Predicting Antibiotics Resistance Genes from Metagenomic Data

Chunyu Zhao

26 December, 2018

Contents

1 CARD (Comprehensive Antibiotic Resistance Database)

1.1 Introduction

CARD is a bioinformatic database of resistance genes, their products and associated phenotypes. As of 06-28-2018, there are 4008 Ontology Terms, 2498 Reference Sequences, 1211 SNPs, 2437 Publications, 2545 AMR Detection Models.

At the core of CARD is the novel **Antibiotic Resistance Ontology (ARO)**, a controlled vocabulary for describing antimicrobial molecules and their targets, resistance mechanisms, genes and mutations, and their relationships.

- The file "aro_index.csv" contains a list of ARO tagging of GenBank accessions stored in CARD.
- The file "aro_categories.csv" contains a list of ARO terms used to categorize all entries in CARD and results via the RGI. These categories reflect AMR gene family, target drug class, and mechanism of resistance.
- The file "aro_categories_index.csv" contains a list a GenBank accessions stored in CARD cross-referenced with the major categories within the ARO. These categories reflect AMR gene family, target drug class, and mechanism of resistance, so GenBank accessions may have more than one cross-reference. For more complex categorization of the data, use the full ARO available at http://card.mcmaster.ca/download.
    - one GenBank accession may have more than one cross_reference.

1.2 Protein Homolog Model

The protein homolog model is an AMR detection model. Protein homolog models detect a protein sequence based on its similarity to a curated reference sequence. A protein homolog model has only one parameter: a curated BLASTP bitscore cutoff for determining the strength of a match. Protein homolog model matches to reference sequences are categorized on three criteria: perfect, strict and loose. A perfect match is 100% identical to the reference sequence along its entire length; a strict match is not identical but the bitscore of the matched sequence is greater than the curated BLASTP bitscore cutoff. Loose matches are other sequences with a match bitscore less than the curated BLASTP bitscore. - Bit-score Cut-off: 600

1.3 R codes parsing CARD

First of all, we used only the CARD homolog models, where under assumptions of curation of the database, the presence of a member of a ARG family is considered a realiable indicator for probable ARG potential. When using the homolog models, we assume that metagenomic reads highly similar to an ARG from a model (having > 95% nucleotide similarity) will confer this functional capacity. [PMID: 30349083]

In **DYNAMIC** study, we are interested in mining which **drugs** the **resistance genes** *confer resistance to*, and further on group by the **AR genes** and **drugs class**.

For beginner like me, it took quite some time to figure out how to extract all these concepts from CARD data, and here are my notes.

The header information in the *protein_fasta_protein_homolog_model.fasta* contains the *ARO accession* of the AR genes (e.g. ARO:3000190).

antibiotic group: to which the gene belong to (e.g. tetO, tetA, dha, or macB)

1.3.1 which drugs the resistance genes *confer resistance to*

We need the ontologyIndex package to exploring the ontology data **aro.obo**, in order to fine which antibiotics the gene confer resistance to.

```
library(tidyverse)
library(pander)
library(ontologyIndex)

card_obo_fp <- "card/card_20180628/card-ontology/aro.obo"
ont.obo <- get_ontology(card_obo_fp, extract_tags = "everything")

## FIRST, get the AROdrug this AROgene`confers resistance to`
ARO.gene <- "ARO:3000190"
drugs <- propagate_relations(ont.obo, ARO.gene, relations = "confers_resistance_to_drug",
                             use_inverse_relations = F, exclude_roots = T)
print(drugs)
```

```
## [1] "ARO:0000051" "ARO:0000069" "ARO:3000152" "ARO:3000528" "ARO:3000667"
## [6] "ARO:3000668"
```

1.3.2 group by the AR genes and drug class

For each drug/antibiotics, to find the drug class it belongs to, we need the **aro_categories.csv**.

antibiotics categories: to which a gene confers resistance to (e.g. macrolides, beta lacmases, or aminoglycosides).

```
## SECOND, look up the `aro_categories` table to find the ARODrugClass information
aro.category <- read_delim("card/card_20180628/card-data/aro_categories.csv", delim="\t")
ARO.drug <- "ARO:0000051"
ans <- get_term_property(ont.obo, property="ancestors", term=ARO.drug, as_names=TRUE)
ARO.drug.class <- lapply(1:length(names(ans)),
                         function(x) aro.category %>% filter(grepl(names(ans)[x], `ARO Accession`))) %>%
  do.call(rbind, .)
pander(ARO.drug.class)
```

ARO Category	ARO Accession	ARO Name
Drug Class	ARO:3000050	tetracycline antibiotic

For each *antibiotic group*, to find the *AMR Gene Family*, we need to *join* the **aro__index.csv** with **aro__categories__index.csv**.

```
## THIRD, give me the AMR Gene Family
aro.index <- read_delim("card/card_20180628/card-data/aro_index.csv", delim="\t")

# cross reference with genbank
aro.category.index <- read_delim("card/card_20180628/card-data/aro_categories_index.csv", delim="\t") %>%
  unique()

aro <- aro.category.index %>%
  left_join(aro.index, by=c("Protein Accession","DNA Accession")) %>%
  select(`ARO Accession`, everything())

aro %>% filter(`ARO Accession` %in% ARO.gene) %>%
  select(- one_of(c("Model Name", "Model ID", "CVTERM ID","DNA Accession", "Protein Accession"))) %>%
  pander(split.table = Inf)
```

