AMR Rules:
Interpretive Standards for AMR Genotypes

ESGEM-AMR Working Group
Technical Guidance for
Defining Interpretive Rule Sets

Background, rationale, objectives and workplan for the ESGEM-AMR Working Group are described here.

Members are encouraged to use their own knowledge and expertise to propose rule sets for their assigned species, as long as they can provide sufficient evidence to justify the rules.

This document lays out technical details regarding the proposed data standard for the rule sets, guidance on technical issues and standards of evidence needed to add rules, and a suggested protocol for populating rule sets.

NOTE: the notation R^{NWT}, R^{WT} is used in these descriptive documents for clarity, to avoid spaces, and in line with the nomenclature of the EUCAST Subcommittee on WGS and Phenotypic AST. However as the interpretive rules need to be encoded in simple text they cannot include superscript, therefore they are written as 'nwt R', 'wt R', etc instead. The two notations have identical meaning, i.e. 'R^{NWT}' should be read interchangeably with 'nwt R' as meaning 'wildtype resistant'.

Authors: Kat Holt, Jane Hawkey, Natacha Couto, ESGEM Executive Committee
Last updated: 16 July 2025

Working Group Objectives

The AMRrules interpretive standards will aim to capture all exceptions to the generalized interpretation of 'presence of gene X' implies 'R^{NWT} or I^{NWT} to drug Y'. Ultimately, they should also differentiate R^{NWT}' from I^{NWT}', and interpret combinations of genes.

However, the initial focus of the ESGEM-AMR Working Group will be on clearly delineating **'wildtype' (core) genotypes underlying 'wildtype' (intrinsic/expected) phenotypes** for clinically relevant bacteria. This is analogous to EUCAST's Expected Resistance and Expert Rules for susceptibility testing, which capture expert knowledge on interpretive rules for AST, and ideally will capture genetic mechanisms behind all expected resistances.

Initial development will work species-by-species, generating **one rule set for each species**, which can be combined together into a single resource, via the AMRrules project.

FAIR and Open Principles

The overall project adheres to FAIR principles to ensure the interpretive standards developed are findable, interoperable, accessible, and reusable.

- Rule sets will be made **publicly available** via open-access repositories under a permissive license (GNU General Public License v3.0).
- They will be **versioned** via numbered releases, and issued with stable document object identifiers (DOIs).

Rule sets and associated tools should be **interoperable with existing resources** for AMR genotype analysis as far as possible, including use of common sequence identifiers, standard gene nomenclature, and data formats.

- The primary reference database for nomenclature and sequence accessions of AMR determinants will be NCBI's Reference Gene Catalog (refgene) and corresponding Reference Gene Hierarchy.
- The primary reference database for nomenclature for drugs and drug classes will be the Comprehensive Antibiotic Resistance Database (CARD) Antibiotic Resistance Ontology.
- Preliminary code has been developed for annotating the gene-level reports output by the AMRFinderPlus tool, which uses NCBI refgene and gene hierarchy for AMR determinants.
- Code will be developed to be interoperable with the hAMRonization format, to facilitate compatibility with the outputs of CARD RGI, ResFinder, and >12 other AMR genotyping tools whose outputs can be readily converted to hAMRonization format. However in order for rule sets to be used with the outputs of tools that do not use NCBI refgene as the primary database, additional code/tools will be needed to harmonize sequence identifiers to those used to specify rule sets.

General Guidance and Priorities

Wildtype phenotypes

The initial focus of ESGEM-AMR will be creating rule sets that clearly delineate **core genes** associated with **'wildtype' phenotypes** for each species, where that phenotype is:

(1) Wildtype / intrinsic / expected resistant (R^{WT})

- This will usually be reflected in the EUCAST Expected Resistance and Expert Rules for the species. Note that where a species has Expected Resistance to a drug, there are no breakpoints or ECOFFS set for that species-drug combination as it is not clinically relevant, although sometimes MIC distribution data is recorded. Where it is not yet recognised that the wildtype phenotype is resistant, there may be an ECOFF and possibly a breakpoint with which to assess resistance from MIC data, and in some cases there may be a genus- or family-level breakpoint.
- Example: Wildtype alleles of core chromosomal gene *bla*SHV in *Klebsiella pneumoniae* hydrolyse penicillins, resulting in the wildtype MIC distribution for ampicillin exceeding the Enterobacterales R breakpoint of >8 mg/L. This is reflected in the EUCAST Expected Resistances (to ampicillin/amoxicillin and ticarcillin), Expert Rules (report as piperacillin resistant, regardless of test result), and lack of MIC distribution data for penicillin or ticarcillin for *Klebsiella pneumoniae*.

(2) Wildtype susceptible (S^{WT})

- This represents species-drug combinations for which the wildtype population has an MIC distribution that is below the S breakpoint, but that contains a core genetic marker that in another species produces an MIC distribution that is above the S or R breakpoint.
- Example: The *oqxAB* operon is a core locus conserved in *Klebsiella pneumoniae* chromosomes, where presence of wildtype alleles is not associated with clinical resistance to any antimicrobials. However when mobilised to other organisms such as *E. coli*, this locus is often hyper-expressed and associated with increased MIC or even clinical resistance to ciprofloxacin and other drugs. This is reflected in EUCAST Expected Resistances in the sense that these do ***not*** record ciprofloxacin as an expected resistance of *Klebsiella pneumoniae*. The wildtype MIC distribution for ciprofloxacin in *K. pneumoniae* mostly falls below the S breakpoints for ciprofloxacin in Enterobacterales (≤0.25).

Acquired phenotypes

Where relevant and supported by sufficient evidence, rules concerning the interpretation of **acquired determinants** may also be included. However in most cases such rules will need to be defined systematically, based on assessment of evidence for AMR determinants both **individually and in combination**. This will require high-volume high-quality genome-phenotype data, which is not yet available for most organisms and would need to be generated. This will be the focus of Phase 2 of the AMRrules initiative.

Format Specification

Rule sets will follow a **standard specification**, based on the one proposed here and developed in this template, although technical details will be further developed and refined by the working group.

Each row defines a rule for interpreting the presence of one **gene** (or specific variant thereof) in one **species**, in terms of its expected genetic context (core or acquired) and its expected effect on **resistance category** for one **drug**.

Each row must have supporting **evidence** for the rule in the form of evidence codes and a PubMed identifier (**PMID**) for a peer-reviewed article, and an explanatory **note** summarizing the interpretation. Where a rule is made based on new unpublished data analysis, the supporting evidence should be submitted for review by the working group, and preprinted or published as soon as possible.

In principle there should be **one row per species / genetic variant / drug combination**, for (i) clarity of interpreting and parsing the rules; and (ii) for clarity of recording evidence for each specific rule and its relevance to a given species, gene and drug. In reality there may be cases where a rule clearly applies across a higher taxonomic level, or across a gene group, or across a drug class; guidance on specifying such rules are discussed below, and will be further developed during the initial development of rule sets.

For any given species, a rule should be included for **every** core gene matching a known resistance determinant (i.e. every core gene that appears in an AMRFinderPlus report), as the phenotypes of core genes should always be discernable from matched genome-phenotype data and/or an assumed link with Expected Resistance phenotypes defined by EUCAST, even in the absence of primary mechanistic literature on the topic.

A rule should also be included for every Expected Resistance phenotype in any given species, even if the genetic mechanism is unknown. This provides a human-readable, curated, comprehensive catalog of the known and unknown mechanisms behind expected resistance. Ensuring expected resistances are encoded in AMRrules assists with downstream phenotype reporting.

In principle, a species' rule set should include a rule for every acquired gene in the refgene database where that gene has been reported in the species, but presence of that gene does not modify the phenotype for the associated drug/class in this organism. However it is anticipated that, for most species, there is not yet sufficient evidence to complete rule sets for all possible acquired genes; and acquired genes that lack functional evidence in this specific organism should not be included (see General Guidance above).

Fields

The specification for each field is available in this spreadsheet template (this guidance document corresponds to version 0.6 of the specification).

Rule ID

Each rule must have a unique 'ruleID', assigned by the curating organism subgroup. A ruleID is prefixed with a 3-letter code that identifies the subgroup, followed by a 4-digit number that is unique within that subgroup. The list of allowed 3-letter codes is in the 'organism subgroup codes' tab of the rule specification template (v0.6).

Txid

The taxonomy ID of the organism to which the rule applies to, normally a species. The txid can be found in the NCBI Taxonomy Database.

Organism

The organism to which the rule applies, which would normally be to the level of species.

