

Annotation of Large Data Sets of Whole Genomes

Daniel Haft

ASM MICROBE, June 22, 2019



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

National Library of Medicine:

repository for the scientific literature.

NCBI produces **GenBank** (as part of INSDC):

repository of record for biological sequence data.

ARCHIVAL !!

PGAP pipeline, RefSeq database:

attempts to connect the literature to sequence data.

CURRENT !!



Goals for PGAP and RefSeq

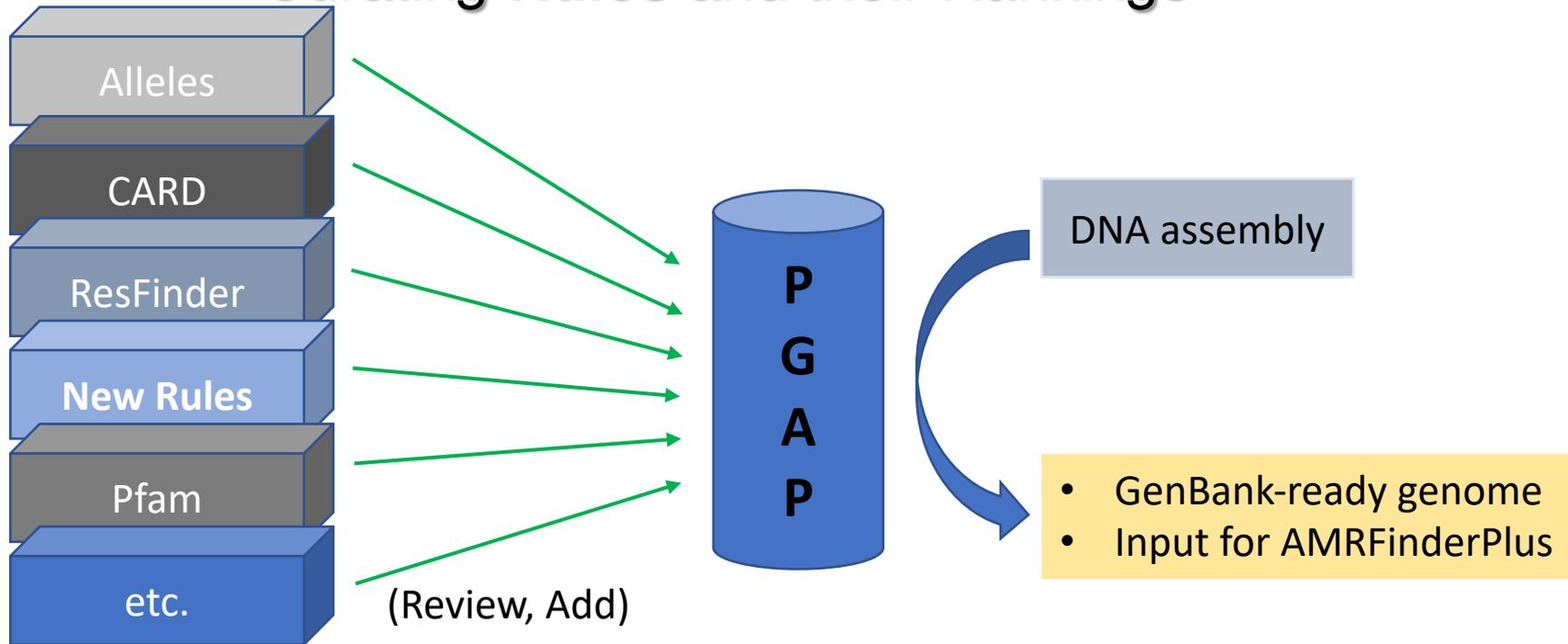
- Be **non-redundant**, **comprehensive**, **consistent**, **well-annotated**
- Find and name every protein
 - **Structure** – Consistency supports meaningful comparative genomics
 - **Function** – Names come from hierarchical evidence; give similar names to similar proteins
- Link **NCBI** objects (sequences) to **NLM** objects (articles)

RefSeq's Strategy for Functional Annotation: **RULES**

- Biocurators, **do not directly annotate any protein!**
- **Instead**, fill a database with **INSTRUCTIONS for PIPELINES**
- Include imported rules; exchange info with partners.
- Evidence is **hierarchical**: the most specific rule wins.
- Make rules **trustworthy, traceable, public, and portable.**



Prokaryotic Genome Annotation Pipeline: Curating **Rules** and their Rankings



What's Under PGAP's Hood?

