OXFORD JOURNAL
OF STUDENT
SCHOLARSHIP

A Comparative Bioinformatic Analysis of Antimicrobial Resistance-Associated Proteins in Bacterial Pathogens

Aarush Sunil Viswanathan (student)
svaarush10@gmail.com

Anu Prabha (mentor)
anuprabha.mandal@gmail.com

ABSTRACT

Antimicrobial resistance (AMR) is perpetuated by proteins that modify drug targets, inactivate antibiotics, reduce intracellular drug concentrations, alter cell membrane structure, or regulate stress responses. We curated and audited a protein-based catalog comparing such determinants across three operational groups: ESKAPE pathogens, common non-ESKAPE bacteria, and rare/understudied bacteria. Reviewed UniProtKB/Swiss-Prot records tagged "antibiotic resistance" were filtered by gene name, classified taxonomically, linked to FASTA files, and represented by MEGA maximum-likelihood trees using the Jones-Taylor-Thornton model and 500 bootstrap replicates. The initial catalog contained 782 entries: 193 ESKAPE, 191 common non-ESKAPE, and 398 rare. After pruning, 562 remained: 83, 120, and 359, respectively. The rare/understudied group retained 90.2% of original entries, compared with 43.0% for ESKAPE. Manual functional annotation was used to categorize the protein records by function for visualization and comparison. Cell wall/envelope proteins dominated ESKAPE (30/83), efflux/transport dominated common non-ESKAPE (47/120), and cell wall/envelope (127/359) and efflux/transport (121/359) dominated the rare/understudied group. Functional categorization was available for 95.2–100% of records, while DrugBank references were scarce, especially in the rare/understudied group (1.1%). Because the trees combined unrelated protein families, they were exploratory only and cannot indicate organism phylogeny or horizontal transfer. This paper uses bioinformatics to bridge gaps in comprehensive comparisons of AMR determinants among ESKAPE, common non-ESKAPE, and rare/understudied bacteria.

Keywords: Biology, Bioinformatics, Antimicrobial resistance, Comparative proteomics, Data curation

INTRODUCTION

The use of antibiotics against infections is not without its downsides, as the same kind of selection that makes antibiotics successful favors the development and spread of antibiotic-resistant variants. An estimate of the AMR burden globally in 2019 put the number of AMR-attributed deaths at 1.27 million while those AMR-associated deaths stood at 4.95 million.[1] By 2021, these figures were revised to 1.14 million attributed deaths and 4.71 million associated deaths, and the projection of an increase in AMR burden remains unless efforts to prevent, diagnose, and treat are improved.[2]

AMR research is focused heavily on, but not limited to, specific organisms that pose a high burden of disease. The acronym ESKAPE (E: *Enterococcus faecium*; S: *Staphylococcus aureus*; K: *Klebsiella pneumoniae*; A: *Acinetobacter baumannii*; P: *Pseudomonas aeruginosa*; E: *Enterobacter* species) was coined to stress the importance of organisms that are "escape" pathogens from numerous antimicrobials.[3] These organisms continue to be therapeutic nightmares due to the combination of natural barriers, rapid acquisition of mobile resistance genes, target modification, antibiotic degrading enzymes, efflux mechanisms, and regulation of adaptive resistance.[4-5] WHO has published their list of priority bacterial pathogens for 2024 which includes, among others, Gram-negative bacteria that are resistant, drug-resistant *Mycobacterium tuberculosis, Pseudomonas aeruginosa, Staphylococcus aureus, Salmonella, Shigella, and Neisseria gonorrhoeae*.[6]

Antimicrobial resistance is not a single molecular phenotypic trait. The drug can be degraded or modified by enzymes; the drug can be altered via mutations or proteins; there can be decreased permeability and active transport to decrease the amount of drug inside the cell; and changes can be made to the cell envelope to modify the interaction of the drug.[7-8] Two-component systems and other regulatory systems can integrate these strategies, and plasmids, transposons, integrons, and other mobile elements can promote their spread among bacteria. The environment and commensal bacteria represent an important source of medically significant mechanisms.[9-13]

It becomes possible to study the resistance-related proteins on a broad range of species through large-scale databases. UniProtKB/Swiss-Prot offers curated manually annotated entries for each protein together with curated names, functions, pathways, and localization information.[14] CARD offers a classification system and curated models of resistance determinants,[15] while BV-BRC includes bacterial genomes, phenotypes, and analysis data and is an extended version of PATRIC platform.[16] Although such databases provide opportunities for large-scale comparative studies, which are hard to conduct experimentally, the research using databases is constrained by several factors including annotation quality, taxonomic coverage, search query language, and the difference between experimentally characterized resistance determinants and simply related proteins.

Research on pangenomes and pathogenomes has revealed that diversity related to resistance is usually found in the accessory genome and that multi-species comparisons are able to find both known and novel

resistance determinants.[17-20] Comparing the reviewed set of AMR-related proteins can therefore provide information on their composition, annotation and sequence information.
Unlike conventional pangenomic studies, which compare gene-content variation across complete genomes and characterize core and accessory genome structure, the present analysis operates at the curated protein level. This enables direct comparison of resistance-associated protein sequences and annotations across taxonomically diverse bacteria, but does not allow for inference of whole-genome gene presence/absence, genomic neighborhood, mobile-element context, or population-level resistance dynamics.

Thus, our objectives were (i) curation of reviewed, gene-named protein sequences associated with AMR; (ii) comparison of taxonomic scope, protein length, functional category, and annotation quality among ESKAPE, common non-ESKAPE, and rare/understudied groups; (iii) exploratory sequence-space visualization using maximum-likelihood tree methodology; and (iv) specification of limitations on biological interpretation arising from the heterogeneity of the protein family set.
In this study, a combined bioinformatics approach has been used to characterize and compare the resistance associated proteins, their associated gene names, functional categories, and pathways in operational categories of ESKAPE, common non-ESKAPE, and rare/understudied pathogens. By doing this, the resulting catalog could be used as a descriptive resource and as a reference for future surveillance-oriented and experimentally validated studies.