The reference taxonomy for confirming species from a genome sequence will be the NCBI Taxonomy Database (note this is a change from specification v0.5, which used GTDB, as this did not allow us to distinguish between important species e.g. *Yersinia pestis vs pseudotuberculosis*, *Burkholderia pseudomallei vs mallei*).

Use prefixes to indicate the taxonomic level to which a name belongs, e.g. the species *Klebsiella pneumoniae* should have the prefix 's__' to indicate species, i.e "s__Klebsiella pneumoniae".

Future considerations:

- How to specify species complex? NCBI does have internal nodes between the level of genus and species, but we will avoid using these as they have no formal status. The AMR R package has this set of micro-organism groups which is used to apply AST interpretation rules.
- Allow a list of species within the species cell? To allow a rule to be specified that applies to a subset or species within a genus. E.g. EUCAST Expected rules are sometimes specified for a set of two or more species. This could also provide a solution to the species complex question. However, you need to be mindful that this complicates the clear provision of evidence to support each rule and its taxonomic scope.

Gene

The primary reference for a gene in AMRrules is its node ID in the NCBI Reference Gene Hierarchy (discussed below). **If an appropriate node ID exists, this should be entered in both the 'gene' and 'node ID' fields and all other gene identifiers are optional** (although it is strongly encouraged to also include NCBI protein or HMM accessions, and a CARD identifier,

if possible, for ease of compatibility). For expected resistance phenotypes where the genetic mechanism is not known just record ‘unknown’ in this field.

If the gene is not in the NCBI hierarchy, the ‘gene’ field should be completed using the gene or allele name that appears in the NCBI refgene database. If it is not in NCBI refgene, use the gene symbol (e.g. 'mexB') - if the gene is present in CARD, use the gene symbol present there, otherwise try to identify the most suitable gene symbol and be sure to include protein/GenBank/HMM/nucleotide accessions and ARO accessions for clarity.

For combinatorial rules

The ‘gene’ column can be a logical expression defining combinations of variants to which an interpretation should apply, in which the objects of the expression are rule identifiers (‘ruleID’).

Here, the ‘ruleID’ can be used as a shorthand label for the variant defined by
‘gene’:’mutation’ (‘variant type’)
specified in the corresponding rule with this ID.

See the section ‘Combinatorial rules’ under ‘Specifying complex AMR variants’ below.

Gene identifiers

The primary reference for a gene in AMRrules is its node ID in the NCBI Reference Gene Hierarchy. **If an appropriate node ID exists, this should be entered in both the ‘gene’ and ‘node ID’ fields and all other gene identifiers listed below are optional** (although it is strongly encouraged to also include NCBI protein or HMM accessions, and a CARD identifier, if possible, for ease of compatibility). **Each rule requires AT LEAST ONE OF node ID, protein, HMM or nucleotide accession.**

node ID

The NCBI Reference Gene Hierarchy has been selected over ‘allele’ or ‘gene family’ name as nomenclature standards are not consistent across AMR gene families; some entries lack ‘allele’ names; and node names clearly demarcate clusters of related variants with shared function for which it makes sense to specify a single rule. Most nodes in the gene hierarchy (~95%) have a one-to-one correspondence with a single ‘gene family’ name and/or ‘allele’ name in refgene, which users will recognise as common names in the literature.

NCBI AMRFinderPlus uses the gene hierarchy but does not by default include the node name in the output file; we recommend users turn this on by setting the parameter `--print_node` when running the tool. However files generated without this switched on can trivially be annotated with node by using the sequence accession to look up the node name, for subsequent rule interpretation.

If a rule applies to all alleles of a given gene (e.g. all *bla*SHV alleles are expected to confer resistance to penicillins, although only some are ESBL), the rule should be specified using the

corresponding node in the Reference Gene Hierarchy (e.g. node: 'blaSHV' is the parent for all numbered *bla*SHV alleles in refgene, hence this is used in the *Klebsiella pneumoniae* specification for wildtype resistance to penicillins in the example template). If no suitable parent node exists, a rule can be specified using * as a wildcard, e.g. 'blaSHV*'.

protein accession

Refseq or GenBank protein accession. Where possible this should match that used in the AMR gene catalog refgene.

HMM accession

HMM accession (for an internal node). This should typically match that used in the NCBI Reference Gene Hierarchy and linked in the AMR gene catalog refgene.

nucleotide accession

RefSeq or GenBank nucleotide accession. This field applies to rRNA genes or promoter variants and is defined by the nucleotide accession and the coordinates defining the gene to which the rules applies (example 'NZ_CP041538.1:1149245-1149489'). Where possible this should match that used in the AMR gene catalog refgene.

ARO accession

CARD Antibiotic Resistance Ontology (ARO) identifier for the gene this rule applies to. Facilitates interoperability with tools that utilise ARO as the primary database rather than refgene. May become essential if ARO develops a gene-drug dictionary that AMRrules would benefit from utilizing. Note there will not always be an ARO for every allele and node in refgene. In this case a representative may be chosen, or the field left blank.

Gene context

Indicates the genomic context of the gene within the specified organism, i.e. whether the gene is 'core' or 'acquired'. Note that a resistance-associated mutation in a core gene (e.g. Ser83Phe in chromosomal GyrA) should be coded as 'core'. A mutation in an acquired gene should be coded as 'acquired'.

The working definition of 'core' gene is present (>90% identity, >90% length) in (i) the chromosome of >95% of genomes of this species (defined as above using NCBI taxonomy); and (ii) the chromosome at >95% isolates that have wildtype AST profiles.

Future considerations:

- When assessing the frequency of a gene within a species, care must be taken with (i) taxonomic definitions (e.g. be aware of species vs species complex etc); and (ii) intraspecies diversity and representativeness of the available genome set.
- E.g. for *Acinetobacter baumannii*, genome databases are dominated by GC1 and GC2 clonal complexes which are highly resistant, such that genes which are core to those clones but lacking from other members of the species may have >95% frequency in the genome database. Rules will need to be developed around this, e.g. dereplicate to

N<100 genomes per lineage (defined as sequence type or mash-distance cluster or similar), with at least N>20 unique lineages represented.

Mutation

Indicate the mutation relative to the gene in 'gene', using the HGVS variant nomenclature and AMRrules syntax guidance.

Typically this will be a protein mutation, for which the correct format is 'p.Ser83Tyr'. Note this differs from the format used in refgene, which would express the same mutation as e.g. '[gene]_S83Y'.
If the rule concerns presence/absence of a gene, the ‘mutation’ column should be ‘-’.

More complex examples are given in the section ‘Specifying complex AMR variants’ and the AMRrules syntax guidance.

Variation type

Indicate the type of variation this rule applies to. Allowed values and their definitions are given in the 'variation type' tab in the AMRrules specification template (v0.6). Most common examples are 'Gene presence detected', 'Protein variant detected', 'Nucleotide variant detected'.

More complex examples of how these terms, combined with mutation syntax, can be used to describe complex AMR variants (including inactivating/truncating mutations, copy number variation, etc) are given in the section ‘Specifying complex AMR variants’ and the AMRrules syntax guidance.

Drug or drug class

Specify EITHER name of the individual drug/s that the rule applies to (in the ‘drug’ field), or the name of the drug class (in the ‘drug class’ field, only if there is evidence that the rule applies to all members of the class).

The reference for drug names and classes is the Antibiotic Resistance Ontology (ARO), developed and hosted by the Comprehensive Antibiotic Resistance Database (CARD). The rule sets should prioritize rules for drugs of clinical relevance to treatment of the species and with EUCAST clinical breakpoints, rather than being exhaustive.

Where a rule applies to multiple drugs, they should be specified in separate rows, with individual references for each gene-drug combination. The exception is if the rule applies to all members of a drug class, in which case it should be defined in a single row in which ‘drug’ is left blank and ‘drug class’ field is completed instead.

Note for beta-lactamases, rules may be specified for individual drugs (preferred; one drug per row, with evidence for each drug) or to the following classes defined in the ARO:

- First-generation cephalosporins
- Second-generation cephalosporins

- Third-generation cephalosporins
- Fourth-generation cephalosporins
- Other cephalosporins and penems

This maps roughly to the current refgene subclasses (which includes groups: beta-lactam, carbapenem, cephalosporin; plus some individual drugs, currently ceftaroline, cefiderocol, inhibitor-combinations).