- 3,167 named AMR **Alleles**
- 22,654 **CDD Domain Architectures**
- 8,148 **BlastRules** (6 levels)
- 16,655 named **HMMs** (8 levels)

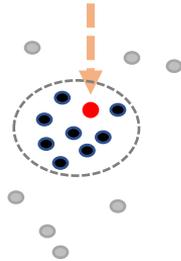
Build a Perfect Annotation Rule and it will define a **Molecular Marker**

- Aim is **0** false-positives, **0** false-negatives.
- Protein profile **H**idden **M**arkov **M**odels (**HMMs**) can do this well.
- For Anti-Microbial Resistance (**AMR**), most of our annotation meet this standard: *if the marker is not detected, it almost certainly is not there.*

A published article ...

NDM-1 is a carbapenemase that confers **resistance** to the antibiotic **meropenem**

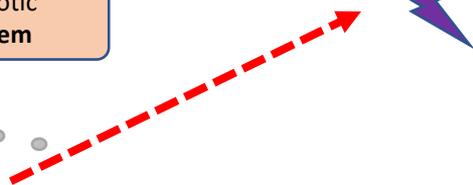
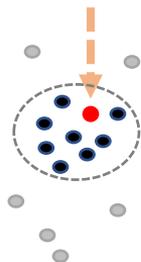
... reports the first of a new **family** of resistance genes ...



A published article ...

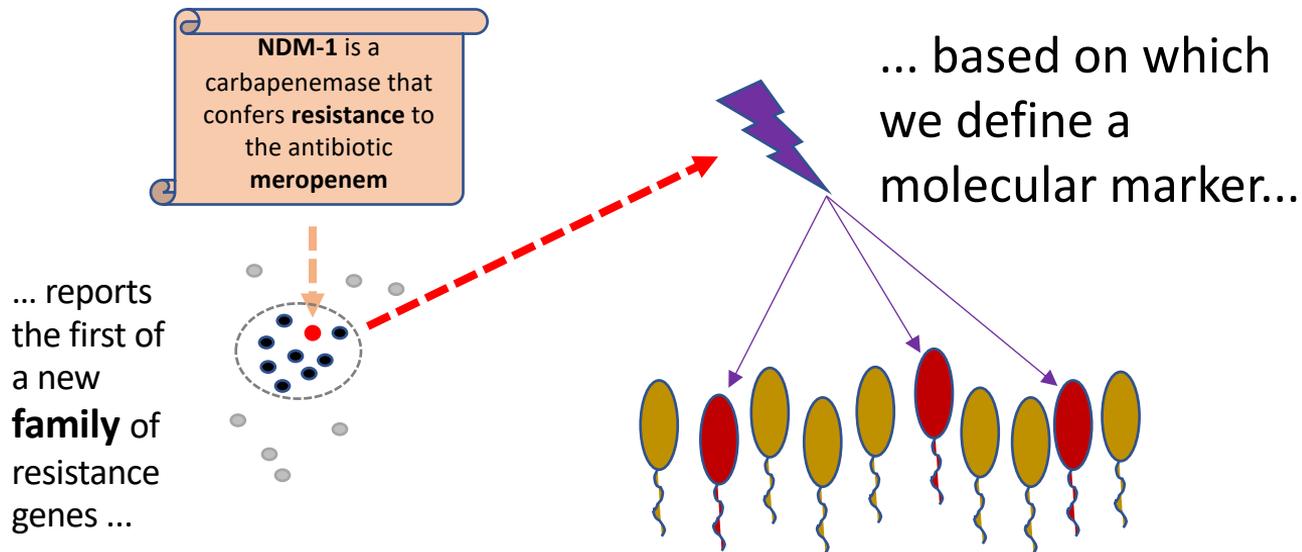
NDM-1 is a carbapenemase that confers **resistance** to the antibiotic **meropenem**

... reports the first of a new **family** of resistance genes ...



... based on which we define a molecular marker...

A published article ...



... which then allows users to know which isolates have **NDM**.

WP_XXXXXXXX proteins are RefSeq proteins

- Based on GenBank / EBI / DDBJ data
- **NCBI owns** the record and is **free to improve it !**

RefSeq database Prefix	Molecule type
NM_, XM_	mRNAs
NR_, XR_	other RNAs
WP_, NP_, XP_, YP_, AP_	Proteins
NC_, NG_, NT_, NW_, AC_, NZ_	DNA

NDM family subclass B1 metallo-beta-lactamase [Acinetobacter lwoffii]

NCBI Reference Sequence: WP_129717965.1

[Identical Proteins](#) [FASTA](#) [Graphics](#)

Go to:

LOCUS WP_129717965 270 aa linear BCT 31-MAY-2019
DEFINITION NDM family subclass B1 metallo-beta-lactamase [Acinetobacter lwoffii].
ACCESSION WP_129717965
VERSION WP_129717965.1
KEYWORDS RefSeq.
SOURCE Acinetobacter lwoffii
ORGANISM [Acinetobacter lwoffii](#)
Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Moraxellaceae; Acinetobacter.
COMMENT REFSEQ: This record represents a single, non-redundant, protein sequence which may be annotated on many different RefSeq genomes from the same, or different, species.

