Research Note: Choice of microbiota database affects data analysis and interpretation in chicken cecal microbiota

Philip M. Campos ,[*,†,‡] Nadia Darwish,[*,†,‡] Jonathan Shao,[†] and Monika Proszkowiec-Weglarz [‡,1]

[*]*Oak Ridge Institute for Science and Education (ORISE) USDA-ARS Research Participation Program, Oak Ridge, TN, USA;* [†]*USDA-ARS, NEA Bioinformatics, Statistics Group, Beltsville, MD, USA; and* [‡]*USDA-ARS, NEA, Beltsville Agricultural Research Center, Animal Biosciences and Biotechnology Laboratory, Beltsville, MD 20705, USA*

ABSTRACT The chicken microbiota is often analyzed to address questions about the effects of diet or disease on poultry health. To analyze the microbiota, bioinformatic platforms such as QIIME 2 and mothur are used, which incorporate public taxonomic databases such as Greengenes, the ribosomal database project (**RDP**), and SILVA to assign taxonomies to bacterial sequences. Many chicken microbiota studies continue to incorporate the Greengenes database, which has not been updated since 2013. To determine whether a choice of database could affect results, this study compared the results of bioinformatic analyses obtained using the Greengenes, RDP, and SILVA databases on a cecal luminal microbiome dataset. The QIIME 2 platform was used to process 16S bacterial sequences and assign taxonomies with Greengenes, RDP, and SILVA. Linear discriminant analysis effect size (**LEfSe**) was performed, allowing for the comparison of taxonomies considered significantly differentially abundant between the three databases. Some notable differences between databases were observed in results, in particular the ability of SILVA database to classify members of the family Lachnospiraceae into separate genera, while these members remained in one group of unclassified Lachnospiraceae through Greengenes and RDP. LEfSe analyses showed that the SILVA database produced more differentially abundant genera, in large part due to the classification of these separate Lachnospiraceae genera. Additionally, the relative abundance of unclassified Lachnospiraceae in SILVA results was significantly lower than in RDP results. Our results show the choice of taxonomic database can influence the results of a microbiota study at the genus level, potentially affecting the interpretation of the results. The use of the SILVA database is recommended over Greengenes in chicken microbiota studies, as more specific classifications at the genus level may provide more accurate interpretations of changes in the microbiota.

Key words: microbiota, Greengenes, RDP, SILVA, broiler chickens

2022 Poultry Science 101:101971
https://doi.org/10.1016/j.psj.2022.101971

INTRODUCTION

The chicken microbiota is often analyzed in studies addressing the effects of diet or disease on the health of poultry. The bacteria that make up the chicken gastrointestinal tract (**GIT**) microbiota can influence nutrient exchange, digestive system physiology, immune system modulation, and pathogen exclusion within hosts (Stanley et al., 2014). The bacterial composition of microbiota may be affected by host-related factors, including age, sex, breed, and the location of the GIT, in addition to environmental factors, including biosecurity level, housing, litter, feed access, and climate (Kers et al., 2018). A healthy gut microbiota may assist in limiting the spread of diseases such as coccidiosis, a parasitic disease that affects poultry production by causing weight loss and reduced efficiency in feed use. The microbiota of the cecum is thought to play a role in response to diseases, with the cecum being associated with the production of polysaccharides and short-chain fatty acids (Stanley et al., 2014).

Developments in next-generation sequencing and bioinformatics have led to more widespread use of these techniques to analyze GIT microbiota data. Microbiota can be characterized using bacterial 16S rRNA, which contains 9 hypervariable regions where similarities and differences between species are determined and highly conserved regions where polymerase chain reaction

Received March 2, 2022.
Accepted May 13, 2022.

Mention of trade name, proprietary product, or specific equipment does not constitute guarantee or warranty by USDA and does not imply its approval to the exclusion of other suitable products.

