Theory and use of bioinformatics tools to detect AMR genes from genomes

Michael Feldgarden

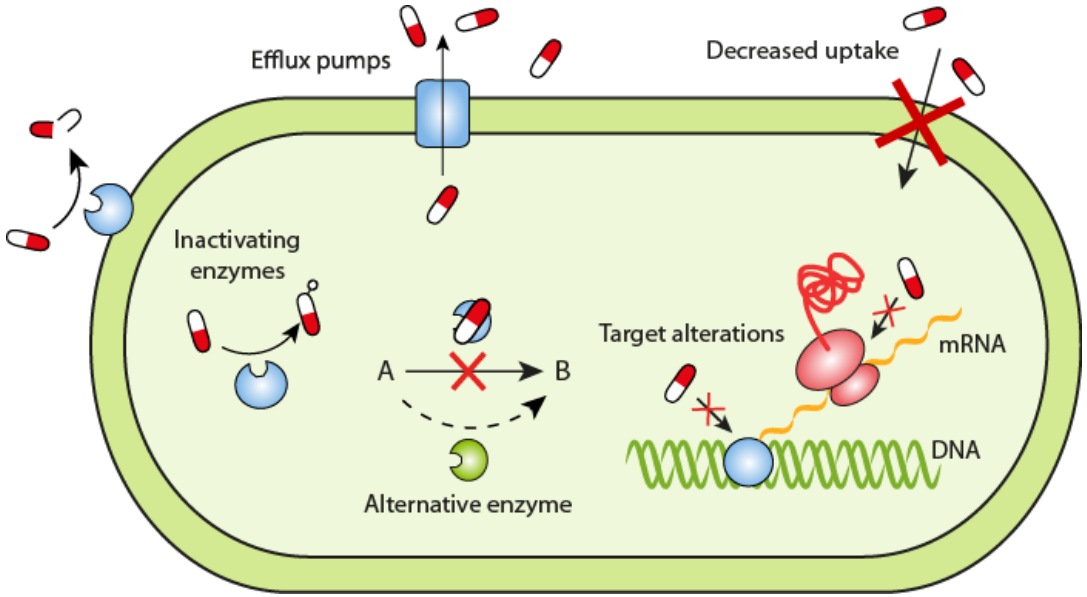
pd-help@ncbi.nlm.nih.gov

U.S. National Library of Medicine
National Center for Biotechnology Information

Why Use These Tools (Align Your Tools with Your Goals)

- Applied uses:
  - Surveillance
    - Often focuses on ‘good’ genes with strong evidence that are known to have an effect
  - Clinical use
    - Edge cases/errors are…**bad**
- Research:
  - Gene discovery
  - Might want to cast a wider, less precise net
- Understand the goals of the tool(s) you are using

Mechanisms of Antibiotic Resistance



- Point mutations (and small insertions/deletions)
- Acquired genes
- *Gene disruption (e.g., IS element insertion)*

https://www.reactgroup.org/toolbox/understand/antibiotic-resistance/resistance-mechanisms-in-bacteria/

Features of Different Tools: Reads vs. assemblies

- Assemblies
  - Assemblers (and annotation tools) can affect results
  - Draft assemblies can ‘squash’ close variants
- Reads
  - ‘Mediocre’ data can be a problem, especially with allelic variants
  - Need to understand how reads are processed, mapped to references
  - Lack of positional information (*where* is the gene?)

Features of Different Tools: Nucleotide databases vs. amino acid databases

- Amino acid describes function
- Nucleotide-based analyses can be faster, but sometimes inaccurate at fine scale
- Many are hybrid (e.g., point mutations of 23S and protein detection)

How Are Genes Detected: BLAST, kmers, and HMMs

- BLAST (and similar methods)
  - Straightforward to implement
  - Easy to understand how it works
  - Nucleotide-based methods
- K-mers
  - Speed–can search large read sets such as microbiome data
  - Usually mechanism-agnostic (for good and bad)
  - Often tied into phenotype prediction
- Hidden Markov Models (HMMs)
  - Alignments of known proteins are used to build HMMs that identify conserved domains of structure and function
  - Typically use protein sequence for speed/computational reasons
  - Based on biological structure, not arbitrary identity thresholds
- Manually curated cutoffs/rules versus One Rule to Bind Them All

