Comparison of OTUs and ASVs in 73 Bird Species from Equatorial Guinea

Darien Capunitan
darien.capunitan@uconn.edu

Recommended Citation
Capunitan, Darien, "Comparison of OTUs and ASVs in 73 Bird Species from Equatorial Guinea" (2018). Master's Theses. 1188.
https://opencommons.uconn.edu/gs_theses/1188

This work is brought to you for free and open access by the University of Connecticut Graduate School at OpenCommons@UConn. It has been accepted for inclusion in Master's Theses by an authorized administrator of OpenCommons@UConn. For more information, please contact opencommons@uconn.edu.
Comparison of OTUs and ASVs in 73 Bird Species from Equatorial Guinea

Darien Capunitan

B.S., University of Connecticut, 2018

A Thesis
Submitted in Partial Fulfillment of the
Requirement for the Degree of
Master of Science
At the
University of Connecticut
2018
Approval Page

Master of Science Thesis

Comparison of OTUs and ASVs in 73 Bird Species from Equatorial Guinea

Presented by

Darien C. Capunitan, B.S.

Major Advisor: ________________________________

Sarah Hird

Associate Advisor: ________________________________

Johann Gogarten

Associate Advisor: ________________________________

Johnathan Klassen

University of Connecticut

2018
Acknowledgements

I would like to thank my major advisor, Sarah Hird, for always giving me guidance and allowing me for being one of the first members of her lab and for overall being awesome.

I would also like to thank the members of the Hird lab for always keeping an amazing working environment. Beth, Kirsten, and Josh are the best.

Thanks to Ryan and Oscar for getting these great samples in the first place and adding to the metadata table.

Thanks to the UCONN MARS facility for processing the samples and a special thanks to Dr. Kendra Maas for answering all my questions.

Thank you to Peter Gogarten and Johnathan Klassen for taking the time to be associate advisors.

Finally I would like to thank my parents for their love and support throughout my existence with a special thanks to Amber and Rodney for always believing in me.
Table of Contents

Approval Page ii
Acknowledgements iii
Table of Contents iv
Abstract vi
Main text 1

Chapter 1: Introduction 1

Chapter 2: Materials and Methods

2.1: Sampling 4
2.2 DNA Extraction and Sequencing 5
2.3 Analyses
  2.3.1 Data Processing: Mothur 6
  2.3.2 Data Processing: DADA2 6
  2.3.3 Data Analysis: Unifrac 7
  2.3.4 Data Analysis: Bray-Curtis 7

Chapter 3: Results

3.1 Alpha diversity 8
3.2 Beta diversity 19

Chapter 4: Discussion 30

Chapter 5: Conclusion 31
Abstract

As scientists discover more information about the communities of bacteria that live on and inside hosts, "the microbiome", a new avenue for understanding the health of humans and animals has opened. There are many analysis pipelines for microbiome data processing, and choice of analytical tools can affect the biological results of the analysis. The current analysis climate does not point toward a single most effective protocol, which hinders comparisons across studies. An important step in microbiome data processing is the assignment of reads into groups of similar organisms. The traditional unit for grouping organisms is the species; however resolving sequences at the species level is not always possible. For this reason, the term “operational taxonomic unit” (OTU) has been used as a proxy for species. Mothur is one popular software package that uses OTUs, as defined by a percent sequence similarity that is set by the user; 97% sequence similarity is frequently cited as the most similar to traditional species.

DADA2 is a newer pipeline that classifies unique sequences as “amplicon sequence variants” or ASVs. ASVs assume biological sequences are present in the sample and can resolve differences in sequence variants at as low a single nucleotide. We compared OTUs and ASVs from a dataset of bird gastrointestinal tracts from Equatorial Guinea. One hundred and forty two samples were taken from 92 birds (73 species) and the V4 region of the 16S rRNA gene was used as the microbial fingerprint (515f and 806R primers). DADA2 produced 4245 ASVs against Mothur’s 9332 OTUs. DADA2 found 507 different genus ranks while Mothur found 932 against the same reference database, Silva version 132. We analyzed alpha and beta diversity in the OTUs and ASVs using the Phyloseq program for R. The data suggests key differences in these analyses.
Procruste analyses (Figures 13 and 16) show significant differences as Mothur data clusters more tightly together. Some samples are also moved substantially across axes.
Chapter 1: Introduction

The microbiome is vital for vertebrate health (Cho et al. 2012). The microbes in the gut aid in digestion, gut immunity and organ development (Diaz Heijtz et al., 2011; Al-ASmakh et al. 2014). There are ten times the number of genes in the human microbiome compared to the human genome, which has further implicated the importance of the microbiome for host development and health. The gut microbiome provides a possible pathway for therapeutic relief of diseases because of its effect on the several health related regions including the gut-brain axis (Schnorr and Bachner 2016). There have already been studies targeting the microbiome for brain, cardiovascular, and several other diseases (Ramezani et al., 2014; Shoemark et al. 2014). The microbiome has also been shown to have a strong relationship with host behavior as both elements can directly affect one another (Ezenwa 2012). The gut microbiome is densely populated and there is substantial intra- and interspecific diversity among hosts of many taxa based on diet, environment, and overall locality (Wang et al. 2013; Spor et al. 2011). Only with continued sampling can scientists begin to understand the importance of the gut microbiome and its roles in host health and evolution.

