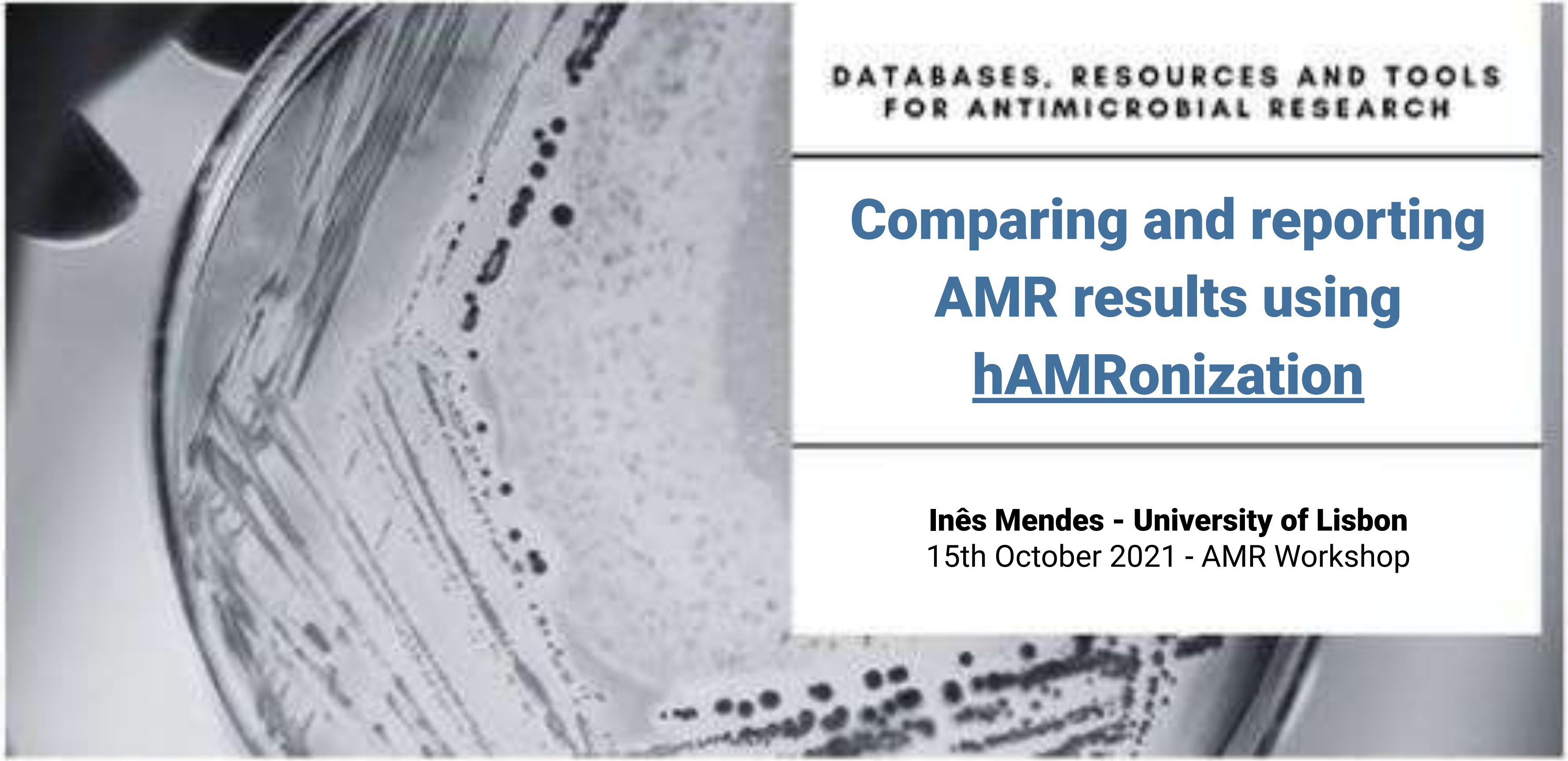


DATABASES, RESOURCES AND TOOLS
FOR ANTIMICROBIAL RESEARCH

Comparing and reporting AMR results using hAMRonization

Inês Mendes - University of Lisbon

15th October 2021 - AMR Workshop

Public Health Alliance for Genomic Epidemiology

jpiamr

CLIMB BIG DATA

PHA4GE - Data Structures Workgroup

Public Health Alliance for Genomic Epidemiology

https://pha4ge.org/

The **Public Health Alliance for Genomic Epidemiology (PHA4GE)** is a global coalition that is actively working to establish consensus standards; document and share best practices; improve the availability of critical bioinformatic tools and resources; and advocate for **greater openness, interoperability, accessibility and reproducibility in public health microbial bioinformatics**.

PHA4GE - Data Structures Workgroup

Dr Emma Griffiths

Chair

Main Goal:

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

Areas of Interest:

Metadata standards, ontologies and conventions; Contextual data harmonization and sharing; Data inputs/outputs, APIs and interoperability; Result reporting and views; Data Security and Encryption; Identity management for role/resource based access

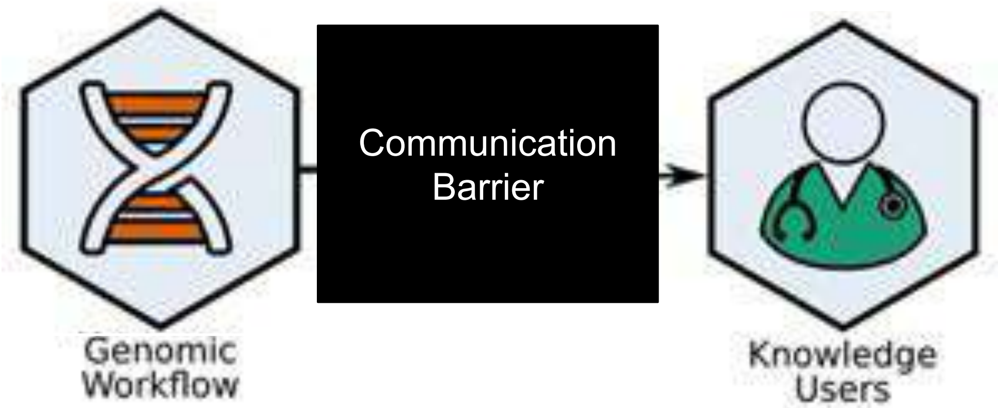
Current projects:

- SARS-CoV-2 Contextual Data Specification
- Gene Detection/AMR Output Specification ⬅

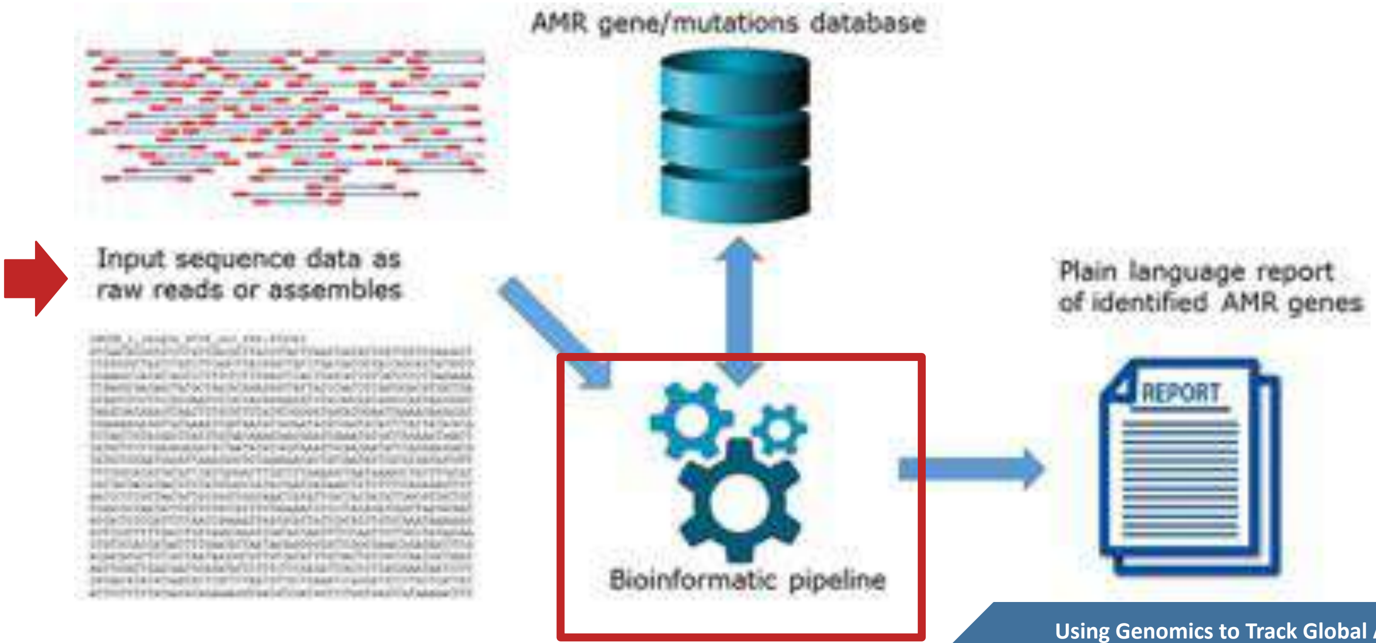
The Antimicrobial Resistance Data Problem