Features of Different Tools: What is reported

- What is reported: closest hit vs. best estimate identification
  - E.g., 99% identical to KPC-2 is *not* KPC-2
  - KPC-2: carbapenemase
  - KPC-33: inhibitor-resistant cephalosporinase (1 nt change from KPC-2)
  - KPC-8: inhibitor-resistant carbapenemase (2 nt changes from KPC-2)
  - Multiple ‘unknown’ KPC proteins: *unknown phenotype*
- Point mutation detection
- ‘Broken gene’ detection (frameshifts, partials, stop codons)
  - Important for porin-based mechanisms
- Descriptions of genes
- Online tools (GUIs)

Things to Look for in a Database

- Is it regularly curated/updated?
- What are the inclusion criteria for genes (and point mutations)?
  - Are only full-length genes included?
    - important for identifying best hit
  - Are start sites are curated?
    - *attC* sites are removed
    - leader peptides verified
- How are gene symbols reported? (hARMonization)
- Are there links to the literature?
- Are possible phenotypes reported?
- **Unfortunately, it's hard to know these things!!**

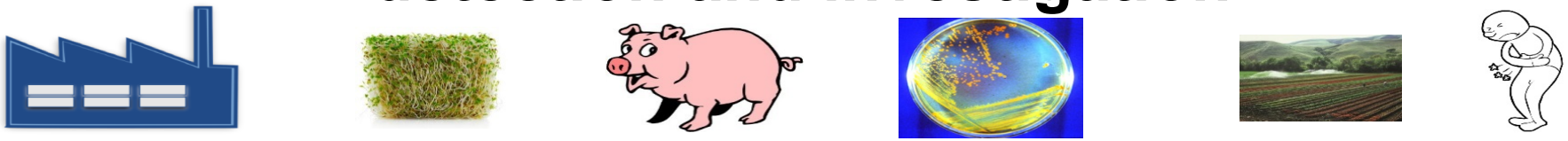
The Big Caveat

- For some organisms, there is a high correlation between genotype-phenotype
  - *Campylobacter, Salmonella,* and *E. coli*, Feldgarden et al., 2019, AAC)
  - 98.4% consistency (more recent analysis suggests >99.7%)
- For others…not so much:
  - Khaledi *et al*. 2020, EMBO
  - Used machine learning and gene expression, still only ~0.9 for some drugs in *P. aeruginosa*
- **Gene expression matters (in some organisms, for some drugs, sometimes) and current tools do not address this***

Common Tools

- ResFinder 4 (CGE)
  - Can use assemblies or reads
  - Nucleotide vs. nucleotide BLAST-based
  - A single identity and a single length threshold
  - Fast
  - Can misassign alleles as closest amino acid hit is not necessarily the closest nucleotide hit
  - Online GUI
- RGI (CARD)
  - Protein database
  - Option for broadening scope to identify novel mechanisms; emphasis on efflux
  - Will accept nucleotide sequence or protein sequence
  - BLAST-based but manual cutoffs
  - Online GUI and ontology
- AMRFinderPlus (NCBI)
  - Protein database
  - Will accept nucleotide sequence or protein sequence
  - Uses BLAST and HMMs to identify AMR genes
  - Manually curated BLAST and HMM cutoffs
  - Explicit partial and internal stop identification
  - No online GUI (but data for >780,000 isolates are available in MicroBIGG-E)

Real time surveillance of pathogens for outbreak detection and investigation



Sampling of clinical/human, food, animal, and environmental bacterial isolates

USDA
PHE
hospitals

GenomeTrakr

State agriculture & food regulatory labs

FSIS & FDA labs food, environmental

PulseNet

State health labs (clinical)