Most of our knowledge of gut microbiomes comes from studies on mammals (Nelson et al., 2013; Muegge et al., 2011; Ley et al. 2008; Colston and Jackson 2016). However, mammals have unique adaptations that affect their microbiomes (e.g., vaginal delivery and milk, Dominguez-Bello et al., 2010; Cho et al., 2012). Although samples from mammals are important microbiome data, the confounding factors that are associated with mammal birth and offspring development are not representative of all animal life. Furthermore, it has not yet been studied how strong the impact of early life mammalian behavior and physiology is on the gut microbiome and how much can be attributed to outside factors (e.g. diet, environment, location).
Birds are a globally distributed class of organisms with great diversity in their diets, behaviors, and ecologies and we know little about their microbiomes. A foundational requirement for comparative studies is the cataloging of microbes from different species’ microbiomes. Sampling diverse, wild species will allow for the description of a “core gut microbiome”. The “core gut microbiome” theoretically consists of species, genera, or families of bacteria that are consistently found among a particular species. Isolating the “core gut microbiome” may illuminate the effect of these bacteria on host health and may be able to distinguish transient bacteria that may overshadow functionally important groups, or at the very least, obscure the true signal. Catalogued “core” species will also help future studies as a point of reference. While current studies have categorized differences between one or a few host species (Benskin et al., 2010; Kreisinger et al., 2015; Xenoulis et al., 2010), studies using many host species can help to build a foundation for understanding the complex relationship between hosts and their diverse environments and ecologies (Hird et al. 2015). A comprehensive detailing of the possible symbiotic relationship between birds and their gut microbiome can be further understood through large sampling studies.

How one analyses a microbiome can affect the results and interpretation of the data. One important step, performed near the beginning of an analysis, is defining the taxonomic unit. This allows scientists to measure and give meaning to microbial data that may not be assigned to species. One of the more popular and widely used measurements is the operational taxonomic unit (Okal & Sneath 1963), or OTU. OTUs classify sequences that are generally related (Schloss et al., 2005). OTUs are needed for comparisons of qualitative data that, without the OTU association, can otherwise not be quantitative. The reliability and reproducibility of the classification of DNA reads into OTUs is important to properly represent data. The stability of
OTUs is of key importance when various clustering methods are applied (He et al. 2015). OTUs have been used extensively through different disciplines, including microbial ecology. There are currently many ways to classify OTUs. “Closed reference” rely on a database and discard sequences that don’t match the database. Closed reference allows an easy way to compare across datasets in which samples are normalized. However, the discarding of data is difficult to support and also creates an overestimation of diversity (Edgar, 2017). "Open reference" OTU picking again rely on a database, however sequences that do not match the reference database are then clustered de novo. Closed and open reference methods can cause significant differences with alpha and beta diversity even with identical data (He et al. 2015). “De novo” methods cluster sequences within their own dataset into groups that are a certain sequence percent similarity; 97% is frequently cited as most representative of "species". This method ensures all of the data is utilized; however since no reference database is used there is an inability to make comparisons across studies. In addition, de novo clustering may take a long time (possibly months) when a large dataset it processed. There are multiple ways to do de novo clustering. OptiClust is the method applied in the popular package Mothur (Westcott et al., 2017). The program Mothur is one of the popular pipelines for microbial community analyses, as it provides all the steps of a microbiome analysis in a single package. Mothur (Schloss 2009) has been cited 8170 times (Google Scholar 1 March 2018).

Recent calls for improved methods of OTU picking (Callahan et al., 2017) have led to the development of a new method of OTU picking, called "amplicon sequence variants" or ASVs and is implemented in the R package DADA2 (Callahan et al., 2016), currently cited 137 times (Google Scholar 1 March 2018). ASVs are an analogous term to OTUs, however it refers to a unique sequence that is assigned to a taxonomic group. Each ASV is then calculated in each
sample and the recorded taxonomy presence is noted. Recently there has been a call for the preferential use of ASVs over OTUs citing greater reproducibility and comprehensiveness (Callahan 2017). There have also been claims that error control is increased by use of ASVs as they can be, “resolved down to the level of single-nucleotide differences over the sequenced gene region” (Callahan 2017). ASVs also allow for an easy way to compare samples across different studies, but they have not been as thoroughly vetted in the literature as OTUs.