METHODS

Data collection and preprocessing

The source dataset was obtained from UniProtKB release 2026_01 using the UniProt keyword “Antibiotic resistance” (KW-0046). The retained source export is dated 8 March 2026, and data collection and database curation continued through 13 March 2026. Because the exact historical UniProt query URL was not retained, the acquisition procedure was reconstructed from the original downloaded UniProt files and accession lists. The working dataset was restricted to reviewed UniProtKB/Swiss-Prot records, where “reviewed” refers to entries tagged as “Reviewed (SwissProt)” within UniProt filter criteria. Records without an annotated gene name were then removed during spreadsheet-based editing.

Protein length was also screened during spreadsheet-based curation using an intended approximate range of 300–899 amino acids. This was a manual quality-control step rather than a criterion implemented in the UniProt retrieval process. As it was not applied as a strict computational cutoff, a small number of records outside the intended range remained in the retained dataset. Organism annotations were then reviewed manually, and non-bacterial entries were removed. The retained UniProt accession IDs were subsequently resubmitted through the UniProt ID Mapping service to organize the records into the three operational groups—ESKAPE, common non-ESKAPE, and rare/understudied—and to add further record-level metadata fields to use in descriptive analysis.

PATRIC/BV-BRC was used only as a secondary validation resource and not to generate the primary dataset or determine inclusion. Selected records were cross-checked against these resources for resistance-related annotation and organism information. Where annotations differed between resources, the curated UniProtKB/Swiss-Prot record was retained as the primary record and the discrepancy was reviewed manually.

Final quality control consisted of double checking reviewed status, gene-name availability, organism assignment, obvious protein-length outliers, and consistency of the available functional annotation before analysis.

Operational taxonomic grouping

Records were allocated into three operational categories that did not overlap with one another.

1. The ESKAPE group consisted of *E. faecium, S. aureus, K. pneumoniae, A. baumannii, P. aeruginosa,* and *Enterobacter* species (Supplementary Table S1).
2. The common non-ESKAPE group included the well-known organisms as specified: *Mycobacterium bovis, Streptococcus pneumoniae, Campylobacter fetus, Campylobacter jejuni, Escherichia coli, M. tuberculosis, N. gonorrhoeae, Helicobacter pylori, and Salmonella* organisms, consisting of various serotypes (Supplementary Table S2).
3. Records for all other bacteria were categorized as rare/understudied (Supplementary Table S3). "Rare/understudied" is thus an operational category and should not be viewed as an epidemiological definition of prevalence.

FASTA retrieval, header normalization and metadata schema

A Python script (Supplementary Code S1) queried UniProt accession numbers in sets of 100 from the UniProt batch REST endpoint, and sequences were written in FASTA format. Two separate scripts retrieved the protein names, gene names, and organisms corresponding to each accession number, and the headers were reformatted into human readable form (Table 1). iTOL label scripts added data on length, pathway, and subcellular localization for visualization purposes (Supplementary Data S2). The accession numbers were considered the stable join key used for data representation and manipulation when needed for different exploratory sequence-tree visualizations.

Table 1. Data schema used for the audited protein catalog.

Field	Type	Operational description
Entry	Text	UniProt accession used as the primary identifier and join key.
Protein names	Text	Curated or provisional protein name.

Field	Type	Operational description
Gene names	Text	Gene symbol(s); the first token was used as the primary gene.
Organism	Text	Species and strain/serotype annotation where available.
DrugBank	Text	DrugBank identifier recorded in the source table.
Pathway	Text	UniProt pathway annotation.
Function	Text	UniProt functional description.
Subcellular location	Text	Cellular localization annotation.
Intramembrane / transmembrane / topological domain	Text	Membrane topology features when annotated.
Length	Integer	Amino-acid sequence length calculated from the supplied FASTA file during audit.

Maximum-likelihood tree construction and visualization

The amino acid sequences for each operational group were aligned in MEGA12 software,[21] using MUSCLE alignment. In the parameters, Gap Penalties were set to -2.90, 0.00, and 1.20, Memory/Iterations to 2048 and 16, and Advanced Options to UPGMA, UPGMA, and 24 (Table 2). Maximum likelihood tree was constructed using Jones-Taylor-Thornton (JTT) substitution model,[22] equal rates across sites, nearest neighbor interchange, and auto neighbor-joining maximum parsimony tree. The bootstrap analysis was performed using 500 replicates (Table 3).[23] The trees were exported into the Newick format (Supplementary Data S1) and viewed in iTOL.[24]

Table 2. MEGA Alignment Explorer MUSCLE alignment options.

Parameter	Setting
Gap Penalties	-2.90
	0.00
	1.20
Memory/Iterations	2048

	16
Advanced Options	UPGMA
	UPGMA
	24

Table 3. MEGA tree-inference settings.

Parameter	Setting
Statistical method	Maximum likelihood
Test of phylogeny	Standard bootstrap
Bootstrap replicates	500
Sequence type	Amino acid
Substitution model	Jones–Taylor–Thornton (JTT)
Rates among sites	Uniform
Gaps/missing data	Use all sites
Heuristic search	Nearest-neighbor interchange
Initial tree	Automatic default (NJ/MP)
Branch-swap filter	None
Threads	4

Pruning

The manual tree-based approach for pruning removed redundant entries at strain level in the current workflow (Figure 1). Highly similar entries representing apparent redundancy at the strain or serotype level were manually collapsed to improve overall visualization. Candidate redundant entries were identified through visual inspection of the inferred sequence-tree relationships along with supporting sequence and metadata information, including organismal context, availability of DrugBank annotation, and historical relevance (Table 4). This procedure was intended to reduce overrepresentation of apparently redundant records rather than explicitly define biologically nonredundant protein families.

As representative selection included manual and metadata-based review, the pruning procedure may introduce selection bias and should not be interpreted as an estimate of protein-family diversity.