Raw Genomic Sequence data

NCBI Submission Portal

Minimal metadata, everything publically accessible

PulseNet National Database

CDC analysis (all raw sequence goes into NCBI

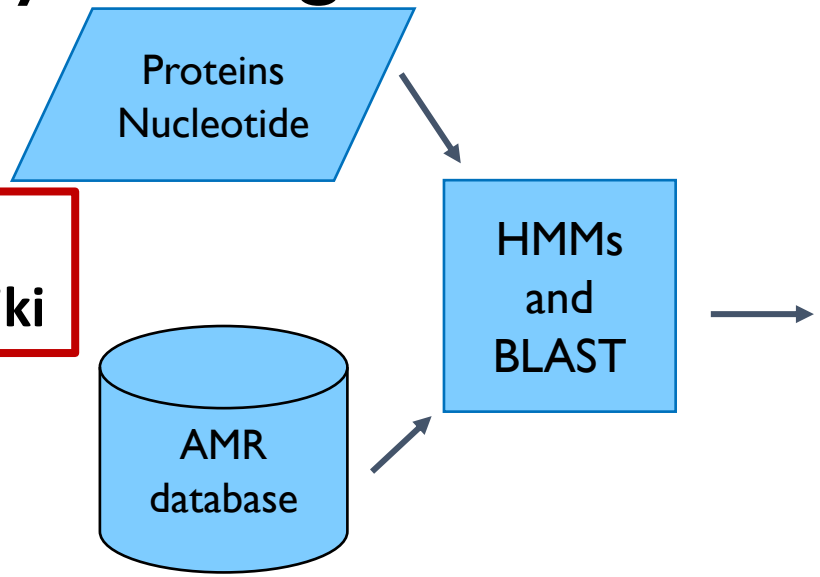
More extensive metadata, analyzed seq data (tools to translate) data shared among PulseNet labs only

Large Scale Requires Concise Information

- hundreds of genomes per day
- can't be 'artisanal'; flipping through multiple columns/rows/tables will not work
- Need *concise, discrete signifier* that conveys appropriate information about genotype (and possibly phenotype)
- That signifier is the *gene symbol*
  - E.g., 99% identical to KPC-2 is *not* KPC-2
  - KPC-2: carbapenemase
  - KPC-33: inhibitor-resistant cephalosporinase (1 nt change from KPC-2)
  - KPC-8: inhibitor-resistant carbapenemase (2 nt changes from KPC-2)
  - Multiple 'unknown' KPC proteins: *unknown phenotype*

AMRFinderPlus Uses a Curated Database, HMMs and BLAST to Identify AMR genes

Available at:
https://github.com/ncbi/amr/wiki



Report on
resistance genes
- integrated into Pathogen Detection Isolate Browser for >952,000 pathogen isolates

"Plus" contains:
716 virulence factors
233 acid, biocide, metal, and heat resistant genes
Optional for users

5,965 resistance proteins
650 HMMs
44 drug classes resisted
~60% beta-lactamases

AMRFinderPlus now finds point mutations!
914 resistance mutations for fifteen taxa including *Campylobacter*, *E. coli*, and *Salmonella*

Building an AMR Database

Domain experts
- Bush and Jacoby (beta-lactamases)
- Marilyn Roberts (MLS/tetracycline)
- Pasteur Institute (beta-lactamases)

Large scale databases
- FDA Center for Veterinary Medicine
- ResFinder
- The C.A.R.D. (~monthly exchanges)

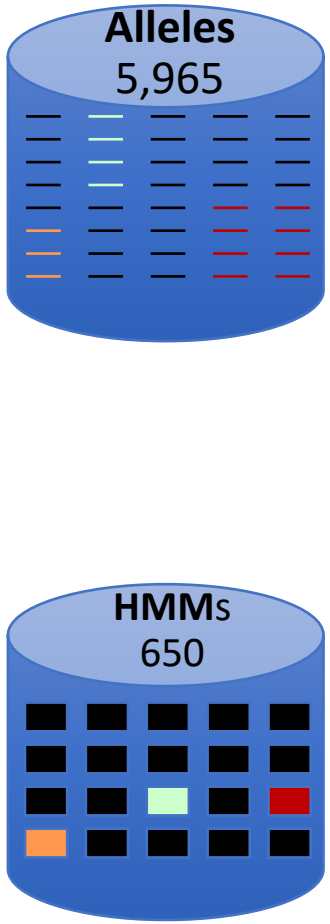
Manual extraction from literature

Ongoing curation of beta-lactamases, Qnr, and MCR

ResFams, TIGRFams, NCBIFams
- *Select*
- *Set cutoffs*

New HMMs
- *Group sequences*
- *Align*
- *Build HMM*
- *Set cutoffs*

allele = unique protein (*blaTEM-1*)
gene = set of related proteins (*sul1*)



Validation

AMRFinderPlus Has a Hierarchical Structure

Similarity to known allele	Protein name	Functional determination
100 % Assign by **BLAST**	KPC-2	*Resistance to carbapenems and other beta-lactam antibiotics*. Epidemiological marker.
98 % Assign by **HMM**	KPC family	HMM score > cutoff of KPC. *Likely resistance to carbapenems and other beta-lactam antibiotics.*
75% Assign by **HMM**	class A beta-lactamase	HMM score > cutoff. *Class A beta-lactamase of unknown specificity*.
23 %	(irrelevant)	**HMM scores < cutoff** prevents false-positive identification as a beta-lactamase. *Not reported*.

