**AMR Prevalence Detection with RGI**

AMR prevalence data was gathered using the Resistance Gene Identifier (RGI), a tool for putative AMR gene detection from submitted sequence data using AMR detection models from CARD. Results from RGI are further categorized through the Antibiotic Resistance Ontology (ARO), e.g. “beta-lactam resistance protein”, “antibiotic inactivation enzyme”, etc.

We used RGI to analyze genomic data for seven common disease-causing gram-negative species: *Acinetobacter baumannii, Burkholderia cepacia, Enterobacter aerogenes, Enterobacter cloacae, Escherichia coli, Klebsiella pneumoniae,* and *Pseudomonas aeruginosa*. For each of these species, we downloaded chromosomal and plasmid assembly sequences from NCBI’s RefSeq database in FASTA format using the Entrez E-Utils BioPython module. We additionally downloaded all unassembled contig sequences from whole genome shotgun projects for each species using the same method (Table 1).

Each chromosome and plasmid sequence file was analyzed individually by RGI, according to its NCBI RefSeq accession, meaning one analysis per chromosome or plasmid. Unassembled contig sequences were analyzed by project, according to the WGS project accession. Results from RGI are output in JSON format, then converted to a tab-separated values table which indicates the top AMR model result for each predicted gene, the relevant ARO categories, and other information (coordinates, accessions, model criteria, etc.).

For each species, the RGI results are then aggregated and parsed together to calculate the best-hit occurrence and percentage for each model across all chromosome, plasmid, and contig sequences. For example, if fifty sequence assemblies were analyzed and blaNDM-1 was predicted in forty separate assemblies, the occurrence would be calculated as 40, and then converted to a percentage based on the total number analyzed, in this case 80%. This calculation was performed twice, once with both perfect and strict best-hit models (Table 2) and then only with perfect best-hit models (Table 3). These data were compiled and annotated with the detected model type and the ARO categories. Note that only protein homolog models and protein variant models currently have detection algorithms in RGI. Occurrence data is not separated by intrinsic or acquired genes, or by chromosomal or plasmid localization, therefore these numbers only indicate total overall occurrence. Reported frequencies have not been corrected for unmeasured clonality of genomic data within NCBI.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Species | Chromosome | Plasmid | Contig (WGS) | Total |
| *A. baumannii* | 55 | 140 | 1764 | 1959 |
| *B. cepacia* | 17 | 4 | 87 | 108 |
| *E. aerogenes* | 6 | 7 | 114 | 127 |
| *E. cloacae* | 17 | 41 | 465 | 523 |
| *E. coli* | 263 | 669 | 4609 | 5541 |
| *K. pneumoniae* | 96 | 395 | 1481 | 1972 |
| *P. aeruginosa* | 92 | 16 | 2144 | 2252 |

**Table 1. Assembled chromosomes, plasmids, and WGS assemblies analyzed.** Species names are as they appear in NCBI; the data were not separated by strain or subspecies. Data columns are labeled with sequence type (replicon or contig) and indicate how many of each type were analyzed.

**Table 2. Estimated prevalence of AMR genes in select gram-negative pathogens as detected by RGI perfect and strict criteria.** Given model names are as they appear in CARD. Data columns are labeled with species name and indicate percent occurrence for each model from analyzed assemblies and contigs. Percent occurrence is calculated by dividing the raw occurrence count for strict and perfect hits by the total number of analyzed genome sequences, plasmid sequences or assemblies for each species. The “ARO Categories” column gives the ARO category information for that model; note that any model may have multiple ARO categories, which are separated by a semicolon. The “CARD Model Type” column indicates if a protein homolog or protein variant model is used for detection with RGI. AMR detection models available in CARD but not detected in any analyzed assembly or contig are not included. Results are not separated by chromosome or plasmid, intrinsic or acquired.

**Table 3. Estimated prevalence of AMR genes in select gram-negative pathogens as detected by RGI perfect-only criteria.** Given model names are as they appear in CARD. Data columns are labeled with species name and indicate percent occurrence for each model from analyzed assemblies and contigs. Percent occurrence is calculated by dividing the raw occurrence count for strict and perfect hits by the total number of analyzed genome sequences, plasmid sequences or assemblies for each species. The “ARO Categories” column gives the ARO category information for that model; note that any model may have multiple ARO categories, which are separated by a semicolon. The “CARD Model Type” column indicates if a protein homolog or protein variant model is used for detection with RGI. AMR detection models available in CARD but not detected in any analyzed assembly or contig are not included. Results are not separated by chromosome or plasmid, intrinsic or acquired.