Rowan Christie
5/7/2020

DNA content in *tetM* genes affects antibiotic resistance in bacteria

One of the leading public health concerns that has arisen in the past few decades is antibiotic resistance. In the scope of this paper, antibiotic resistance is defined as the ability of bacteria to survive or grow in inhibitory antibiotic concentrations. Bacteria are becoming increasingly resistant to antibiotics, leading to greater mortality (Cosgrove 2003). This problem is recognized by several international public health organizations as a danger to public health, which agree that it is essential to study to monitor risks to public health (Cassini 2019). To do this, genetic markers of resistance are analyzed from bacteria grown in the presence of antibiotics.

One such gene that has been found to be of concern is the *tetM* gene, which encodes tetracycline resistance (Doenhoefer 2012). Tetracyclines are a common antibiotic that inhibit bacterial growth by binding to elongating ribosomes to prevent or slow delivery of GTP, ternary complex EF-Tu, and aminoacylated tRNA to the A-site (Wilson 2009). However, bacteria that are tetracycline resistant often possess ribosome protection proteins that bind to the ribosome and prevent the antibiotic from reaching the binding site (Chopra 2001). These bacteria have risen in number in the past few decades with the use of this antibiotic, increasing tetracycline resistance determinants and limiting the drug (Roberts 2005). As a result, there have been many studies that have been conducted that have sequenced the *tetM* gene, many of which are available in NCBI - a search of the NCBI Nucleotide database with the keyword *tetM* shows that there are over 36000 strains that have been submitted with this gene. Alongside these sequences, many of these studies also share other data on these strains such as minimum inhibitory concentration (MIC), which confers the level of antibiotic resistance, source of bacteria, location, and antibiotic testing methods. Due to this wealth of information, it is possible to do a meta-analysis, which is a statistical procedure for compiling data with multiple variables from multiple studies to identify common effects and determine possible reasons for variation in the data. By synthesizing the data across multiple studies, we can get a sense of the bigger picture by capturing a much bigger scope and variation of data, rather than data from a single study, as results often vary between different studies.

There have been other studies that have performed meta-analysis in from an antibiotic resistance context and have found significant results which have informed public health policies (Bell 2014; Mather 2019; Zhen 2019). For example, a meta-analysis performed in China helped prove that antibiotic resistance is often correlated with significantly longer hospital stays and costs, bringing awareness to the issues antibiotic resistance causes (Zhen 2019). This proves that meta-analysis can be a powerful tool for studying antibiotic resistance genes.

I aim to explore how, in regards to bacteria with *tetM* genes, the phylogeny, MIC, guanine-cytosine (GC) composition, source (such as cattle or fermented food) of bacteria and

other possible variables are related. How does MIC respond to percent identity, GC % content or amino acid % in *tetM* genes in different bacterial strains? Which species of bacteria with the tetM gene have the highest antibiotic resistance? Which bacterial sources lead to the highest antibiotic resistance? Are more recent bacterial species/strains more dangerous? Conducting a meta-analysis of sequences with the tetM gene will enable me to collect the data I need to discover the answers to my questions.

First, I created a spreadsheet outlining the variables I wanted to collect data on, which included taxon, country, bacterial source, and MIC. I searched the NCBI Nucleotide database using the keyword *tetM*, which returned over 36000 records. For each item returned, I recorded the number of records found at the time, NCBI item number (the order the sequence was in the list of items returned), date, accession number, strain title and link to strain sequence, study title and link to study (if there was a corresponding study that could be found), and whether or not the study included MIC data on the strain. If not, I excluded the strain from my analysis, although I left it in the spreadsheet and literature review as evidence of my workflow. I also explained how I found the MIC data in the corresponding studies in the literature review, where I detailed how each search was conducted. For studies that did measure MIC for the specific strain, I recorded taxon, growth conditions, antibiotic susceptibility testing methods, coding strand (complete or partial), species, strain, base pairs, country of origin, date, source of bacteria, and MIC for the specified antibiotic. Here is the spreadsheet. Currently, I have looked at 96 records, and there are 37 records from 8 studies I plan to include. However, I still feel that an analysis of this scale would benefit from additional data to strengthen its accuracy, and would more clearly demonstrate the potential effect of factors such as bacterial source, since each source listed has very few sample sizes. I plan to continue going through the NCBI records and add them to the spreadsheet. For now, I have conducted some calculations with the observations I currently do have.