[1]Corresponding author: monika.weglarz@usda.gov

(**PCR**) primers can be attached. Despite the usefulness of 16S rRNA sequencing, differences in DNA isolation, library preparation, and PCR methodology, such as the difference in the choice of PCR primers in chicken microbiota studies (Darwish et al., 2021), can introduce biases that affect the representation of bacterial groups.

In addition to biases introduced during sequencing, choices made during bioinformatic processing may also influence taxonomic composition results. Bioinformatic processing is typically performed using an open-source platform such as QIIME 2 (Quantitative Insights Into Microbial Ecology) (https://qiime2.org/) or mothur (https://mothur.org/), which involves a step where taxonomy is assigned to bacterial sequences found in samples. Taxonomic assignments are reliant on public databases such as Greengenes (https://greengenes.secondgenome.com/), the ribosomal database project (**RDP**) (http://rdp.cme.msu.edu/), and SILVA (https://arb-silva.de/). The majority of studies on the chicken microbiota rely on classifications by the Greengenes database, which poses an advantage in allowing for the comparison of results between different studies. However, there may be a disadvantage in the continued use of Greengenes in future studies, as the Greengenes database was last updated in August 2013, potentially leading to studies presenting less accurate results than if they used RDP or SILVA, both of which have released updated versions in 2020.

Of the three databases, SILVA is the largest based on 16S taxonomies, followed by RDP and lastly Greengenes (Balvočiūtė and Huson, 2017). SILVA is commonly used as a reference database in modern microbiota studies of other systems, for example, the gut microbiota of humans and other animals, as well as non-animal systems such as soybean and soil microbiota. The aim of this study was to compare taxonomic classifications of bacteria and relative abundance results from poultry cecal microbiota by the Greengenes, RDP, and SILVA databases. This study uses a cecal luminal microbiota dataset to demonstrate the differences in results that may occur from choosing one database over the other.

MATERIALS AND METHODS

Animal care, experimental design, DNA isolation, and DNA sequencing procedures were performed as described in Campos et al. (2022). The 16S rRNA gene sequences determined in this study were deposited in the NCBI Sequence Read Archive database (SRA accession #PRJNA736980). The QIIME 2 platform (https://qiime2.org/) version 2020.11 was used to perform microbiome bioinformatics. Demultiplexed, paired-end sequence data from 48 samples was denoised with DADA2 via the q2-dada2 plugin using a quality cutoff of 25. Feature classifiers for each database were trained with q2-feature-classifier fit-classifier-naive-bayes using the Greengenes 13_8 97% OTUs reference sequences and taxonomy, the RDP Release 11 unaligned Bacteria 16S reference sequences and taxonomy, and the SILVA 138 99% OTUs reference sequences and taxonomy. The SILVA reference sequences and taxonomy were obtained as pre-formatted files that were processed using RESCRIPt from the QIIME 2 data resources page (https://docs.qiime2.org/2022.2/data-resources/). The RESCRIPt process involves filtering sequences based on length of the amplicon and associated species, ambiguous nucleotide content, and/or homopolymers, and dereplicating sequences, a method that removes duplicate sequences that may be assigned different taxonomies, resulting in fewer inconsistencies and improved processing (Robeson et al., 2021). While running RESCRIPt manually may introduce bias depending on selected parameters (e.g., if the minimum length cutoff for an amplicon is too long, the dataset may be biased towards bacteria that have been fully sequenced), the files provided by the QIIME 2 developers are a way of standardizing this process. Taxonomy was assigned to amplicon sequence variants (**ASVs**) using the q2−feature−classifier classify−sklearn naïve Bayes taxonomy classifier.

To analyze relative abundance data produced by each database, feature tables were collapsed to the genus taxonomic level via q2-taxa, where ASV counts were normalized by total sum scaling normalization. Centered-log ratio transformation was applied to relative abundance data in R 4.0.3 (https://r-project.org/). Statistical analysis was performed using ANOVA and Tukey's Honest Significant Difference (**Tukey HSD**) test to determine significance between relative abundance data of taxonomic groups in the Greengenes, RDP, and SILVA databases. Where Greengenes, RDP, and SILVA produced the same taxonomic classification (e.g., results from both databases contained unclassified Lachnospiraceae), relative abundances of the same classification were compared. In other cases, identical ASVs were found to be classified as different taxa (e.g., Greengenes classified a group of ASVs as *Faecalibacterium* while SILVA classified this group as *Subdoligranulum*). Feature IDs representing DNA sequences were cross-referenced between the QIIME 2 taxonomy results of the three databases to confirm that different classifications were being produced by the databases for identical DNA sequences, and the relative abundances of these classifications were compared.