Large Scale Requires Concise Information

- hundreds of genomes per day
- can't be 'artisanal'; flipping through multiple columns/rows/tables will not work
- Need *concise, discrete signifier* that conveys appropriate information about genotype (and possibly phenotype)
- That signifier is the *gene symbol*

The Utility of HMMs: 'Beta-lactamases' in GenBank

- Examined **GenBank** protein sequences that had 'beta-lactamase' in product name and not described as partial or synthetic constructs:
  - Only **11%** of sequences (108,386/1,030,160) appear to be beta-lactamases
  - Only **20%** of unique proteins (27,682/137,297) appear to be beta-lactamases

- Examined 21 putative metallo-β-lactamases from metagenomic data that had been functionally characterized:
  - AMRFinder correctly identified the 18 functional metallo-β-lactamases
  - AMRFinder correctly did not call the 3 non-functional proteins as beta-lactamases

Berglund *et al.* 2017. Identification of 76 novel B1 metallo-β-lactamases through large-scale screening of genomic and metagenomic data. Microbiome 5:134

- Nayfach et al. 2021: used RGI, ResFinder, and AMRFinderPlus to confirm viral beta-lactamases (only ~0.5% of putative beta-lactamases appear to be beta-lactamases)

Using AMRFinderPlus

- Optimal use is with nucleotide sequence, protein sequence, and a .gff file
- The AMRFinderPlus database (Reference Gene Catalog) curation is linked to NCBI's curation of PGAP
  - Proteins will be called the correct length
- Can detect species-specific point mutations and genes
- Optionally, can detect virulence genes and stress response genes
- Easy to install using Bioconda (*good for bioinformatics in general*)

Using AMRFinderPlus: some command line options

```
amrfinder (-p <protein_fasta> | -n <nucleotide_fasta) [options]
```

Example:

```
amrfinder --nucleotide /home/feldgard/test.nuc.fa --output
/home/feldgard/test.nuc.tsv
```

More complex example:

```
amrfinder --nucleotide /home/feldgard/test.nuc.fa \
          --protein /home/feldgard/test.protein.fa \
          --gff /home/feldgard/test.gff \
          --output /home/feldgard/test.nuc.tsv \
          --organism Escherichia \
          --plus
```

genome sequence

set of annotated proteins

describes gene location

output file

organism flag (optional)

scope (optional virulence and stress resistance gene detection)

Two examples:

- The good: *S. enterica* SAMN05201855

```
amrfinder --protein GCA_006697045.2_ASM669704v2_protein.faa\
--nucleotide GCA_006697045.2_ASM669704v2_genomic.fna \
--gff GCA_006697045.2_ASM669704v2_genomic.gff \
--output GCA_006697045.2.tsv \
--organism Salmonella \
--plus
```

https://www.ncbi.nlm.nih.gov/biosample/SAMN05201855

The bad: *P. aeruginosa* SAMN17616831

```
amrfinder --protein GCA_016905405.1_ASM1690540v1_protein.faa \
--nucleotide GCA_016905405.1_ASM1690540v1_genomic.fna \
--gff GCA_016905405.1_ASM1690540v1_genomic.gff \
--output GCA_016905405.1.tsv \
--organism Pseudomonas_aeruginosa \
--plus
```

https://www.ncbi.nlm.nih.gov/pathogens/isolates/#SAMN17616831

S. enterica SAMN05201855

Resistance phenotype	AMR genes
ampicillin	*blaTEM-1*
gentamicin	*aac(3)-IId*
tetracycline	*tet(A), tet(B)*

No resistance genes found that confer resistance to 11 susceptible phenotypes.
(also 1 streptomycin resistance gene, though streptomycin was not tested)

P. aeruginosa SAMN17616831

Resistance phenotype	AMR genes
amikacin	**????**
aztreonam	*blaGES-2*
cefepime	*blaGES-2*
ceftolozane-tazobactam	**???**
ciprofloxacin	gyrA_T83I, parC_S87L
gentamicin	*aac(3)-I, aac(6')-Ib4*
imipenem-relebactam	**????**
imipenem	*blaGES-2*
levofloxacin	gyrA_T83I, parC_S87L
meropenem-vaborbactam	**????**
meropenem	*blaGES-2*
piperacillin-tazobactam	*blaGES-2*
tobramycin	**????**