Here, we conduct a direct comparison of calling taxonomic units as ASVs (with DADA2) and 97% OTUs (with Mothur) using 92 birds belonging to 73 species from Equatorial Guinea. We compare OTU/ASV abundance and diversity, taxonomic ranks, and a variety of other variables. We also include similarities and dissimilarities of taxonomy and beta diversity (using Bray Curtis and Unifrac distances). A Procruste analysis was conducted to quantify the dissimilarities between ASVs and OTUs using their respective PCOA and NMDS plots. Although quantifying the differences between ASVs and OTUs does not correlate to a preference for one or the other, it will highlight the significance of using different pipelines on “identical” data. These data are also the first representation of the microbiome from these 73 bird species.

Chapter 2: Materials and Methods

2.1 Sampling

All birds were collected from 8 sites in Equatorial Guinea from 5 January 2016 to 2 February 2016 (Table 2). Mist nets were used to catch birds, which were then euthanized with thoracic compression and immediately prepared as museum specimens. Intestinal tracts were tied off, then removed and stored in 100% ethanol within approximately 15 minutes of death. The Dirección General de Protección y Guardería Forestal and the Universidad Nacional de Guinea Ecuatorial provided specimen collection and export permits.
2.2 DNA Extraction and Sequencing

DNA was extracted from the luminal contents of the intestines using Qiagen Power Fecal kits following the standard protocol. Some intestines were subsampled (samples labeled with suffixes A, B, C, D). Blanks (n=7) were extracted at the same time as the samples as negative (kit) controls.

Extractions were sequenced at the MARS facility at the University of Connecticut on a MiSeq. Negative (kit) controls from the PCRs were also sequenced (n=7). PCR replicates were also sequenced for a subset of the samples; these samples are labeled with the suffixes X, Y. A total of 142 samples were sequenced.

Quant-iT PicoGreen kit were used to quantify DNA extractions. Amplification of the V4 region of the 16S rRNA gene was done using 30 ng of extracted DNA as template. The V4 region was amplified using 515F and 806R with Illumina adapters and dual indices (8 basepair golay on 3’ (Caporaso 2012), and 8 basepair on the 5’ (Kozich et al., 2013)). The PCR reaction was incubated at 95°C for 3.5 minutes, the 30 cycles of 30 s at 95.0°C, 30 s at 50.0°C and 90 s at 72.0°C, followed by final extension as 72.0°C for 10 minutes. PCR products were pooled for quantification and visualization. PCR products were normalized based on the concentration of DNA from 250-400 bp then pooled. The pooled PCR products were cleaned using the Mag-Bind RxnPure Plus (Omega Bio-tek) according to the manufacturer’s protocol. The cleaned pool was sequenced on the MiSeq using v2 2x250 base pair kit (Illumina, Inc).

2.3 Analyses

2.3.1 Data Processing: Mothur

Raw sequences were processed and analyzed in Mothur version 1.39.5 (Schloss et al., 2009) according to the standard MiSeq protocol.
Quality control, trimming and de-noising were performed as previously outlined (Kozich, et al., 2013). All sequences were aligned to the Silva reference alignment database version 128 and filtered so that they all overlapped with no overhangs. Sequences were clustered into OTUs based on 97% similarity, and up to the genus-level taxonomic affiliation of each OTU was identified according to current taxonomy (Silva v128). The Mothur script can be found in the Appendix A.

2.3.2 Data Processing: DADA2

Raw sequences were processed through DADA2 pipeline according to the DADA2 pipeline tutorial 1.2 in R (Callahan et al., 2016). The DADA2 R script can be found in the Appendix B. Filtering was completed according to the DADA2 protocol based on the number of ambiguous bases, a minimum quality score, and the expected errors in a read (Callahan 2016). The filtered fastq files were dereplicated and output as unique sequences with their corresponding abundance. Denoising and merging were completed according to the DADA2 pipeline tutorial, while error parameters removed chimeras (Callahan 2016). All sequences were aligned to the current Silva reference alignment database (Silva v128), the same as was used for the Mothur analysis.

2.3.3 Data Analysis: Unifrac

Unifrac (Lozupone and Knight, 2005) distance matrices were created between bird gut samples to compare and represent individual gut microbial communities. Distances are calculated based on how much of the branch length is unique or shared in a common phylogenetic tree between the microbial gut communities. Unifrac distances were calculated in Mothur and DADA2. A DADA2 script (Appendix B) was applied to DADA2’s seqtab.nochim file. R programs DECIPHER and phanghorn were used to create alignment and tree roots.
respectively (Wright 2015). The tree was then merged with the otu and tax table into a single phyloseq object. Ordinations of weighted and unweighted Unifrac were created with non-metric multidimensional scaling (NMDS) distance matrixes. Mothur tree was created in Mothur under MiSeq SOP. The stability.tre file was merged with existing Mothur shared, constaxonomy, and mapping files in R with import.mothur command through phyloseq. Again a singular phyloseq object was created and ordinations were calculated with NMDS distance matrixes to create weighted and unweighted Unifrac plots. Scripts can be found in Appendix C.

