Certificate In Applied Statistics

Travis Williams

Friday, March 20, 2020 at 13:00 CDT

Contents

Basic Document Information

Last Compiled on:
Friday, March 20, 2020 at 13:00 CDT

Version.string:
R version 3.6.2 (2019-12-12)

Rstudio Version:
1.2.5033

R nickname:
Dark and Stormy Night

Language:
R

Platform:
x86_64-apple-darwin15.6.0

Contents and Confidentiality Statment

This file contains the analysis, discussion, and visualization of the data generated by the original research of Travis J. Williams for the research pertaining to the useage of cloacal swabs to evaluate cecal content microbiota. All information, data, analysis, and visualizations are the intellectual property of the Athrey lab, Travis J. Williams, Dr. Giridhar Athrey, Texas A&M University, and thus are confidential and proprietary information and may not be reproduced without permission. For reproduction permission please contact the Dr. Giridhar Athrey.

Research Goal

The gut microbiota is an ecological community of commensal and non-commensal microorganisms. In broiler chickens (*Gallus gallus domesticus*), a chicken that has been selected specifically for meat production, the gut microbiota has a central role in health and performance and are found throughout the entire length of the broiler's gastrointestinal tract (GIT), although most research concentrates on the organs within the lower sections: the small intestine (duodenum, jejunum, and ileum), large intestine, cecum, and cloaca. The ceca are a vital site of functional activity as they retain nearly 10^{11} microbial cells per gram and are an important location for fluid resorption via the translocation of urea from the urodeum and the fermentation of carbohydrates. As a consequence, the ceca are the most sampled gut segment in chicken gut microbiota studies. A standard experimental method of microbiota analysis in poultry involves the invasive sampling of the ceca, following euthanasia, which prevents longitudinal studies of the same experimental animals.

The cecal microbiota in chicken show broad similarities with lower large-intestinal microbiota, and the cloaca abuts the large-intestine. Due to their proximity to the large intestine and ceca, cloacal (or vent) swabs are an alternative, non-invasive method used on domestic, migratory, or endangered bird species where invasive sampling may not be permitted. Because of this non-invasive aspect, cloacal swabs are used frequently as a diagnostic tool for assaying and monitoring disease-causing agents such as *Salmonella spp.*, Avian Influenza, Coccidiosis, and *Campylobacter coli*. Thus, it is critical to determine if and how representative cloacal microbiota are of cecal microbiota. If cloacal microbiotas approximate the cecal microbiota, it would enable non-invasive longitudinal studies. On the other hand, if cloacal microbiota is not a reliable proxy for cecal microbiota occurrence and abundance, then its utility for assessing avian microbiota would be limited.

Our study focuses directly on the resolution and repeatability of microbiota patterns. To resolve the reciprocity of cecal and cloacal microbiotas, we used a paired sample approach to compare cecal and cloacal microbiota communities sampled from the same individuals. Based on previously published works, we hypothesized that cloacal microbiota is not representative of cecal microbiota from the same individuals. We used 16S rRNA sequence-based analysis of α and β diversity of the communities between the two sampling methods and two dietary treatments. Each paired sample was analyzed using the MOTHUR pipeline along with the Phyloseq, Vegan, and other packages on the R platform.

Bioinformatic Pipeline for Microbiota Evaluation

Resultant fastq files from sequencing were processed using MOTHUR software *v. 1.39.5*. Briefly, paired-end reads were joined using the *make.contigs* command. We aligned the sequences to the SILVA database *v. 132* and removed chimeric sequences using the UCHIME program *v. 4.2.40*. Low abundance Operational Taxonomic Units (OTU's) and Singletons were removed from analysis with the *split.abund* command using cutoff=1. Chloroplast, Mitochondria, Eukaryota, and other unknown sequences were removed from the dataset using the *remove.lineage* command. Total OTU's were then generated at the species level (0.03) and then classified using the *classify.otu* and *dist.seqs* commands, respectively. OTU tables and other output from MOTHUR were further analyzed on the R platform *v. 3.6.2* using the Phyloseq *v. 1.28.0* and Vegan *v. 2.5-6* packages. We generated plots using the ggplot2 package *v. 3.3.0*.

Statistical Tests for α and β Diversity:

To assess whether the abundance structures, ignoring taxonomic composition, are comparable between cloacal and cecal communities, we performed non-parametric tests on the relative abundance and cumulative distribution functions of the paired cecal and cloacal swab datasets based on the Bray-Curtis distance using the Phyloseq and ggplot2 packages. We used the two-sample Kolmogorov-Smirnov (KS) test using the ks.test function in R {base} to assess whether the microbiota populations (cecal vs. swabs) are from the same distribution. The two-sample ranked location-scale tests of Cucconi and Lepage were implemented using the nonpar package *v. 0.1-2* using the *cucconi.test* and *lepage.test* commands, respectively. The Cucconi test is a ranked test that assesses whether the locations and scales of the two population distributions are equal while the Lepage test is a ranked location-scale test that combines the Ansari-Bradley test for scale and the Wilcoxon-Mann-Whitney test for location.