Future considerations re beta-lactamases:

- Recommend focusing on specifying rules for individual drugs but prioritizing those for which EUCAST specifies breakpoints and that are on the essential medicines list. For Enterobacterales, this would mean 23 drugs in 10 classes, see summary. One might propose specifying rules for these 10 classes in future.
- The solution proposed above is to allow the specification of rules for the five groups of drugs that are defined in ARO and map broadly to different enzyme activities, meaning it will often be logical to specify a single rule that links a genetic determinant to the drug group. Resistance to drugs outside these groups, and inhibitor combinations, still need to be specified individually.

Phenotype

Specify either ‘wildtype’ or ‘nonwildtype’, to indicate whether members of this species with this gene are expected to fall in the wildtype or nonwildtype part of the MIC distribution. This is equivalent to identifying whether the MIC is expected to fall below or above the ECOFF, if one is defined. If the gene is a core gene, the expected phenotype should generally be 'wildtype', unless the rule refers to a specific variant of the core gene for which there is evidence of a nonwildtype phenotype.

Clinical category

Indicate the categorization (S/I/R) associated with this gene, for members of this species, using the breakpoint standard indicated. If the drug this rule applies to appears on the EUCAST Expected Resistances list for this organism, and the gene is a core gene, the expected phenotype should be 'wildtype' and the category should be ‘R’. If the gene is identified as a core gene but the drug does not appear on the Expected Resistances list for this organism, there should be strong evidence from literature and/or matched genome/phenotype data to support the assignment of ‘R’.

Breakpoint

The breakpoint (preferably MIC) that was used to define the expected phenotype category. For categorization as 'R', breakpoint should be given in the form 'MIC >X [units]'; for categorization as 'S', use 'MIC <=X [units]'; for categorization as 'I', use 'I >X [units]'. This is to provide clarity as to the definition used, as breakpoints can differ between standards and change over time.

If the rule is defined on the basis of an ECOFF, indicate the threshold used in the same manner as for a breakpoint.

If it is an Expected (intrinsic) resistance, the breakpoint is irrelevant (and usually undefined) so enter 'not applicable'.

Breakpoint standard

The AST phenotyping standard used to interpret this rule. In the format '[Name] [version] ([year])', e.g. 'EUCAST v14.0 (2024)' or 'EUCAST ECOFF (January 2024)' (as EUCAST ECOFFs are not versioned, indicate month and year). This is to facilitate including different interpretations using different AST standards, such as EUCAST, CLSI, veterinary standards, ECOFFS, etc and for clarity as to the definition used as breakpoints can change over time.

If it is an Expected (intrinsic) resistance, there will not typically be a breakpoint, in this case indicate the version of the expert rules e.g. 'Expected resistant phenotypes v 1.2 (13 January, 2023)'.

Breakpoint condition

This indicates the specific conditions for this breakpoint, if it is relevant. For example, in EUCAST, cefuroxime for *E. coli* has breakpoints for iv and oral administration.

PMID

PubMed identifier for the 'best' peer-reviewed research article that provides evidence that this gene is associated with this phenotype category in this species.

Where a rule is made based on new unpublished data analysis, the supporting evidence should be submitted for review by the working group, and preprinted or published as soon as possible.

Future considerations:

- Would be useful to establish evidence codes, or confidence levels, to be associated with rule sets. This would support a range of use cases, including allowing filtering to high-confidence rules where desired, and distinguishing high vs low confidence in downstream reports or inferred antibiograms.

Evidence code

Indicate the nature of the evidence supporting the rule, using standardised terms from the Evidence & Inclusion Ontology (ECO). More than one can be listed, and curators should include all forms of evidence available to support the rule. In principle any ECO codes can be used, but in most cases it will be most appropriate to choose from the subset listed in the 'evidence codes' tab of the AMRrules template (also provided as a dropdown selection in the main data entry tab of this spreadsheet). The source for each type of evidence should be given in the 'PMID' field.

Currently suggested ECO terms:

- ECO:0001091 knockout phenotypic evidence
- ECO:0000012 functional complementation evidence
- ECO:0001113 point mutation phenotypic evidence

- ECO:0000024 protein-binding evidence
- ECO:0001034 crystallography evidence
- ECO:0000005 enzymatic activity assay evidence
- ECO:0000042 gain-of-function mutant phenotypic evidence
- ECO:0007000 high throughput mutant phenotypic evidence
- ECO:0001103 natural variation mutant evidence
- ECO:0005027 genetic transformation evidence
- ECO:0000020 protein inhibition evidence
- ECO:0006404 experimentally evolved mutant phenotypic evidence
- ECO:0000054 double mutant phenotype evidence

Evidence grade

Indicate the expert curators' overall assessment of the level of support provided by all evidence considered:

- 'High' indicates the curators are confident in the categorisation, and believe that the likelihood that the effect will be substantially different from this is low.. This grade should be used when experimental evidence provides strong support for the interpretation of this gene/variant in this species for this drug. If there is statistical geno/pheno evidence available, it supports this interpretation.
- 'Moderate' indicates the curators believe that the categorisation most likely reflects the true effect, and the likelihood that the effect will be substantially different is moderate.. There is good evidence to support the interpretation of this gene/variant in this species for this drug, but there is some uncertainty (e.g. lack of direct evidence in this organism although evidence from related organisms is convincing; or there is good statistical geno/pheno evidence but no experimental evidence of mechanism).'Low' indicates the curators believe that the categorisation might not reflect the true effect, and the likelihood that the effect will be substantially different is high. This grade should be used when there is evidence supporting a link between this gene/variant and this drug, but the interpretation in this species is unclear (e.g. lack of evidence in this organism or related organisms; statistical geno/pheno evidence is lacking, or does not support a clear effect; or there are trustworthy but conflicting reports).
- 'Very low' indicates the curators have no confidence that the categorisation reflects the true effect, and the likelihood that the effect will be substantially different is high. There is no trustworthy evidence as to the effect in this organism, or there is conflicting evidence. The categorical interpretation is based on assumptions made from unrelated organisms and may be wrong.

Evidence limitations

Indicate the expert curators' assessment of the key limitation(s) of the available evidence:

- lacks evidence for this species
- lacks evidence for this genus
- lacks evidence for this allele
- lacks evidence of the degree to which MIC is affected
- low clinical relevance

- unknown clinical relevance
- statistical geno/pheno evidence but no experimental evidence
- conflicting evidence
- lacks formal breakpoints

Rule curation note

A short explanatory note describing the mechanism and/or reasoning for the rule. Note this is intended to document the justification for setting the rule, not necessarily something to propagate into annotation of downstream genotype reports or genome reports.

Proposed - not yet implemented: Explanatory note

Several people have raised the need for a more clinically or epidemiologically informative ‘note’ field that would be suitable to propagate into a downstream genome report designed for clinical or public health knowledge users, e.g. flagging clinical considerations such as when the presence of an intrinsic beta-lactamase does not usually imply carbapenem resistance, but clinicians should be made aware that upregulation or copy number increases can occur that would result in carbapenem resistance. (Note the ‘rule curation note’ field is not intended for this purpose, but is rather intended to document the reasoning behind the rule.)

Specifying complex AMR variants

The ESGEM-AMR Data & Tools group reviewed examples of complex variants submitted by Working Group members on 5 September 2024, and developed the following guidance and examples illustrating how these variants can be specified using the AMRrules fields ‘gene’, ‘mutation’ and ‘variation type’. This information is also included in the syntax guidance in GitHub.