2.3.4 Data Analysis: Bray-Curtis

Ordinations of Bray-Curtis were calculated for Mothur and DADA2. In phyloseq, each sample was square root-transformed to bacterial phylum where a percentage was calculated to create a pairwise distance matrix. The two-dimensional positions of the samples were more similar the closer they were to one another and more different the farther they were apart. Wisconsin distances (Bray-Curtis) were calculated from previously created phyloseq objects where an OTU table, taxonomy table, and mapping file were merged together. Ordinations were first calculated and then plotted against the mapping file to visualize trends. Scripts can be found in Appendix C.

Chapter 3: Results

3.1 Alpha diversity

Taxonomy bar charts were first created to note differences among taxonomic ranks according to Table 1. All samples can be seen at the class level for Mothur (Figure 1) and DADA2 (Figure 2). The most abundant 12 bacterial classes were isolated for better visualization. Bacilli, Bacterodia, and Mollicutes were most abundant among samples. Differences between Mothur and DADA2 can be seen at the class level among all samples, however this becomes
more evident when the first 8 samples are plotted at the phylum level (Mothur in Figure 3 and DADA2 in 4). Despite an identical dataset, Mothur and DADA2 assign different relative abundances of bacterial phyla against an identical reference dataset (Silva v128). For example, the third (*Eurillas latirostris*) and sixth (*Cecropis abyssinica*) samples from the left showed a substantial increase in Proteobacteria, on between using OTUs (Figure 3) compared to the ASVs (Figure 4).

Firmicutes and Proteobacteria were the most abundant phyla among samples. Mothur and DADA2 both detected 29 phyla against Silva version 128. However, at lower taxonomic ranks the two methodologies deviate from one another, as can be seen in Table 1. Mothur has a noticeable increase in classes, orders, and genus compared to DADA2. This becomes evident in the alpha diversity plots of Mothur (Figure 5) and DADA2 (Figure 6). Observed OTUs/ASVs, Shannon and Simpson indexes produced similar plots (Figure 5). Mothur also has 9332 OTUs compared to 4245 ASVs in DADA2 (Table1). Despite the larger number of OTUs compared to ASVs, the same trends can be seen in observed OTUs/ASVs alpha diversity test (Figure 6). Although some discrepancies can be noted, it appears alpha diversity was not affected by which program was used.
Table 1: Comparative table of different taxonomy classes, OTUs/ASVs and total sequences between Mothur and DADA2. Both pipelines’ taxonomies are based off the reference database Silva version 128.

<table>
<thead>
<tr>
<th>Taxa</th>
<th>Mothur</th>
<th>DADA2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phylum</td>
<td>29</td>
<td>29</td>
</tr>
<tr>
<td>Class</td>
<td>75</td>
<td>58</td>
</tr>
<tr>
<td>Order</td>
<td>150</td>
<td>102</td>
</tr>
<tr>
<td>Genus</td>
<td>932</td>
<td>507</td>
</tr>
<tr>
<td>OTUs/ASVs</td>
<td>9332</td>
<td>4245</td>
</tr>
<tr>
<td>Mars</td>
<td>SpecimenID</td>
<td>Sex</td>
</tr>
<tr>
<td>------</td>
<td>------------</td>
<td>-----</td>
</tr>
<tr>
<td>401</td>
<td>401</td>
<td></td>
</tr>
<tr>
<td>567</td>
<td>567</td>
<td></td>
</tr>
<tr>
<td>650</td>
<td>650</td>
<td></td>
</tr>
<tr>
<td>651</td>
<td>651</td>
<td></td>
</tr>
<tr>
<td>652</td>
<td>652</td>
<td></td>
</tr>
<tr>
<td>653</td>
<td>653</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Metadata table of all Equatorial Guinea samples. MARS refers to unique sample names assigned by MARS facility, used to track samples through analysis. Samples were collected from 6 different sites and longitudes and latitudes are listed for specificity. Diet assignments of c, o, and h refer to carnivorous, omnivorous and herbivorous respectively.
<table>
<thead>
<tr>
<th>MARS</th>
<th>Proveid</th>
<th>Sex</th>
<th>Provincia</th>
<th>Locality</th>
<th>Elevation</th>
<th>Lat</th>
<th>Long</th>
<th>Diet</th>
<th>Order</th>
<th>Family</th>
<th>Genus</th>
<th>Species</th>
<th>Common_Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>458a</td>
<td>Green-backed Shrike</td>
<td>Male</td>
<td>West Nams</td>
<td>Oudna, 1 km to the Universidad Americana de Guayaquil</td>
<td>3.361</td>
<td>8.662</td>
<td>18</td>
<td>1</td>
<td>2016</td>
<td>Passeriformes</td>
<td>Monarchidae</td>
<td>Terpsiphone rufiventer</td>
<td>Red_bellied_paradise_flycatcher</td>
</tr>
<tr>
<td>459</td>
<td>Green-headed Shrike</td>
<td>Male</td>
<td>West Nams</td>
<td>Oudna, 1 km to the Universidad Americana de Guayaquil</td>
<td>3.361</td>
<td>8.662</td>
<td>18</td>
<td>1</td>
<td>2016</td>
<td>Passeriformes</td>
<td>Monarchidae</td>
<td>Terpsiphone atrocaudata</td>
<td>Black-crested_flycatcher</td>
</tr>
<tr>
<td>460</td>
<td>Green-winged Pytilia</td>
<td>Male</td>
<td>West Nams</td>
<td>Oudna, 1 km to the Universidad Americana de Guayaquil</td>
<td>3.361</td>
<td>8.662</td>
<td>18</td>
<td>1</td>
<td>2016</td>
<td>Passeriformes</td>
<td>Monarchidae</td>
<td>Terpsiphone atrocaudata</td>
<td>Black-crested_flycatcher</td>
</tr>
</tbody>
</table>