- Multiple missing mechanisms
- Could be efflux
- AMRFinderPlus screens for these resistance mechanisms, but could be novel mechanisms

Conclusions

- Prediction can be very accurate for some organisms
  - E.g., most Enterobacterales (Feldgarden et al., 2019)
- Some bug-drug combinations are challenging
  - New phenotypes often are inadequately understood
  - Porins (the broken gene problem)
- *Pseudomonas* and *Acinetobacter* are hard
  - Khaledi *et al*. 2020, EMBO
  - Used machine learning and gene expression, still only ~0.9 for some drugs in *P. aeruginosa*
- Use the appropriate tool for your needs
  - Methods matter
  - Database quality matters
  - What output do you need?

NCBI Resources

AMRFinderPlus:
https://github.com/ncbi/amr/wiki

Reference HMM Catalog:
https://www.ncbi.nlm.nih.gov/pathogens/hmm/

Reference Gene Catalog
https://www.ncbi.nlm.nih.gov/pathogens/isolates/refgene/

Isolate Browser:
https://www.ncbi.nlm.nih.gov/pathogens/isolates

MicroBIGG-E
https://www.ncbi.nlm.nih.gov/pathogens/microbigge/

Reference Gene Hierarchy
https://www.ncbi.nlm.nih.gov/pathogens/genehierarchy/

Questions: pd-help@ncbi.nlm.nih.gov

Acknowledgements

Richa Agarwala
Victor Ananiev
Azat Badretdin
Slava Brover
Joshua Cherry
Jinna Choi
Vyacheslav Chetvernin
Robert Cohen
Michael DiCuccio
Boris Fedorov
Michael Feldgarden
Lewis Geer
Renata Geer
Dan Haft
Lianyi Han
Avi Kimchi
Michel Kimelman
William Klimke
Alex Kotliarov
Valerii Lashmanov
Aleksandr Morgulis
Eyal Moses
Chris O'Sullivan
Arjun Prasad

Edward Rice
Kirill Rotmistrovskyy
Alejandro A. Schaffer
Nadya Serova
Stephen Sherry
Sergey Shiryev
Martin Shumway
Oleg Shutov
Douglas Slotta
Alexandre Souvorov
Tatiana Tatusova
Francoise Thibaud-Nissen
Igor Tolstoy
Lukas Wagner
Hlavina Wratko
Chunlin Xiao
Alexander Zasypkin
Eugene Yaschenko
Mingzhang Yang

David Lipman
James Ostell
Kim Pruitt

CDC
FDA/CFSAN
GenFS
USDA-FSIS
PHE/FERA
NARMS
NIHGRI
NIAID
WRAIR
Broad
Wadsworth/MDH
Vendors: PacBio, Illumina, Roche

pd-help@ncbi.nlm.nih.gov

This research was supported by the Intramural Research Program of the NIH, National Library of Medicine. http://www.ncbi.nlm.nih.gov
National Center for Biotechnology Information – National Library of Medicine – Bethesda MD 20892 USA

Features of Different Tools

- Reads vs. assemblies
  - Assemblies
    - Assemblers (and annotation tools) can affect results
    - Assemblies can ‘squash’ close variants
  - Reads
    - ‘Mediocre’ data can be a problem, especially with ‘allelic’ variants
    - Need to understand how reads are processed, mapped to references
    - Lack of positional information (*where* is the gene?)
- Nucleotide databases vs. amino acid databases
  - Amino acid describes function
  - Nucleotide-based analyses can be faster, but sometimes inaccurate at fine scale
- What is reported: closest hit vs. best estimate identification
  - E.g., 99% identical to KPC-2 is *not* KPC-2
- Point mutation detection
- ‘Broken gene’ detection (frameshifts, partials, stop codons)

Features of Different Tools

- BLAST, kmers, and HMMs
  - BLAST (and similar methods)
    - Straightforward to implement
    - Easy to understand how it works
    - Nucleotide-based methods
  - K-mers
    - Speed
    - Usually mechanism-agnostic (for good and bad)
    - Often tied into phenotype prediction
  - Hidden Markov Models (HMMs)
    - Alignments of known proteins are used to build HMMs that identify conserved domains of structure and function
    - Typically use protein sequence for speed/computational reasons
    - Based on biological structure, not arbitrary identity thresholds
- Manually curated cutoffs/rules versus One Rule to Bind Them All
- Descriptions of genes
- Online tools (GUIs)