Syntax for ‘mutation’ column

This follows HVGS, including

- Gene and protein start sites are position 1 (there is no position 0)
- Ranges are specified using x_y; for insertions the coordinates are specified as inclusive_exclusive, otherwise ranges are inclusive_inclusive
- Unknown ranges are specified with parentheses, (x_y). E.g. p.(1_100)insGlyAsp means an insertion of 2 amino acids (Gly and Asp) anywhere between codons 1 and 100 inclusive (as opposed to a replacement of amino acids 1 through 100 with GlyAsp, which would be expressed as p.1_100delinsGlyAsp).
- Coordinates are specified relative to the reference sequence of a protein (p) or coding sequence (c)
    - Coordinates upstream of coding sequence are specified relative to the start site, with a hyphen, e.g. ‘c.-35’ indicates 35 bp upstream of start codon
- Mutations in protein and DNA are specified differently, e.g.
    - p.Ser83Tyr: change to protein sequence from Ser to Tyr at codon 83
    - c.25C>T: change to nucleotide coding region from C to T at nucleotide position 25
- Stop codons are specified (in both DNA and protein variants) as *
- Following IUPAC, X signifies any amino acid, N signifies any DNA base
- ^ (caret) is used as "or", e.g. p.(Gly719Ala^Ser)
- The letters ‘inv’ indicate the inverse (i.e. reverse complement) of a sequence
- Repeat sequences are specified as sequence[N] where N is the number of copies of the repeat

Syntax specific to AMRrules

- AMRrules requires amino acids be specified as three-letter codes (whereas HGVS allows single-letter or three-letter codes)
- In HGVS you must specify the reference sequence explicitly using a sequence accession, followed by ‘:’ and then the mutation, e.g. ‘NF000285.3:p.Gly238Ser’. In AMRrules the gene is specified in separate column/s (‘gene’, ‘refseq accession’, ‘ARO accession’) and should not be repeated in the mutation column. So the above rule should be coded as:
    - gene = blaSHV
    - refseq accession = NF000285.3
    - ARO accession = ARO:3000015
    - mutation = p.Gly238Ser

- In AMRrules, insertion sequences (IS) should be labeled with their IS name as per ISfinder, as many do not have their own sequence accessions in refseq. E.g. insertion of ISAba125 should be specified as 'ins[ISAba125]', and insertion in reverse orientation to the gene to which the rule applies should be specified as 'ins[ISAba125:inv]'.
- In AMRrules, rules intended to apply when a gene is present in a minimum of N copies can be specified using the [N] syntax to indicate the minimum repeat/copy number of the whole coding sequence, as c.[N].
    - Note this syntax does not convey any information about the location of the copies, i.e. c.[2] simply indicates that there are at least 2 copies of the gene detected in the genome, whether they are tandem repeats or in different replicons such as one in the chromosome and one in a plasmid.
- In HGVS, the presence of multiple alleles (i.e. heterozygous) is specified as a colon-separated list of allelic variants e.g. '[allele1];[allele2].
    - In AMRrules, rules that apply to variation in a multi-copy gene can be specified in this way, with each allele explicitly stated.
    - Alternatively if the rule applies when a minimum of N copies of the gene carry the mutation (e.g. mutation in ≥3 copies of 23S rRNA resulting in resistance to azithromycin), this can be abbreviated using the [N] syntax to indicate the minimum repeat/copy number, as c.[allele][N] or p.[allele][N], e.g. 'c.[2045A>G][3]'.
- In AMRrules, rules that apply to 'low frequency variants', i.e. when a minimum fraction of reads, X, support presence of the allelic variant in a sequenced population, the minimum fraction can be specified by extension of the syntax for copy number, as [X]. E.g. 'p.[Ala94Gly][0.13]' (example from the *Mycobacterium tuberculosis gyrA* gene).
    - To put another way, in AMRrules the repeat syntax [X] is interpreted as a minimum copy number if X is an integer, and as a minimum read fraction if X is a double/float between 0 and 1.

Examples of complex variants

ID	gene	mutation	variation type	drug	category
KPN0001	blaSHV	-	Gene presence detected	ampicillin	wt R
KPN0008	gyrA	p.Ser83Tyr	Protein variant detected	ciprofloxacin	nwt I
KPN0009	parC	p.Ser80Ile	Protein variant detected	ciprofloxacin	nwt I
KPN0010	ompK36	c.25C>T	Nucleotide variant detected	meropenem	nwt S
KPN0011	ompK36	p.114_115insGlyAsp	Protein variant detected	meropenem	nwt I
KPN0012	mgrB	p.(1_100)	Gene truncation detected	colistin	nwt R
KPN0013	qnr	-	Gene presence detected	ciprofloxacin	nwt I
NGO0001	mtrR	-	Inactivating mutation detected	macrolides	nwt R
KPN0014	mgrB	p.Glu30*	Protein variant detected	colistin	nwt R
ECO0001	ampC	c.-11C>T	Promoter variant detected	ceftriaxone	nwt R
ECO0002	ampC	c.-14_-13insGT	Promoter variant detected	ceftriaxone	nwt R
ACI0001	blaOXA-58	c.(-35_1)ins[ISAba125:inv]	Promoter variant detected	ceftriaxone	nwt R
NGO0002	23S rDNA	c.[2045A>G][3]	Nucleotide variant detected in multi-copy gene	azithromycin	nwt R
ECO0003	blaTEM	c.[3]	Gene copy number variant detected	piperacillin+tazobactam	nwt R
MTC0001	gyrA	p.[Ala94Gly][0.13]	Low frequency variant detected	ciprofloxacin	nwtR

- **p.Ser83Tyr**: change to protein sequence from Ser to Tyr at codon 83
- **c.25C>T**: change to nucleotide coding region from C to T at nucleotide position 25
- **p.114_115insGlyAsp**: change to protein sequence, with an insertion of amino acids Gly and Asp between codons 114 and 115
- **p.(1_100)**: truncation (of any kind) anywhere in the first 100 amino acids of the protein sequence
- **c.-11C>T**: change to nucleotide sequence from C to T, 11 bases upstream of the start site for the gene.
- **c.-14_-13insGT**: insertion of nucleotides GT between positions -14 and -13, upstream of the start site of the gene
- **c.(-35_1)ins[ISAba125:inv]** insertion of ISAba125, in reverse orientation (:inv), anywhere between 35 bases upstream of the start site, and the start of the gene coding sequence
- **c.[2045A>G][3]**: substitution of A to G at position 2045 of the gene. This mutation must occur in minimum 3 copies
- **c.[3]**: gene needs to be present with a minimum of 3 copies
- **p.[Ala94Gly][0.13]** protein variant is present in >13% of reads

Specifying combinatorial rules

Combinatorial rules are defined using logical expressions in the 'gene' column, where the objects of the expression are rule identifiers ('ruleID') that can be used as shorthand labels for the variants defined by 'gene':'mutation' ('variant type') specified in the corresponding rules. The 'variation type' should be specified as 'Combination'.

- Each rule must have a unique 'ruleID', assigned by the curating subgroup and prefixed with a 3-letter code that identifies the subgroup.
- E.g. in the table below, 'KPN0008' can be used in a logical expression in the 'gene' column to demarcate 'gyrA:p.Ser83Tyr', and KPN0013 can be used to demarcate 'qnr (Gene presence detected)'.
- So, the combination of these two variants can be specified as 'KPN0008 & KPN0013', which expands to 'gyrA:p.Ser83Tyr & qnr (Gene presence detected)'.

Rules must be specified explicitly if the effect of the combination is NOT the same as the 'most resistant' (in terms of exceeding breakpoints, R > I > S; or deviation from wt, nwt > wt) predicted category of the component rules. E.g. in the table below:

- The individual rules KPN0008 and KPN0009 solo each have expected category 'nwt I', but in combination we expect 'nwt R', so we need to specify the rule for the combination 'KPN0008 & KPN0009'.
- The expected category for genomes meeting rule KPN0002 (i.e. carrying core gene oqxA, => wt S) in addition to rule KPN0008 (i.e. with an acquired gyrA mutation, => nwt I) is nwt I. This is the same, not greater, than one of the component rules (KPN0008) so we do not need to specify the combination explicitly.

Note this means the combination must be specified explicitly if the combined effect is LESS resistant than the 'most resistant' component, e.g. in this example from TB, deletion in one gene renders the resistance mutation in another gene irrelevant so the combination must be specified.

See further discussion below under 'Open issues -> Combinatorial rules -> General principles'.

ruleID	gene	mutation	variation type	drug	category
KPN0002	oqxA	-	Gene presence detected	ciprofloxacin	wt S
KPN0008	gyrA	p.Ser83Tyr	Protein variant detected	ciprofloxacin	nwt I
KPN0009	parC	p.Ser80Ile	Protein variant detected	ciprofloxacin	nwt I
KPN0013	qnr	-	Gene presence detected	ciprofloxacin	nwt I
KPN0051	KPN0008 & KPN0009	-	Combination	ciprofloxacin	nwt R
KPN0052	(KPN0008 | KPN0009) & KPN0013	-	Combination	ciprofloxacin	nwt R

Rule set development

Drafting of rule sets

It is recommended that a copy of the rule template be created for each organism, to be shared by subgroup members who are working together on that specific organism. The mode of collaboration between members assigned to the same organism is left to the individual members, but we suggest creating a directory in the AMRrulesCuration GitHub repository for each organism to share files.

Key literature supporting each rule should be recorded in the 'PMID' field and a short description of the logic for each rule added to the 'note' field of the rule template. In addition, evidence for rules should be documented in a short narrative (e.g. in a Google doc), explaining the logic followed and any data analysis undertaken to develop and test the rules.