In order to calculate percent identity, I used the NCBI BLAST with the blastn program, which searches nucleotide subjects against a nucleotide query. I needed a reference query to input into BLAST, so I decided to use the reference *tetM* gene from the Comprehensive Antibiotic Resistance Database (CARD), which was from a Staphylococcus aureus strain. The subjects for the BLAST were the accession numbers of strains from the spreadsheet that had MIC data. After the BLAST had finished running, I copied the results for the percent identity into the spreadsheet, sorting by accession number to ensure the values were correctly lined up.

I calculated the CG content of the sequences in R by using the *genbankr* library from the package Bioconductor, which allowed me to access the sequences from their accession numbers (Becker 2020). Usually, the sequences in NCBI will have DNA that is not part of the *tetM* coding strand. I subsetted the sequence data to only include the coding strand that encoded tetM. I used the Bioconductor function *alphabetFrequency()* to determine the frequency of C and G letters, allowing me to calculate percent GC content.

I also wanted to examine the phylogeny of the *tetM* genes, to see how MIC and bacterial species were related to each other. To do this, I needed to create a phylogenetic tree. First, I used the msa package to perform the multiple sequence alignment with the msa() function (Bodenhofer 2015). I used the default model, ClustalW. To convert the multiple sequence alignment to a PhyDat class to make it a usable tree format, I used the msaConvert() function with the parameter type equal to "phangorn::phyDat." To analyze the phylogeny, I used the Phangorn package, which allows for testing of several models and comparing phylogeny using Maximum Likelihood, Maximum Parsimony, and distance methods (Schliep 2017).

To develop an estimation of the phylogenetic tree, I created a distance matrix using the dist.ml() function, reducing sequence alignments into a matrix of pairwise distances. The Jukes Cantor model (1969) was used for our distance matrix in this case, and it is the simplest model of evolution that assumes that a base that mutates has an equal chance of being substituted by any base (Cantor 1969). With this distance matrix, we can test different algorithms on it to determine the best fit for the data.

There are several different ways to estimate phylogenetic trees from distance matrices — two of them are the Unweighted Pair-Group Method with Arithmetic Averaging (UPGMA) and Neighbor-Joining. UPGMA will assume that the rates of evolution are the same so that distances are additive and ultrametric, while Neighbor-Joining can allow for unequal rates of evolution and only requires that distances are additive (Sokal 1958; Saitou 1987). We tested both on our distance matrix using the functions *upgma()* and *NJ()*. To determine which algorithm best fit the data, I calculated the parsimony score. The algorithm with the maximum parsimony will usually be the best fit for the data. To measure the parsimony score, I ran the *parsimony*() function for each algorithm result. The UPGMA algorithm returned a parsimony score of 329 while the Neighbors-Joining algorithm returned a parsimony score of 318. Thus, UPGMA algorithm should be the best fit for the data due to having the maximum parsimony score.

As I wanted the tree to be optimized as much as possible, I aimed to run the Maximum likelihood model. This model helps determine the parameters of the distribution that can best describe the data, meaning that this model is useful for optimizing the branch length and the tree topology. To do this, I ran the *pml()* function with the UPGMA algorithm result as the first parameter and the tree structure as the second parameter to compute the likelihood of a given tree and create a fitted model. This fitted model was then run in the optim.pml() function as the first parameter, and for the other parameters, model was equal to "JC" (Jukes Cantor model 1969) model and rearrangement was equal to "stochastic."