Two statistical tests were performed in R to evaluate the α and β diversity between sampling locations and amongst dietary treatments within each sampling location. We used the two-sample Wilcoxon Signed-Rank test to compare the Chao1 and Inverse Simpson (InvSimpson) α diversity measures. To compare β diversity, we used the Permutational Multivariate Analysis of Variance (PERMANOVA) using the "Adonis" function of the Vegan package with 9999 permutations. In addition to PERMANOVA, we compared β diversity in MOTHUR using HOMOVA, AMOVA and unifrac.weighted. We applied the weighted unifrac test to investigate the probability that two or more communities have the same structure by chance. These three species-level non-parametric tests were computed using the Yule and Clayton measure of dissimilarity average phylogenetic distances. The statistical significance of all comparisons was assessed at α=0.05.

Hypothesis Tests

$$H_O : \text{Microbtiota}_{(\alpha \text{ and } \beta \text{ Diversity Cloacal Swabs})} = \text{Microbtiota}_{(\alpha \text{ and } \beta \text{ Diversity Cloacal Swabs})}$$

$$H_{A1} : \text{Cloacal Swabs}_{(\alpha \text{ or } \beta \text{ Diversity})} \neq \text{Cecal Content}_{(\alpha \text{ or } \beta \text{ Diversity})}$$

$$H_{A2} : \alpha \text{ Diversity}_{(CloacalSwabs)} \neq \alpha \text{ Diversity}_{(CecalContent)}$$

$$H_{A3} : \beta \text{ Diversity}_{(CloacalSwabs)} \neq \beta \text{ Diversity}_{(CecalContent)}$$

Data Filtering and Pre-Processing

The raw data from sequencing generated a total of 694,559 reads, with an average of 18,278 reads per sample. Total read depth per sample was limited to an arbitrary minimum of 3,005 to ensure adequate read depth in any given sample, and thus was the cutoff for inclusion in further analysis. One cloacal swab library out of the 40 total libraries sequenced was excluded as a result of this read threshold and to retain the paired nature of our analysis, the corresponding cecal content sample was excluded. We proceeded with nineteen paired samples for which both cloacal swab and cecal data was found. The cecal samples had an average of 22,656 reads (IQR 20,312 - 24,515 reads), whereas the cloacal swabs had an average of 13,900 reads (IQR 6,902 – 20,366 reads). The average Good's coverage for all thirty-eight samples was 99.57%, (Standard Deviation = 0.305%), showing that the retained datasets had adequate sequence coverage to sample OTUs. Q-Q plots of the Good's Coverage Index values are given in Figures 1A and 1B, summary statistics for both the cecal and cloacal datasets are given in Table 1, as well as Rarefaction curves summaries (Figure 2 and Figure 3) are presented to visualize the quality and normality of the data that passed our pre-processing parameters.

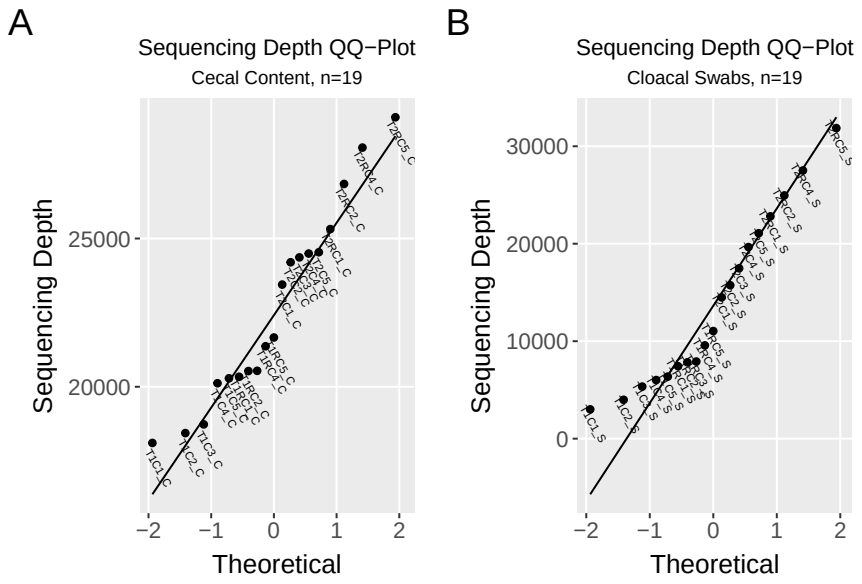

Figure 1: QQ-Plots of sequencing depth organized by Method where (A) represents the Cecal Content Samples and (B) represents the Cloacal Swab Samples.

Table 1: Descriptive statistics of the sequencing data for the nineteen Cecal Content (C) and nineteen Cloacal Swabs (S).