ARO Accession	AMR Gene Family	Drug Class	Resistance Mechanism	Model Sequence ID	ARO Name
ARO:3000190	tetracycline-resistant ribosomal protection protein	tetracycline antibiotic	antibiotic target protection	4234	tetO

1.4 Antibiotics used in our study

Table 3: antibiotics used and their related drug classes

ARODrug	DrugName	ARO.DrugClass	DrugClassName
ARO:3000689	metronidazole	ARO:3004115	nitroimidazole antibiotic
ARO:0000054	amoxicillin	ARO:3000008	penam
ARO:0000036	ciprofloxacin	ARO:0000001	fluoroquinolone antibiotic
ARO:0000028	vancomycin	ARO:3000081	glycopeptide antibiotic
ARO:0000069	doxycycline	ARO:3000050	tetracycline antibiotic
ARO:3000158	azithromycin	ARO:0000000	macrolide antibiotic
ARO:0000058	cefazolin	ARO:0000032	cephalosporin
ARO:3000329	sulfamethoxazole	ARO:3000282	sulfonamide antibiotic
ARO:3000188	trimethoprim	ARO:3000171	diaminopyrimidine antibiotic
ARO:3000517	rifaximin	ARO:3000157	rifamycin antibiotic
ARO:0000059	cefepime	ARO:0000032	cephalosporin
ARO:3000641	cefalexin	ARO:0000032	cephalosporin
ARO:0000065	clarithromycin	ARO:0000000	macrolide antibiotic
ARO:0000066	clindamycin	ARO:0000017	lincosamide antibiotic

2 Literature review

2.1 Developmental dynamics of the preterm infant gut microbiota and antibiotic resistome

We assembled 2004 contigs containing 794 AR genes as annotated using **Resfams**, identifying extensive resistance to b-lactams, amhenicols, tetracyclines, and polymyxins.

To extand functional AR gene analysis to all shotgun-sequenced preterm infant gut microbiomes, we used **ShortBRED** to generate short unique markers for all AR gene families identified in functional selections and AR-specific gene databases.

Relative abundance of antibiotic resistance genes was calculated using ShortBRED. - Shotgun reads were mapped to the resulting AR-specific markers and normalized across samples to generate AR-gene profiles for all infant gut metagenomes. - RPKM: reads per kilobase of reference sequence per million sample reads - RGI(contigs level): RGI relies on open reading frame detection.

2.2 DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data

- On the other hand, for short metagenomic reads, a stricter identity constraint of ~80% is recommended [20, 29] to avoid a high false positive rate.
- In principle, the best hit approach works well for detecting **known** and **highly conserved** categories of ARGs but may fail to detect novel ARGs or those with low sequence identity to known ARGs [19,30].

2.2.1 ARG annotation of CARD and ARDB

The ARDB and CARD databases both contain information to aid in the classification of ARGs, including the antibiotic category to which a gene confers resistance (e.g. macrolides, beta lactamases, or aminoglycosides) and the antibiotic group to which the gene belongs.

Thus, a total of **102 antibiotics** that were found in the ARDB and CARD databases were further consolidated into **30 antibiotics categories**.

2.3 Technical details

2.3.1 PMID: 23877117

- All **non-redundant genes** in each metagenomic data set were aligned with these resistance proteins using BLASTx with a E-value threshold of 1e-10 and query coverage of at least 70%.

2.3.2 PMID: 25003965

- a protein was called an AR protein if it had > 80% amino acid identify over 85% of the length of the target sequence.
- ORFs: > 90 amino acids (Figure 1)

2.3.3 PMID: 27411009

- Vancomycin, but not amoxicilin, decreased bacterial diversity and reduced Firmicutes involved in short-chain fatty acid and bile acid metabolism, concomitant with altered plasma and/or fecal metabolite concentrations.
- VANCO decreased the relative abundance of mainly Gram-positive bacteria of the Firmicutes phylum. Along the most strongly affected groups were genus-like groups that contain known butyrate-producing species from Clostridium clusters IV and XIVa, such as Coprococcus eutactus, Faecalibacterium prausnitzii, and Anaerostipes caccae, as well as species involved in BA dehydroxylation such as Clostridum leptum. Conversely, Gram-negative Proteobacteria, members of Clostridium cluster IX and VANCO-resistant Gram-positive Bacilli such as Lactobacillus plantarum and Enterococcus, showed increased relative abundance after VANCO treatment.

- VANCO inhibits GA conversion and SCFA production. This was accompanied by an increase of fecal primary BAs. (soga)

2.4 Reference

PMID: 27572443: Developmental dynamics of the preterm infant gut microbiota and antibiotic resistome

PMID: 30001517: Strain-Level Analysis of Mother-to-Child Bacterial Transmission during the First Few Months of Life

PMID: 30349083: Recovery of gut microbiota of healthy adults following antibiotic exposure

PMID: 29391044: DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data