

hAMRonization

Improving the Utility of Antimicrobial Resistance Gene Prediction Tools Using the PHA4GE AMR Gene Detection Specification

Catarina Inês Mendes, Finlay Maguire, Dan Fornika, Alexander Manuele, Elizabeth Culp, Simon H Tausch, Amogelang R Raphenya, Brian Alcock, Michael Feldgarden, Gregory H Tyson, Josefina Campos, Adam A Witney, David M Aanensen, Allison Black, Emma Hodcroft, Lee S Katz, Paul E Oluniyi, Idowu B Olawoye, Ruth E Timme, Ana Tereza R Vasconcelos, Duncan R MacCannell, Emma J Griffiths, Andrew J Page, on behalf of the **Public Health Alliance for Genomic Epidemiology (PHA4GE) consortium**



AMERICAN
SOCIETY FOR
MICROBIOLOGY



Public Health Alliance for
Genomic Epidemiology

PHA4GE Data Structures Workgroup

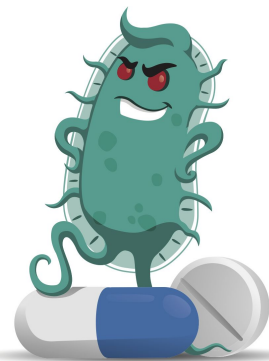


Develop, share, promote **data standards and standardized data** models to improve data harmonization, integration, transformation, sharing, and control/security.

The Antimicrobial Resistance Data Problem

Antimicrobial resistance (**AMR**) represents a growing public health crisis of global scope and relies on the **detection of AMR genes from genomic data**.

A large number of **different bioinformatic tools** have been developed to perform this task but **differ in supported inputs**, search algorithm, parameterisation, underlying reference databases and **output formats**.



```
$ head abricate/report.tsv
#FILE SEQUENCE START END STRAND GENE COVERAGE COVERAGE_MAP GAPS %COVERAGE %IDENTITY DATABASE ACCESSION PRODUCT RESISTANCE
GCF_010120755.1_ASM1012075v1_genomic.fna NZ_CP039729.1 656118 657620 + eat(A) 1-1503/1503 ===== 0/0 100.00 100.00 ncbi NG_047762.1 ABC-F type ribosomal protection protein Eat(A) PLEUROMUTILIN
GCF_010120755.1_ASM1012075v1_genomic.fna NZ_CP039729.1 2163439 2163987 - aacA-ENT1 1-549/549 ===== 0/0 100.00 100.00 ncbi NG_052371.1 aminoglycoside 6'-N-acetyltransferase AMINOGLYCOSIDE
GCF_010120755.1_ASM1012075v1_genomic.fna NZ_CP039729.1 2564038 2565516 - msr(C) 1-1479/1479 ===== 0/0 100.00 98.92 ncbi NG_048003.1 ABC-F type ribosomal protection protein Msr(C) MACROLIDE
GCF_010120755.
GCF_010120755.
GCF_010120755.
GCF_010120755.
```

```
$ head amrfind
Protein identi
Identity to r
NA NZ_CP0
GyrA NA
NA NZ_CP0
ion protein Ea
NA NZ_CP0
subunit A Par
NA NZ_CP0
ansferase
NA NZ_CP0
ion protein Ms
NA NZ_CP0
A
NA NZ_CP0
A NA
NA NZ_CP0
NA NZ_CP0
resistance hi
```

```
$ head ariba/r
#ariba_ref_nam
ge_ref_ct
AAC_3_IId.300
AAC_6_Ib4.30
87R NONSYN
AAC_6_Ib4.30
102L NONSYN
AAC_6_Ib4.30
490T
BRP_MBL_30012
n-like antibio
CRP_3000518.AP
YN 85
CTX_M_15.30018
DHA_7_3002136.
Escherichia_co
611.4 0
```

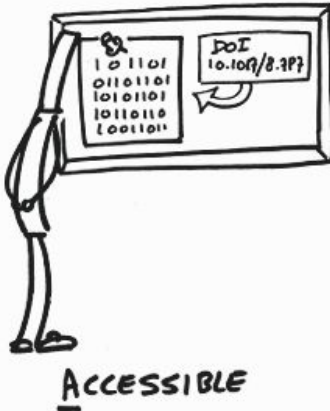
```
$ head srst2/S
Sample DB
SRR10313716
SRR10313716
SRR10313716
SRR10313716
SRR10313716
SRR10313716
SRR10313716
SRR10313716
SRR10313716
SRR10313716
SRR10313716
SRR10313716
SRR10313716
SRR10313716
SRR10313716
ResFinder bltA bltA17_29.370 33.96 495npzindel 0.432 474 0.107 307 502 bltA17_1_FJ490230; FJ490230; blaTemtopr III
SRR10313716 ResFinder sul1 sul1_11 92.857 62.65 63holes edge0.0 0.0 882 0.034 227 1889 sul1_11_DQ914960; DQ914960; sulphonamides
SRR10313716 ResFinder sul1 sul1_2 100.0 68.151 trun 0.0 927 0.029 183 1044 sul1_2_CP002151; CP002151; sulphonamides
SRR10313716 ResFinder blaTEM blaTEM-1_1 100.0 117.002 0.0 861 0.032 258 1394 blaTEM-1_1_JF910132; JF910132; betalactamase
SRR10313716 ResFinder Qnr-S QnrS2 97.869 79.482 58snp2indel13holes 9.006 657 0.1 476 1997 Qnr-S2_1_JF261185; JF261185; quinolone
```