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

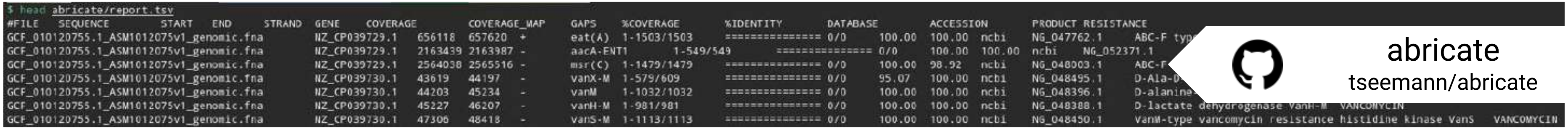
A large number of **different bioinformatic tools** have been developed to perform this task but **differ in supported inputs**, search algorithm, parameterisation, underlying reference databases and **output formats**.



Communication Barrier



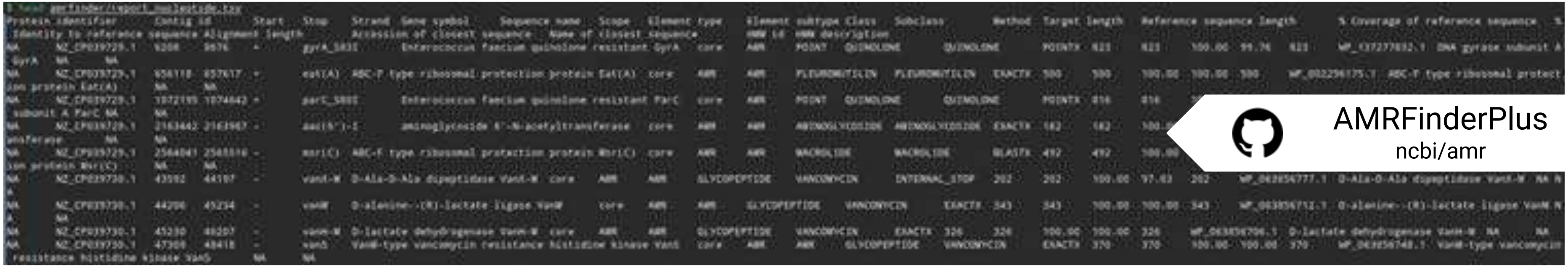
Using Genomics to Track Global Antimicrobial Resistance
Hendriksen et al, Front. Public Health, 2019



```
$ 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 typ
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
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
GCF_010120755.1_ASM1012075v1_genomic.fna        NZ_CP039730.1   43619   44197   -       vanX-M  1-579/609       =============== 0/0     95.07   100.00  ncbi    NG_048495.1     D-Ala-D
GCF_010120755.1_ASM1012075v1_genomic.fna        NZ_CP039730.1   44203   45234   -       vanM    1-1032/1032     =============== 0/0     100.00  100.00  ncbi    NG_048396.1     D-alanine
GCF_010120755.1_ASM1012075v1_genomic.fna        NZ_CP039730.1   45227   46207   -       vanH-M  1-981/981       =============== 0/0     100.00  100.00  ncbi    NG_048388.1     D-lactate dehydrogenase VanH-M  VANCOMYCIN
GCF_010120755.1_ASM1012075v1_genomic.fna        NZ_CP039730.1   47306   48418   -       vanS-M  1-1113/1113     =============== 0/0     100.00  100.00  ncbi    NG_048450.1     VanM-type vancomycin resistance histidine kinase VanS   VANCOMYCIN
```

abricate

tseemann/abricate

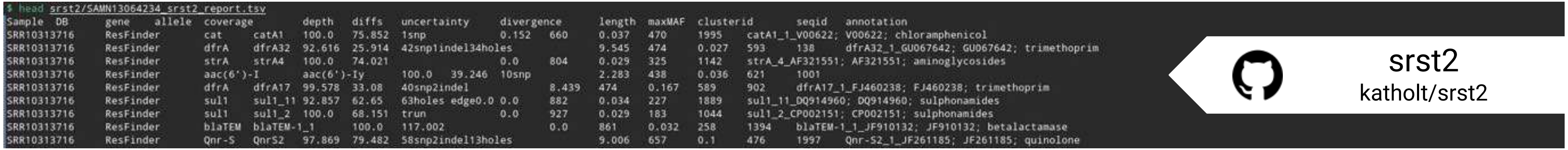


AMRFinderPlus

ncbi/amr

ariba

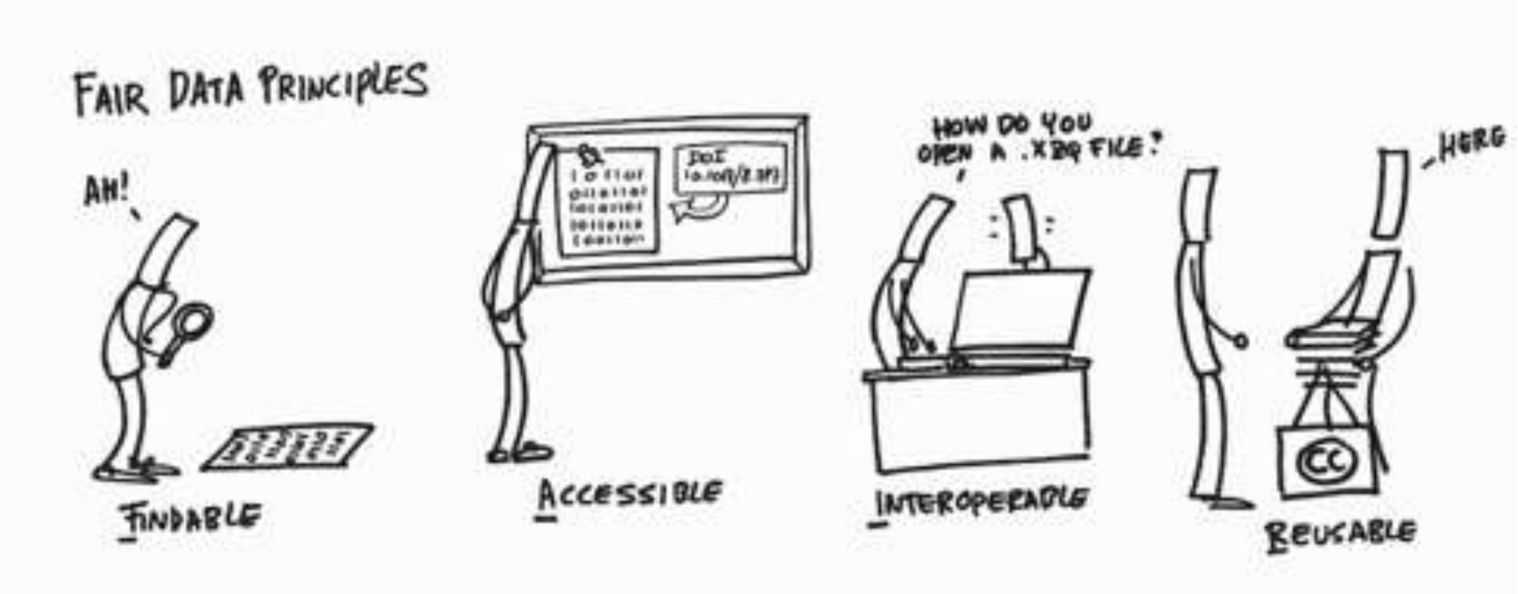
sanger-pathogens/ariba



```
$ head srst2/SAMN13064234_srst2_report.tsv
Sample  DB      gene    allele  coverage        depth   diffs   uncertainty     divergence      length  maxMAF  clusterid       seqid   annotation
SRR10313716     ResFinder       cat     catA1   100.0   75.852  1snp    0.152   660     0.037   470     1995    catA1_1_V00622; V00622; chloramphenicol
SRR10313716     ResFinder       dfrA    dfrA32  92.616  25.914  42snp1indel34holes              9.545   474     0.027   593     138     dfrA32_1_GU067642; GU067642; trimethoprim
SRR10313716     ResFinder       strA    strA4   100.0   74.021          0.0     804     0.029   325     1142    strA_4_AF321551; AF321551; aminoglycosides
SRR10313716     ResFinder       aac(6')-I       aac(6')-Iy      100.0   39.246  10snp           2.283   438     0.036   621     1001
SRR10313716     ResFinder       dfrA    dfrA17  99.578  33.08   40snp2indel             8.439   474     0.167   589     902     dfrA17_1_FJ460238; FJ460238; trimethoprim
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
```