The tree structure of the maximum likelihood optimization was run using the *ggtree* package to visualize the tree (Yu 2020; Yu 2018; Yu 2017). Ggtree has many useful features for visualization, including annotation, which is why it is the visualization tool of choice. I adjusted node size and color based on MIC and species data.

To determine how MIC responded to CG, percent identity, and year, I calculated the correlation between the MIC for tetracycline, percent identity, percent GC content, and year using the *cor()* in R. I also used the *t.test()* function to determine the significance of relationships. While MIC and percent identity ($r^2 = 0.2859238$) and MIC and year ($r^2 = -0.4358124$) were not strongly correlated, MIC and percent CG content were ($r^2 = -0.859289$). This correlation was also significant (t = -4.3667, df = 36, p-value = 0.0001021). I then decided to create a figure representing this relationship using the ggplot2 package in R (Figure 1) (Wickham 2016).

I find correlation between CG content and MIC quite interesting because it implies that CG content in *tetM* genes drives MIC. So for example, if you knew of a bacteria species that had a *tetM* gene with lower CG content on average, you could predict that it would have higher tetracycline MIC, and thus be harder to inhibit with tetracycline. I explored this with Figure 2, and noticed that there were some species, such as *Neisseria gonorrhoeae* that tended to have a higher GC content, but had lower MIC, supporting the prediction that species with higher GC have lower MIC, meaning they are less antibiotic resistant. So, just from this Figure 2, I can reasonably infer that *Neisseria gonorrhoeae* is unlikely to be highly resistant to tetracycline, due to its higher GC content. With this information, we could potentially identify many species that are more likely to be dangerous due to their CG content in their tetM gene. Figure 3 shows the phylogenetic relationship between the tetM genes, and their corresponding tetracycline MIC through node size. What is interesting is that some *tetM* genes in *Neisseria gonorrhoeae* strains appear to be more distantly related than *tetM* genes in other species, such as *Lactobacillus sakei*. This could imply that horizontal transfer is occurring between different bacteria species. Different clades may also have significant differences in tetracycline MIC - Clade 1 and 3 seem to have lower MIC than Cade 2, according to the node size. Clade 4 also indicates strains with greater MIC than strains in Clade 1 and 3. This could suggest that *tetM* gene phylogeny and MIC are related.

Overall, this analysis indicates that there is valuable information to be gained in researching the differences in tetM genes and their corresponding MIC. There was a strong correlation found between CG content in tetM genes and tetracycline MIC, and the phylogeny of the tetM strain may have a significant impact on how resistant a strain is. The next steps I would like to take with this analysis would be calculate the percentage of each amino acid translated and see how that might correlate with MIC. I am very interested to see this, considering how CG content was found to have a strong relationship with MIC. I would also like to look at other potential factors such as methods, year, location, local climate, source as well. Additionally, I would also like to keep adding data to strengthen this analysis and help me explore the effects of factors such as bacterial source.