VANCOMYCIN

sequence %
ase subunit A
somal protect
oisomerase IV
6'-N-acetyltr
somal protect
VanX-M NA N
ligase VanM N
NA NA
pe vancomycin

ref_ctg_chan
0 W
0 S
G
and bleomyci
K29T NONS
g.1 2822

FAIR DATA PRINCIPLES



The FAIR Guiding Principles for scientific data management and stewardship
Wilkinson et al, Scientific Data, 2016

Standardized AMR Gene Detection Output

AMR Prediction Tools

Lots of software with different output formats.

hAMRonization Workflow

Run any dataset through the set of tools.

hAMRonization Package

Take tool-specific output and reformat to standardized output.

Standardized Output Specification

Enables consistent comparison of AMR genes detected by different tools.

hAMRonization package automates conversion
Tools can be built on standardized output



The hAMRonization Workflow



https://github.com/pha4ge/hAMRonization_workflow

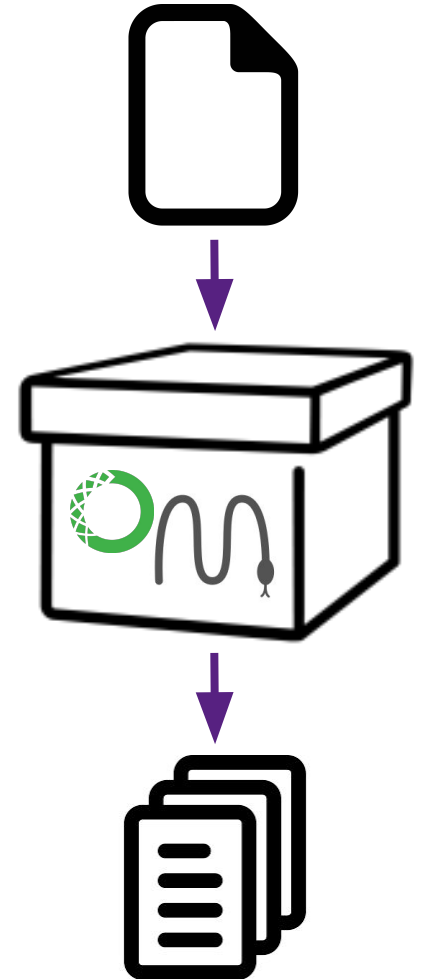
Snakemake workflow to run **14 different species-agnostic AMR gene detection tools** selected tools on a dataset, installing **fixed versions** of the tools from **conda** on execution.

Ensures:

- Reproducibility
- Scalability

Tool inclusion criteria:

- Specificity for AMR gene detection
- Open-source



The hAMRonization Package



<https://github.com/pha4ge/hAMRonization>

Biopython-compatible parser and **command-line utility** automatically transformation of reports from 14 different species-agnostic AMR gene detection tools into “**hAMRonization**”-compatible reports.

Validation and programmatic use of the **specification** was facilitated via the development of **JSON** and **SALAD** schemata. **Validation** of the parsing utilities ensured by unit testing.



The hAMRonization Package



<https://github.com/pha4ge/hAMRonization>

An example from abricate:

```
> hamronize abricate <report> --reference_database_version <version> \  
--analysis_software_version <version> --format json
```

Applies hAMRonization specification to output from abricate (OUTPUT.tsv)

Combining all the reports:

```
> hamronize summarize --summary_type interactive <list of reports>
```

Summary output formats:

- TSV
- Json
- Interactive HTML



AMERICAN
SOCIETY FOR
MICROBIOLOGY



The hAMRonization Package



<https://github.com/pha4ge/hAMRonization>

Public Health Alliance for Genomic Epidemiology

Search Show Only Genomes With Hits Restore Results

	abricate: config 0	amrfinderplus: config 0	csstar: config 0	resfinder.py: config 0	staramr: config 0
ERR873305	11 hits	10 hits	10 hits	6 hits	9 hits
ERR873306	11 hits	10 hits	10 hits	6 hits	9 hits
ERR873307	11 hits	10 hits	10 hits	6 hits	9 hits
ERR873308	11 hits	10 hits	10 hits	6 hits	9 hits
ERR873309	16 hits	14 hits	14 hits	6 hits	11 hits
ERR873310	11 hits	10 hits	10 hits	6 hits	9 hits
ERR873311	11 hits	10 hits	10 hits	6 hits	9 hits
ERR873312	11 hits	10 hits	10 hits	6 hits	9 hits