srst2

katholt/srst2



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

1.

The **lack of standardizatio**n 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**.

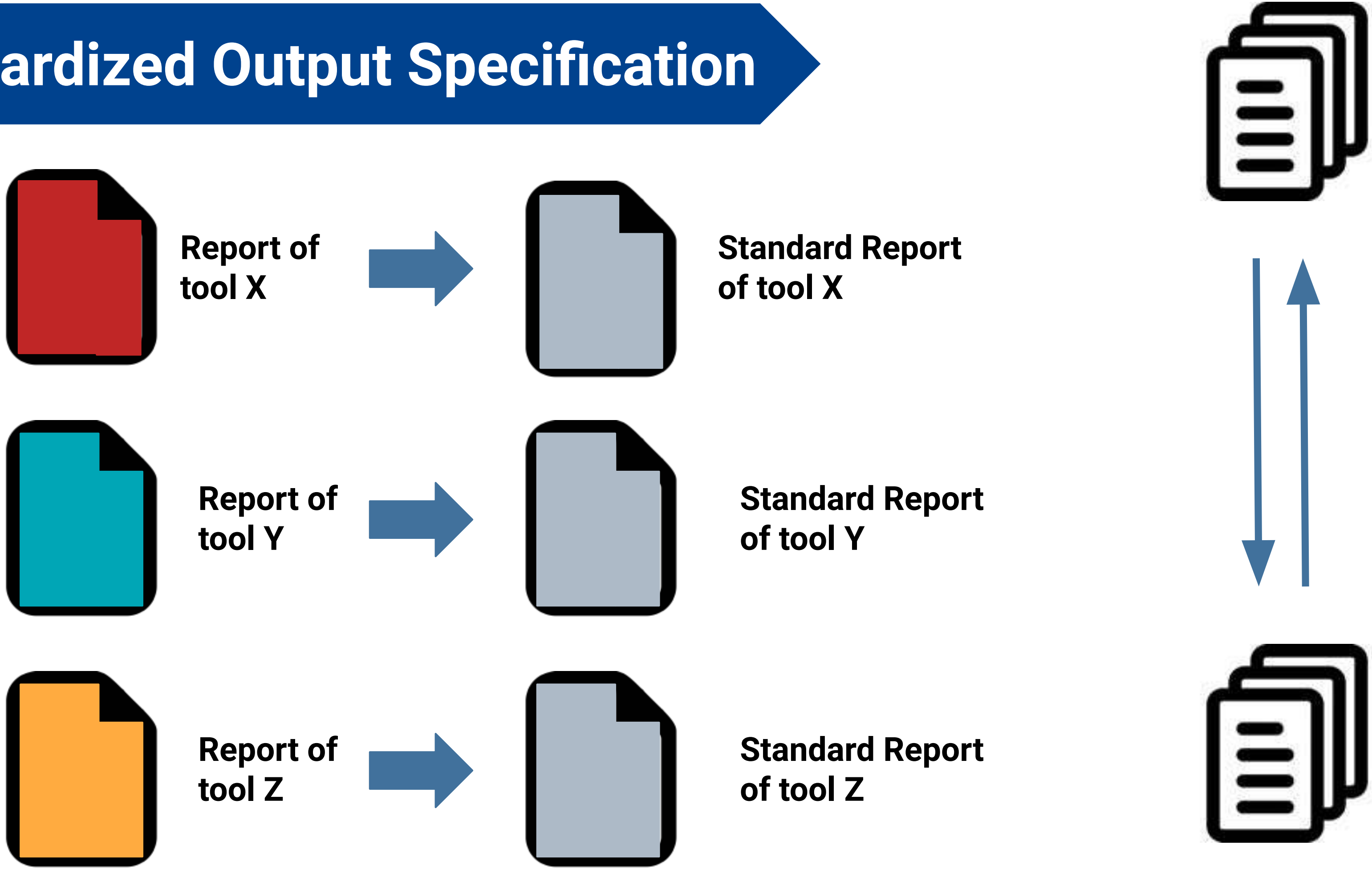
Standardized Output Specification

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**

Standardized Output Specification



Report of tool X

Standard Report of tool X

Report of tool Y

Standard Report of tool Y

Report of tool Z

Standard Report of tool Z

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.