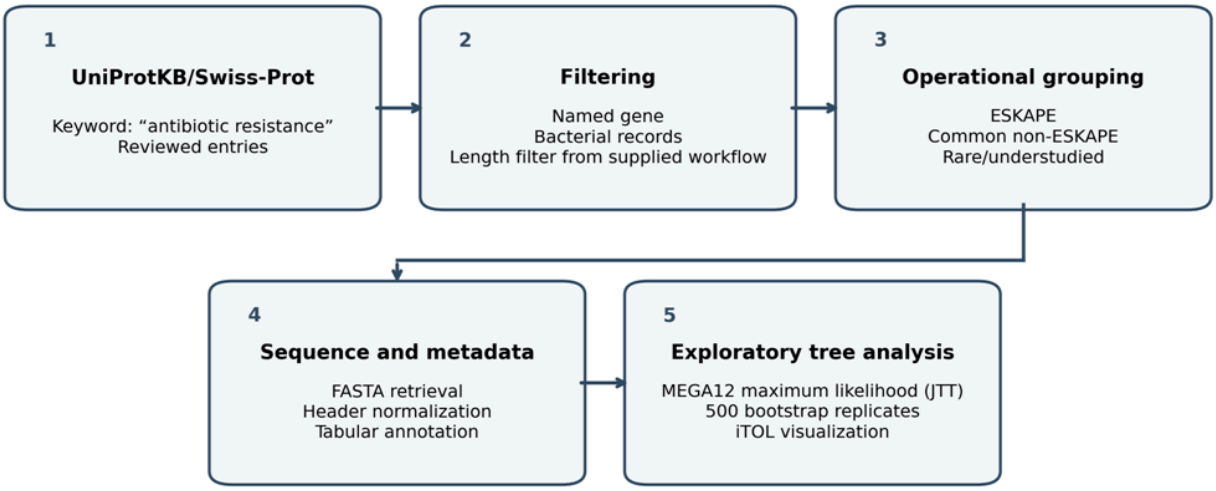

Figure 1. Reviewed UniProtKB/Swiss-Prot records associated with the keyword "antibiotic resistance" were filtered, organized into three operational groups, linked to FASTA and metadata records, and analyzed using the supplied MEGA12/iTOL workflow.

Table 4. Pruning decision rules.

Situation	Rule
Same protein, same species, strains judged to be similar upon manual review	Collapse representatives
DrugBank-linked vs non-linked redundant records	Prefer DrugBank-linked record if available
No DrugBank-linked representative	Retain historically relevant representative(s)

Situation	Rule
Different bacterial species	Do not collapse simply because protein names are identical
Different serotypes/strains	Decide redundancy at sequence/species level, prune case-by-case
Y. pestis strain-level duplicates	Retain general *Y. pestis* representative, not branches with *Y. pestis subsp. X* or *Y. pestis (strain X)*.

Functional categorization and descriptive analysis

Functional categories were assigned manually to each pruned entry during iTOL annotation. Each protein was reviewed using its available protein name, gene name, pathway, function, subcellular-location information, and other supporting annotation information, being assigned to one or more plausible functional categories. These categories included beta-lactamase; efflux/transport; cell wall/envelope; regulatory/sensor; DNA-associated; RNA/translation; mobile genetic element; specific resistance; metabolism/biosynthesis; and other/unclassified. Some proteins could reasonably fit more than one category and were therefore annotated in multiple categories during the iTOL editing process. In the final iTOL visualization and category counts, the category displayed for a multiply annotated protein reflected the most recently created or assigned annotation rather than a predefined ranking or computational hierarchy. (Supplementary Data S3). The resulting classification was thus treated as descriptive outputs of annotation rather than definitive biological classifications.

RESULTS AND DISCUSSION

Data acquisition

The source catalog comprised 782 bacterial proteins: 193 ESKAPE, 191 common non-ESKAPE and 398 rare/understudied. The sample sizes in each dataset were 83, 120, and 359 for the total of 562 sequences (Table 5, Figure 2). The proportions of retained sequences were very different for each group: 43.0%, 62.8% and 90.2% for ESKAPE, common non-ESKAPE and rare/understudied, respectively (Table 5). Here and throughout the results, “rare/understudied” refers to the operational category defined by

exclusion from the ESKAPE and common non-ESKAPE groups and should not be interpreted as an epidemiological measure of incidence or prevalence. Therefore, pruning was mostly applied to the ESKAPE group while collapsing was relatively little needed for the taxonomically broad residual group.

Table 5. Dataset composition after pruning (Supplementary Table S4a).

Dataset	Source n	Pruned n	Retained %	Species labels	Primary genes	Protein names	Length range	Median length
ESKAPE	193	83	43.0	6	61	65	300–889	420
Common non-ESKAPE	191	120	62.8	12	88	97	301–964	474
Rare/understudied	398	359	90.2	154	144	162	300–914	462

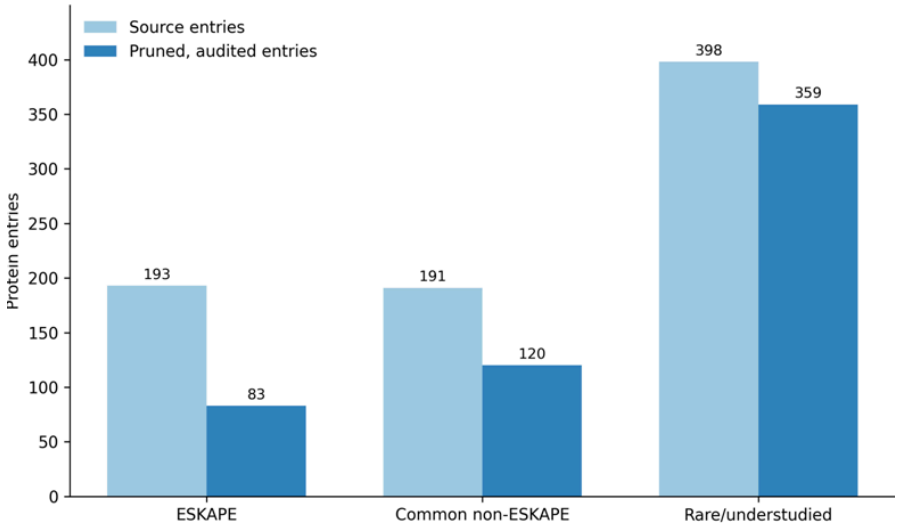

Figure 2. Dataset size before and after pruning.

Taxonomic breadth, gene representation and protein length

The ESKAPE representative dataset had six species-level labels, 61 unique primary genes, and 65 unique protein names. The common non-ESKAPE dataset had 12 species-level labels, 88 unique primary genes, and 97 unique protein names. The rare/understudied dataset contained 154 species-level labels, 144

unique primary genes, and 162 protein names. These figures represent catalog breadth rather than abundance or genetic diversity.