Members may wish to consider publishing this narrative, either stand-alone or as part of umbrella article/s describing the project; this will be discussed at the working group progress meetings.

Automated rule validation

Before submitting rule sets for review, the subgroup members should validate their rules by running the AMRrulevalidator Python package developed by Jane Hawkey. The script will validate the rules file and will provide a summary of checks that passed or failed.

Review of rule sets

Once a rule set is ready for review, the file (in TSV format) should be added to the /draftrules directory in the AMRrulesCuration GitHub repository. The working group Chair/s will review them for format and content (including the linked supporting evidence), and the group's Lead Bioinformatician will review and test using available genome data.

Release of approved rule sets

Once reviewed and approved, new rule sets will be added to the master rule set in the main branch of the AMRrulesCuration github repository and copied to the main AMRrules Python package, and a versioned release will be made that includes the new rules together with release notes recording the addition of the organism and the names of all contributors.

NOTE: As we aim for AMR rules to be widely adopted, it may be necessary to add steps to the review and release cycle to meet the needs of stakeholder organizations such as CARD or EUCAST.

Initial Protocol: Wildtype phenotypes

This protocol should be considered a starting point for the working group, and will be developed and refined as the group attempts to populate rules.

This protocol focuses on wildtype phenotypes, as the first priority of the working group. Where sufficient data is available to propose rules for acquired resistance, these may be recorded. However such rules will not be included in the first release of the AMR rules, as inclusion criteria and standards of evidence need to be defined first.

Wildtype resistant

1. Add a row for each drug on the EUCAST 'Expected resistance' list for the organism

- Make sure to use the right (generic, not brand) name for the drug, this should be the name of the drug as it appears in the Expected Resistance list, but please check this matches exactly an entry in the ARO. If it does not, flag for review.

2. If you already know the gene/s responsible for a trait, then:

- Confirm whether the gene meets the definition of 'core', and complete the 'context' field.
- Locate the gene/alleles in refgene to find the corresponding node ID/s to enter in the 'gene' field.
- Review whether there is a suitable internal node in the reference gene hierarchy (e.g. for fosA in *Klebsiella pneumoniae*, the wildtype chromosomal alleles correspond to fosA5, fosA6, fosA9, fosA10 which descend from the common ancestor node fosA5_fam, therefore the correct node to specify this rule is 'fosA5_fam'). If it is not clear which alleles/nodes capture the right gene, flag that additional analysis will be needed to resolve this. If there is no single node to describe the rule, add separate rules for each node to which you think it applies (one row for each).
- Review the supporting literature to find those articles that provide the best evidence for this gene being responsible for wildtype (intrinsic) resistance to the drug in this organism.
- If you are sure there is sufficient evidence that this specific gene supports the interpretation of 'wt R' using current breakpoints, record this as the interpretation. If there is any uncertainty, flag this for review.
- Record the PubMed identifiers for the relevant citations in the PMID column. If possible try to add one citation per unique node, however if the same citation supports multiple alleles/nodes that is permissible.
- Write a short explanatory 'note' if needed to support the interpretation and explain any points of potential confusion.

3. If you don't know the core gene responsible:
i) Conduct some literature searches to check if this is known;
ii) Check the 'Existing Curations' resources below; and
iii) Review the output of AMRfinderplus on genomes from 'wildtype' organisms
(susceptible to all drugs besides those on 'Expected resistance' list) to try to identify 'intrinsic' determinants associated with each expected resistance.

- This depends on availability of genome data with matched antibiogram data.
    - **Public AST data** can be found via NCBI Pathogens AST Browser and BV-BRC.
    - **Consistently assembled public genomes** for a given species can be easily downloaded as a single tarball from the AllTheBacteria FTP site (described here). Note that their team is in the process of running AMRFinderplus on all genomes, check the github for progress which may save you running AMRFinderplus yourself.
    - Note that you should not rely on the genotype calls in the NCBI Pathogens browser, as these calls are currently not re-run with new updates of AMRFinderplus, so the results include calls from a mix of versions and differences between genomes may reflect differences in the version that was run rather than differences in the actual genotypes.
    - If downloading assemblies from other sources, species should be confirmed by reference to the NCBI Taxonomy.
    - Consideration should be given to the quality and source of genome data and AST data, and the diversity of the genome data.
- A suggested rule of thumb to define a gene as a core gene candidate for expected resistance is n=50 genomes with wildtype antibiogram, including >20 unique lineages (defined by unique MLST lineage, or clustering based on mash distance). Ideally the isolates should also be diverse in terms of geography and source. Specific criteria will be developed by the working group as we go, please just try to maximize the diversity of data you include and be sure to record details of these parameters for review and discussion later.
- MICs based on microbroth dilution are the preferred source of AST data. If you need to combine multiple sources of data from different laboratories, instruments or methods take care to compare the MIC distributions and check for batch effects.
- If you don't have enough genomes with known wildtype antibiograms, it may be useful to explore all genomes (regardless of antibiogram availability or result) in order to identify core genes that may explain expected resistance.
- AMRfinderplus should be run afresh on the genome assemblies, using the latest version and parameters (including the organism option, if this organism is supported by AMRFinderPlus).
  Example command:
  ```
  amrfinder -n genomeA.fasta --plus --print_node --organism Acinetobacter_baumannii --name genomeA > genomeA_amr_results.txt
  ```
- (Note the values in the 'AMR genotypes' field in NCBI Pathogens online portal cannot be relied upon for our purposes, as it is not rerun each time the databases are updated, and so there can be inconsistencies in the outputs for different genomes that are added years apart.)
- Map the detected core determinants to the Expected resistances using refgene class/subclass and PMIDs (in the AMRfinderplus output) and your own literature searches.
- Complete the rules, one per gene/node, as outlined above.

4. If you still have unexplained Expected resistances, i.e. where you can't identify a core gene responsible in the AMRfinderplus, some options for further exploration of potential mechanisms are:

- ResFam search of wildtype genomes
- CARD RGI search of wildtype genomes

Wildtype susceptible

1. Review the output of AMRfinderplus on a set of organisms with 'wildtype' antibiograms (i.e. no resistances beyond drugs with Expected resistance). See above for where to find genomes with matched AST data and to run AMRfinderplus.

2. If intrinsic determinants are reported by AMRfinder that don't map to an Expected resistance, these may require a 'wt S' rule. It is unlikely they reflect true 'wt R' that is not included in the Expected resistance.

- First confirm the gene is core, across a diverse set of genomes, considering those with antibiograms as well as the broader population.
- For those with antibiograms, examine the MIC distribution for isolates whose genomes have no other known determinants for this drug. Compare this with the clinical breakpoints to confirm it's clear the wildtype phenotype is susceptible.
- Review the reference MIC/DD distributions at EUCAST. Is the organism's wildtype distribution shifted relative to close neighbours? I.e. Does it have a higher ECOFF? (e.g. fosfomycin wildtype distribution for *Klebsiella pneumoniae* is higher than for *E. coli* due to core gene *fosA*).

3. Review literature evidence for the gene's link to resistance and associated mechanism

- Is there any evidence for this specific organism, or is the evidence for resistance in other organisms only?
- Is its resistance linked to mobile elements in those other organisms?

4. If these steps identify no evidence of resistance, record the interpretation as wt S and add the relevant supporting PMIDs and notes.

Open issues

Which drugs

should be considered/included for a given organism?

- If Expected resistance is recorded for drug X, the rule set should aim to include and provide interpretation for the core gene/s responsible for this resistance.
- If a gene is associated with a large drug class, consideration should be given as to which drugs in that class are relevant to include individual rules for. The existence of EUCAST clinical breakpoints is a good indication of whether the drug is relevant to treatment. Priority should also be given to drugs recommended for global surveillance by WHO GLASS, and those on the WHO Global Essential Medicines List (or national lists and treatment guidance).

Defining resistance

Multiple breakpoints

If there are multiple breakpoints for the same drug-bug combination for different clinical conditions? E.g. for Enterobacterales, the ciprofloxacin R breakpoint is >0.125 for meningitis, and R >0.5 for indications other than meningitis.

- Specify a separate rule for each breakpoint, indicating the breakpoint in the 'breakpoint' field (e.g. 'MIC >0.125 mg/L') and the indication in the 'breakpoint condition' field (e.g. 'Non-meningitis').

No breakpoint?