Search Results

Total hits: 0

Genomes with hits: 0

Tools with hits: 0

Differential results: 0

Selected

Compare Clear

Public Health Alliance for Genomic Epidemiology

Search Show Only Genomes With Hits

	abricate: config 0	amrfinderplus: config 0	csstar: config 0	resfinder.py: config 0	staramr: config 0
ERR873305	11 hits	10 hits	10 hits	6 hits	9 hits
ERR873306	11 hits	10 hits	10 hits	6 hits	9 hits
ERR873307	11 hits	10 hits	10 hits	6 hits	9 hits
ERR873308	11 hits	10 hits	10 hits	6 hits	9 hits
ERR873309	16 hits	14 hits	14 hits	6 hits	11 hits
ERR873310	11 hits	10 hits	10 hits	6 hits	9 hits
ERR873311	11 hits	10 hits	10 hits	6 hits	9 hits
ERR873312	11 hits	10 hits	10 hits	6 hits	9 hits

Search Results

Total hits: 0

Genomes with hits: 0

Tools with hits: 0

Differential results: 0

Selected

Compare Clear

Public Health Alliance for Genomic Epidemiology

Search Show Only Genomes With Hits

	abricate: config 0	amrfinderplus: config 0	csstar: config 0	resfinder.py: config 0	staramr: config 0
ERR873305	11 hits	10 hits	10 hits	6 hits	9 hits
ERR873306	11 hits	10 hits	10 hits	6 hits	9 hits

Search Results

Total hits: 130

Genomes with hits: 43

Tools with hits: 6

Differential results: 43

Selected

Compare Clear

The hAMRonization Package



<https://github.com/pha4ge/hAMRonization>

Public Health Alliance for Genomic Epidemiology

Search Show Only Genomes With Hits Restore Results

	abricate: config 0	amrfinderplus: config 0	csstar: config 0	resfinder.py: config 0	staramr: config 0
ERR873305	11 hits	10 hits	10 hits	6 hits	9 hits
ERR873306	11 hits	10 hits	10 hits	6 hits	9 hits
ERR873307	11 hits	10 hits	10 hits	6 hits	9 hits
ERR873308	11 hits	10 hits	10 hits	6 hits	9 hits
ERR873309	16 hits	14 hits	14 hits	6 hits	11 hits
ERR873310	11 hits	10 hits	10 hits	6 hits	9 hits
ERR873311	11 hits	10 hits	10 hits	6 hits	9 hits
ERR873312	11 hits	10 hits	10 hits	6 hits	9 hits

Search Results

Total hits: 0

Genomes with hits: 0

Tools with hits: 0

Differential results: 0

Selected

Compare Clear

Public Health Alliance for Genomic Epidemiology

fosfomycin Show Only Genomes With Hits

	abricate: config 0	amrfinderplus: config 0	csstar: config 0	resfinder.py: config 0	staramr: config 0
ERR873305	11 hits aac(6)-29a aph(3)-Ib blaOXA-395 blaPDC-158 blaPDC-55 blaVIM-2 catB7 cmiB1 crpP fosA-354827590 sul1	10 hits aac(6)-29a aph(3)-Ib blaOXA-395 blaPDC-3 blaVIM-2 catB7 crpP fosA qacEdelta1 sul1	10 hits aac(6) aph(3) bcrl blaOXA blaPAO blaVIM catB7 crpP sul1	6 hits aac(6)-29a blaVIM-2 catB7 crpP fosA sul1	9 hits aac(6)-29a aph(3)-Ib blaOXA-50 blaPAO blaVIM-2 catB7 crpP fosA sul1