To determine whether the choice of database could affect other analyses that use relative abundance data as input, the Linear Discriminant Analysis Effect Size (**LEfSe**) algorithm was performed on data from each database using the Huttenhower Galaxy Server (Galaxy version 1.0, http://huttenhower.sph.harvard.edu/galaxy/). The LEfSe analysis is used to identify taxa with significant differential abundance between groups of 2 or more biological conditions. In the case of this study, the cecal luminal microbiota from 2 groups were compared, with the treatment group including 24 chickens infected with 1.0×10^4 oocysts of the parasitic disease *Eimeria tenella*, and the control group including 24 chickens sham infected with water. Default parameters, including a 0.05 alpha value for the Kruskal-Wallis test and a 2.0

threshold on the logarithmic LDA score for discriminative features, were selected for the analyses.

RESULTS AND DISCUSSION

Bacterial Abundance Overview

The data utilized in this analysis was as reported in Campos et al. (2022), with 4,198,119 reads remaining after sequence quality control on 48 cecal luminal samples, an average of 87,461 reads per sample, an average read length of 428 bp per sample, a total of 521 unique ASVs observed overall, and an average of 148 ASVs observed per sample. At the genus level, the 10 most abundant taxa were *Faecalibacterium*, unclassified Lachnospiraceae (2 separate classifications), [*Ruminococcus*] (names in square brackets are contested names), *Escherichia, Lactobacillus, Oscillospira*, unclassified Clostridiales, *Ruminococcus*, and *Butyricicoccus* according to Greengenes. According to RDP, the 10 most abundant taxa were unclassified Lachnospiraceae (one classification), *Gemmiger, Escherichia-Shigella, Lactobacillus, Anaerobacterium, Butryicicoccus, Clostridium IV*, unclassified Ruminococcaceae, unclassified Bacillales, and *Coprobacillus*. According to SILVA, the ten most abundant taxa were unclassified Lachnospiraceae (one classification), *Subdoligranulum, Escherichia-Shigella*, [*Ruminococcus*] (torques group), *Lactobacillus, Clostridia UCG-014, Eisenbergiella, Erysipelatoclostridium, Butryicicoccus*, and unclassified Oscillospiraceae.

One discrepancy observed in these results that could affect the interpretation of chicken microbiota analyses is the difference in a taxon's name for the same ASV. For example, sequences in our study were classified as *Faecalibacterium* by Greengenes at the genus level were named as *Gemmiger* by RDP and as *Subdoligranulum* by SILVA. These sequences had same average relative abundance of 15.8% in each set of results, with only the taxonomy assigned differing. Recent genetic analysis has suggested *Faecalibacterium prausnitzii* are a separate group from *Gemmiger/Subdoligranulum* (Fitzgerald et al., 2018). Given the RDP and SILVA databases have separate classifications available for *Faecalibacterium* but did not identify any bacteria in our study as *Faecalibacterium*, the Greengenes classification of *Faecalibacterium* may be outdated. Though the *Faecalibacterium* and *Gemmiger/Subdoligranulum* groups are closely related and both are known for butyrate production, *Faecalibacterium prausnitzii* is specifically known for anti-inflammatory effects (Sokol et al., 2008), which could impact the interpretation of results.