- Consider any relevant notes in the EUCAST Breakpoints Table, consult EUCAST guidance documents, and check whether the organism has Expected Resistance.
- Note that if EUCAST Breakpoints Table has a 'dash' instead of numerical values it indicates the microbe can be reported resistant without further testing. The same is true if Expected Resistance is recorded for this drug in this organism. If either or both of these is true, it implies the interpretation should be 'wt R' IF there is clear evidence that the wildtype resistance is functionally attributable to this specific gene.
- If the above do not apply (ie there is no evidence of wildtype resistance) but there is an ECOFF, consider specifying a rule based on the ECOFF instead. Indicate the ECOFF in the 'breakpoint' field, and record 'ECOFF [date]' in the 'breakpoint_standard'. If a gene does not push the MIC distribution above the ECOFF the correct categorization is 'wt S'. If the gene pushes the MIC distribution above the ECOFF but there are no breakpoints, the correct categorization is 'nwt'; it cannot be 'nwt R' as there is no definition of R.
- In the absence of an ECOFF or any other guidance, it may be suitable to define an unofficial (internal-use) ECOFF based on available MIC distributions and interpret against this; this would need to be documented in the 'note' column and fully described

in the narrative description (including showing the distribution, and the source of the data).
- Record the lack of formal breakpoints in the 'Evidence limitations' field, as 'lacks formal breakpoints'.

MIC distribution vs breakpoint

What to do if a genetic determinant shifts an MIC distribution upwards, but not so high that the tail of the distribution exceeds a breakpoint or ECOFF? What if the breakpoint cuts the MIC distribution in half? What fraction of the distribution needs to exceed an 'R' breakpoint to interpret the gene as 'nwt R'?

Note this generally relates to the effect of **acquired** resistance determinants, as opposed to resolving the effect of core genes and understanding wildtype phenotypes, which is Phase 1 and the priority for the ESGEM-AMR working group. Resolving these issues conclusively will require large amounts of high-quality AST data with matched genome data, as part of Phase 2.

There are inherent difficulties in creating rules based on the effect of a gene on MIC, as what we observe is a distribution of MIC values not a single value. Sources of variation include natural biological variation, and technical variation in measuring the MIC. Notably, MICs are measured via doubling dilutions, they are not measured as continuous variables, so the underlying distribution is not directly observed. Ideally, MIC distributions would be bimodal with two peaks corresponding to wildtype susceptible (responsive to drug treatment) and non-wildtype resistant (likely to fail treatment even on high exposure), and breakpoints would be set such that they fall in clear troughs between those two peaks. However the reality is not always so clear, particularly as we are observing the distribution of discretised measures of MIC rather than the underlying distribution.

In principle, a rule defining the expected impact of a gene X on a MIC category for drug Y (e.g. gene X => nwt R) could be defined in a number of ways, with respect to how the breakpoint relates to the observed MIC distribution (for isolates that carry gene X but no other known acquired determinants for drug Y). Possible rules could be:

1. Any of the distribution exceeds the breakpoint;
2. The expected value exceeds the breakpoint (in the statistical sense of the expected value of a random variable, which equals the mean value), or
3. The majority (or entirety) of the distribution exceeds the breakpoint.

In practice, requiring the entirety of the distribution exceed a breakpoint is impractical as there is likely to be a lower tail of the observed MIC distribution that falls below the breakpoint due to technical artefacts (e.g. assay variability, errors in the sequence data, or changes during culture steps between AST and DNA extraction for sequencing). Basing a rule on the upper extreme of the distribution is also problematic as this too could be driven by technical artefacts, or by rare unknown resistance determinants.

Requiring a fraction of the distribution to exceed the breakpoint avoids the issue of extreme values in the tails of the MIC distribution. The question then becomes what fraction. A required fraction of half or greater is equivalent to requiring that the expected value exceeds the breakpoint. This is attractive as it is simple to define and does not require parametrization of the distribution, which is problematic given the non-continuous nature of MIC data. A pragmatic solution is therefore to require only half the observed values to exceed the breakpoint in order to define the phenotype category, but to record details of the observed fraction as an indicator of confidence in the categorization. In practice, it is expected that investigative analyses of real, large, MIC distributions for multiple drugs and bugs in Phase 2 will help bring clarity to this issue and facilitate the definition of clear guidance and confidence metrics for interpretation of acquired genes.

- The proposed rule of thumb is to set rules based on the expected value of the MIC associated with the gene (ie >half the distribution exceeds the breakpoint).
    - If the median MIC of isolates carrying gene X and no other known determinants of Y exceeds the R breakpoint, the interpretation is ‘nwt R’.
    - If the median MIC exceeds the S breakpoint but not the R breakpoint, then the categorization is ‘nwt I’.
    - If the median MIC exceeds the ECOFF but not the S breakpoint, the categorization is ‘nwt S’.
    - If the median MIC falls below they ECOFF the categorization is ‘wt S’
- The ‘note’ field can be used to record when the rule is borderline, e.g. if the fraction of the distribution exceeding the breakpoint is less than 90%.

Combinatorial rules

How to capture combinatorial rules, i.e. where a combination of determinants is required to achieve a phenotype?

Notably, this mainly relates to categorizing the effect of **acquired** resistance determinants (as opposed to resolving the effect of **core** genes and understanding wildtype phenotypes, which is the first priority for the working group) and will require large amounts of high-quality AST data with matched genome data to properly define.

General principles

- The additive assumption should hold unless specified otherwise; i.e.
    - If there is a rule ‘gene1 => nwt R for drug X’ this is assumed to hold regardless of other rules specified for drug X, or for any other drug.
    - It is assumed that resistance to combined drug X+Y requires resistance to drug X and drug Y. If there are rules ‘gene1 => nwt R for drug X’ and ‘gene2 => nwt R for drug Y’, resistance to combined drug X+Y is assumed.

Combination drugs

- For simple drug combinations such as trimethoprim-sulfamethoxazole, the assumption is that resistance requires a determinant conferring resistance to each component (e.g. a

determinant conferring trimethoprim resistance plus a determinant conferring sulfamethoxazole resistance). Therefore, separate rules should be provided for each component drug, and inference of resistance to the combination can be trivially made by checking for resistance to both component drugs.

- If a gene were identified that conferred resistance to the combination drug, AND each of its constituent components individually, this could be handled by indicating two rules, one for each component drug (e.g. gene => trimethoprim R and one for gene => sulfamethoxazole R). Again, resistance to the combination can be trivially inferred from resistance to both component drugs.
- If a gene were identified that conferred resistance to the combination drug and JUST ONE OF its constituent components, this could be handled by indicating two rules, one for gene => combination and one for gene => [component drug].

Drug + inhibitor combinations

- In some cases, resistance is conferred by a single gene, e.g. a beta-lactamase that evades inhibition to hydrolyze the drug in the presence of drug+inhibitor. The correct way to specify this is a single rule specifying ‘gene => drug+inhibitor nwt R’.
- If separate resistance determinants are known for the drug and the inhibitor, the assumption is that these work additively such that presence of both confers resistance to the combination. In this case each ‘gene => drug’ or ‘gene => inhibitor’ rule should be specified separately.
- Sometimes separate resistance determinants are known for the drug and the inhibitor, but these do not always work additively, e.g.:
    - ‘gene1 => drug nwt R’
    - ‘gene2 => drug nwt R’
    - ‘gene3 => inhibitor nwt R’
    - ‘gene1+gene3 => drug+inhibitor wt S’
    - ‘gene2+gene3 => drug+inhibitor nwt R’.
- In this case, the individual rules should be specified along with any **exceptions** to the additive assumption, i.e.:
    - ‘gene1+gene3 => drug+inhibitor wt S’ should be specified, as it differs from the additive effect assumed from the individual gene1 and gene3 rules (which would be ‘drug+inhibitor wt R’)
    - It is not essential to specify a rule that ‘gene2+gene3 => drug+inhibitor nwt R’ as this is the assumption from the rules ‘gene2 => drug nwt R’ and ‘gene3 => inhibitor nwt R’, although it may be helpful to make this explicit and provide evidence.