~~##Evidence-For-Name-Assignment-START##~~

Evidence Category :: HMM
Evidence Accession :: NF000259.2
Evidence Source :: NCBI FAM

~~##Evidence-For-Name-Assignment-END##~~

New!

Details

NCBI accession	NF000259.2
Source identifier	NCBIFAM NF000259.2
Product name ?	NDM family subclass B1 metallo-beta-lactamase
Label ?	blaNDM
Gene symbol	blaNDM
Family type ?	exception
HMM length ?	270 aa
Sequence cutoff ?	580
Domain cutoff ?	580
Number of RefSeq protein hits ?	19

Protein hits

HMM NF000259.2 hits 19 RefSeq proteins above the sequence cutoff (580) and domain cutoff (580). It is used to name 9 of these proteins. The other 10 proteins derive their names from higher precedence annotation evidence.

 Filters [v](#)
[Download](#)

Accession ▲ ▼	Protein name	Named by	Organism	Sequence score ▲ ▼	Domain score ▲ ▼	Length (aa) ▲ ▼	RefSeq assemblies ▲ ▼	Coverage
WP_004201164.1	subclass B1 metallo-beta-lactamase NDM-1	NG_049326.1	Bacteria	643.5	643.3	270	752	
WP_023408309.1	subclass B1 metallo-beta-lactamase NDM-5	NG_049337.1	Enterobacterales	641.1	640.9	270	314	
WP_032495672.1	subclass B1 metallo-beta-lactamase NDM-9	NG_049341.1	Enterobacteriaceae	642.6	642.4	270	67	
WP_032495622.1	subclass B1 metallo-beta-lactamase NDM-7	NG_049339.1	Enterobacteriaceae	641.9	641.8	270	42	
WP_063860861.1	subclass B1 metallo-beta-lactamase NDM-4	NG_049336.1	Enterobacteriaceae	642.9	642.7	270	19	

There is *(in principle)* no Competition among Functional Annotation Pipelines

- Rules of different types interleave by having different precedences.
- To import a new collection, assign an appropriate new precedence.
- Behave like a part of a rule-building consortium.

Rules are meant to circulate. NCBI imports rules. NCBI shares its rules by FTP. The entire pipeline is now available as PGAP-X.

Families within Families

beta-lactamase (conceptual)

class A beta-lactamase (HMM:NF033103)

metallo-beta-lactamase (HMM:NF012229)

subclass B1 metallo-beta-lactamase (HMM:NF033088)

NDM family subclass B1 metallo-beta-lactamase (HMM:NF000259)

subclass B1 metallo-beta-lactamase **NDM-1** (allele)

subclass B1 metallo-beta-lactamase **NDM-2** (allele)

subclass B1 metallo-beta-lactamase **NDM-3** (allele)

VIM family subclass B1 metallo-beta-lactamase (HMM:NF012100)

SPM family subclass B1 metallo-beta-lactamase (HMM:NF012150)

subclass B2 metallo-beta-lactamase (HMM:NF033087)

subclass B3 metallo-beta-lactamase (HMM:NF033105)

class C beta-lactamase (HMM:NF033085)

class D beta-lactamase (conceptual)

class D beta-lactamase (main branch) (HMM:NF012161)

class D beta-lactamase (other branch) (HMM:NF000270)



Ranking Evidence using Assigned Precedence

Precedence Score	Evidence Type	Rule Accession	Annotation
100	Exact Allele	blaNDM-1	subclass B1 metallo-beta-lactamase NDM-1
96	BlastRuleIS	NBR001104	IS6-like element IS26 family transposase
95	BlastRuleException	NBR006109	3-deoxy-7-phosphoheptulonate synthase AroF
80	HMM (equivalog)	TIGR00034	3-deoxy-7-phosphoheptulonate synthase
60	CDD architecture	10870091	N-acyl homoserine lactonase family protein
55	HMM (subfamily)	NF000282.2	multidrug efflux RND transporter permease subunit
30	HMM (domain)	PF00801.18	PKD domain-containing protein
0	legacy naming system	WP_018364973	bacteriocin BlpK
0	-	-	hypothetical protein

RefSeq is prioritizing your Favorite Bacterial Biomarkers

- Anti-Microbial Resistance (AMR)
- Virulence
- Named transposases
- Serological markers
- Enzymes (for pathway reconstruction)
- ESKAPE proteome
- *(and by request)*

Help Maintain Sense and Clarity in Protein Naming

- For AMR proteins where NCBI assigns alleles, submit sequences to NCBI. Authors and GenBank contributors, please **DO NOT GUESS!**
- For other curated AMR families, contact the traditional expert.
- For remaining AMR, do not introduce allele numbering without a very good reason.
- Engineer protein product names with care. Function, Process, Family
- **Always provide a protein accession in your publications.** WP_XXXXXXXX series is ideal – same translated protein no matter what nucleotide source.