Comparisons of Relative Abundance

The relative abundance values of the most abundant taxa were compared between results from the 3 databases where possible (Figure 1). The relative abundance

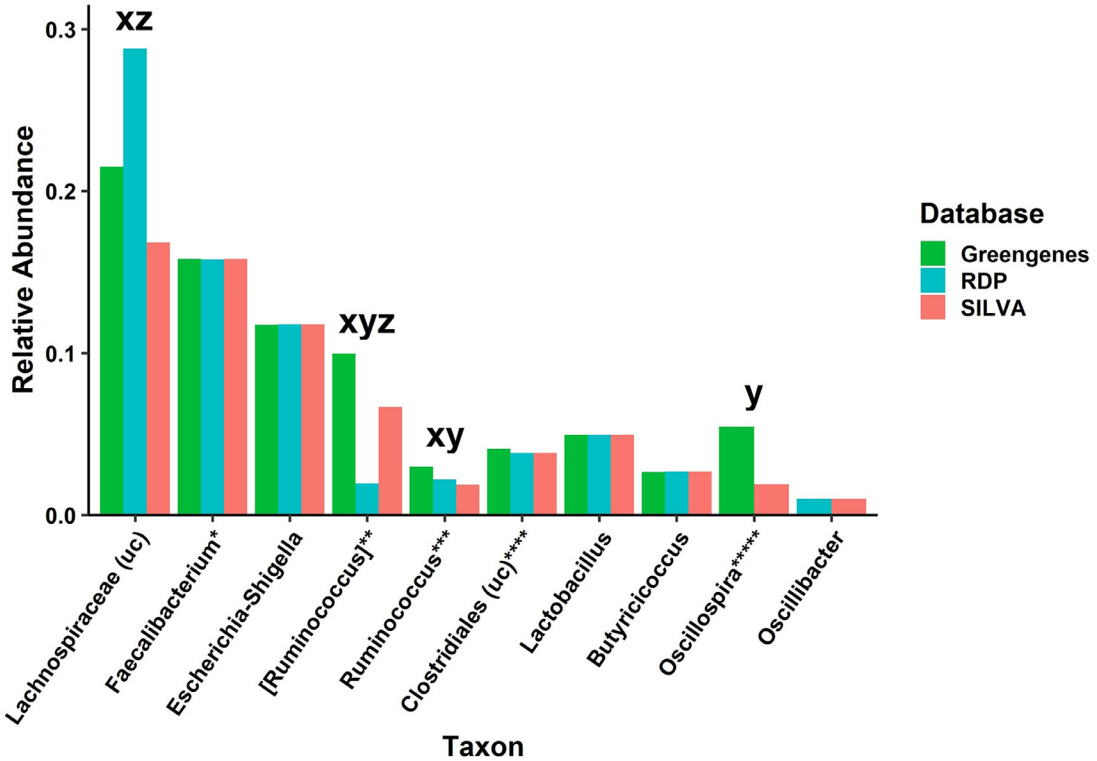

Figure 1. Relative abundance of taxonomic groups compared between the Greengenes, RDP, and SILVA databases. * = *Faecalibacterium* (Greengenes) was classified as *Gemmiger* in RDP and *Subdoligranulum* in SILVA. ** = [*Ruminococcus*] (Greengenes) was compared with *Ruminococcus 2* from RDP and [*Ruminococcus*] *torques group* from SILVA. *** = *Ruminococcus* (Greengenes) was compared with unclassified Ruminococcaceae from RDP and Ruminococcaceae incertae sedis from SILVA. **** = Unclassified Clostridiales (Greengenes) was compared with *Anaerobacterium* from RDP and *Clostridia UCG-014* from SILVA. ***** = *Oscillospira* (Greengenes) was compared with unclassified Oscillospiraceae from SILVA. Significant differences ($P < 0.05$) are denoted with: x = between Greengenes and RDP, y = between Greengenes and SILVA, and z = between RDP and SILVA.