11
| ID  | SampleID | Sex | Provincia | Locality         | Elevation | Lat   | Long   | Day | Month | Year | Diet     | Order     | Family    | Genus       | Species       | Common_Name     |
|-----|----------|-----|-----------|------------------|-----------|-------|--------|-----|-------|------|----------|-----------|-----------|-------------|--------------|----------------|------------------|
| 1   | 301      | Male| Bioko     | Pico Basile      | 2670      | 3.595 | 8.769  | 1   | 2     | 2016 | o        | Passeriformes| Zosteropidae| Zosterops  | brunneus     | Fernando_Po_speirops |
| 2   | 300      | Female| Bioko      | Pico Basile      | 2670      | 3.595 | 8.769  | 1   | 2     | 2016 | o        | Passeriformes| Zosteropidae| Zosterops  | brunneus     | Fernando_Po_speirops |

... Additional rows for other samples...

12
Figure 1: Relative abundance chart of the top 12 bacterial classes in Mothur of all samples. There are a total of 142 samples and 75 classes (data not shown) found through the DADA2 pipeline. X-axis refers to bird genus and species name while the legend refers to abundant bacterial classes.
Figure 2: Relative abundance chart of the top 12 bacterial classes in DADA2 of all samples. There are a total of 142 samples and 75 classes (data not shown) found through the DADA2 pipeline. X-axis refers to bird genus and species name while the legend refers to abundant bacterial classes.
**Figure 3:** Relative abundance chart of bacterial phylum in Mothur. These are the first 8 samples with the top 8 phylum. X-axis is bird genus and species names with identical species representing samples taken/dissections performed on the same bird *. Legend refers to abundant bacterial phylum.
Figure 4: Relative abundance chart of bacterial phylum in DADA2. These are the first 8 samples with the top 8 phylum. X-axis is bird genus and species names with identical species representing samples taken/ dissections performed on the same bird *. Legend refers to abundant bacterial phylum.
**Figure 5:** Comparison of estimated number of OTUs (top) to ASVs (bottom) across bird orders with diet category highlighted. Shannon and Simpson indexes also included. A-Apodiformes, Ch-Charadriiformes, Cl-Columbiformes, Cr-Coraciiformes, Cu-Cuculiformes, Pa-Passeriformes, and Pc-Piciformes.
**Figure 6:** Comparison of estimated number of OTUs (left) to ASVs (right) across bird orders with diet category highlighted. A-Apodiformes, Ch-Charadriiformes, Cl-Columbiformes, Cr-Coraciiformes, Cu-Cuculiformes, Pa-Passeriformes, and Pc-Piciformes.
3.2 Beta diversity

Bray-Curtis distances were calculated between all pairwise sample comparisons in both Mothur and DADA2 processed data. NMDS plots of Bray-Curtis were plotted against several variables including bird order, diet, and overall locality (Mothur results Figures 7, 8, 9; DADA2 results Figures 10, 11, 12). No obvious trends on both the Mothur and DADA2 NMDS plots can be seen through the variables mentioned or others not shown. However, variations can be seen in all Mothur vs DADA2 plots that may be visualized under the variables tested. Procrustes analyses were run to show these variations. Figure 13 shows a Procrustes analysis of the Mothur vs DADA2 NMDS plots where arrows show movement of Mothur data points to DADA2. There is a significant difference between Mothur and DADA2 in which DADA2 points are drawn across a larger axis and various samples are scattered on opposing ends depending on which program was used.