Lengths of proteins differed within each group (Figure 3). Proteins in the ESKAPE set varied in length between 300 and 889 amino acids, with the median value being 420. For common non-ESKAPE proteins, lengths varied between 301 and 964 amino acids, with a median value of 474. For rare/understudied proteins, lengths varied between 300 and 914 amino acids, with a median of 462.

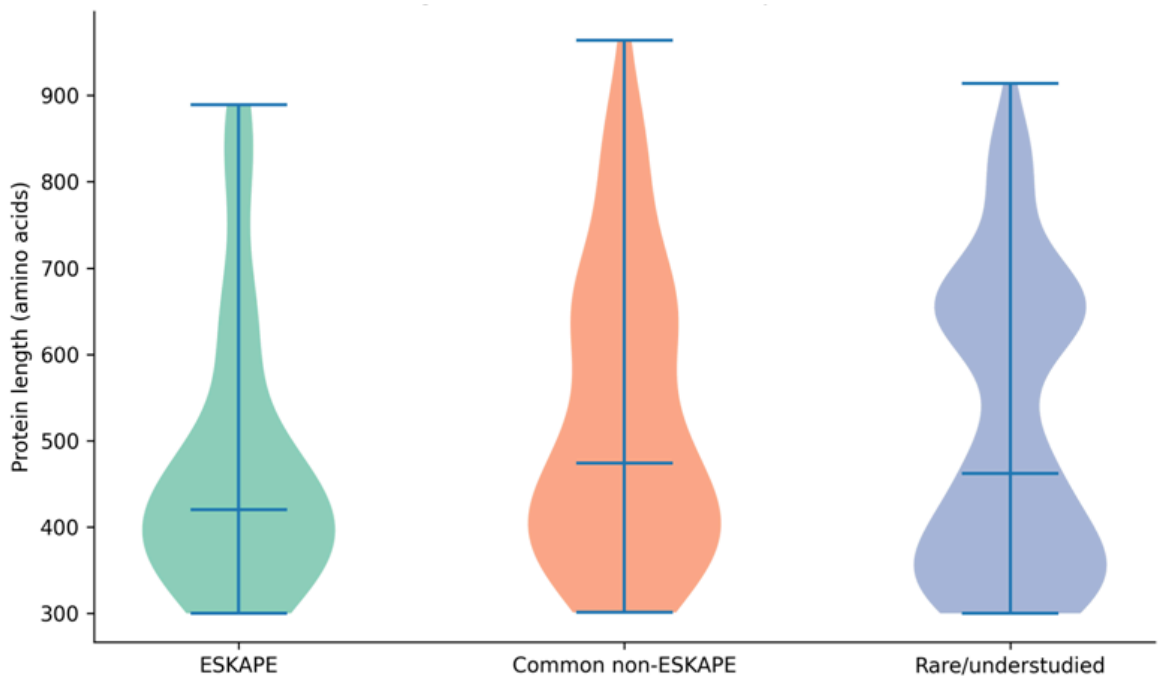

Figure 3. Protein-length distributions in pruned datasets. Violin width represents the empirical density of sequence lengths; horizontal bars mark the median and extrema. The distributions are descriptive and reflect the query and curation workflow rather than the proteomes of the represented organisms.

Functional composition of the ESKAPE dataset

Proteins associated with the cell wall/envelope accounted for 30 out of 83 (36.1%), while efflux/transport proteins made up 15 out of 83 (18.1%). The other major categories include DNA-related proteins (11; 13.3%). Other proteins in the database included RNA/translation proteins (7), resistance-specific proteins (6), regulatory/sensor proteins (5), beta-lactamases (4), metabolic/biosynthetic proteins (1), and four others/unclassified (Table 6, Figure 4). These results are in line with the importance of cell envelope modifications, beta-lactam antibiotic targets, drug transporters, and target modifications in ESKAPE resistance mechanisms; however, this database cannot be used to estimate the prevalence of resistance.

The most significant proportional reduction in the number of representatives in the source sets came from the ESKAPE category, with the number dropping from 193 to 83 representatives. The comments provided suggested that some of the redundancy was due to the presence of multiple strains of *S. aureus*. In

addition, qualitative analysis revealed the presence of multifunctional proteins, such as ArnB, NagZ, and Fem-family proteins, which could be classified into multiple categories.

Functional composition of common non-ESKAPE pathogens

Proteins associated with efflux/transport made up the biggest category in the non-ESKAPE common set (47/120; 39.2%), followed by cell wall/envelope proteins (22; 18.3%), RNA/translation proteins (12; 10.0%), DNA-related proteins (11; 9.2%), beta-lactamases (9; 7.5%), regulatory/sensor proteins (8; 6.7%), and proteins of specific resistance (7; 5.8%) (Table 6, Figure 4). Many entries related to each other within the *Salmonella* serotypes and *M. tuberculosis* appeared in the provided tree and were key candidates for representatives. There were twenty entries with DrugBank identification with the highest absolute and proportional coverage among the three groups (Supplementary Table S4b).

Functional composition of rare or understudied taxa

The rare/understudied group held 359 out of 398 audited source entries. The cell wall/protein envelope group (127; 35.4%) and the efflux/transport proteins (121; 33.7%) together made up 69.1%. The rest of the records included RNA/translation proteins (39), specific resistance proteins (22), beta-lactamases (18), regulatory/sensor proteins (13), DNA-associated proteins (10), metabolic/biosynthetic proteins (4), mobile element proteins (3), and two others (Table 6, Figure 4).

The large number of kept records suggests a good distribution of different species: many proteins that seemed similar appeared to be from different species and thus were not regarded as redundant strain records.

Yersinia was represented by a prominent cluster of entries close to each other in the tree provided. If there were several strain or serotype records of *Y. pestis* that were redundant to each other, the generalized entry on the species level was retained according to the provided pruning criterion.
The rare/understudied group had more provisional names for proteins compared to ESKAPE (though the total number of them was still relatively small - 17 out of 359).

Table 6. Functional categories displayed in the final iTOL annotation for the pruned datasets (Supplementary Table S4c).