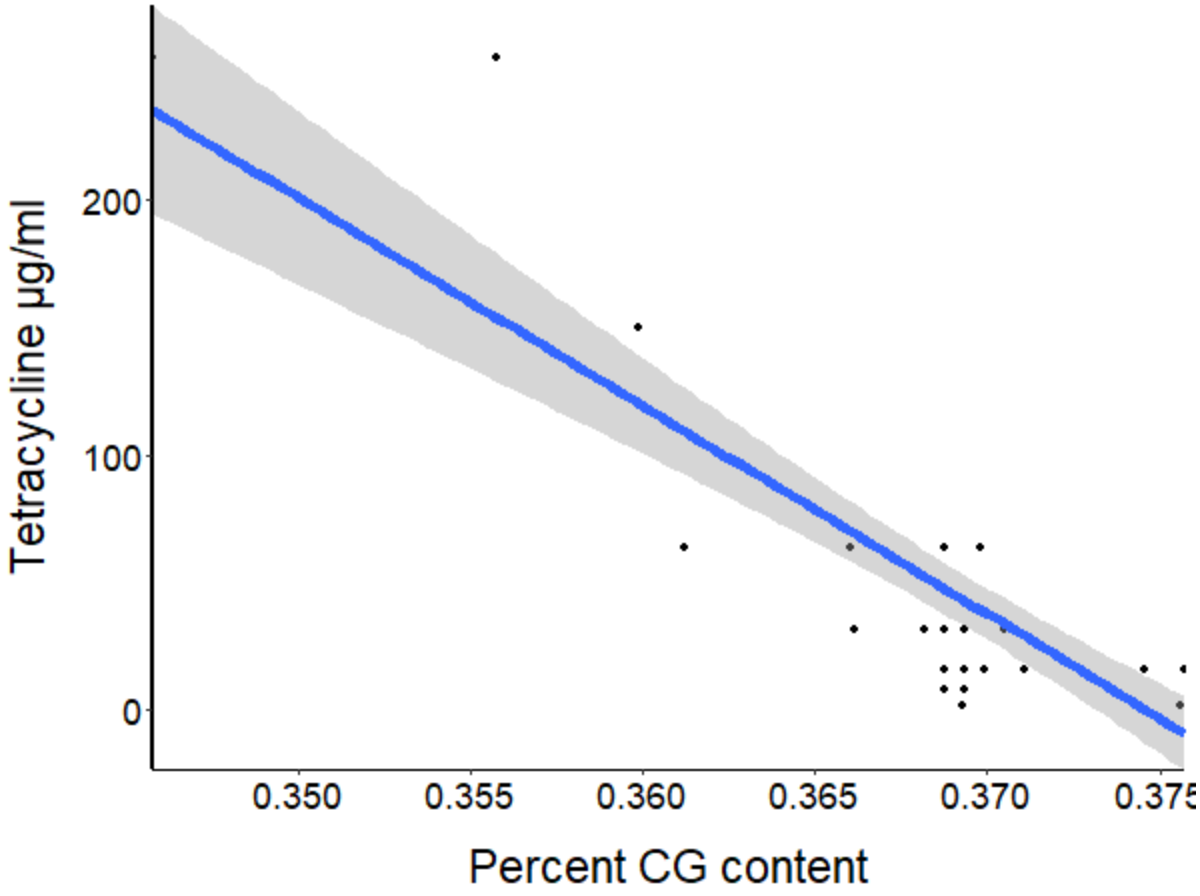

Figure 1. Relationship between MIC for tetracycline µg/ml and percent GC content in tetM genes. The blue line represents the trendline, which shows the MIC as a function of percent GC content. The gray area represents the 95% confidence level interval.