of unclassified Lachnospiraceae differed between databases ($P < 0.001$), with RDP having a significantly higher ($P < 0.001$) average (28.8%) in comparison with both Greengenes (21.5%) and SILVA (16.8%), though Greengenes was not considered significantly higher than SILVA ($P > 0.05$). SILVA producing the lowest percentage is best explained by SILVA's ability to separate the Lachnospiraceae family into many genera that are not present in Greengenes or RDP, including *Eisenbergiella, Sellimonas, Shuttleworthia, Lachnoclostridium*, and *Tyzzerella*. Although the Greengenes and SILVA results were not considered significantly different, the gap of 4.7% in relative abundance could also be accounted for by the additional genera found by SILVA. These results are one example where relative abundance numbers for certain taxa may differ based on the taxonomic database. Considering that Lachnospiraceae was the most abundant family on average in our samples, the choice of database could affect the interpretation of the results. Lachnospiraceae has important implications in research of the effects of diet and disease on poultry, with members of the Lachnospiraceae family being known for the production of short-chain fatty acids (**SCFAs**), of which certain SCFAs such as butyrate are thought to improve weight gain in chickens challenged with *E. maxima* infection (Hansen et al., 2021). Using the Greengenes or RDP databases would still allow for a general interpretation of the potential links between Lachnospiraceae and disease, however, the more specific classifications from SILVA would allow for a more nuanced interpretation. In our study, *Eisenbergiella* decreased in abundance in infected chickens compared to the control, and a previous study has shown reductions in *Eisenbergiella* are associated with reduced production of metabolic products such as butyrate (Luo et al., 2018). As butyrate's importance has been shown, the specificity SILVA provides at the genus level in this case would be beneficial in identifying specific taxonomic groups that impact poultry health.

In many cases, there was agreement between the three databases in relative abundance and taxonomy, such as with *Escherichia* (11.8%), *Lactobacillus* (4.9%), and *Butyricicoccus* (2.7%), however, there were additional examples where database choice resulted in differences that may be of concern in chicken microbiota studies. Significant differences between Greengenes's [*Ruminococcus*] (10.0%), RDP's *Ruminococcus 2* (1.9%), and SILVA's [*Ruminococcus*] torques group (6.7%) could be explained by some ASVs from Greengenes's [*Ruminococcus*] instead being classified as unclassified Lachnospiraceae by RDP and as either Lachnospiraceae (uncultured or unclassified) or *Sellimonas* by SILVA. *Oscillospira* in Greengenes (5.5%) was significantly higher than unclassified Oscillospiraceae in SILVA (1.9%; $P < 0.001$), explained by Greengenes's *Oscillospira* being classified as *Oscillibacter, Flavonifractor, Intestimonas, Colidextribacter*, uncultured Oscillospiraceae, or unclassified Oscillospirales by SILVA, and no appropriate group was found for comparison within the RDP database as RDP classified these ASVs as *Oscillibacter, Flavonifractor, Pseudoflavonifractor, Intestimonas*, or unclassified Ruminococcaceae. Some bacterial groups did not differ ($P > 0.05$) in abundance between the three databases but were classified as different taxa. For example, an unclassified Clostridiales in Greengenes (4.1%) appeared to be more specifically classed as *Clostridia UCG-014* (3.8%) in SILVA, while being classed as *Anaerobacteria* (3.8%) in RDP.

LEfSe Analyses

With the Greengenes database, 12 genera and the order Burkholderiales were determined as more abundant in birds infected with *Eimeria tenella*, while 13 genera were more abundant in control birds (Figure 2A). With the RDP database, 10 genera and the order Burkholderiales were determined as more abundant in infected birds, while 16 genera and a phylum of unclassified bacteria were more abundant in control birds (Figure 2B). With the SILVA database, 9 genera and the order Burkholderiales were determined as more abundant in infected birds, while 25 genera were more abundant in control birds (Figure 2C). Again, the greater number of specific genera being identified by the SILVA database, including genera in the Lachnospiraceae family, appeared to be a factor in our results by increasing the number of genera determined as differentially abundant by LEfSe. With the SILVA database, LEfSe identified some of these genera, which could provide additional insight towards understanding the effects on infection on chicken microbiota.