Displayed functional category	ESKAPE n	ESKAPE (%)	Common non-ESKAPE n	Common non-ESKAPE (%)	Rare/understudied n	Rare/understudied (%)
Beta-lactamase	4	4.8	9	7.5	18	5.0

Displayed functional category	ESKAPE n	ESKAPE (%)	Common non-ESKAPE n	Common non-ESKAPE (%)	Rare/understudied n	Rare/understudied (%)
Cell wall/envelope	30	36.1	22	18.3	127	35.4
DNA-associated	11	13.3	11	9.2	10	2.8
Efflux/transport	15	18.1	47	39.2	121	33.7
Metabolic/biosynthetic	1	1.2	2	1.7	4	1.1
Mobile genetic element	0	0.0	1	0.8	3	0.8
Other/unclassified	4	4.8	1	0.8	2	0.6
RNA/translation	7	8.4	12	10.0	39	10.9
Regulatory/sensor	5	6.0	8	6.7	13	3.6
Specific resistance	6	7.2	7	5.8	22	6.1

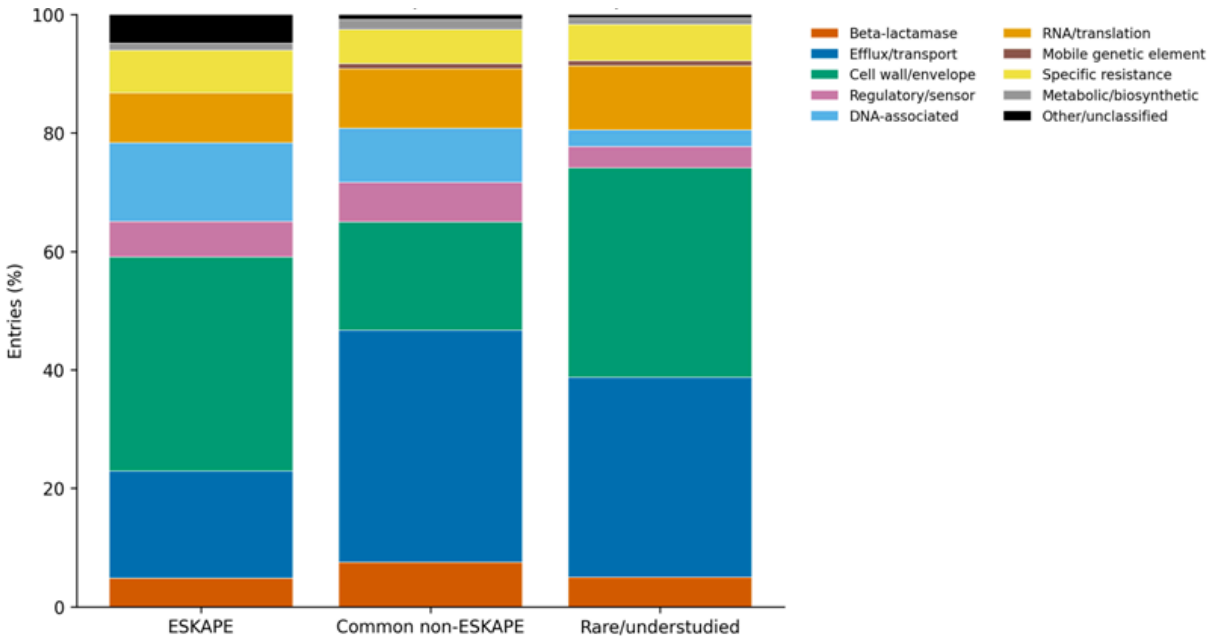

Figure 4. Functional categories represented in the final iTOL annotation. Categories were assigned manually based on protein annotation and supporting biological information. Some proteins could plausibly belong to multiple categories; for such proteins, the category displayed in the final iTOL visualization reflected the most recently created/assigned annotation. Counts describe the curated catalog, not mechanism frequencies in bacterial populations.

Annotation completeness and evidence density

Functional information was found for 79/83 ESKAPE entries (95.2%), all 120 common non-ESKAPE entries, and 346/359 rare/understudied entries (96.4%) (Figure 5). Subcellular-location information was found for 62.7%, 84.2%, and 71.9%, respectively. Pathway information was less abundant (26.5%, 18.3%, and 30.4%, respectively). Information about transmembrane locations was most abundant for the common and rare categories (48.3% and 52.4%), as expected given the substantial role of efflux/transport systems in drug resistance. There was no information about any intramembrane features in the provided tables.

Linkage to DrugBank data was sparse: 6/83 ESKAPE entries (7.2%), 20/120 common non-ESKAPE entries (16.7%), and 4/359 rare/understudied entries (1.1%). The significant shortage in the latter category is in line with expectations of the disconnect between taxonomically broad resistance biology and drug-development annotation. This does not imply limited therapeutic relevance, rather sheds light on limitations of scope and curation history of the source databases (Supplementary Table S4b).