Matching Article Content to Protein Sequences

Descriptions inferred from article ...		
nicotine dehydrogenase subunit L		
nicotine dehydrogenase subunit S		
(S)-6-hydroxynicotine oxidase		
6-hydroxypseudooxynicotine oxidase		
6-hydroxy-3-succinoylpyridine 3-monooxygenase		
2,5-dihydropyridine 5,6-dioxygenase		
maleamate amidohydrolase		

Matching Article Content to Protein Sequences

Descriptions inferred from article were matched, with difficulty, to protein accessions	... that the author submitted to GenBank . The annotation as submitted is shown.
nicotine dehydrogenase subunit L	KZB95048.1	hypothetical protein AVM11_17640
nicotine dehydrogenase subunit S	KZB95064.1	(2Fe-2S)-binding protein
(S)-6-hydroxynicotine oxidase	KZB95041.1	hypothetical protein AVM11_17600
6-hydroxypseudooxynicotine oxidase	KZB95055.1	hypothetical protein AVM11_17690
6-hydroxy-3-succinoylpyridine 3-monooxygenase	KZB95066.1	para-nitrophenol 4-monooxygenase
2,5-dihydropyridine 5,6-dioxygenase	KZB95052.1	leucyl aminopeptidase
maleamate amidohydrolase	KZB95053.1	carbamoylsarcosine amidase

Your Protein Characterization on 1,000,000 Genomes

- Reviewers #1, #2, and #3 are **not** the most important readers of your paper
- Shout-out to **PaperBLAST** – see the importance of citing protein accessions
- Article → Annotation Rule → Named protein with provenance → Citations
- Contact NCBI about your collection of perfectly named model proteins

write to info@ncbi.nlm.nih.gov w/ cc to daniel.haft@nih.gov

Acknowledgements

This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.

RefSeq Software Developers:

- Mike DiCuccio
- Slava Chetvernin
- Azat Badretdin

RefSeq Biocurators:

- Kathleen O'Neill
- **Daniel.Haft (at nih.gov)**
- Wenjun Li
- CDD group

Pathogen Detection Team:

- Michael Feldgarden
- Arjun Prasad
- Slava Brover
- Martin Shumway
- Bill Klimke

RefSeq and CDD Leadership:

- Kim Pruitt
- Francoise Thibaud-Nissen
- Aron Marchler-Bauer

The NCBI team would like to learn more about workflows and **user needs for large scale analyses of microbial genomes** to support our users better.

We would be especially interested in anyone interested in doing **large-scale data analyses using cloud-based resources**.

If you are interested in participating and would be willing to **share your email address, contact our personnel at booth #443**.

Links

<https://submit.ncbi.nlm.nih.gov/subs/genome/>

Have NCBI annotate your genome for submission to GenBank.

<ftp.ncbi.nlm.nih.gov/hmm/>

HMMs from NCBI

<ftp.ncbi.nlm.nih.gov/pub/blastrules/>

BlastRules from NCBI

https://www.ncbi.nlm.nih.gov/genome/annotation_prok/evidence/TIGR03969/

An NCBI HMM web page (navigate there from **WP_136245628.1**, *e.g.*)

BlastRules: A New Opportunity for Collaboration

a.k.a.

Please Send Us

Biomarker Names and Accessions!

#proteins	type	name	gene	pubmed
CAQ30329.1	BlastRuleException	virulence-associated protein	VapJ vapJ	18606735
CAQ30332.1	BlastRuleException	virulence-associated protein	VapK vapK	18606735
CAQ30337.1	BlastRuleException	virulence-associated protein	VapM vapM	18606735
CAQ30339.1	BlastRuleException	virulence-associated protein	VapB vapB	18606735
CAQ30394.1	BlastRuleException	virulence-associated protein	VapL vapL	18606735
CAQ30399.1	BlastRuleException	virulence-associated protein	VapH vapH	18606735
CAQ30407.1	BlastRuleException	virulence-associated protein	VapA vapA	18606735
CAQ30409.1	BlastRuleException	virulence-associated protein	VapC vapC	18606735
CAQ30410.1	BlastRuleException	virulence-associated protein	VapD vapD	18606735
CAQ30416.1	BlastRuleException	virulence-associated protein	VapE vapE	18606735

PMID:18606735: an article with 28 citations in PMC, about virulence plasmids from *Rhodococcus equi*, a soil bacterium and a mostly opportunistic animal pathogen.