Comparing the Greengenes and SILVA Databases

In addition to database choice, it should be noted that % identity at which the database is clustered to is another factor that may introduce bias. Many chicken microbiota studies, including those published in the past 4 years, have continued to utilize the Greengenes 97% identity database, and this widespread use allows for comparison between studies. Our study compared the Greengenes database at 97% identity to SILVA at 99% identity (the recommended option), possibly introducing bias through the % identity difference that could affect certain sensitive classifications. However, our results show that Greengenes produced the same results as SILVA in many cases, despite the difference in % identity. Most importantly, in the cases where results were not the same, there were fundamental differences in the databases due to SILVA being more up to date, such as the presence of additional genera in the Lachnospiraceae family in SILVA or updates in taxonomy because of recent phylogenetic studies, both of which changing the % identity would have no influence. SILVA has been shown in other studies to be a larger database compared to Greengenes (Balvočiūtė and Huson, 2017), therefore, chicken microbiota researchers using

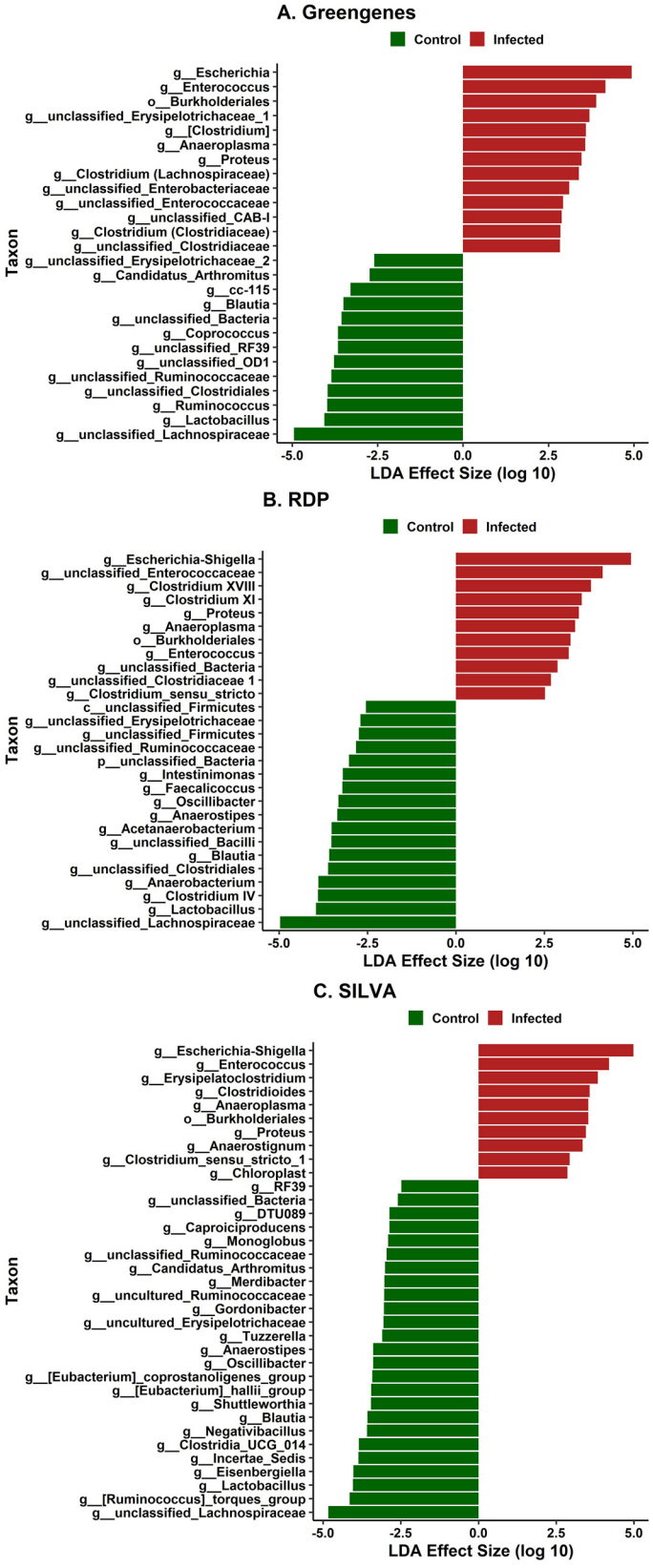

Figure 2. Linear Discriminant Analysis effect size (LEfSe) using the (A) Greengenes, (B) RDP, and (C) SILVA databases to determine differentially abundant bacteria in the chicken cecal luminal microbiota of *Eimeria tenella* infected birds and control birds.