	C_Reads	C_Goods_Coverage	S_Reads	S_Goods_Coverage
Sample Size	19	19	19	19
Minimum	18,106	99.61%	3,005	98.70%
1st Quartile	20,312	99.70%	6,902	99.15%
Median	21,662	99.72%	11,043	99.49%
Mean	22,656	99.72%	13,900	99.41%
3rd Quartile	24,515	99.74%	20,366	99.72%
Maximum	29,088	99.86%	31,862	99.86%
IQR	4,204	0.043%	13,464	0.576%
Range	10,982	0.25%	28,857	1.16%

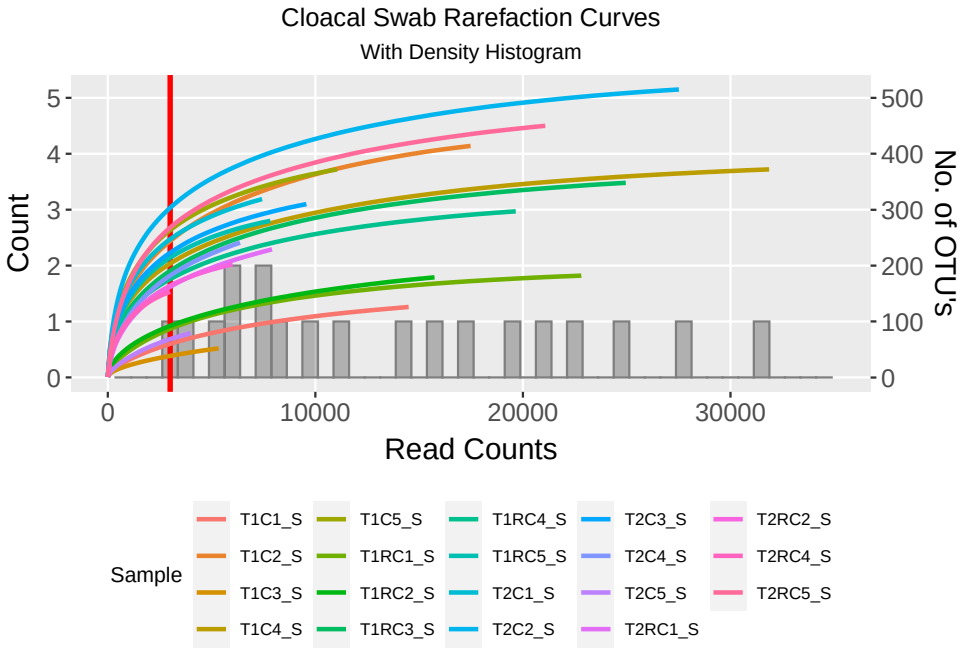

Figure 2: Rarefaction curves for the cloacal swab samples, where curve color designates each of the nineteen individual samples. The red vertical line at 3005 indicates the minimum read count threshold necessary for further analysis. The histogram in the background displays the number of rarefaction curves (samples) that terminate at a specific read count depth.

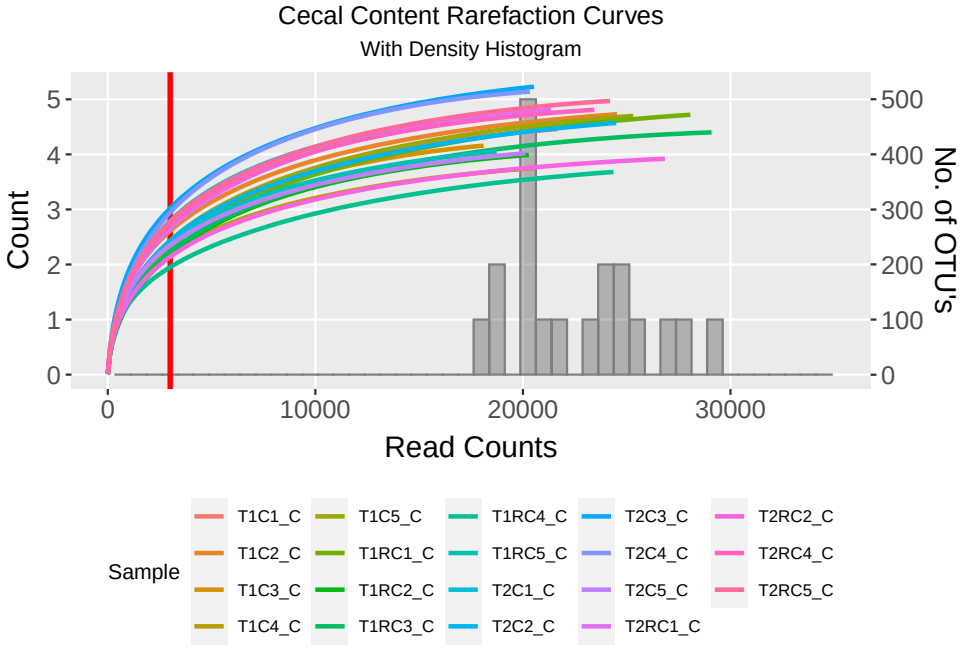

Figure 3: Rarefaction curves for the cecal content samples, where curve color designates each of the nineteen individual samples. The red vertical line at 3005 indicates the minimum read count threshold necessary for further analysis. The histogram in the background displays the number of rarefaction curves (samples) that terminate at a specific read count depth.

