANK

PDB Structures

- How long will it take to predict structures?
- How expensive will it be?
- Computerome? DTU HPC?
- Testing and benchmarking different tools
  - **Collabfold**
  - Foldseek
  - Alphafold

ColabFold

- User friendly and accessible
  - MMseqs2 integration
  - Faster than Alphafold

HMM Profiles

- “How similar proteins we should put together in each HMM for PanRes2?”
- Testing what NCBI AMRFinderPlus and ResFams did and how they build their profiles
- Different approaches?
    - Try to find a percentage of identity to cluster and then predict?
    - Predict a subset of structures, cluster and then create HMMs

Thank you!