Data included both sample replicates, where individual intestines were sequentially extracted, and PCR replicates from a single extraction. Replicate data showed interesting trends between the two different programs and between the replicates themselves. Sample replicates yielded different bacterial species and amounts (Figure 14, 15), but were largely similar (see * Figure 3, 4). Figures 14 and 15 highlight the variation of replicates. Some groups of replicates cluster together as seen by the connected black lines, which is to be expected. However there are several instances of replicates that cluster away from its other replicate partners. A Procrustes analysis was run on the replicates-only datasets from Mothur and DADA2 (Figure 16). Again arrows refer to the movement of Mothur points to DADA2. The trends noted in Figure 13 are more evident with the smaller sample size.
Figure 7: NMDS plot of Mothur data with Bray-Curtis distance applied. Figure legend labels are used as reference only and are not meant to show particular trends or groupings. Colored labels of sample orders refer to classified bird orders. Bird orders are Apodiformes-orange, Charadriiformes-yellow, Columbiformes-green, Coraciformes-cyan, Cuculiformes-blue, Passeriformes-purple, Piciformes-pink.
**Figure 8:** NMDS plot of DADA2 data with Bray-Curtis distance applied. Figure legend labels are used as reference only and are not meant to show particular trends or groupings. Colored labels of sample orders refer to classified bird orders. Bird orders are Apodiformes-orange, Charadriiformes-yellow, Columbiformes-green, Coraciformes-cyan, Cuculiformes-blue, Passeriformes-purple, Piciformes-pink.
**Figure 9:** NMDS plot of Mothur data with Bray-Curtis distance applied. Figure legend labels are used as reference only and are not meant to show particular trends or groupings. Colored labels of diet labeled c-carnivorous, h-herbivorous, and o-omnivorous refers to diet of particular bird.
**Figure 10:** NMDS plot of DADA2 data with Bray-Curtis distance applied. Figure legend labels are used as reference only and are not meant to show particular trends or groupings. Colored labels of diet labeled c-carnivorous, h-herbivorous, and o-omnivorous refers to diet of particular bird.
Figure 11: NMDS plot of Mothur data with Bray-Curtis distance applied. Colored labels of refer to sampling locality of bird.
Figure 12: NMDS plot of DADA2 data with Bray-Curtis distance applied. Colored labels of refer to sampling locality of bird.
**Figure 13:** Procrustes analysis of DADA2 and Mothur NMDs plots under Bray-Curtis distance application. Arrows refer to movement of Mothur data plot points to Dada2 points.
Figure 14: NMDS plot of Mothur replicate data with Bray-Curtis distance applied. Colored labels of refer to bird order. Colored lines refer to sample replicates where back lines connect dissected samples from the same bird and red lines connect PCR replicates in the same bird.
Figure 15: NMDS plot of DADA2 replicate data with Bray-Curtis distance applied. Colored labels of refer to bird order. Colored lines refer to sample replicates where back lines connect dissected samples from the same bird and red lines connect PCR replicates in the same bird.
**Figure 16**: Procrustes analysis of DADA2 and Mothur replicates NMDs plots under Bray-Curtis distance application. Arrows refer to movement of Mothur data plot points to Dada2 points.
Chapter 4: Discussion

How scientists analyze data matters. The defining of the taxonomic units is an extremely vital and foremost step. New data analysis pipelines come out regularly as the data science field continues to grow. As the importance of statistical analysis pipelines becomes more evident, there is an ever-developing need to become more standardized in our methods. While new analytical tools may find niches among those who have specific needs, an overpopulation of methods creates a lack of uniformity and adds difficulty to comparing across studies. Determining which tools work best will allow scientists to more easily share their data and create an environment where other scientists can more readily check others’ statistical analyses and apply similar methods to their own data. Ideally, all microbiome figures, tables, and other types of data representation would be directly comparable.

The differences between Mothur’s 97% OTUs and DADA2’s ASVs were clear (Table 1). Despite the same number of phyla being detected in both methods, there is significant deviation of taxonomic classes that escalates at lower taxonomic ranks. The Procrustes analysis of the Mothur and DADA2 NMDS plots shows differences in the plotting of OTUs and ASVs. The microbiome is a growing scientific field, as scientists look for a greater understanding of these unique systems. Data analysis pipelines are key in this field to interpret biological results. The determination of use of OTUs or ASVs could have a larger impact on gut microbiome data than initially anticipated. Reproducibility between runs or entire studies could have significant effects on large studies such as the Human Microbiome Project.