Broad Differences Between Cecal and Cloacal Microbiota Members

Table 2: Summary of Phylogenetic Information.

	All_Samples	Cloacal_Swabs	Cecal_Content
Sample Size	38	19	19
Number of OTU	1790	914	1626
Number of Family	88	82	60
Number of Phylum	11	11	10

Overall, the thirty-eight samples yielded 1790 OTU's assigned to 88 Families (Table 2). The top three families based on relative abundance were *Lactobacillaceae* (11.409%, Phylum Firmicutes), *Ruminococcaceae* (9.979%, Phylum Firmicutes), and *Peptostreptococcaceae* (9.817%, Phylum Firmicutes). The nineteen cecal content samples yielded 1626 total OTU's from 60 total Families (Table 2) with *Ruminococcaceae* (22.163%, Phylum Firmicutes), *Barnesiellaceae* (11.954%, Phylum Bacteroidetes), and *Rikenellaceae* (8.126%, Phylum Bacteroidetes) as the top three families in the cecal content samples based on relative abundance (Table 3).

Table 3: Top 3 Families in the Cecal Content samples (relative abundance >=0.02).

OTU	Abundance	rabund	Phylum	Family
Otu0002	0.2873480	0.2216329	Firmicutes	Ruminococcaceae
Otu0001	0.1549853	0.1195409	Bacteroidetes	Barnesiellaceae
Otu0003	0.1053493	0.0812564	Bacteroidetes	Rikenellaceae

The nineteen cloacal swab samples yielded 914 total OTU's from 82 total Families (Table 2). The top three relative abundant families in the cloacal swab samples did not show a uniform distribution. The top three most abundant families in the cloacal samples (Table 4) were *Lactobacillaceae* (16.588%, Phylum Firmicutes), *Peptostreptococcaceae* (10.104%, Phylum Firmicutes), and *Pasteurellaceae* (8.931%, Phylum Proteobacteria). It is noteworthy that *Peptostreptococcaceae* was only present in five out of nineteen cloacal swab samples and the family *Lactobacillaceae* were present only in six of the nine T2 cecal content samples.

Table 4: Top 3 Families in the Cloacal Swab samples (relative abundance >=0.02).

OTU	Abundance	rabund	Phylum	Family
Otu0004	0.3775477	0.1658836	Firmicutes	Lactobacillaceae
Otu0009	0.2299580	0.1010370	Firmicutes	Peptostreptococcaceae
Otu0016	0.2032565	0.0893051	Proteobacteria	Pasteurellaceae

Twenty-five families were represented in the cloacal swab samples with twenty families represented in the cecal content samples, as shown in Figure 4A and Figure 4B, respectively. There were eleven families shared cecal and cloacal samples (Table 5), with nine families (*Atopobiaceae*, *Bacteria_unclassified*, *Bifidobacteriaceae*, *Christensenellaceae Clostridiales_unclassified* (OTU 0087), *Clostridiales_vadinBB60_group* (OTU 0045), *Coriobacteriaceae*, *Gastranaerophilales_fa* (OTU 0010), and *Helicobacteraceae* (OTU 0005)) unique to cecal content and fourteen families (*Actinobacteria_unclassified*, *Actinomycetaceae*, *Clostridiales_unclassified* (OTU 0096), *Clostridiales_vadinBB60_group* (OTU 0037), *Corynebacteriaceae*, *Enterobacteriaceae*, *Enterococcaceae*, *Gastranaerophilales_fa* (OTU 0014), *Helicobacteraceae* (OTU 0021), *Mollicutes_RF39_fa*, *Pasteurellaceae*, *Peptostreptococcaceae*, *Planococcaceae*, and *Staphylococcaceae*) unique to cloacal swabs.

Table 5: Shared OTU's between all Cloacal Swab and Cecal Content Samples.