Search Results

Total hits: 44

Genomes with hits: 41

Tools with hits: 4

Differential results: 41

Selected

fosA-354827590 from genome 0, tool 0

fosA from genome 0, tool 1

fosATR from genome 0, tool 2

Compare Clear

Public Health Alliance for Genomic Epidemiology

fosfomycin Show Only Genomes With Hits

	ERR873305 abricate: config 0	ERR873305 amrfinderplus: config 0	ERR873305 csstar: config 0
input_file_name	ERR873305	ERR873305	ERR873305
gene_symbol	fosA-354827590	fosA	fosATR
gene_name	FosA family fosfomycin resistance glutathione transferase	FosA family fosfomycin resistance glutathione transferase	fosATR
reference_database_id	ncbi	NCBI Reference Gene Database	RefGANNOT
reference_database_version	2020-Apr-19	2020-03-20.1	2020-Nov-05
reference_accession	NG_047883.1	WP_003082280.1	fosATR
analysis_software_name	abricate	amrfinderplus	csstar
analysis_software_version	1.0.1	3.6.10	2.0.0
sequence_identity	98.53	98.52	98.529
config_id	griBUGSERR873305_36	griBUGSERR873305_36	griBUGSERR873305_36
query_start_aa			
query_stop_aa			
query_start_nt	8371	8374	

Standardized Output Specification

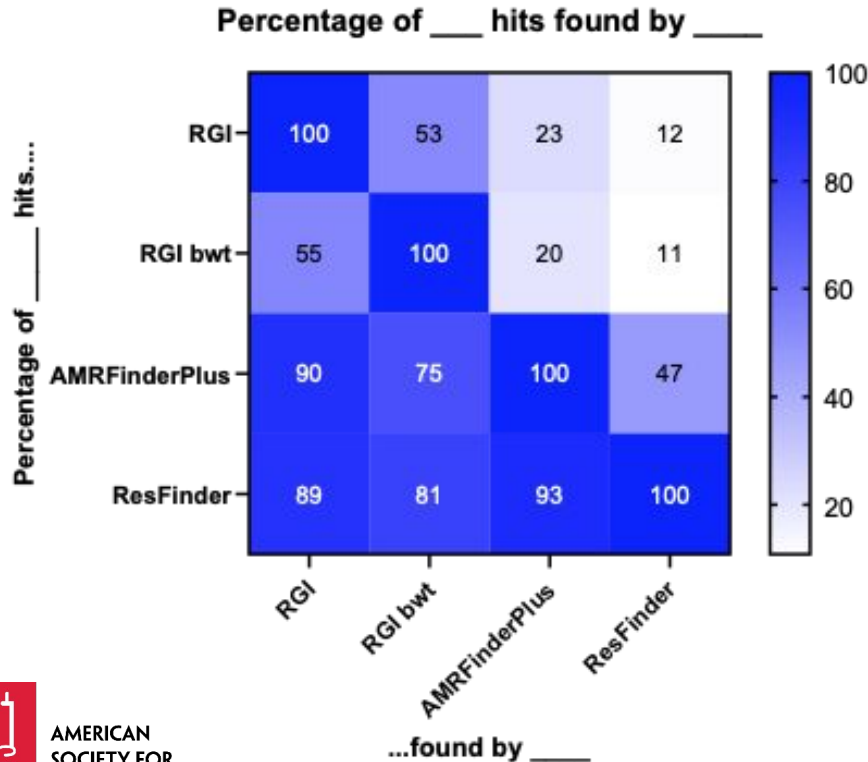
Data **specification** for the detection of AMR genes based on **harmonisable, publicly available, community standards** implementable via a **report template** with a **standardized list of labels and terms** and respective **ontology identifiers**.

Mandatory terms:

- Input File Name
- Gene Symbol
- Gene Name
- Sequence Identity
- Reference Database ID
- Reference Database Version
- Reference Accession
- Reference Sequence Length
- Target Sequence Length
- Analysis Software Name
- Analysis Software Version



89 *Klebsiella pneumoniae* sequence data



RGI: BLASTp + CARD

RGI bwt (beta): Bowtie2 + CARD

AMRFinderPlus: BLASTx + BARRG DB


ResFinder: BLASTn + ResFinder DB

	RGI			
	RGI	bwt	AMR	Res
Total # unique gene names	140	223	70	46


87 *Pseudomonas aeruginosa* sequence data

Isolates (ST=111) associated with Metallo-Carbapenemases resistance outbreak in the United Kingdom


- Most isolates (73) carried VIM-2, conferring the resistance phenotype
- Multiple tools finding the same genes increases evidence, which increases confidence



The **lack of standardization** in the reporting of **AMR gene detection** greatly hinders the comparison of results across the public health sector. The myriad of options available for this purpose highlights a grave **interoperability** problem.



We developed a **standardized data specification** to improve data harmonization and interoperability, implemented in a **Biopython-compatible parser** and **command-line utility**.



The hAMRomization allows the **dissemination of results to stakeholders in a single consistent format**, allowing not only the comparison of tools and databases, but the validation of results through multiple detection algorithms.

Thank You!

Questions? Contact Us.

Catarina Inês Mendes

cimendes@medicina.ulisboa.pt

@ines_cim

PHA4GE Data Structures Workgroup

datastructures@pha4ge.org

PHA4GE Consortium

help@pha4ge.org

@pha4ge