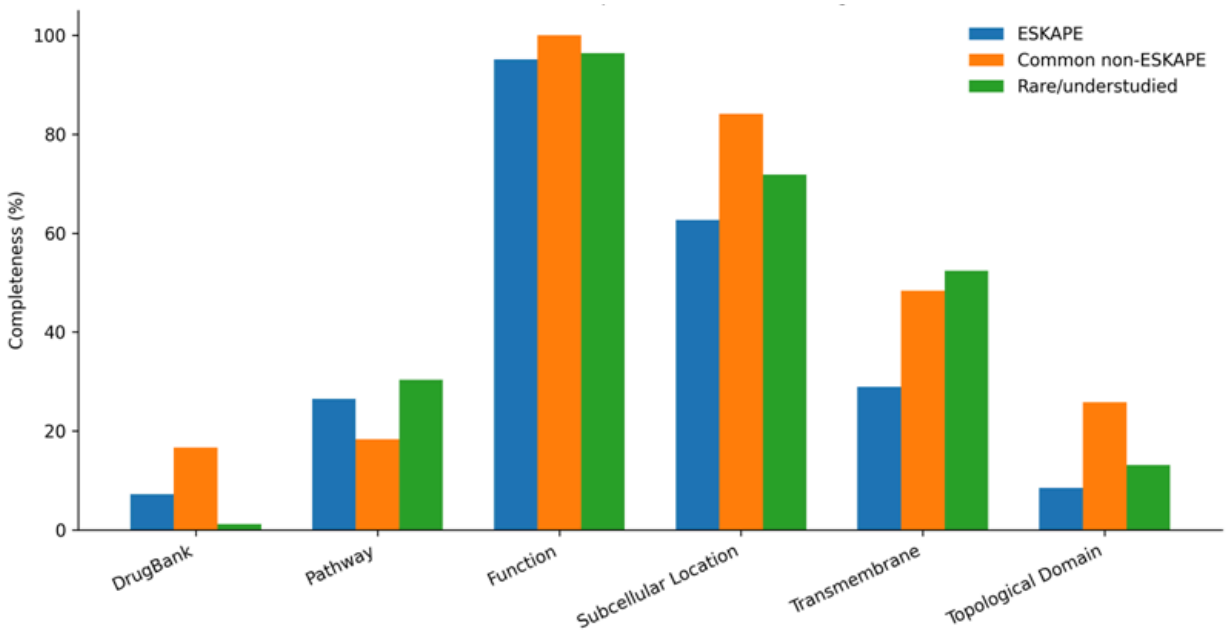

Figure 5. Metadata completeness. Percentages represent non-empty fields in the pruned datasets. DrugBank, pathway, function, localization, transmembrane, and topological-domain fields were inherited from the tables (supplementary tables); missingness reflects absence of annotation rather than absence of the biological feature.

Exploratory tree structure and bootstrap support

The aligned trees had 83 ESKAPE, 120 common non-ESKAPE, and 359 rare/understudied tips (Figure 6; Supplementary Figures S1–S3). The Newick files included 152, 167 and 390 support values for internal nodes, respectively. Medians for bootstrap proportions were 0.190, 0.332 and 0.674. Nodes supported to at least 0.70 were 56, 65, and 188; nodes supported to at least 0.95 were 40, 54, and 150. The greater median support in the rare trees is not necessarily an indication that those taxa or mechanisms of resistance have a better documented common evolutionary history. It might just indicate blocks of similar sequences, sequence composition, alignment properties, or topology in the provided dataset. Bootstrap support should be interpreted as confidence in the reproducibility of the inferred branch under the specified alignment and model, rather than as evidence of biological homology or common ancestry across the heterogeneous protein dataset.

Since all groups include unrelated protein families—enzymes, transporters, regulators, ribosomal proteins, cell wall proteins, and others—it would be impossible for one global alignment to assure the positional homology of all columns. The provided trees show the patterns of similarity identified through the provided workflow.

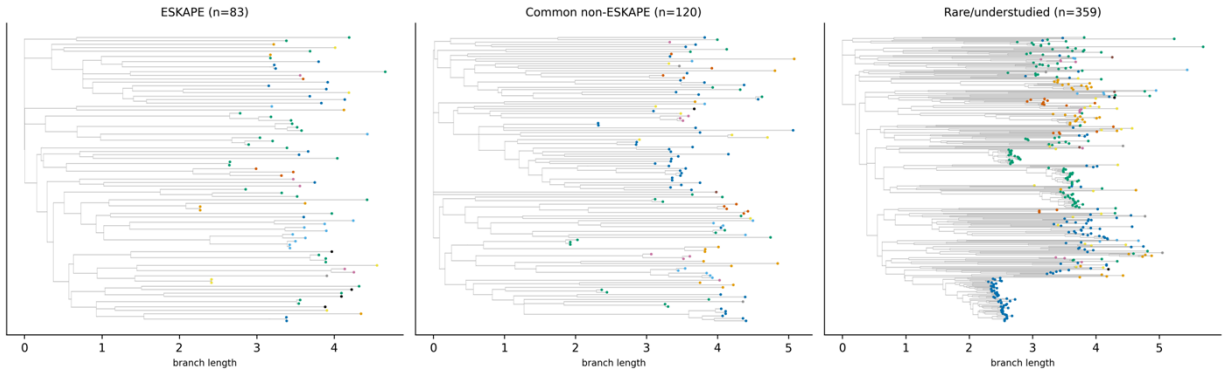

Figure 6. Exploratory global sequence trees. Colored terminal points correspond to the functional categories displayed in the final iTOL annotation. The trees span heterogeneous protein families and must not be interpreted as species phylogenies or family-specific evolutionary histories.

Redundancy in ESKAPE and the consequences of research intensity

The ESKAPE subset kept only 43.0% of its original entries. The information indicates that the *S. aureus* strains are the primary contributors to repeats and quasi-repeats. It is possible because the amount of scientific work on the study of methicillin resistance, cell wall synthesis, glycopeptide resistance, efflux, and target modifications in *S. aureus* strains is very high.[4-5] This case demonstrates one feature of databases: a larger number of entries may reflect greater research and annotation intensity rather than a greater number of molecular resistance mechanisms. Representation filtering makes visualization easier, but may also remove strain level variations that could be of biological interest. A more rigorous future comparative analysis would require pruning apparently redundant data via explicit sequence clustering with the identity and coverage threshold rather than manual representative selection.

Efflux and cell-envelope biology as recurrent themes

Transport/efflux proteins and cell-wall/envelope proteins were the most commonly appearing categories within the catalog. These categories contain multiple levels of function. Transport systems may be broadly resistant through the extrusion of multiple compound classes, whereas porins and envelope proteins modify permeability. Cell-wall and envelope proteins contain antibiotics, resistance elements, and pathway components that may become modified to influence the susceptibility to beta-lactams, glycopeptides, and polymyxins.[4-5, 7-8] Examples include MexA/OprM-type transport proteins, MacB-family transporters, penicillin-binding proteins, Mur enzymes, Fem proteins, Arn pathway enzymes, and Van system enzymes. The commonality of these proteins across all taxonomic groups suggests the biological recurrence of transport and envelope remodeling as resistance mechanisms, but no distinction can be made between constitutive housekeeping proteins and expressed resistance proteins.

Annotation completeness and translational bias

The high completeness of functional annotations can be attributed to the intentional inclusion of only reviewed Swiss-Prot entries,[14] leading to increased annotation quality but also introducing significant curation bias since proteins from well-understood organisms and pathways are usually reviewed, have names, experimental characterization, and links to external databases. On the other hand, the low coverage by DrugBank in the rare/understudied category (Supplementary Table S4b) is most likely caused by a lack of pharmacological curation and not low mechanistic significance. Pathway fields, on the other hand, were included in less than one-third of the records in any of the groups.