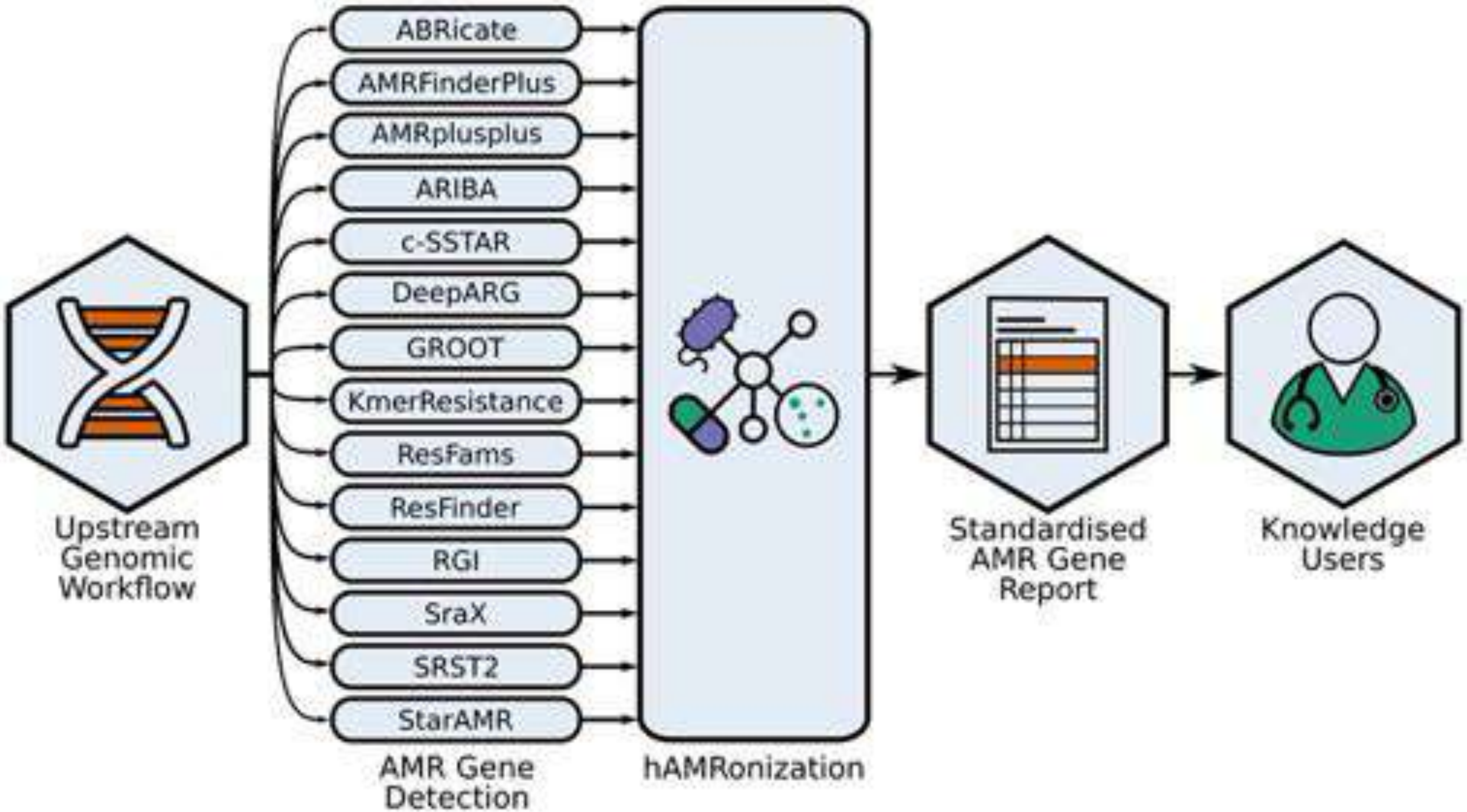
hAMRonization Package

Take tool-specific output and reformat to standardized output.

Standardized Output Specification

Enables consistent comparison of AMR genes detected by different tools.

The hAMRonization package automates conversion to a standardized output where tools can be built on.


Upstream Genomic Workflow
ABRicate
AMRFinderPlus
AMRplusplus
ARIBA
c-SSTAR
DeepARG
GROOT
KmerResistance
ResFams
ResFinder
RGI
SraX
SRST2
StarAMR
AMR Gene Detection
hAMRonization
Standardised AMR Gene Report
Knowledge Users

2.

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

The hAMRonization Workflow

https://github.com/pha4ge/hAMRonization_workflow

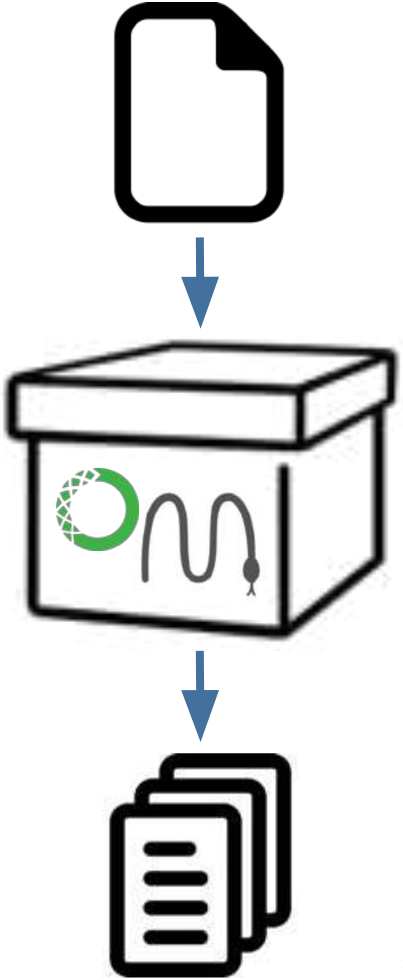
Snakemake workflow to run **14 different species-agnostic AMR gene detection tools** selected tools on a dataset, installing **fixed versions** of the tools from **conda** on execution.

Ensures:

- Reproducibility
- Scalability

Tool inclusion criteria:

- Specificity for AMR <u>gene</u> detection
- Open-source



Dr Finlay Maguire

Dr Adam Witney

Dr Simon Tausch

14

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.


Dr Finlay Maguire


Inês Mendes


Alex Manuele


Amos Raphenya

And Counting!

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:

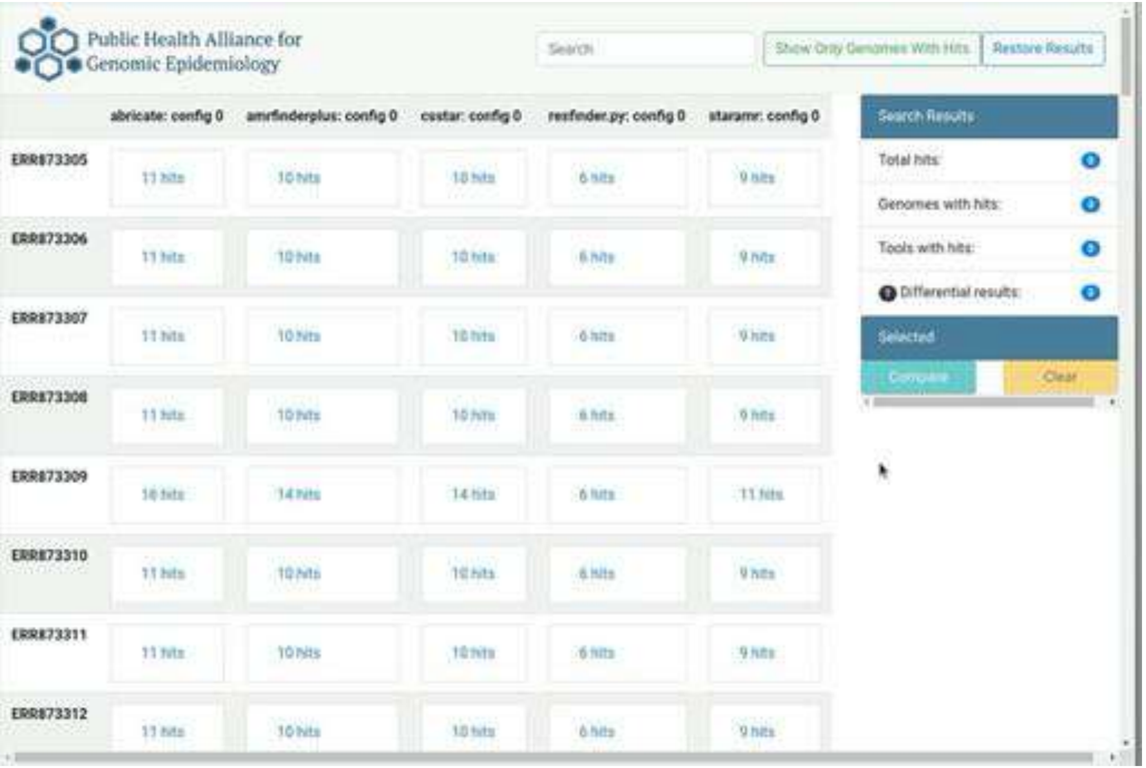
```
> hamronize summarize --summary_type interactive <list of reports>
```

Summary output formats:

- TSV
- Json
- Interactive HTML

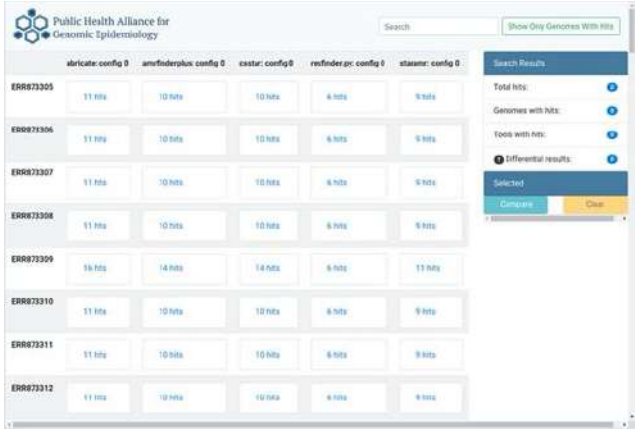
The hAMRonization Package

https://github.com/pha4ge/hAMRonization



Public Health Alliance for Genomic Epidemiology

	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



Public Health Alliance for Genomic Epidemiology

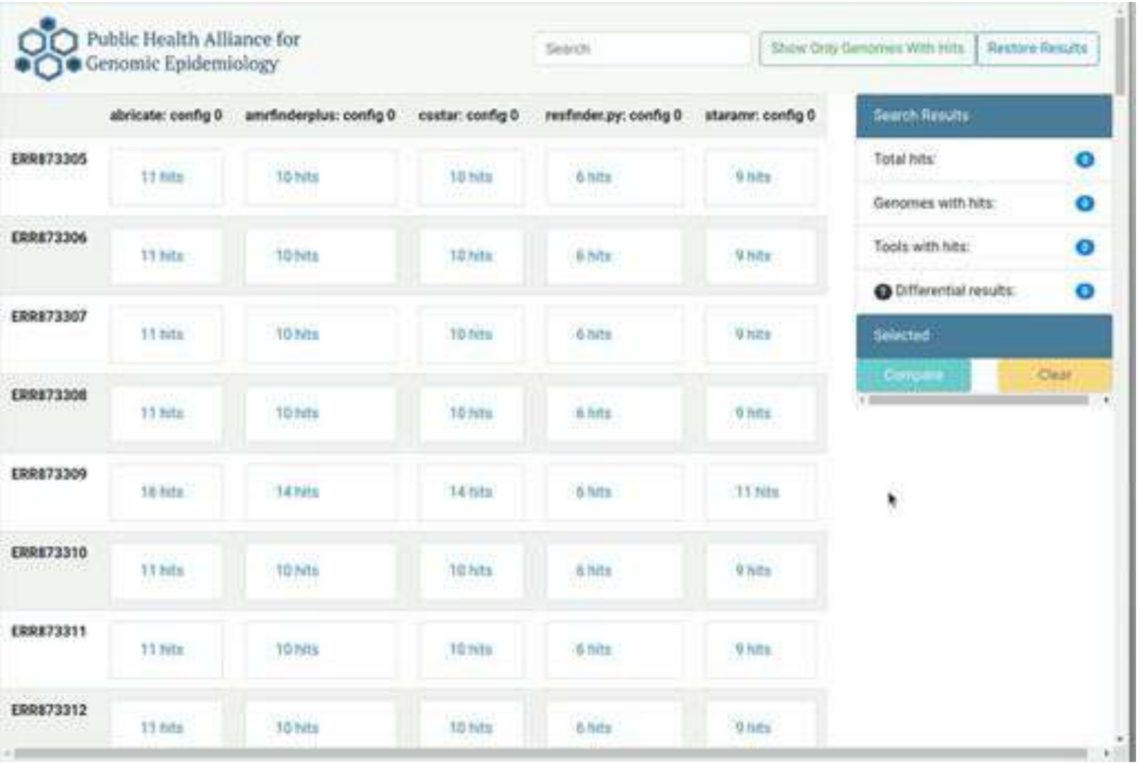
	abricate: config 0	amrfinderplus: config 0	csstar: config 0	resfinder.py: config 0	staramr: config 0
ERR873305	11 hits	10 hits	10 hits	6 hits	9 hits
ERR873306	11 hits	10 hits	10 hits	6 hits	9 hits
ERR873307	11 hits	10 hits	10 hits	6 hits	9 hits
ERR873308	11 hits	10 hits	10 hits	6 hits	9 hits
ERR873309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

Public Health Alliance for Genomic Epidemiology

https://hamronization-demo.herokuapp.com

The hAMRonization Package

https://github.com/pha4ge/hAMRonization



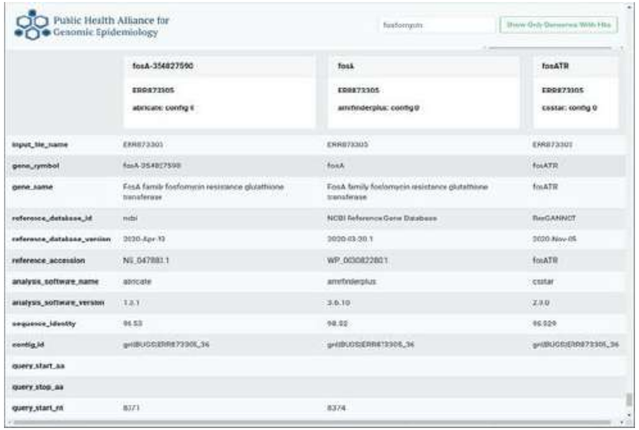
	abricate: config 0	amrfinderplus: config 0	csstar: config 0	resfinder.py: config 0	staramr: config 0
ERR873305	11 hits	10 hits	10 hits	6 hits	9 hits
ERR873306	11 hits	10 hits	10 hits	6 hits	9 hits
ERR873307	11 hits	10 hits	10 hits	6 hits	9 hits
ERR873308	11 hits	10 hits	10 hits	6 hits	9 hits
ERR873309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

	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
	aac(6')-29a	aac(6')-29a	aac(6')	aac(6')-29a	aac(6')-29a
	aph(3')-IIb	aph(3')-IIb	aph(3')	blaVIM-2	aph(3')-IIb
	blaOXA-395	blaOXA-395	bcr1	catB7	blaOXA-50
	blaPDC-158	blaPDC-3	blaOXA	crpP	blaPAO
	blaPDC-55	blaVIM-2	blaPAO	fosA	blaVIM-2
	blaVIM-2	catB7	blaVIM	sul1	catB7
	catB7	crpP	catB7		crpP
	cmlB1	fosA	crpP		fosA
	crpP	qacEdelta1	fosATR		sul1
	fosA-354827590	sul1	sul1		
	sul1				
ERR873306					

fosA-354827590 from genome 0, tool 0

fosA from genome 0, tool 1

fosATR from genome 0, tool 2



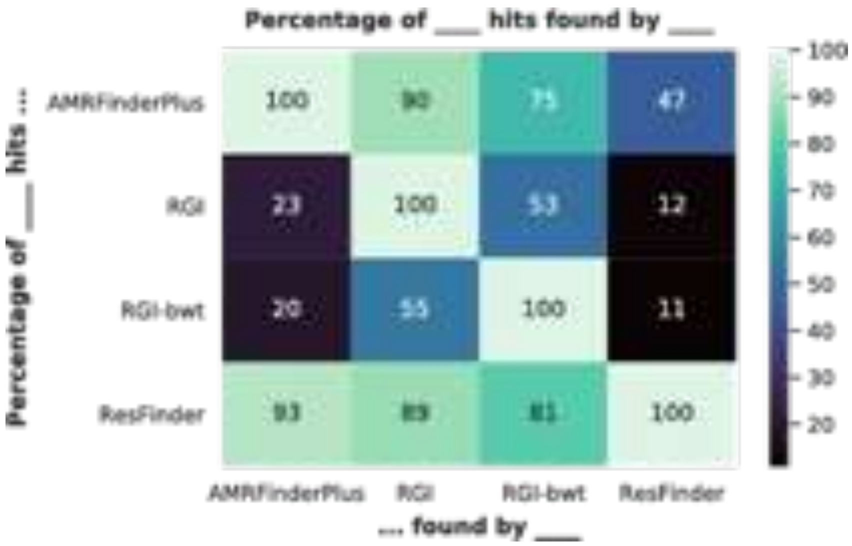
	fosA-354827590 ERR873305 abricate: config 0	fosA ERR873305 amrfinderplus: config 0	fosATR ERR873305 csstar: config 0
input_file_name	ERR873305	ERR873305	ERR873305
gene_symbol	fosA-354827590	fosA	fosATR
gene_name	FosA family fosfomycin resistance glutathione transferase	FosA family fosfomycin resistance glutathione transferase	fosATR
reference_database_id	ncbi	NCBI Reference Gene Database	ResGANNOT
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	96.53	98.02	96.529
contig_id	gnl|BUCC|ERR873305_36	gnl|BUCC|ERR873305_36	gnl|BUCC|ERR873305_36
query_start_aa			
query_stop_aa			
query_start_nt	8371	8374	

https://hamronization-demo.herokuapp.com

3.

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.