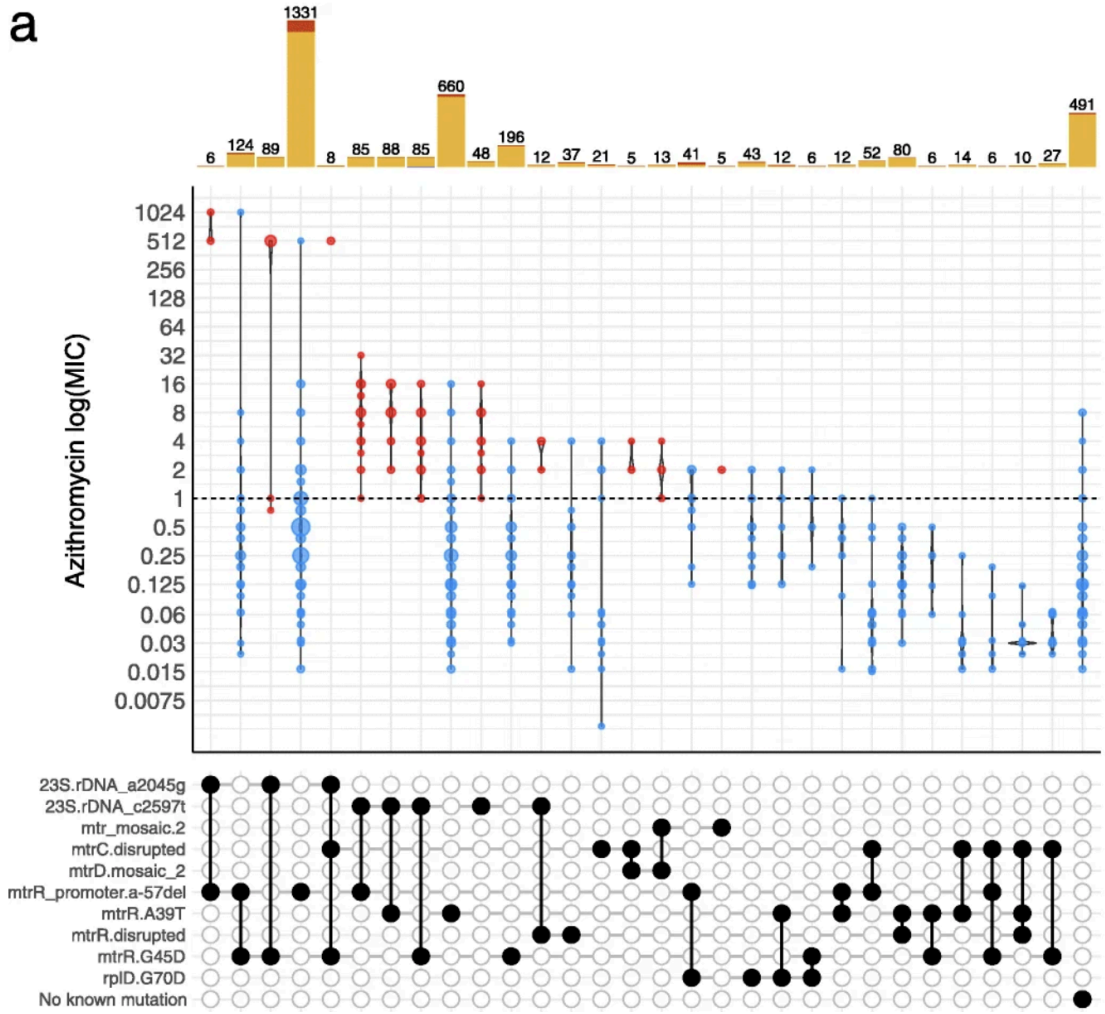
Single drug, multiple determinants

Some genes have a small or no effect on their own, but in combination can shift an MIC substantially.

- If the combined effects are simply additive (e.g. ‘gene1 => wt S’, ‘gene2 => nwt R’, ‘gene1+gene2 => nwt R’), and the ‘most resistant’ category applies, then there is no need to specify a combinatorial rule.

- If the combined effect is not additive (‘gene1 => wt S’, ‘gene2 => wt S’, ‘gene1+gene2 => nwt R’) then a separate rule should be given for the combination.
- Note that to minimise the need for specifying large numbers of combinatorial rules, curators should attempt to identify an internal node in the gene hierarchy that captures the rule accurately, rather than specifying each individual allele. E.g. a single rule ‘GyrA_S83F + qnr => ciprofloxacin nwt R’ is preferable to multiple rules for ‘GyrA_S83F + qnrS1’, ‘GyrA_S83F + qnrB1’, ‘GyrA_S83F + qnrB2’ (assuming of course this specification is accurate).

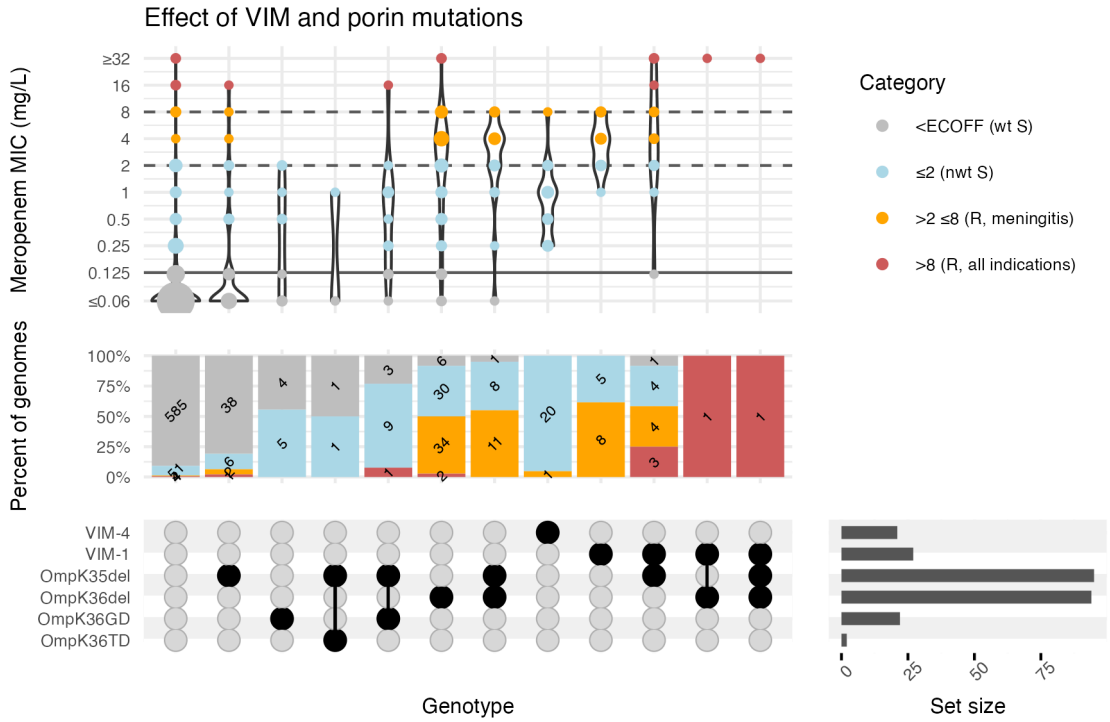
Example 1: resistance in *N. gonorrhoeae* is mostly mutational, and combinations are often required and need to be specified individually, see e.g. Figure 3 here (reproduced below):



Example 2: *van* operon, which includes several genes, including multiple that are required for vancomycin resistance in e.g. *Enterococcus faecium* and multiple that aren’t. Therefore presence of a single gene is not sufficient to define a rule set, and rules would need to specify

the combination/s that are known to reliably result in resistance (and potentially specify those that are known to retain susceptibility, for clarity).

Example 3: combination of carbapenemase enzymes with porin mutations.
Some carbapenemases (e.g. VIM and OXA variants) raise meropenem MIC in *Klebsiella pneumoniae* above the ECOFF (0.125 mg/L), but not above the S breakpoint (2 mg/L). In combination with porin mutations the MIC distributions shift further upwards, sometimes above the S/R breakpoint for meningitis (R >2) and sometimes above the R breakpoint for other indications (R >8). See figure below, which is produced from the meropenem MIC data from EUSCAPE mapped to carbapenemase and porin mutations called by Kleborate in the corresponding genomes.



- In this case, following the guidance above and based on the most common effect shown in the plot, we would need to specify:
  - ompK35-deletion => ***wt S***
  - ompK36GD => above ECOFF but below S => ***nwt S***
  - ompK36-deletion => above S (2) but below R (8)
    => ***nwt R*** for meningitis, ***nwt I*** for other indications
  - VIM-4 => above ECOFF but below S => ***nwt S***
  - VIM-1 => above S (2) but below R (8)
    => ***nwt R*** for meningitis, ***nwt I*** for other indications
  - VIM-1 + ompK36-deletion => above R (8)
    => ***nwt R*** for all indications
  - Combinations with ompK35-deletion + ompK36-deletion don't need to be specified because they obey the additive assumption

    - Some combinations are not observed so cannot be defined from this data
- Note that carbapenemases that raise MIC >8 in the absence of any porin mutations (such as KPC, NDM, IMP) only require a single rule to assign 'wt R' for each indication.