Interpretation of the exploratory sequence-tree visualization

The maximum-likelihood phylogenetics relies on the assumption that the aligned columns are homologous characters undergoing evolution according to an explicitly specified model. Any errors in the multiple-sequence alignment process significantly impact the inferred branch lengths and topology.[25] In our case, the set comprises sequences from unrelated families with distinct structure, domains, length and evolutionary histories. Global alignment may misalign non-homologous residues and a JTT tree built from such an alignment cannot be considered as phylogeny of organisms or a particular protein family. Instead, it could be more indicative of similarity of the protein/function.[22,25]

Bootstrap analysis provides resampling stability conditioned on the chosen alignment, model and data.[23] Those highly supported branches in the provided trees should be viewed as stable elements of this analysis and not as evidence for common ancestry or horizontal gene transfer. The tree nodes are helpful in visualizing sequence space organization and identifying possible family clusters for further study. A common observation that is expected in functional similarity in proteins across bacteria is that the same protein from different bacteria clusters closer than the bacteria themselves. This is expected due to the common domain and active site structures needed for the protein to perform its function.

The study indicates that AMR-associated proteins cover a broad taxonomic and mechanistic spectrum apart from the usual ESKAPE strains. The ESKAPE strains showed the most strain-level overlap, while the rare/understudied taxa exhibited the broadest species coverage and maintained the largest percentage of source records. Efflux/transport and cell wall/envelope biology were recurring categories for all groups, but the DrugBank linkage was significantly reduced among rare taxa. The trees and tables serve as an accessible exploratory tool; however, biologically sound evolutionary inference will demand further study on homologous family segregation, alignment, genomic context, and phenotype validation.[26]

LIMITATIONS

Several limitations affect interpretation of this catalog. Restricting the analysis to reviewed UniProtKB/Swiss-Prot records favors proteins, organisms, and pathways that have received greater research and experimental attention. Differences among groups in functional annotation, pathway information, subcellular localization, and DrugBank linkage may therefore partially reflect research and database curation history rather than biological differences. In addition, manual representative selection

during pruning involved investigator judgement and may introduce selection bias. Because the operational groups were not normalized for taxonomic sampling, catalog counts should not be interpreted as estimates of population-level resistance prevalence or biological diversity. Because the sequence-space trees contain unrelated protein families, their branch structure is descriptive and cannot be used to determine organism phylogeny, common ancestry, or horizontal gene transfer. Finally, because functional categorization was performed manually and the final displayed category for multiply annotated proteins depended on the order of iTOL annotation, the category counts should be interpreted as descriptive visualization outputs rather than definitive biological classifications.

In addition, the study does not analyze clinical isolates, antimicrobial-susceptibility phenotypes, patient outcomes, treatment responses, or longitudinal surveillance data.

CONCLUSION

This study curated and audited a database of 562 antimicrobial resistance proteins across ESKAPE pathogens, commonly encountered non-ESKAPE bacteria, and rare/understudied bacteria. The term "rare/understudied" is used here as an operational dataset classification and does not imply epidemiological prevalence or incidence. The ESKAPE group showed greater strain-level redundancy in results compared to wider taxonomic representation in the rare/understudied group, and across all three groups, cell wall/envelope and efflux/transport proteins were shown to be the most frequently represented functional categories in the dataset. Functional categorization for each entry was almost complete, however DrugBank linkage was sparse; only 1.1% of determinants in the rare/understudied group were linked to DrugBank entries. These results show that research is focused primarily on commonly encountered pathogens, and they highlight gaps in pharmacological annotation within the rare/understudied group. This comparative bioinformatics approach serves as a tool to inform future research into the topic, and the resulting catalog may support hypothesis generation and provide a reference resource for future surveillance-oriented and experimentally validated studies.

AI ACKNOWLEDGEMENT

During the preparation of this work, the authors used ChatGPT in order to edit the exploratory sequence-tree visualization for visual clarity with node coloring based on protein categories. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

ACKNOWLEDGEMENTS

The authors thank CreatED for connecting the student author with the research mentor.

REFERENCES

1. Murray, C. J. L.; Ikuta, K. S.; Sharara, F.; Swetchinski, L.; *et al.* Global burden of bacterial antimicrobial resistance in 2019: A systematic analysis. *Lancet* **2022**, *399* (10325), 629–655. https://doi.org/10.1016/S0140-6736(21)02724-0.
2. GBD 2021 Antimicrobial Resistance Collaborators. Global burden of bacterial antimicrobial resistance 1990–2021: A systematic analysis with forecasts to 2050. *Lancet* **2024**, *404* (10459), 1199–1226. https://doi.org/10.1016/S0140-6736(24)01867-1.
3. Rice, L. B. Federal funding for the study of antimicrobial resistance in nosocomial pathogens: No ESKAPE. *J. Infect. Dis.* **2008**, *197* (8), 1079–1081. https://doi.org/10.1086/533452.
4. Miller, W. R.; Arias, C. A. ESKAPE pathogens: Antimicrobial resistance, epidemiology, clinical impact and therapeutics. *Nat. Rev. Microbiol.* **2024**, *22* (10), 598–616. https://doi.org/10.1038/s41579-024-01054-w.
5. De Oliveira, D. M. P.; Forde, B. M.; Kidd, T. J.; Harris, P. N. A.; *et al.* Antimicrobial resistance in ESKAPE pathogens. *Clin. Microbiol. Rev.* **2020**, *33* (3), e00181-19. https://doi.org/10.1128/CMR.00181-19.
6. World Health Organization. *WHO Bacterial Priority Pathogens List, 2024: Bacterial Pathogens of Public Health Importance to Guide Research, Development and Strategies to Prevent and Control Antimicrobial Resistance*; World Health Organization: Geneva, 2024. ISBN 978-92-4-009346-1.
7. Blair, J. M. A.; Webber, M. A.; Baylay, A. J.; Ogbolu, D. O.; Piddock, L. J. V. Molecular mechanisms of antibiotic resistance. *Nat. Rev. Microbiol.* **2015**, *13*, 42–51. https://doi.org/10.1038/nrmicro3380.
8. Wright, G. D. Bacterial resistance to antibiotics: Enzymatic degradation and modification. *Adv. Drug Deliv. Rev.* **2005**, *57* (10), 1451–1470. https://doi.org/10.1016/j.addr.2005.04.002.
9. Crofts, T. S.; Gasparrini, A. J.; Dantas, G. Next-generation approaches to understand and combat the antibiotic resistome. *Nat. Rev. Microbiol.* **2017**, *15* (7), 422–434. https://doi.org/10.1038/nrmicro.2017.28.
10. Hu, Y.; Gao, G. F.; Zhu, B. The antibiotic resistome: Gene flow in environments, animals and human beings. *Front. Med.* **2017**, *11* (2), 161–168. https://doi.org/10.1007/s11684-017-0531-x.
11. Forsberg, K. J.; Reyes, A.; Wang, B.; Selleck, E. M.; Sommer, M. O. A.; Dantas, G. The shared antibiotic resistome of soil bacteria and human pathogens. *Science* **2012**, *337* (6098), 1107–1111. https://doi.org/10.1126/science.1220761.
12. Mahnert, A.; Moissl-Eichinger, C.; Zojer, M.; Bogumil, D.; *et al.* Man-made microbial resistances in built environments. *Nat. Commun.* **2019**, *10*, 968. https://doi.org/10.1038/s41467-019-08864-0.
13. Munk, P.; Brinch, C.; Møller, F. D.; Petersen, T. N.; *et al.* Genomic analysis of sewage from 101 countries reveals global landscape of antimicrobial resistance. *Nat. Commun.* **2022**, *13*, 7251. https://doi.org/10.1038/s41467-022-34312-7.
14. The UniProt Consortium. UniProt: The Universal Protein Knowledgebase in 2025. *Nucleic Acids Res.* **2025**, *53* (D1), D609–D617. https://doi.org/10.1093/nar/gkae1010.