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

Elizabeth Culp

McMaster University

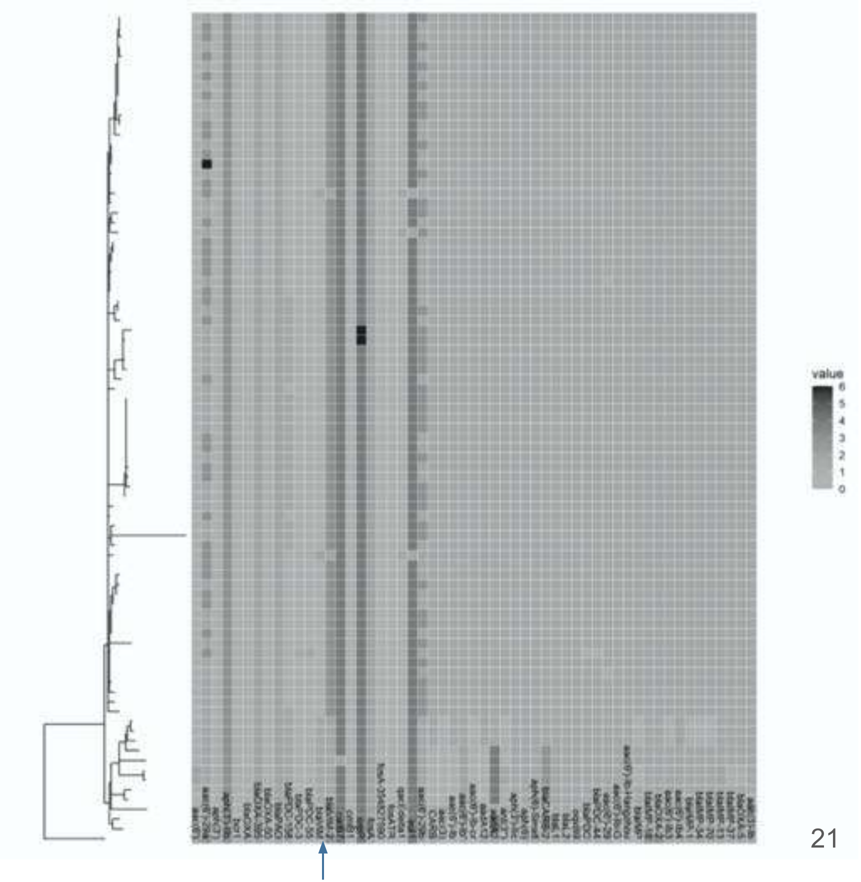
87 *Pseudomonas aeruginosa* sequence data

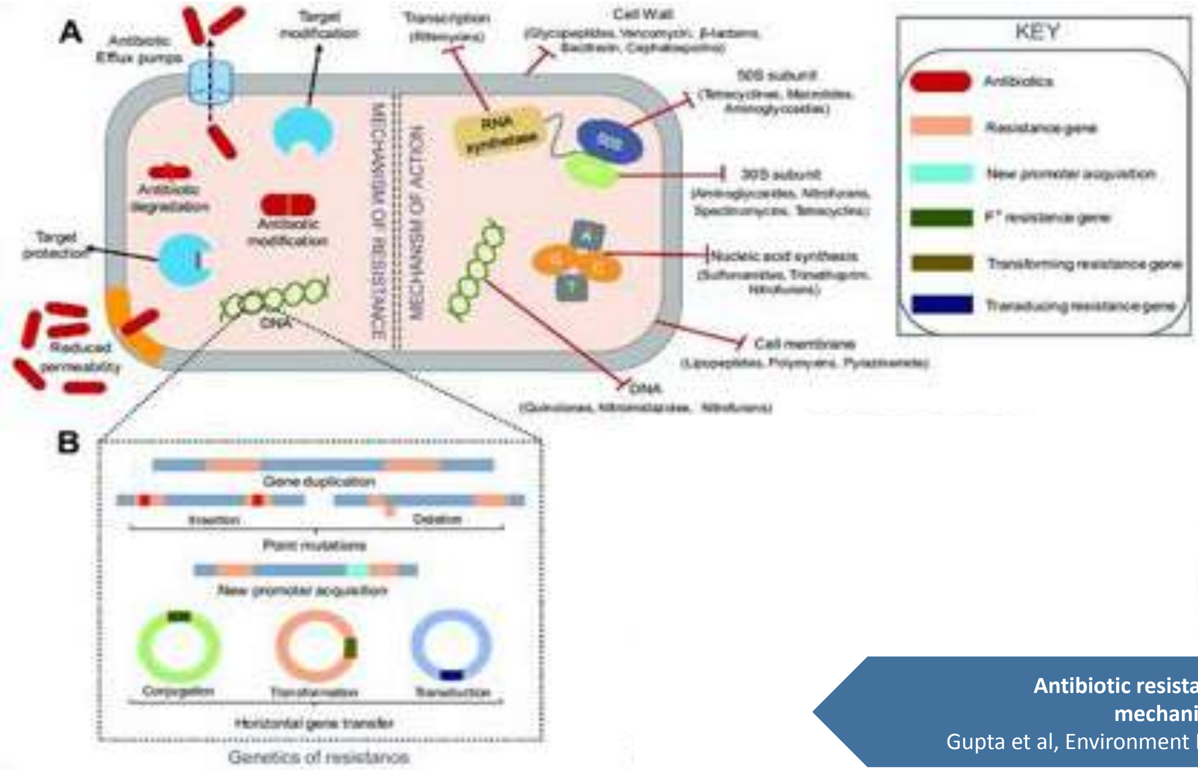
Isolates (ST=111) associated with Metallo-Carbapenemases resistance outbreak in the United Kingdom

- Most isolates (73) carried *VIM-2*, conferring the resistance phenotype
- Dataset run on **abricate, amrfinderplus, csstar, resfinder** and **staramr**
- Multiple tools finding the same genes increases evidence, which increases confidence

Dr Adam Witney

St George's
University of London





Antibiotic resistance dissemination mechanisms and pathways

Gupta et al, Environment International, 2020


CAUTION
WORK
IN PROGRESS

Standardized Output Specification

The hAMRonization Package

https://github.com/pha4ge/hAMRonization

Gene Detection Standard

\+

Mutation Detection Standard

Additional terms:

- **Genetic Variation Type**
- Predicted Phenotype
- Predicted Phenotype Confidence Level
- Nucleotide mutation
- Nucleotide mutation interpretation
- Protein mutation
- Protein mutation interpretation
- Frequency of variant

Standardized Output Specification

The hAMRonization Package

https://github.com/pha4ge/hAMRonization

HGVS
HUMAN GENOME
VARIATION SOCIETY

Sequence variant nomenclature system implemented for genetic changes and denoting sequence coordinates.

Due to the vast types of genetic changes that might occur, nomenclature is not easily understood. Programmatic interface to provide an interpretation is required.

https://varnomen.hgvs.org/recommendations/general/
https://github.com/conmeehan/laymansHGVS

The TB-Profiler case study

TBProfiler database (tbdb) incorporates **1541 individual mutation/resistance** associations across 16 anti-tuberculosis drugs

TB Profiler

https://github.com/jodyphelan/tbdb/

- **Genetic Variation Type: protein_mutation**
- Nucleotide mutation: c.1349C>T
- Nucleotide mutation interpretation:
  This is a subst found in rpoB at position 1349 where the reference has a C and the sample has a T