Wild host microbiomes will always be vital for understanding the many factors that shape this complex system. Only through a wild host gut microbiome can one begin to catalogue the bacteria that live there and the features that contribute to its formation. There is difficulty
sampling most wild hosts without invasive procedures, however the collaboration with natural history museums can allow scientists to increase the practical use of specimens that are already being collected. Birds play a particular role in the understanding of gut microbiomes. Birds have a symbiotic relationship with microorganisms that can likely be traced back before the Aves lineage was created. From an evolutionary standpoint, the genetic relationship between birds and their gut microbiome could illuminate various questions about microbiomes in other systems. The global distribution of birds also allows for the research of the effects of diet, climate, and environment on the gut microbiome.

The data used for pipeline comparisons is unpublished data of bird gut samples from Equatorial Guinea. The samples are diverse among bird species, orders, location, sex, and diet. The gut sampling process also proved diverse as they were cut at varying sections along the gastrointestinal tract. There are key differences among samples that are directly related to where the samples were taken along the gastrointestinal tract. For this reason, sampling location uniformity may be a larger issue than initially anticipated. Standardization at this initial step must be taken into consideration to ensure similar regions are sampled. We have yet to fully understand the impact of sampling different regions of bird microbiomes. Until variation along the gastrointestinal tract has been further described on the micro-scale we must continue sampling similarly across studies.

Chapter 5: Conclusion

The highlight of OTU/ASV differences do not equate to a preference of one over another, but rather displays their differences in analyzing a particular dataset. Our dataset has displayed double the amount of taxonomic units using 97% OTUs compared to ASVs. This highlights the biological differences between these two methods. While it may be true that many tools fulfill a
specific scientist’s needs, the differences among them matter. For these reasons, understanding your own dataset and data processing protocol is vital whenever taxonomic units are required.
Appendix A: Mothur R Script

rm(list=ls())

setwd("/Users/dariencapunitan/Desktop/bird_poop") # set working directory with full file path
library(vegan) # call package
library(ggplot2)
library(dplyr)
library(scales)
library(grid)
library(reshape2)
library(Matrix)
library(phyloseq)

theme_set(theme_bw())

sharedfile = "stability.TWO.shared"
taxfile = "stability.TWO.cons.taxonomy"
mapfile = "EQ_map_forRCollab.csv"

mothur_data <- import_mothur(mothur_shared_file = sharedfile, mothur_constaxonomy_file = taxfile)
map <- read.csv(mapfile, row.names = c(1), sep=".", header=TRUE)
head(map)
map <- sample_data(map)

moth_merge <- merge_phyloseq(mothur_data, map)
moth_merge

colnames(tax_table(moth_merge))
colnames(tax_table(moth_merge)) <- c("Kingdom", "Phylum", "Class", "Order", "Family", "Genus")
Appendix B: DADA2 R Script

```r
rm(list=ls())

setwd("/Users/dariencapunitan/Desktop/HirdEG_fastq_Dada")

#source("https://bioconductor.org/biocLite.R")
#biocLite("dada2")

#biocLite("devtools")
#library("devtools")
#devtools::install_github("benjjneb/dada2")
library(tibble)
library(dada2); packageVersion("dada2")
library(ShortRead); packageVersion("ShortRead")
library(ggplot2); packageVersion("ggplot2")

path <- "/Users/dariencapunitan/Desktop/HirdEG_fastq_Dada"

fns <- list.files(path)

fastqs <- fns[grepl("\.fastq\$", fns)]
fastqs <- sort(fastqs) # Sort ensures forward/reverse reads are in same order
fnFs <- fastqs[grepl("_R1", fastqs)] # Just the forward read files
fnRs <- fastqs[grepl("_R2", fastqs)] # Just the reverse read files
sample.names <- sapply(strsplit(fnFs, "_"), `[`, 1)
# Specify the full path to the fnFs and fnRs
fnFs <- file.path(path, fnFs)
fnRs <- file.path(path, fnRs)

plotQualityProfile(fnFs[[174]])
```

34
plotQualityProfile(fnRs[[174]])

# Make directory and filenames for the filtered fastqs
filt_path <- file.path(path, "filtered")
if(!file_test("-d", filt_path)) dir.create(filt_path)
filtFs <- file.path(filt_path, paste0(sample.names, ".F_filt.fastq.gz"))
filtRs <- file.path(filt_path, paste0(sample.names, ".R_filt.fastq.gz"))

# Filter
for(i in seq_along(fnFs)) {
  fastqPairedFilter(c(fnFs[i], fnRs[i]), c(filtFs[i], filtRs[i]),
                    truncLen=c(240,160),
                    maxN=0, maxEE=c(2,2), truncQ=2, rm.phix=TRUE,
                    compress=TRUE, verbose=TRUE)
}
derepFs <- derepFastq(filtFs, verbose=TRUE)
derepRs <- derepFastq(filtRs, verbose=TRUE)