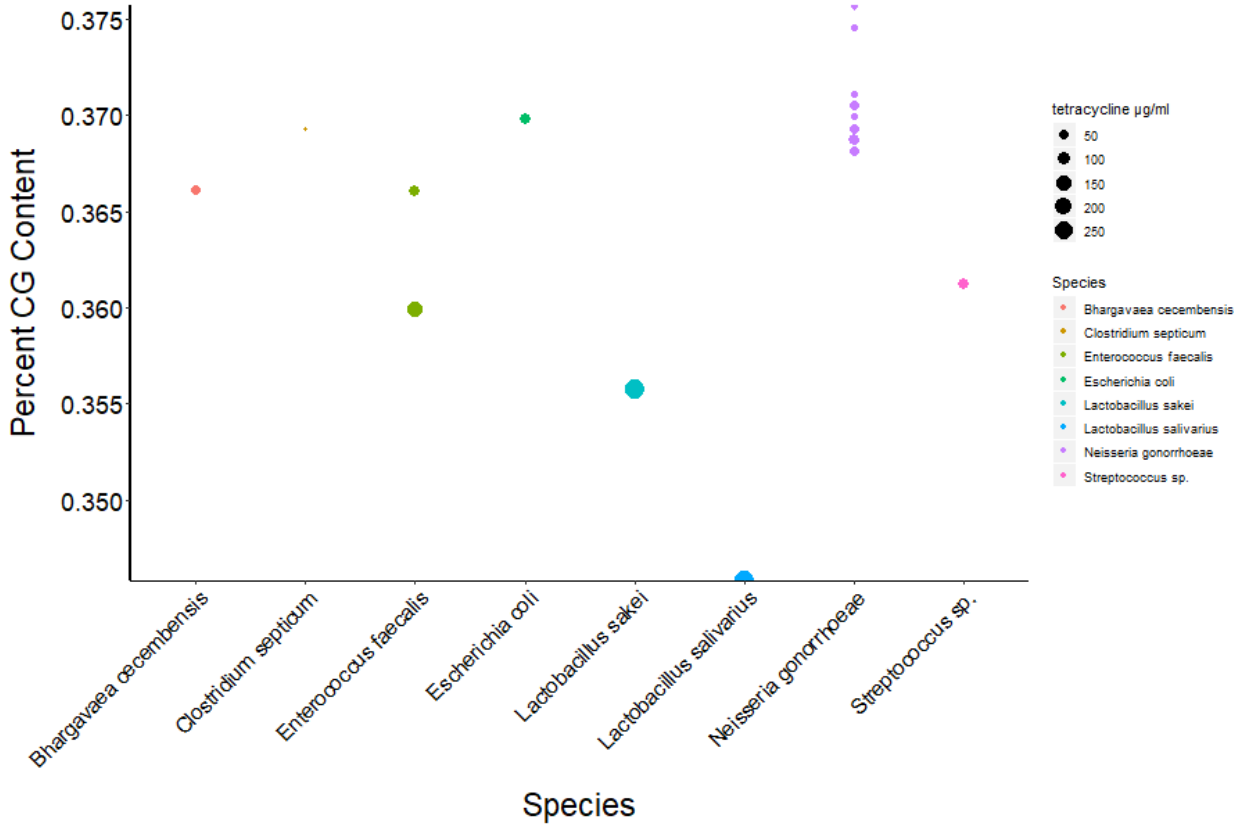

Figure 2. Comparison of percent CG content and tetracycline µg/ml for each bacterial species with the *tetM* gene. The size of the dots represent the level of MIC, while the color indicates the species.

