Overview

The process begins with one fasta file and multiple GenBank files. The fasta file is the local download of the PlasmidFinder database reference in our paper (Card et al. 2019). The GenBank files come from the Entrez search strategy also described in the paper. The ultimate output is a CSV file and text-based report file for each input GenBank file. The CSV file contains basic information (e.g., plasmid length), the incompatibility group(s) the plasmid best aligns to, accession numbers of identical plasmids, some gene/function annotation based on key term searches of the GenBank file's CDS regions, and some other metadata extracted from the GenBank files. The text-based report is a file containing various information and statistics about each group of plasmids from the various input GenBank files. We also generated a tree to help visualize the identical plasmids.

Our process occurs stepwise, with most steps requiring the output from the previous steps. As our project developed, additional steps were inserted or modified. While most steps do depend on the output of the previous step(s), the order is in many instances arbitrary. The code is published online in this GitHub repository (https://github.com/ridgelab/plasmidCharacterization). A detailed description of the steps with their respective inputs and outputs is also maintained in the repository. Accordingly, we describe here a conceptual overview and note a few key things about the data. Each output CSV file requires the following input processed from the "raw" input data (in no particular order): (a) a list of identical plasmids for each accession, (b) extracted metadata from the GenBank files, (c) gene/function annotations extracted from the GenBank files, and (d) a list of incompatibility groups. Each output statistics file is created based on each CSV file just described.

Identical Plasmids. First, a blast database was created with makeblastdb. Each plasmid sequence (which would have to be extracted from the GenBank files) is aligned with blastn to each other plasmid sequence in a pairwise fashion. Hits were kept only if the percent identity was >=98%. Plasmids were considered identical if the hits covered >=98% of both the query and the subject sequence. We created a tree using a simple distance metric to help visualize the identical plasmids. The distance metric is the sum of the query and subject covered bases divided by the sum of the length of the query and subject sequences. The Newick formatted tree was made from the distance matrix using the makeNewick.py script from CAM (Miller et al. 2019) and is available on GitHub at https://github.com/ridgelab/cam. makeblastdb and blastn are part of the BLAST+ Suite (Altschul et al. 1990; Camacho et al. 2009).

GenBank Metadata. The sequencing technology used to sequence each plasmid was identified with GNU AWK. The remaining metadata was also obtained from the GenBank files using GNU AWK. The remaining data points are as follows: country of origin for the plasmid, isolation source for the plasmid, plasmid collection data, and source organism.

GenBank Annotations. This is by far the most complicated part of the process. First, search regions were extracted from the GenBank files. The search regions were the function, gene, note, and product sections of the CDS features. We then identified matches in these regions to key terms (these key terms were obtained as described in our paper). The search occurred under the following strategy:

The search terms are each part of one or more categories. Each can belong to multiple categories, but only if the categories are subsets of each other. Five principal categories exist, two of which have subcategories. The category structure is as follows:

- Antimicrobial Resistance
  - Beta-lactamase
    - Beta-lactamase Special
- Toxin/Antitoxin System
- DNA Maintenance/Modification
  - DNA Maintenance/Modification Special
- Mobile Genetic Elements
- Hypothetical Genes

The strategy could be described as top-to-bottom, in-to-out; i.e., Antimicrobial Resistance is more important that Toxin/Antitoxin System and Beta-lactamase Special is more important than Beta-lactamase and Antimicrobial Resistance. The reason these are shown nested instead of simply above their parents is because a match for a Beta-lactamase Special search term will increment the count for not only itself, but also its parents. If no matches are found, the CDS region being searched is classified as "Other". Some CDS regions will never be searched for these terms if they first match a term in a special "Ignored" category. Provided a CDS region is not to be ignored, it will be searched with Beta-lactamase Special terms, then Beta-lactamase terms, then Antimicrobial Resistance Terms, then Toxin/Antitoxin System terms, and so-forth, until a match is found (thus halting the search on this CDS region) or no more search terms remain, in which case it is assigned to the "Other" category. All CDS regions are converted to lowercase before being searched as described. These terms are listed, with their associated Python regular expressions, in the supplement of our paper.

Incompatibility Groups. The incompatibility fasta sequences were downloaded from the PlasmidFinder database as previously described. This was turned into a database using makeblastdb. Each plasmid sequence was then aligned to the database using blastn and hits were retained only if the percent identity was >=80%. Hits were further dropped if the subject (the sequences in the database) coverage was <60%. The “best” hits were then used to determine which incompatibility group(s) applied to each plasmid. “Best” is defined as the result(s) with the highest percent identity and those that have percent identities within only 1 percent of the highest one. makeblastdb and blastn are part of the BLAST+ Suite (Altschul et al. 1990; Camacho et al. 2009).

A comment on data availability

The version of the PlasmidFinder database that we downloaded is no longer available. Accordingly, we release the fasta file we downloaded for reproducibility purposes. However, we advise a fresh download for any new experiments. This file may be found in the repository at

the following path: `data/original_incompatibility_groups/incompatibility.fasta`. Similarly, many GenBank files have been updated since our download on 1 March 2018. We likewise release the versions we downloaded here for reproducibility purposes. However, we recommend fresh downloads of these files for new analyses. A script (labelled as "Step 0") is released with the online code repository for such a purpose. Please note that additional plasmids could (and should) be included now if the Entrez search strategy were to be re-done. The script would not reflect such changes as it downloads the specific GenBank groupings we used with accession numbers, completely ignoring the Entrez strategy. This is appropriate for reproducing our results, but it would probably not be ideal for a future study.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic Local Alignment Search Tool. Journal of Molecular Biology **215**: 403-410. doi:10.1016/S0022-2836(05)80360-2.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. 2009. BLAST+: architecture and applications. BMC Bioinformatics **10**: 421. doi:Artn 421\nDoi 10.1186/1471-2105-10-421.

Card, G., Pickett, B., Ridge, P., and Robison, R. 2019. Characterization of Carbapenem-Resistance Plasmids. In press.

Miller, J.B., McKinnon, L.M., Whiting, M.F., and Ridge, P.G. 2019. CAM: An alignment-free method to recover phylogenies using codon aversion motifs. In Press.