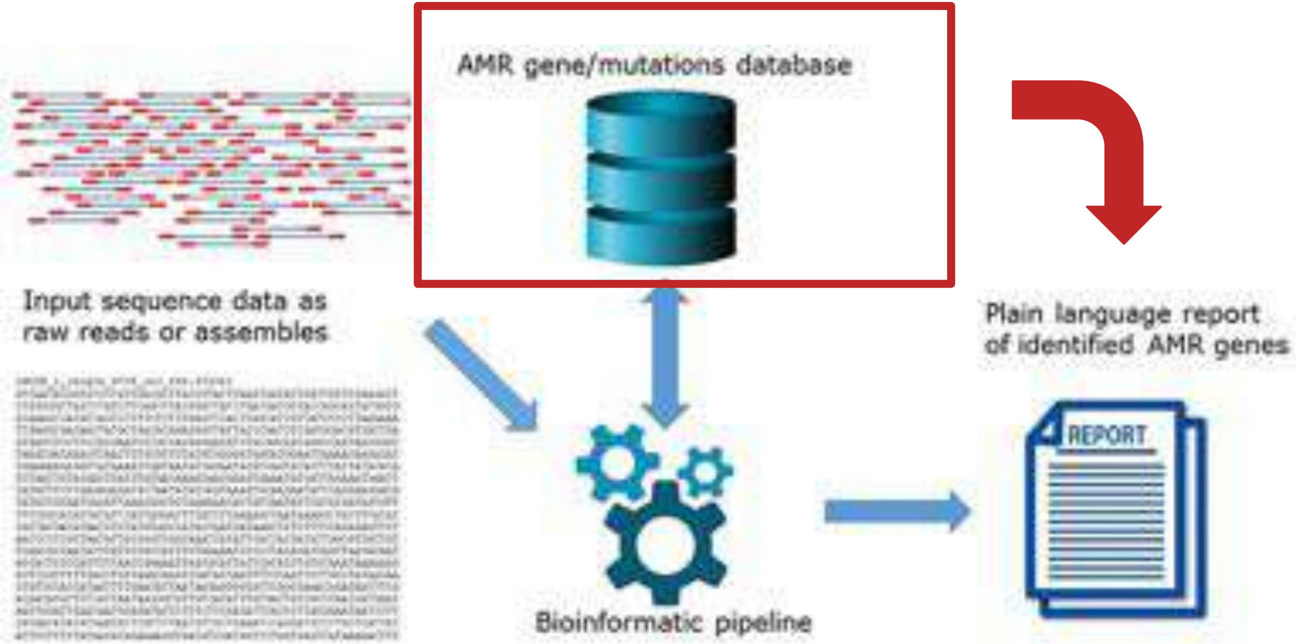
- Protein mutation: p.Ser450Leu
- Protein mutation interpretation:
  This is an amino acid subst found in rpoB at position 450 where the reference has a Serine and the sample has a Leucine
- Frequency of variant: 0.57


Jody Phelan


Conor Meehan

4.

Resistance phenotypes can be not only plasmid-mediated, or due to the presence of whole genes, but also chromosomal mutations. The ability to detect **not just gene absence/presence**, but **more granular changes** is of uttermost importance for different public health analyses.



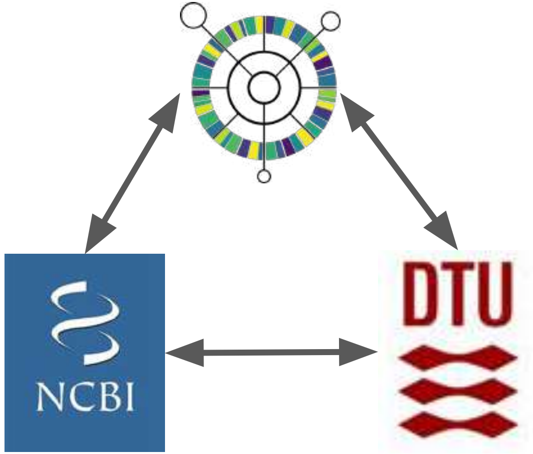
Using Genomics to Track Global Antimicrobial Resistance
Hendriksen et al, Front. Public Health, 2019

AMR Database Harmonization

Charm😍Db

(pronounced 'charmed' /tʃɑː(r)md/)

https://gitlab.com/antunderwood/chamredb




Anthony Underwood


Inês Mendes

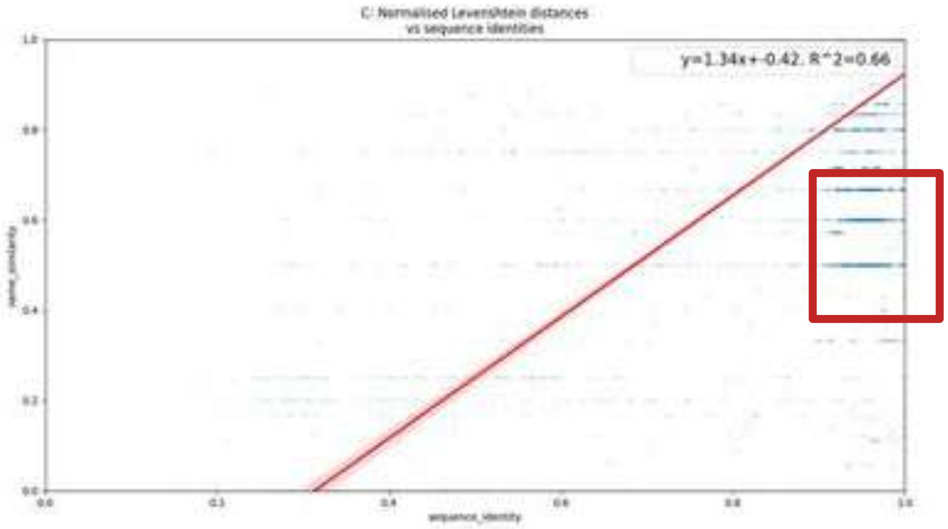

Alex Manuele

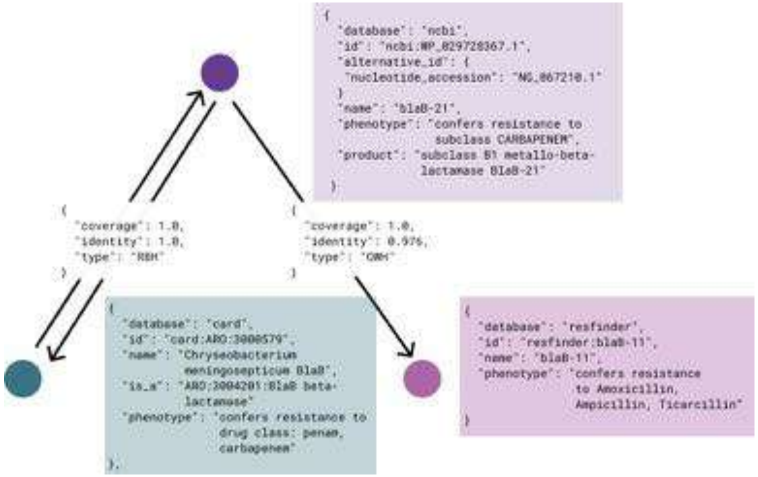

Dr Adam Witney


Trestan Pillonel


Varun Shamanna

AMR Database Harmonization





```
{
  "database": "ncbi",
  "id": "ncbi:WP_029728367.1",
  "alternative_id": {
   "nucleotide_accession": "NG_067210.1"
  }
  "name": "blaB-21",
  "phenotype": "confers resistance to
                subclass CARBAPENEM",
  "product": "subclass B1 metallo-beta-
              lactamase BlaB-21"
 }
```

```
{
  "coverage": 1.0,
  "identity": 1.0,
  "type": "RBH"
}
```

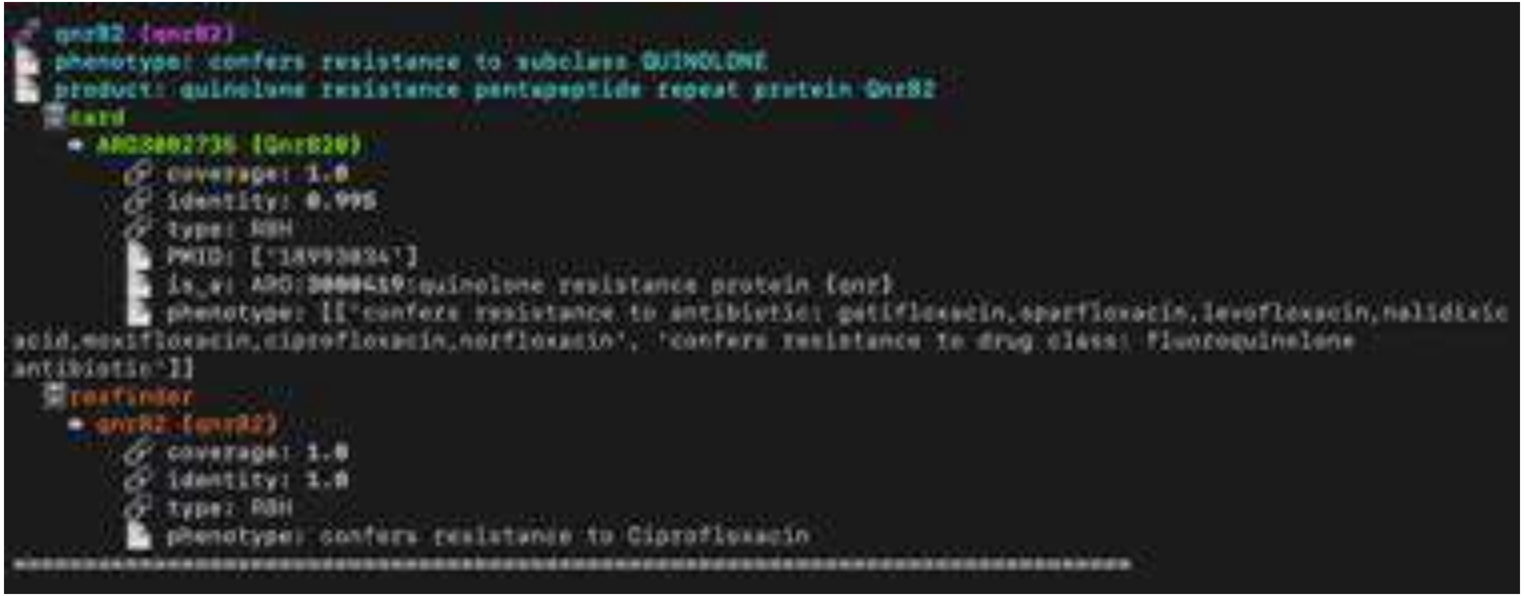
```
{
  "coverage": 1.0,
  "identity": 0.976,
  "type": "OMH"
}
```

```
{
  "database": "card",
  "id": "card:ARO:3000579",
  "name": "Chryseobacterium
           meningosepticum BlaB",
  "is_a": "ARO:3004201:BlaB beta-
           lactamase"
  "phenotype": "confers resistance to
                drug class: penam,
                carbapenem"
},
```

```
{
  "database": "resfinder",
  "id": "resfinder:blaB-11",
  "name": "blaB-11",
  "phenotype": "confers resistance
                to Amoxicillin,
                Ampicillin, Ticarcillin"
}
```