Greengenes at any % identity are recommended to consider the use of SILVA.

In conclusion, taxonomic classifications and relative abundance numbers of certain taxonomic groups may be affected by the choice of reference database during bioinformatic processing of chicken microbiota. In genus-level analyses, database choice can affect the number of differentially abundant taxa between treatment groups, further affecting the interpretation of results. Although Greengenes is commonly used in

chicken microbiota studies, future studies should consider the use of the SILVA database to avoid outdated taxonomic classifications and produce greater specificity in results at the genus level.

ACKNOWLEDGMENTS

The work was funded by in house USDA-ARS CRIS # 8042-31000-108-00D.

DISCLOSURES

The authors declare no conflict of interest.

REFERENCES

Balvočiūtė, M., and D. H. Huson. 2017. SILVA, RDP, Greengenes, NCBI and OTT — how do these taxonomies compare? BMC Genomics. 18:114.

Campos, P. M., K. B. Miska, S. Kahl, M. C. Jenkins, J. Shao, and M. Proszkowiec-Weglarz. 2022. Effects of Eimeria tenella on cecal luminal and mucosal microbiota in broiler chickens. Avian Dis. 66:39–52.

Darwish, N., J. Shao, L. L. Schreier, and M. Proszkowiec-Weglarz. 2021. Choice of 16S ribosomal RNA primers affects the microbiome analysis in chicken ceca. Sci. Rep. 11:1–15.

Fitzgerald, C. B., A. N. Shkoporov, T. D. Sutton, A. V. Chaplin, V. Velayudhan, R. P. Ross, and C. Hill. 2018. Comparative analysis of Faecalibacterium prausnitzii genomes shows a high level of genome plasticity and warrants separation into new species-level taxa. BMC Genomics. 19:1–20.

Hansen, V. L., S. Kahl, M. Proszkowiec-Weglarz, S. C. Jiménez, S. F. Vaessen, L. L. Schreier, M. C. Jenkins, B. Russell, and K. B. Miska. 2021. The effects of tributyrin supplementation on weight gain and intestinal gene expression in broiler chickens during Eimeria maxima-induced coccidiosis. Poult. Sci. 100:100984.

Kers, J. G., F. C. Velkers, E. A. Fischer, G. D. Hermes, J. A. Stegeman, and H. Smidt. 2018. Host and environmental factors affecting the intestinal microbiota in chickens. Front. Microbiol. 9:235.

Luo, L., M. Hu, Y. Li, Y. Chen, S. Zhang, J. Chen, Y. Wang, B. Lu, Z. Xie, and Q. Liao. 2018. Association between metabolic profile and microbiomic changes in rats with functional dyspepsia. RSC Adv. 8:20166–20181.

Robeson, M. S., D. R. O'Rourke, B. D. Kaehler, M. Ziemski, M. R. Dillon, J. T. Foster, and N. A. Bokulich. 2021. RESCRIPt: reproducible sequence taxonomy reference database management for the masses. PLoS Comput. Biol. 17:e1009581.

Sokol, H., B. Pigneur, L. Watterlot, O. Lakhdari, L. G. Bermúdez-Humarán, J.-J. Gratadoux, S. Blugeon, C. Bridonneau, J.-P. Furet, and G. Corthier. 2008. Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. Proc. Natl. Acad. Sci. U. S. A. 105:16731–16736.

Stanley, D., R. J. Hughes, and R. J. Moore. 2014. Microbiota of the chicken gastrointestinal tract: influence on health, productivity and disease. Appl. Microbiol. Biotechnol. 98:4301–4310.