How Bacterial Diversity Impacts the Hatching Success of European Birds

Sonya Utecht, University of Arkansas for Medical Sciences, Little Rock, AR

Electronic Supplementary Materials available on [GitHub]

Abstract

It is hypothesized that bacterial diversity on wild bird eggshells could have an impact on hatching success. To study the impact of bacteria on bird clutch survival several researchers in Spain monitored over 600 nests and took samples for analysis [1]. Among the data collected were over 600 rRNA sequences that this study is reanalyzing [1]. The original work's clear methodology highly encourages reproduction [1]. The final conclusion is that bacterial diversity does not impact hatching success, however good scientific writing can still allow even negative results to be useful to the field.

Introduction

The bacteria on eggshells may have an important role in the survival and evolutionary adaptations of birds and bacteria [1]. We know that incubation reduces bacterial density while also increasing the temperature and increasing hatching success [1]. But many questions are still unanswered. How does incubation change the community assembly? Do the taxa of the bacteria have an association with hatching success? How have bacteria and bird species evolved together? Several studies have looked into these ideas, but have conflicting results [1]. This study hopes to add to the literature available as we try to understand microbial and animal evolution.

In addition, since this is a reproduction using data gathered by another set of researchers, I hope to help promote the importance of designing studies and writing papers with the goal of reproduction and replication at the forefront. As science is in a state of crises over poor replicability [3][4], researchers must work to make elaboration on the steps of a study a true focus. Due to the careful method descriptions in the original paper, not only can evolutionary microbiologists and ornithologists learn something, but other scientific fields, especially those using genetic sequencing techniques, can see good examples of what makes for a clear and reproducible study. In support of this, I have included electronic supplementary materials with the exact commands that I ran during my analysis so future researchers can easily verify and replicate my steps.

I hypothesized that bacterial diversity will have a noticeable relationship with hatching success, and that certain bacterial taxa will have higher appearance rates in clutches with lower success rates. 16S rRNA Sequence data was obtained from the *the European Nucleotide Archive* [2] from the paper *"Bacterial density rather than diversity correlates with hatching success across different avian species"* [1].

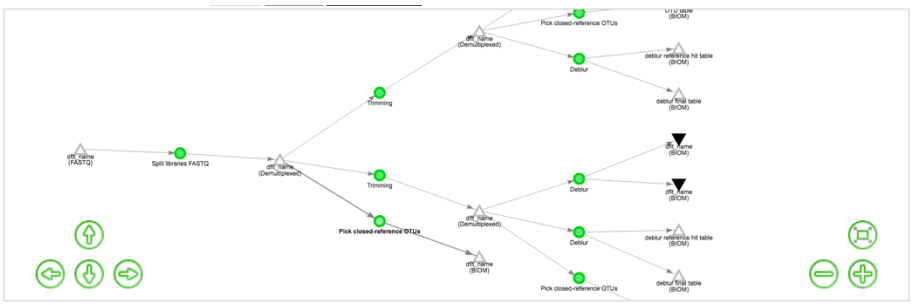
The authors approached the hypothesis that bacterial assembly and density impacts hatching success [1]. To research this, they monitored 600 nest boxes and several wild nests in the Hoya de Guadix plateau of Spain [1]. Nests were visited daily from ~2 days before first hatching until ~3 days after the end of hatching [1]. Swabs were gathered at the beginning of incubation and after hatching was complete [1]. Length and width off each egg was measured with a caliper after swabbing [1]. Data from any nests where no birds hatched was thrown out, as the nest may have been abandoned or unfertile [1]. The collected swabs were used for culturing in agar plates which were used to estimate the bacterial density [1]. Fingerprinting was performed with ARISA and 16s rRNA sequencing was performed with HiSeq Illumina which were used to study bacterial community assembly [1]. The final dataset included 157 clutches from 17 bird species and 609 rRNA sequences [1] [5].

Methods

Imported into QIIME2

Sequence data was first obtained from *Qiita[5]*. The Qiita repository [5] is a fantastic tool for improving scientific replicability that can include extremely useful provenance graphics that show the pipeline the authors ran through and highlights many of their parameters that are used. When originally working with the Qiita [5] dataset however, I ran into several issues because the data was not demultiplexed. After several attempts to demultiplex according to the original paper's methods I moved to using a previously demultiplexed dataset.

After a trial run on a data subset to work out any issues in the planned pipeline, demultiplexed sequence data was downloaded from *the European Nucleotide Archive* [2] onto the UAMS HPC. Data was then imported into QIIME2 v2019.1[6]. The original study used QIIME v1.9 [1][7].