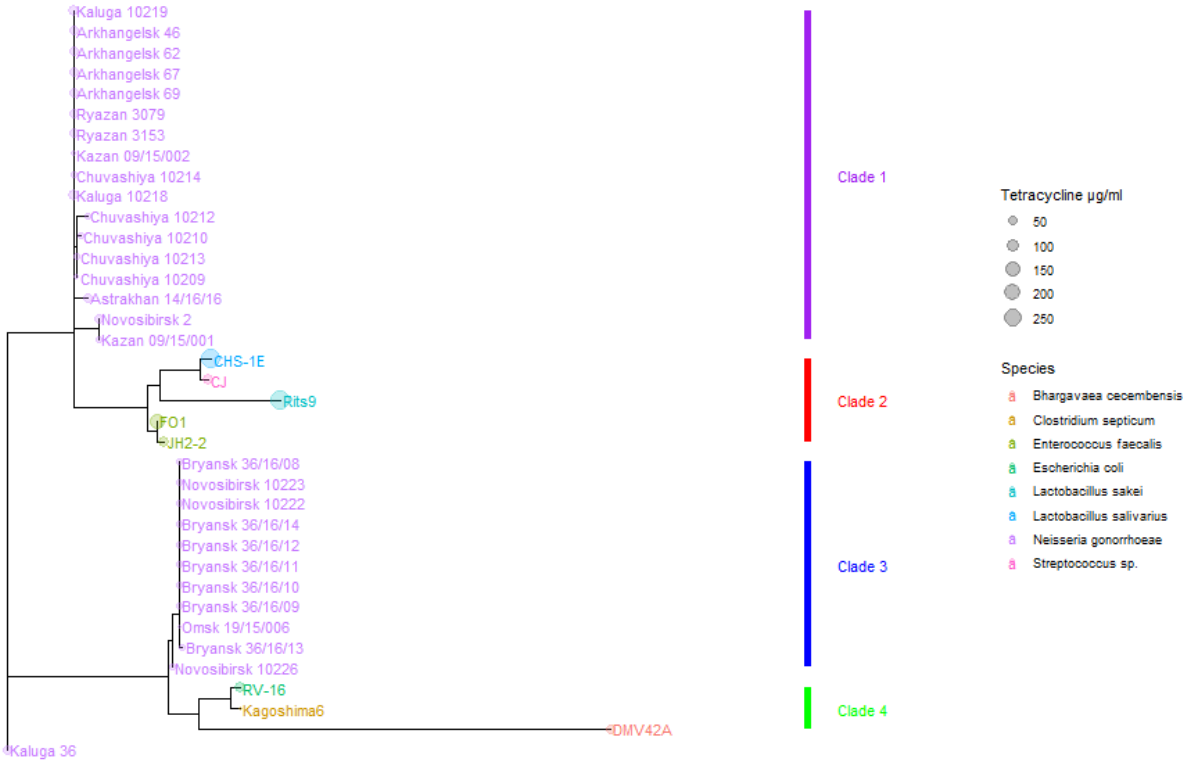

Figure 3. Phylogenetic tree of *tetM* genes in various bacteria species. Size of the nodes represents the MIC for tetracycline µg/ml, color of the nodes represents the bacterial species of the node, and node label represents the name of the strain.

References

Becker G, Lawrence M (2020). genbankr: Parsing GenBank files into semantically useful objects. R package version 1.17.0.

Bell, B. G., Schellevis, F., Stobberingh, E., Goossens, H., & Pringle, M. (2014). A systematic review and meta-analysis of the effects of antibiotic consumption on antibiotic resistance. BMC infectious diseases, 14, 13. https://doi.org/10.1186/1471-2334-14-13

Bodenhofer U, Bonatesta E, Horejs-Kainrath C, Hochreiter S (2015). "msa: an R package for multiple sequence alignment." Bioinformatics, 31(24), 3997–3999. doi: 10.1093/bioinformatics/btv494.

Cassini, A. et al. (2019) Attributable deaths and disability adjusted life-years caused by infections with antibiotic-resistant bacteria in the EU and the European Economic Area in 2015: a population-level modelling analysis. Lancet. Infect. Dis. 19, 56–66.

Chopra, I. & Roberts, M. (2001) Tetracycline antibiotics: Mode of action, applications, molecular biology, and epidemiology of bacterial resistance. Microbiol Mol Biol Rev 65(2):232–260.

Cosgrove, S. E. & Carmeli, Y. (2003). The impact of antimicrobial resistance on health and economic outcomes. Clin. Infect. Dis. 36, 1433–1437.

Wilson, D. N. (2009) The A-Z of bacterial translation inhibitors. Crit Rev Biochem Mol Biol 44(6):393–433.

Roberts, M. C. (2005) Update on acquired tetracycline resistance genes. FEMS Microbiol Lett 245(2):195–203.

Doenhoefer, A., Franckenberg, S., Wickles, S., Berninghausen, O., Beckmann, R., & Wilson, D. (2012). Structural basis for TetM-mediated tetracycline resistance. *PNAS*, *109*(42), 16900–16905. doi: 10.2210/pdb3j25/pdb

Haubert, L., Cunha, C. E. P. D., Lopes, G. V., & Silva, W. P. D. (2018). Food isolate Listeria monocytogenes harboring tetM gene plasmid-mediated exchangeable to Enterococcus faecalis on the surface of processed cheese. *Food Research International*, *107*, 503–508. doi: 10.1016/j.foodres.2018.02.062