15. Alcock, B. P.; Huynh, W.; Chalil, R.; Smith, K. W.; *et al.* CARD 2023: Expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Res.* **2023**, *51* (D1), D690–D699. https://doi.org/10.1093/nar/gkac920.
16. Olson, R. D.; Assaf, R.; Brettin, T.; Conrad, N.; *et al.* Introducing the Bacterial and Viral Bioinformatics Resource Center (BV-BRC): A resource combining PATRIC, IRD and ViPR. *Nucleic Acids Res.* **2023**, *51* (D1), D678–D689. https://doi.org/10.1093/nar/gkac1003.
17. Hyun, J. C.; Monk, J. M.; Palsson, B. O. Comparative pangenomics: Analysis of 12 microbial pathogen pangenomes reveals conserved global structures of genetic and functional diversity. *BMC Genomics* **2022**, *23*, 7. https://doi.org/10.1186/s12864-021-08223-8.
18. Hyun, J. C.; Monk, J. M.; Szubin, R.; Hefner, Y.; Palsson, B. O.; *et al.* Global pathogenomic analysis identifies known and candidate genetic antimicrobial resistance determinants in twelve species. *Nat. Commun.* **2023**, *14*, 7690. https://doi.org/10.1038/s41467-023-43549-9.
19. Tettelin, H.; Masignani, V.; Cieslewicz, M. J.; Donati, C.; *et al.* Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial pan-genome. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102* (39), 13950–13955. https://doi.org/10.1073/pnas.0506758102.
20. Medini, D.; Donati, C.; Tettelin, H.; Masignani, V.; Rappuoli, R. The microbial pan-genome. *Curr. Opin. Genet. Dev.* **2005**, *15* (6), 589–594. https://doi.org/10.1016/j.gde.2005.09.006.
21. Kumar, S.; Stecher, G.; Suleski, M.; Sanderford, M.; *et al.* MEGA12: Molecular Evolutionary Genetic Analysis version 12 for adaptive and green computing. *Mol. Biol. Evol.* **2024**, *41* (12), msae263. https://doi.org/10.1093/molbev/msae263.
22. Jones, D. T.; Taylor, W. R.; Thornton, J. M. The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **1992**, *8* (3), 275–282. https://doi.org/10.1093/bioinformatics/8.3.275.
23. Felsenstein, J. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **1985**, *39* (4), 783–791. https://doi.org/10.1111/j.1558-5646.1985.tb00420.x.
24. Letunic, I.; Bork, P. Interactive Tree of Life (iTOL) v6: Recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Res.* **2024**, *52* (W1), W78–W82. https://doi.org/10.1093/nar/gkae268.
25. Morrison, D. A. Multiple sequence alignment: In pursuit of homologous DNA positions. *Genome Res.* **2007**, *17* (2), 133–144. https://doi.org/10.1101/gr.5232407.
26. Feldgarden, M.; Brover, V.; Haft, D. H.; Prasad, A. B.; *et al.* Validating the AMRFinder Tool and Resistance Gene Database by using antimicrobial resistance genotype–phenotype correlations in a collection of isolates. *Antimicrob. Agents Chemother.* **2019**, *63* (11), e00483-19. https://doi.org/10.1128/AAC.00483-19.

SUPPLEMENTARY MATERIALS

Supplementary Figure S1. Full ESKAPE exploratory sequence tree (PDF and PNG).
Supplementary Figure S2. Full common non-ESKAPE exploratory sequence tree (PDF and PNG).

Supplementary Figure S3. Full rare/understudied exploratory sequence tree (PDF and PNG).
Supplementary Table S1. ESKAPE row-level protein catalog.
Supplementary Table S2. common non-ESKAPE row-level protein catalog.
Supplementary Table S3. rare/understudied row-level protein catalog.
Supplementary Table S4a. Dataset summary.
Supplementary Table S4b. Annotation completeness.
Supplementary Table S4c. Functional category counts.
Supplementary Data S1. Newick trees with normalized accession tips.
Supplementary Data S2. iTOL label annotation files.
Supplementary Data S3. iTOL protein-category annotation files.
Supplementary Code S1. Reproducible, descriptive analysis, figure generation, and tree correction script.
All files - Sunil Viswanathan_Prabha_Supplementary files

AUTHOR INFORMATION

Aarush Sunil Viswanathan is a 12th grader at Liberty High School in Lake Saint Louis, Missouri, USA. His research interests include genetic and microbial engineering. He conducted this research under the mentorship of Anu Prabha from IIT Bombay, and he plans to pursue Chemical and Biomolecular Engineering after graduation.

Dr. Anu Prabha holds a Ph.D. in Bioinformatics from IIT Bombay. Her work focuses on developing automated, data-driven healthcare solutions using bioinformatics, data science, and programming. Her doctoral research involved sequence and structure analysis and computational fold-specific annotation methods.