AMR Database Harmonization

```
$ chamredb query -d ncbi -i qnrB2
```



```
 qnrB2 (qnrB2)
 phenotype: confers resistance to subclass QUINOLONE
 product: quinolone resistance pentapeptide repeat protein QnrB2
   card
    ARO:3002735 (QnrB20)
        coverage: 1.0
        identity: 0.995
        type: RBH
        PMID: ['18993834']
        is_a: ARO:3000419:quinolone resistance protein (qnr)
        phenotype: [['confers resistance to antibiotic: gatifloxacin,sparfloxacin,levofloxacin,nalidixic
acid,moxifloxacin,ciprofloxacin,norfloxacin', 'confers resistance to drug class: fluoroquinolone
antibiotic']]
   resfinder
    qnrB2 (qnrB2)
        coverage: 1.0
        identity: 1.0
        type: RBH
        phenotype: confers resistance to Ciprofloxacin
================================================================================
```

AMR Database Harmonization

```
$ hamronize abricate report.tsv --reference_database_version db_v_1 \
--analysis_software_version tool_v_1 --format json
```

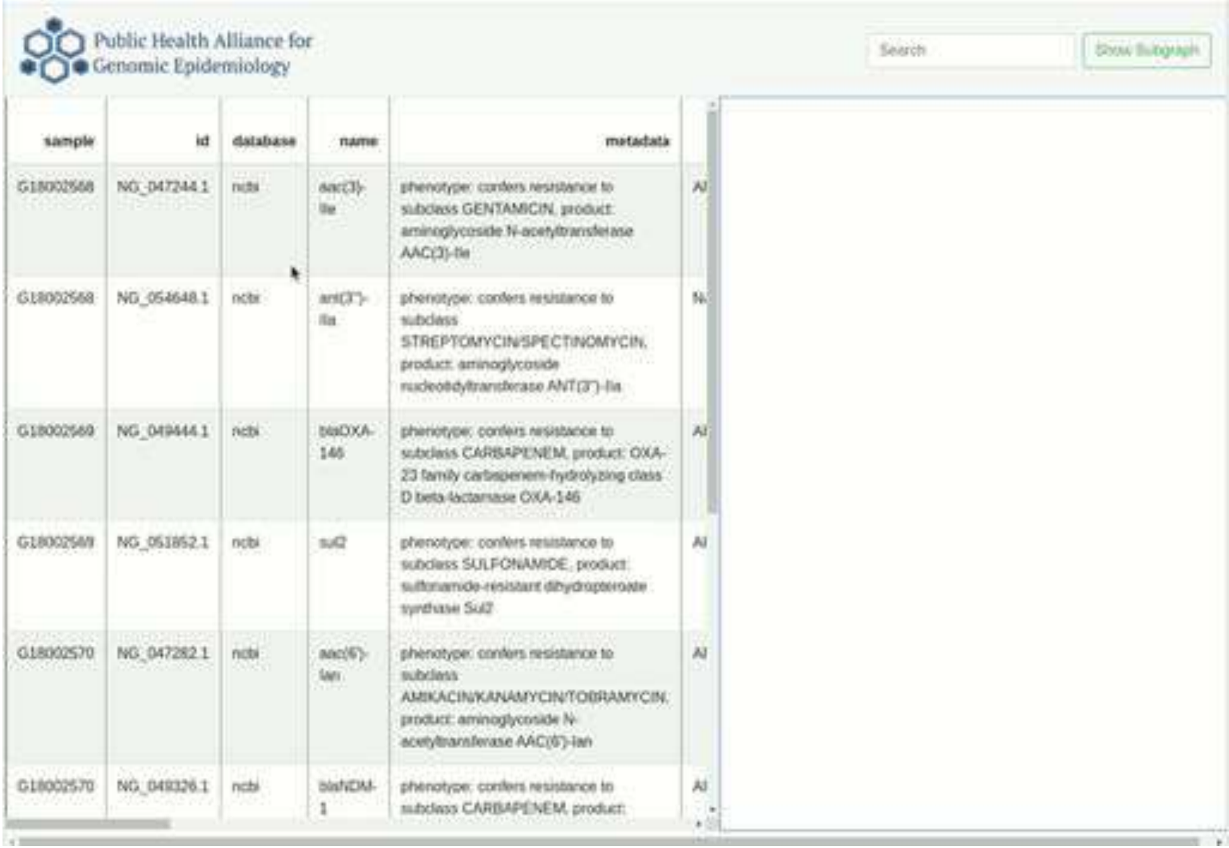
…

```
$ hamronize summarize -o hamronize_summary.json -t json *.json
```

```
$ chamredb query -j hamronize_summary.json -o hamronize_summary.tsv
```

AMR Database Harmonization

Charm😍Db



sample	id	database	name	metadata
G18002568	NG_047244.1	ncbi	aac(3)-IIe	phenotype: confers resistance to subclass GENTAMICIN, product: aminoglycoside N-acetyltransferase AAC(3)-IIe
G18002568	NG_054648.1	ncbi	ant(3'')-IIa	phenotype: confers resistance to subclass STREPTOMYCIN/SPECTINOMYCIN, product: aminoglycoside nucleotidyltransferase ANT(3'')-IIa
G18002569	NG_049444.1	ncbi	blaOXA-146	phenotype: confers resistance to subclass CARBAPENEM, product: OXA-23 family carbapenem-hydrolyzing class D beta-lactamase OXA-146
G18002569	NG_051852.1	ncbi	sul2	phenotype: confers resistance to subclass SULFONAMIDE, product: sulfonamide-resistant dihydropteroate synthase Sul2
G18002570	NG_047282.1	ncbi	aac(6')-Ian	phenotype: confers resistance to subclass AMIKACIN/KANAMYCIN/TOBRAMYCIN, product: aminoglycoside N-acetyltransferase AAC(6')-Ian
G18002570	NG_049326.1	ncbi	blaNDM-1	phenotype: confers resistance to subclass CARBAPENEM, product:

5.

A unified global picture requires not only a **common ground** for the comparison of results from different **tools**, but a way to compare the **databases** that are used to generate those results.

15th October 2021
AMR Workshop

Inês Mendes
cimendes@medicina.ulisboa.pt
Twitter: @ines_cim

PHA4GE Data Structures Workgroup
datastructures@pha4ge.org

PHA4GE Consortium
help@pha4ge.org
https://pha4ge.org/
Twitter: @pha4ge

Public Health Alliance for Genomic Epidemiology

jpiamr

CLIMB BIG DATA