OTU	Cloacal_Swab_Relabund	Cecal_Content_Relabund	Phylum	Family
Otu0001	0.0789118	0.1195409	Bacteroidetes	Barnesiellaceae
Otu0002	0.0707640	0.2216329	Firmicutes	Ruminococcaceae
Otu0003	0.0409090	0.0812564	Bacteroidetes	Rikenellaceae
Otu0004	0.1658836	0.0218070	Firmicutes	Lactobacillaceae
Otu0007	0.0389352	0.0400790	Bacteroidetes	Bacteroidaceae
Otu0011	0.0163858	0.0323466	Bacteroidetes	Tannerellaceae
Otu0012	0.0534946	0.0721904	Firmicutes	Lachnospiraceae
Otu0013	0.0113538	0.0252862	Firmicutes	Acidaminococcaceae
Otu0019	0.0368682	0.0205850	Epsilonbacteraeota	Campylobacteraceae
Otu0026	0.0163389	0.0155916	Proteobacteria	Burkholderiaceae
Otu0043	0.0148655	0.0202546	Bacteroidetes	Marinifilaceae

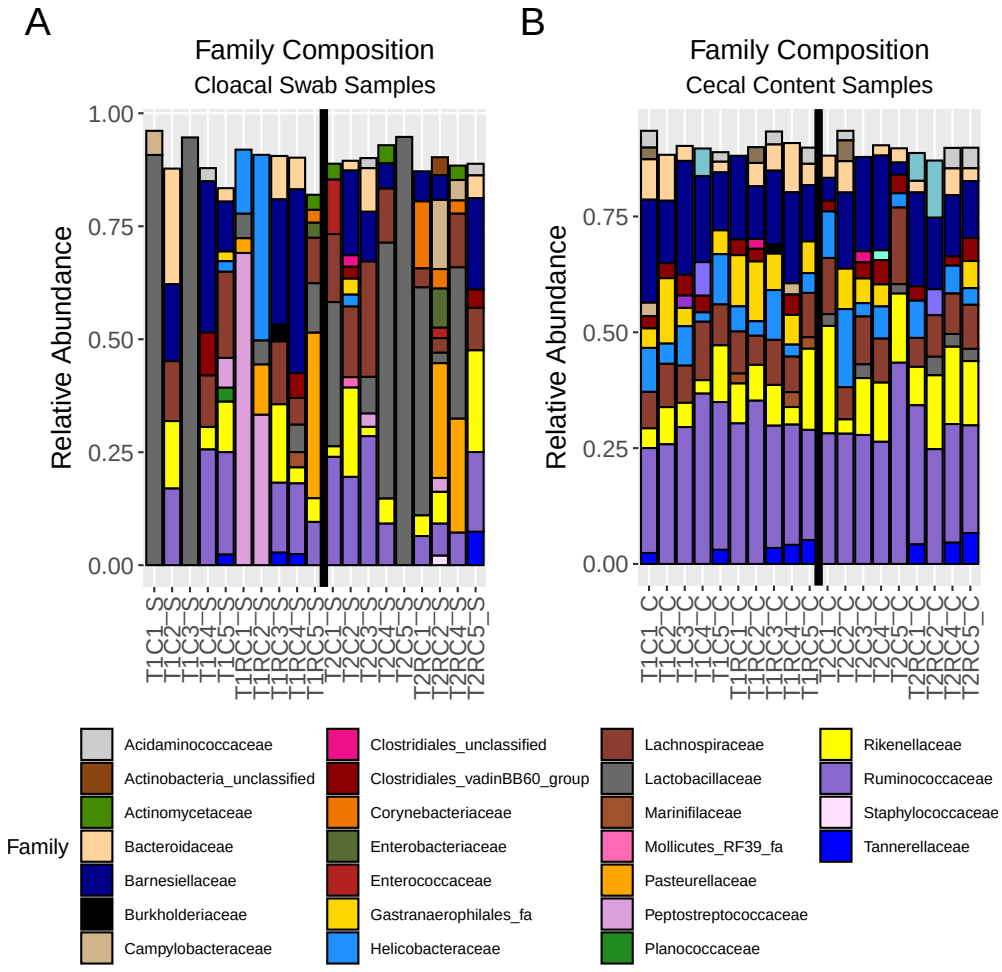

Figure 4: The relative abundance of families observed for each cloacal swab (A) and cecal content sample (B) that has >= 3005 reads; n=19 for each method. The vertical black line divides samples belonging to the two dietary treatments.

α Diversity Analysis

Cecal Content vs. Cloacal Swabs

We compared microbial species richness and diversity of the cecal content and cloacal swabs using the Chao1 and Inverse Simpson estimators (Figure 5). Both the Wilcoxon Signed-Rank test (W=4, p-value = 2.67e-05) and paired t-test (t=5.7938, p-value = 1.724e-05) showed highly significant differences in the Chao1 index between the cecal content and cloacal swabs, with the highest richness observed in the cecal samples. A higher Chao1 value indicates a higher number of low abundance taxa, e.g., singletons. The higher value in cecal samples, suggests that more rare taxa were captured in cecal samples. However, the Inverse Simpson species diversity estimator was not different between cloacal and cecal samples based on a Wilcoxon Signed-Rank test (W=48, p-value = 0.06021) and a paired t-test (t=1.9445, p-value = 0.06763). Similar to the Chao1 findings, the cecal content had higher microbial diversity, compared to the cloacal swabs. As the Inverse Simpson index estimates the richness weighted by the proportional abundance of taxa present within samples, the non-significance suggests that the two types did not differ in their internal weighted abundances.