Existing Curations

AbritAMR reporting logic

Copy of refgenes database with 'enhanced_class' and 'enhanced_subclass' appended: https://github.com/MDU-PHL/abritamr/blob/master/refgenes_2024-02-23.csv

See Supplementary Information in the AbritAMR Paper for a description of the logic behind the 'enhanced subclass' column, and also description of the 'reporting logic' which includes:

1. Beta-lactamase classes that are assigned 'not reportable' for particular organisms as they are considered intrinsic resistance:
- MBL carbapenemase blaL1 or blaL2 (not reportable in *Stenotrophomonas maltophila*)
- OXA-51 family carbapenemase (not reportable in *Acinetobacter baumannii, calcoaceticus, nosocomialis, pittii*)
- AmpC beta-lactamase (not reportable in *E. coli/Shigella*)

2. Certain drugs/genes that are assigned 'reportable' for particular organisms only:
- Oxazolidinone & phenicol resistance (reportable only in *Enterococcus, Staphylococcus aureus, Staphylococcus argenteus*)
- Vancomycin resistance (only *van* A, B, C, D, E, G, L, M reportable)
- Methicillin resistance (*mecI* or *mecR* not reportable)

ResFinder phenotype resources

ResFinder has undertaken curation of genes and mutations to drugs, for selected organisms.

List of in-silico panels used in ResFinder (list of drugs covered per organism):
https://bitbucket.org/genomicepidemiology/resfinder_db/src/master/phenotype_panels.txt

- Campylobacter
- Campylobacter jejuni
- Campylobacter coli
- Enterococcus faecalis
- Enterococcus faecium
- Escherichia coli
- Mycobacterium tuberculosis
- Salmonella
- Salmonella enterica

- Staphylococcus aureus

Resfinder Mapping of drug to class:
https://bitbucket.org/genomicepidemiology/resfinder_db/src/master/antibiotic_classes.txt
Resfinder Phenotype panels (mapping gene to drug, not organism specific):
https://bitbucket.org/genomicepidemiology/resfinder_db/src/master/phenotypes.txt

Example rows:

Gene_accession no.	Class	Phenotype	PMID	Mechanism of resistance	Notes
aac(6')-Ib-cr_1_DQ303918	Aminoglycoside, Quinolone	Tobramycin, Dibekacin, Amikacin, Sisomicin, Netilmicin, Fluoroquinolone, Ciprofloxacin	unpublished	Enzymatic modification	MIC of ciprofloxacin does not always increase above ECOFF PMID 16369542
blaOXA-51_1_DQ385606	Beta-lactam	Amoxicillin, Ampicillin, Imipenem	16940137, 15649299, 16630258	Enzymatic inactivation	Class D;OXA-51-like;Natural in Acinetobacter baumannii;;Phenotype described in Acinetobacter baumannii if simultaneous presence of insertion sequence ISAba1
blaOXA-66_1_AY750909	Beta-lactam	Unknown Beta-lactam	15760431	Enzymatic inactivation	Class D;OXA-51-like;Natural in Acinetobacter baumannii;;

Resfinder Phenotype panels (mapping mutation to drug):
https://bitbucket.org/genomicepidemiology/pointfinder_db/src/master/campylobacter/phenotypes.txt

Organisms with mutations in DB:
- Campylobacter
- Enterococcus faecalis
- Enterococcus faecium
- Escherichia coli
- Helicobacter pylori
- Klebsiella
- Mycobacterium tuberculosis
- Neisseria gonorrhoeae
- Salmonella
- Staphylococcus aureus

Example rows for Klebsiella:

#Gene_accession	Type Gene	Mutation ID	Codon_pos	Ref_nuc	Ref_codon	Res_codon	Class	Phenotype	PMID	Mechanism of resistance	Notes

ompK36_1_Z3 3506.1	AA	ASP-49	49	GAC	D	S		Cephalosporins	25245001		Under development. Phenotype should be used with caution.
acrR_1_AJ318 073.1	AA	ALA-20	20	GCT	A	T	Quinolone	Fluoroquinolone	12936981		Under development. Phenotype should be used with caution.
ramR_1_KY46 5996.1	AA	ALA-19	19	GCG	A	V	Tetracycline	Tigecycline	28533243		Under development. Phenotype should be used with caution.
gyrA_1_AF05 2258.1	AA	SER-83	83	TCG	S	L	Quinolone	Ciprofloxacin	22633335		Under development. Phenotype should be used with caution.
ompK36_1_Z3 3506.1	AA		131	-	ins	SG		Cephalosporins	25245001		Under development. Phenotype should be used with caution.
acrR_1_AJ318 073.1	NUC		382	-	del	CAGGC CCAGC GGCAG	Quinolone	Norfloxacin, Ciprofloxacin	12936981		Under development. Phenotype should be used with caution.

Example rows for Salmonella:

#Gene_ accession	TypeG ene	Mutation ID	Codon_ pos	Ref _nuc	Ref_ codon	Res_codon	Class	Phenotype	PMID	Mechanism of resistance
pmrA_1_CP0 55130.1	AA	GLY-15	15	GGG	G	R	Polymyxin	Colistin	##########	Target modification
pmrB_1_CP0 51284.1	AA	ARG-81	81	CGC	R	C,H	Polymyxin	Colistin	##########	Target modification
gyrA_1_MH93 3946.1	AA	ALA-67	67	GCC	A	P	Quinolone	Nalidixic acid, Ciprofloxacin	7492118	Target modification
acrB_1_CP00 0026.1	AA	ARG-717	717	CGG	R	Q,L	Macrolide	Azithromycin	31730615	Target modification
16S-rrsD_1_C P049983.1	NUC	CYS-1065	1065	C	C	T	Aminoglycoside	Spectinomycin	12402084	Target modification

Pathogenwatch AMR libraries

Pathogenwatch have curated AMR libraries for the following species:

Species	NCBI Code
Campylobacter	194
Enterococcus faecium	1352
Neisseria gonorrhoeae	485
Staphylococcus aureus	1280
Salmonella Typhi	90370
Streptococcus pneumoniae	1313
Vibrio cholerae	666

The libraries include one file per species, (e.g. this one for *Streptococcus pneumoniae*) and comprise:

- A 'genes' list, with the gene name, nucleotide sequence, and coverage/% identity required to report as detected. Core genes with known resistance SNPs are also included.
- A 'mechanisms' list, listing the phenotype (resistant or intermediate) associated with individual genes, variants or combinations thereof.

There are also separate libraries organised not by species, but by gene class:

- gram_neg_carbapenemases.toml
- gram_neg_esbl.toml
- gram_neg_colistin.toml (mcr alleles)

Notes on curation and curators are in the github repository, and papers showing validation against phenotype are listed in the Pathogenwatch documentation.

CARD Prevalence

When users search CARD for a specific gene/allele, the 'Resistomes' tab will display the prevalence of the gene per species, estimated from NCBI Chromosome, NCBI Plasmid, and NCBI WGS. This can be helpful to explore which organisms a gene is core to and to what extent it is plasmid-borne and disseminated to other organisms.

E.g. searching for FosA6, which is intrinsic (S^{WT}) in *K. pneumoniae* yields:

Prevalence: protein homolog model (view sequences)

Species	NCBI Chromosome	NCBI Plasmid	NCBI WGS
Escherichia coli	0%	0%	0.05%
Klebsiella aerogenes	4%	0%	3.95%
Klebsiella pneumoniae	84.97%	0.01%	50.11%
Klebsiella quasipneumoniae	91.6%	0%	68.03%
Salmonella enterica	0%	0%	0.01%

E.g. searching for FosA2, which is intrinsic in the *Enterobacter cloacae* complex:

Prevalence: protein homolog model (view sequences)

Species	NCBI Chromosome	NCBI Plasmid	NCBI WGS
Enterobacter asburiae	87.1%	0%	64.82%
Enterobacter cancerogenus	83.33%	0%	85.71%
Enterobacter chengduensis	100%	0%	84%
Enterobacter cloacae	92.86%	0%	64.22%
Enterobacter hormaechei	87.05%	0%	58.83%
Enterobacter kobei	90.91%	0%	63.76%
Enterobacter roggenkampii	69.77%	0%	52.88%
Escherichia coli	0%	0%	0.01%
Klebsiella pneumoniae	0%	0%	0%
Providencia stuartii	62.5%	0%	50%
Salmonella enterica	0.06%	0%	0.01%