# Appendix 3c: ARG Databases, Categories and Names of ARGs

## CARD

CARD consists of a standard reference database (2,631 DNA sequences obtained from high-confidence sources, e.g. sequenced genomes) and a database of variant DNA sequences (149,075), essentially instances of the reference sequences observed "in the wild" (the WildCARD). We obtained licensing for use of these resources from McMaster University, and downloaded the databases (release labelled 20th March 2020) and RGI software (version 5.1.0) on 18th May 2020.

## The DeepARG database

DeepARG provides no version number, but we used the edition dated Feb 2020. Its database of 12,279 protein sequences is designated by the authors as Version 2.

## Database Comparison

Prior to performing the analysis, we performed a comparison of the sequence content of the databases to provide perspective on the results.

### Categories of ARGs

Because the number of different ARGs is large, it is useful to report not just frequencies of each ARG but also categories. The CARD database makes extensive use of the Antibiotic Resistance Ontology (ARO) developed alongside it, and classifies each ARG in three ways: **gene family** (this is still quite fine-grained, and so there are a large number of categories); **drug class**; and **resistance mechanism**. The latter two classifications apply one or more terms to each ARG (often many in the case of resistance mechanism), which presents difficulties for reporting frequencies of each category (some ARGs will contribute to two or more, and so overall totals are not strictly additive).

DeepARG's classification ("**DeepARG class**") is simpler and has relatively few terms. We also identified what may be inconsistencies in the classification, since some gene names (for which there are multiple reference sequences) are not always placed in the same class.

In summary, we consider the DeepARG classification to be less systematic than the ARO used by RGI/CARD. We therefore used the ARO-based RGI results as the classification in this study.

### Names of ARGs

We have also found cases of identical sequences in the two databases which have been assigned a different ARG name in each. Some of these differences may be issues of nomenclature.

Attempting to harmonise ARG names between the two is beyond the scope of this project, and would in effect require curation of ARG sequences. We have therefore used the reference data as-is.