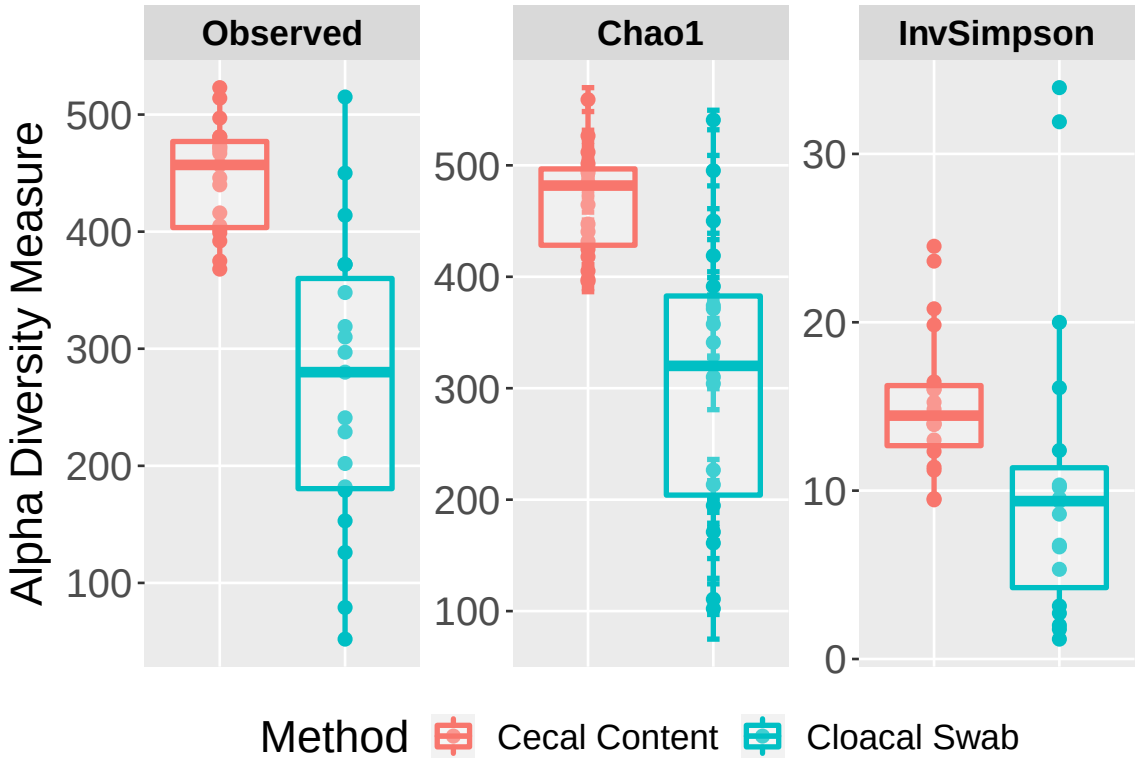

Figure 5: Boxplots of the observed, Chao1, and Inverse Simpson (InvSimp) alpha diversity indices for the comparison of cecal content and cloacal swab samples.

Dietary Treatment Comparison Within Sampling Methods

To assess whether cecal or cloacal swabs captured differences between dietary treatments, we performed richness and diversity analyses, comparing the two diets. Figure 6A shows comparisons within cecal data, and Figure 6B shows cloacal swab comparisons. The Wilcoxon Signed-Rank test on the Chao1 estimator returned non-significant p-values for the cecal content treatments (W = 24, p-value = 0.09472) and cloacal swab treatments (W = 40, p-value = 0.7197). The Chao1 values were higher for T2 than T1 in both the cecal content and cloacal swab methods. Similarly, the Inverse Simpson index was not different between diets based on cecal samples (W=40, p-value = 0.7197) or the cloacal swab data (W = 27, p-value = 0.1564). In summary, neither the cecal nor the cloacal swab samples showed significant differences between dietary

treatments based on the Chao1 and Inverse Simpson indices.