Kim, Y., Jun, L., Park, S., Yoon, S., Chung, J., Kim, J., & Jeong, H. (2007). Prevalence of tet(B) and tet(M) genes among tetracycline-resistant Vibrio spp. in the aquatic environments of Korea. *Diseases of Aquatic Organisms*, *75*, 209–216. doi: 10.3354/dao075209

Mather, M. W., Drinnan, M., Perry, J. D., Powell, S., Wilson, J. A., & Powell, J. (2019). A systematic review and meta-analysis of antimicrobial resistance in paediatric acute otitis media. *International Journal of Pediatric Otorhinolaryngology*, *123*, 102–109. doi: 10.1016/j.ijporl.2019.04.041

Perreten, V., Kollöffel, B., & Teuber, M. (1997). Conjugal Transfer of the Tn916-like Transposon TnFO1 from Enterococcus faecalis Isolated from Cheese to Other Gram-positive Bacteria. *Systematic and Applied Microbiology*, *20*(1), 27–38. doi: 10.1016/s0723-2020(97)80045-8

Jukes, T. H., Cantor CR (1969). Evolution of Protein Molecules. New York: Academic Press. pp. 21–132.

Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution. 4 (4): 406–425. doi:10.1093/oxfordjournals.molbev.a040454

Sasaki, Y., Yamamoto, K., Tamura, Y., & Takahashi, T. (2001). Tetracycline-resistance genes of Clostridium perfringens, Clostridium septicum and Clostridium sordellii isolated from cattle affected with malignant edema. *Veterinary Microbiology*, *83*(1), 61–69. doi: 10.1016/s0378-1135(01)00402-3

Schliep, K., Potts, A. J., Morrison, D. A., Grimm, G. W. (2017), Intertwining phylogenetic trees and networks. Methods in Ecology and Evolution, 8: 1212--1220. doi:10.1111/2041-210X.12760

Sokal, M. (1958). A statistical method for evaluating systematic relationships. University of Kansas Science Bulletin. 38: 1409–1438.

Shaskolskiy, B., Dementieva, E., Leinsoo, A., Petrova, N., Chestkov, A., Kubanov, A., … Gryadunov, D. (2018). Tetracycline resistance of Neisseria gonorrhoeae in Russia, 2015–2017. *Infection, Genetics and Evolution*, *63*, 236–242. doi: 10.1016/j.meegid.2018.06.003

Thumu, S. C. R., & Halami, P. M. (2012). Presence of erythromycin and tetracycline resistance genes in lactic acid bacteria from fermented foods of Indian origin. *Antonie Van Leeuwenhoek*, *102*(4), 541–551. doi: 10.1007/s10482-012-9749-4

You, Y., Hilpert, M., & Ward, M. J. (2013). Identification of Tet45, a tetracycline efflux pump, from a poultry-litter-exposed soil isolate and persistence of tet(45) in the soil. *Journal of Antimicrobial Chemotherapy*, *68*(9), 1962–1969. doi: 10.1093/jac/dkt127

Yu, G. (2020). Using ggtree to Visualize Data on Tree-Like Structures. Current Protocols in Bioinformatics, 69(1), e96. doi: 10.1002/cpbi.96,

Yu, G., Lam, T. T., Zhu, H., & Guan, Y. (2018). Two methods for mapping and visualizing associated data on phylogeny using ggtree. Molecular Biology and Evolution, 35, 3041-3043. doi: 10.1093/molbev/msy194

Yu, G., Smith, D., Zhu, H., Guan, Y., & Lam, T. T. (2017). ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution, 8, 28-36. doi: 10.1111/2041-210X.12628

Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.

Zhen, Lundborg, Sun, Hu, & Dong. (2019). The Clinical and Economic Impact of Antibiotic Resistance in China: A Systematic Review and Meta-Analysis. *Antibiotics*, *8*(3), 115. doi: 10.3390/antibiotics8030115