# Name the derep-class objects by the sample names
names(derepFs) <- sample.names
names(derepRs) <- sample.names

dadaFs.lrn <- dada(derepFs, err=NULL, selfConsist = TRUE, multithread=TRUE)

errF <- dadaFs.lrn[[1]]$err_out
dadaRs.lrn <- dada(derepRs, err=NULL, selfConsist = TRUE, multithread=TRUE)
errR <- dadaRs.lrn[[1]]$err_out

plotErrors(dadaFs.lrn[[1]], nominalQ=TRUE)

dadaFs <- dada(derepFs, err=errF, multithread=TRUE)
dadaRs <- dada(derepRs, err=errR, multithread=TRUE)
mergers <- mergePairs(dadaFs, derepFs, dadaRs, derepRs, verbose=TRUE)
# Inspect the merger data.frame from the first sample
head(mergers[[1]])

seqtab <- makeSequenceTable(mergers[names(mergers) != "Mock"])
dim(seqtab)
table(nchar(getSequences(seqtab)))

seqtab.nochim <- removeBimeraDenovo(seqtab, verbose=TRUE)
dim(seqtab.nochim)
sum(seqtab.nochim)/sum(seqtab)
taxa <- assignTaxonomy(seqtab.nochim, "silva_nr_v128_train_set.fa")
unname(head(taxa))

library(phyloseq); packageVersion("phyloseq")
library(ggplot2); packageVersion("ggplot2")
library(data.table)
library(reshape)
theme_set(theme_bw())

#import mapping file similar to mothur import
mapfile = "EQ_map_forRCollab.csv"
map <- read.csv(mapfile, row.names = c(1), sep="", header=TRUE)
head(map)
map <- sample_data(map)

#importing changed taxa file
taxfile = "DadaTaxa2.csv"
taxa2 <- read.csv(taxfile, row.names = c(1), sep="","", header=TRUE)
head(taxa2)
taxa2 <- sample_data(taxa2)

# Construct phyloseq object (straightforward from dada2 outputs)
ps <- phyloseq(otu_table(seqtab.nochim, taxa_are_rows=FALSE),
               tax_table(taxa))
ps
moth_merge <- merge_phyloseq(ps, map)
moth_merge
Appendix C: Diversity Calculations

#alpha diversity
plot_richness(moth_merge, x="Order_Abrev", measures=c("Observed", "Shannon", "Simpson"), color="Diet") + theme_bw()

#Beta diversity
#NMDS/Bray
ord.nmds.bray <- ordinate(moth_merge, method="NMDS", distance="bray")
plot_ordination(moth_merge, ord.nmds.bray, color="Sample_Order", title="Bray NMDS X")

#Tree for Unifrac
source("https://bioconductor.org/biocLite.R")
biocLite("DECIPHER")
library(DECIPHER)
seqs <- getSequences(seqtab.nochim)
names(seqs) <- seqs # This propagates to the tip labels of the tree
alignment <- AlignSeqs(DNAStringSet(seqs), anchor=NA, verbose=FALSE)

#construct a neighbor-joining tree, and then fit a Generalized time-reversible with Gamma rate variation
#maximum likelihood tree using the neighbor-joining tree as a starting point.

library(phangorn) #choose no when from compilation
phangAlign <- phyDat(as(alignment, "matrix"), type="DNA")
dm <- dist.ml(phangAlign)
treeNJ <- NJ(dm) # Note, tip order != sequence order
fit = pml(treeNJ, data=phangAlign)
fitGTR <- update(fit, k=4, inv=0.2)
fitGTR <- optim.pml(fitGTR, model="GTR", optInv=TRUE, optGamma=TRUE, rearrangement = "stochastic", control = pml.control(trace = 0))
detach("package:phangorn", unload=TRUE)
Appendix D: Procrustes Analysis

pro <- procrustes(Mothur.nmds, Dada.nmds)

pro

summary(pro)

plot(pro)
Appendix E: Unifrac Analysis

```r
ps_dada <- phyloseq(otu_table(tax_table, taxa_are_rows=FALSE),
                    sample_data(moth_merge),
                    tax_table(taxa), phy_tree(fitGTR$tree))

ps_dada

ord.pcoamothur.unifrac <- ordinate(ps_dada, method="NMDS", distance="unifrac")

plot_ordination(ps_dada, ord.pcoamothur.unifrac, color="Diet", title="Uni Un NMDS")
```
References


Westcott, S. L. & Schloss, P. D. OptiClust, an improved method for assigning amplicon-based sequence data to operational taxonomic units. mSphere 2, 1–11 (2017).