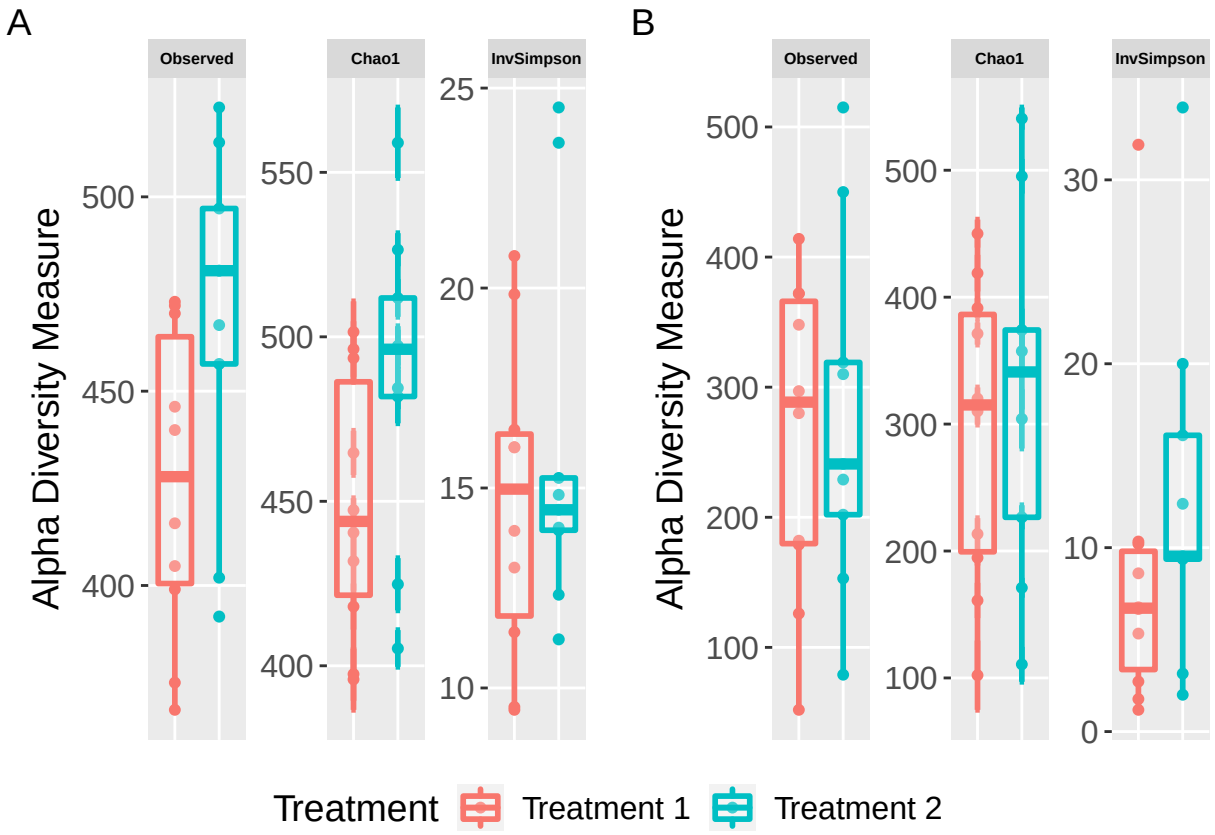

Figure 6: (A) Boxplot of the Observed, Chao1, and Inverse Simpson (InvSimpson) measurements for alpha diversity of cecal content samples by Treatment. (B) Boxplot of the Observed, Chao1, and Inverse Simpson (InvSimpson) measurements for alpha diversity of cloacal swab samples by Treatment.

β Diversity Statistical Analysis

The PERMANOVA analysis showed significant difference (F. Model = 8.3319, R2 = 0.18794, p-value = 0.0001) in the centroids and dispersion between cecal and cloacal microbiota. This difference of community structure was further supported with significant results from the HOMOVA (BValue = 3.6193, p-value <0.001) and Weighted Unifrac (WScore = 0.750238, WSig = <0.001) tests. On the whole, the results from the PERMANOVA, HOMOVA, and Weighted Unifrac tests all show that the microbiota communities inferred from cloacal swabs are different from cecal microbiota.

Next, we investigated whether differences in the diet treatments (T1 and T2) elicit differences in communities (β diversity) inferred using cecal versus cloacal samples. We found that the cecal content sampling method detected differences in β diversity between the diets, but this difference was not observed in the cloacal swab sampling method. The results from PERMANOVA showed that the cecal content was different between dietary treatments (F. Model = 2.1281, R2 = 0.11125, p-value = 0.0249). This observation is supported by significant HOMOVA (Bvalue = 0.806737, p-value = 0.024) and AMOVA (Fs = 2.6899, p-value = 0.044) test results. Similarly, the Weighted Unifrac also returned a significant difference (WScore = 0.639013, WSig <0.001). Overall the cecal data microbiotas were significantly different between diets.

In contrast, and all but the Weighted Unifrac test returned non-significant. The PERMANOVA (F. Model = 0.99558, R2 = 0.05532, p-value = 0.4124), HOMOVA (Bvalue = 0.0240581, p-value = 0.218), and AMOVA (Fs = 1.00622, p-value = 0.416) all show the cloacal swab microbiota were not different between the two

diets. In contrast, the Weighted Unifrac comparison of microbiotas between diets using cloacal swabs was significantly different (WScore =0.620973, WSig <0.001).

Cumulative Distribution Function Analysis

The cumulative distribution functions of the relative abundances between cecal and cloacal microbiotas were significantly different. The KS (D=0.11995, p-value= 9.745e-08) and Cucconi tests (C=426.916, p-value<1.e-05) were highly significant. The results of the Kolmogorov-Smirnoff and Cucconi tests on the cumulative distribution function further demonstrate that both the location and the scales of the cecal content and cloacal swab distributions of relative abundances are highly different.

Further comparison of the dietary treatments within and between sampling methods using the KS, Lepage, and Cucconi tests yielded no significant differences (Table 6). These results are surprising, considering the differences observed using cecal samples. However, as these tests focused on the overall abundance distributions, while ignoring the taxonomic representation, these results make sense.

Table 6: Summary of the statistical test used in comparing the geometric mean distributions of dietary treatments within methods. Dietary comparisons are listed in the first column, with the statistical tests for the specified comparison in the adjacent columns. Each cell contains the respective test statistic and p-value for that comparison.

	KS_Test	Lepage_Test	Cucconi_Test
Cecal Content: T1 vs. T2	D=0.2569; p-value=0.5494	L=1.1637; p-value=0.5625	C=0.346; p-value=0.705
Cloacal Swab: T1 vs. T2	D=0.1773; p-value=0.8199	L=1.853; p-value=0.4033	C=0.733; p-value=0.523

Discussion

In this study, we showed that the microbiota identified from cloacal swabs are not representative of the cecal microbiota, and therefore not a suitable approach to sampling the microbial communities of the lower gastrointestinal tract. This result was highly surprising, given that the cecal and large intestine microbiotas are alike by week five in chicken. Not only were the cloacal communities limited in their resemblance to cecal communities, the patterns of presence-absence as inferred by richness estimates were also significantly different. These findings suggest that there is a high degree of stochasticity to taxa sampled from the cloaca.

High variability of cloacal microbiota

While the factors influencing fecal microbiota differences from cecal communities (external conditions, environmental microbiota) are expected, the cloacal swab dissimilarities and variability are more surprising. It is not clear if the cloaca of chicken is colonized, unlike other parts of the GIT. While numerous surveys of cloacal microbiota exist in the literature, in wild birds, the cloacal microbiota is often the only locus for characterizing gut microbiota as euthanasia may not be an option. However, our results show that the taxonomic composition and community profiles obtained from cloacal swabs can be highly random, with little consensus even when collected under controlled conditions. We found lower richness and diversity of microbial taxa in the cloaca, compared to the cecal microbiota. However, our data showing significantly higher richness in the ceca, suggests that host genetics are not driving lower richness or diversity in the cloaca. The variability of cloacal swab data was revealed only in contrast with the paired cecal datasets.

Our analysis leads us to advocate extreme caution when inferring lower GIT microbiota patterns from cloacal swabs of birds. Cloacal swabs are used routinely to assess infection status in domesticated, pet, and wild bird species. In the majority of these cases, targeted assays (RT-PCR) use swab samples for the detection of pathogenic species. In these cases, we rely on the sensitivity of the assays provides valuable information for treatment or containment of pathogens, especially in poultry operations. While our data

show high variability in the representation of taxa in cloacal samples, the sensitivity of RT-PCR approaches may allow lower detection thresholds. However, the reciprocity of taxon representation with cloacal 16S rRNA sequencing and targeted PCR methods needs to be determined experimentally.

Resolution of microbial community differences between diets

In our analysis of microbiota between the two dietary treatments, we found that neither cecal nor cloacal samples were able to differentiate between diets. Both these sample types appear to be equivalent in their inability to differentiate between diets. However, we emphasize that this equivalency exists aside from the fact that the cecal and cloacal communities were highly dissimilar. Also, while the cecal samples were similar due to the overlapping distributions between diets, the similarity of cloacal swabs is driven by the high variability across all cloacal samples (Figure 6B). Additionally, the housing environment, rather than dietary protein source, is known to be a more significant factor driving cecal microbiota differences. In this present study, all the chickens were raised in the same barn (across replicate pens), where they were provided with the same bedding material and water source. Therefore, the high variability among cloacal samples, all collected in a controlled environment, represents, in our opinion, the high variability inherent to cloacal samples.

Conclusions

In this study, we showed that cloacal swabs do not faithfully approximate either the α and β diversity of cecal samples, based on paired samples. Therefore, cloacal swabs are unsuitable for assessing lower GIT microbiota in birds. While the high variability of cloacal microbiota has been reported previously, our study provides experimental evidence to capture the randomness of cloacal microbiota, in contrast with the consistency of cecal samples. One of the consequences of our finding is that cloacal samples, akin to fecal samples, are not suitable or reliable for longitudinal studies of gut microbiota patterns in birds. Finally, the high inter-individual variation of cloacal swab data warrants an experimental assessment of their reliability for targeted diagnostic methods.

Appendix

All the data and metadata files, protocols, shell script, and R code (in an PDF format generated via RMarkdown) have been deposited on Figshare and can be accessed at <10.6084/m9.figshare.11819655>.