Job 8c9482fb-cb17-4377-a814-2edf20de80d6

Command:	Pick closed-reference OTUs (QIIMEq2 1.9.1) OTU picking using a closed reference approach
Status:	success
Job parameters:	
reference-tax:	/databases/gg/13_8/taxonomy/97_otu_taxonomy.txt
similarity:	0.97
reference-seq:	/databases/gg/13_8/rep_set/97_otus.fasta

Figure 1: Qiita Provenance graph showing job parameters[5]

Deblur Trim Length 100

After data was imported, it was filtered [17] and the Deblur tool was used for denoising. Deblur was chosen over DADA2 because it was the tool used in the original analysis [1][5][8][9]. The trim length 100 was specified in the Qiita pipeline [1][5][8].

```
qiime quality-filter q-score \
 --i-demux single-end-demux.qza \
 --o-filtered-sequences demux-filtered.qza \
 --o-filter-stats demux-filter-stats.qza

qiime deblur denoise-16S \
  --i-demultiplexed-seqs demux-filtered.qza \
  --p-trim-length 100 \
  --p-jobs-to-start 36 \
  --o-representative-sequences rep-seqs-deblur.qza \
  --o-table table-deblur.qza \
  --p-sample-stats \
  --o-stats deblur-stats.qza
```

Closed Reference OTU picking (Greengenes [10])

Next Closed Reference OTU picking was preformed with Greengenes. This was based off data in the project page on Qiita [1] [5]. The paper used Open Reference OTU picking with Greengenes v10_13 using a method combing closed and open reference picking styles [1] [10] [11]. However, this was not clearly detailed in the paper with any parameters, while Qiita

specifically listed closed reference along with the similarity and reference parameters [5]. I downloaded the Greengenes file set and imported the files in to QIIME2, and then ran the QIIME2 commands for Closed Reference OTU Picking [6] [10] [14].

```
qiime vsearch cluster-features-closed-reference \
  --i-table table-deblur.qza \
  --i-sequences rep-seqs-deblur.qza \
  --i-reference-sequences gg_otu_97.qza \
  --p-perc-identity 0.97 \
  --p-threads 0 \
  --o-clustered-table table-cr-97.qza \
  --o-clustered-sequences rep-seqs-cr-97.qza \
  --o-unmatched-sequences unmatched-cr-97.qza
```

Filtering of mitochondria, chloroplasts, archaea, and low frequency OTUs

At this point data was transferred to another machine, a 2017 Macbook Pro on High Sierra v10.13.6 running QIIME2 v2018.11[6]. Next the Archaea, Chloroplasts, Mitochondria, and low frequency OTUs were removed. QIIME2 included a way to remove specific taxa with just the parameters of archaea, chloroplasts, and mitochondria, however for OTU frequency a parameter needed to be chosen [6]. The paper specified the frequency they removed was .005% and under of total OTU frequency. To calculate this for my remaining dataset I looked at the visualization of the data to get the total frequency and multiplied by .00005 for the approximate frequency to drop.

```
!qiime taxa filter-table \
  --i-table hpc/output/table-cr-97.qza \
  --i-taxonomy hpc/output/taxonomy.qza  \
  --p-exclude mitochondria,chloroplast,archaea \
  --o-filtered-table hpc/output/table-no-mitochloroarch.qza
```

```
!qiime taxa filter-seqs \
  --i-sequences hpc/output/rep-seqs-cr-97.qza \
  --i-taxonomy hpc/output/taxonomy.qza  \
  --p-exclude mitochondria,chloroplast,archaea \
  --o-filtered-sequences hpc/output/seqs-no-mitochloroarch.qza
```

```
!qiime feature-table filter-features \
  --i-table hpc/output/table-no-mitochloroarch.qza \
  --p-min-frequency 220 \
  --o-filtered-table hpc/output/filtered-table.qza
```

Align to tree / Core Metrics / Taxa Barplot

Before filtering there were over 3,000 OTUs. After final filtering was complete I had ~800 OTUs remaining, compared to the original authors ~550 OTUs [1]. It was a larger gap than I had hoped for, but not unusable. For analysis, I aligned the data to a tree[15][16], created PCoA plots

with the core metrics command, and created a taxonomic barchart. Only a select subset of these can be included in this paper, however the electronic supplementary materials contain the full set of analytic graphics.

Beta Diversity

A box and whisker visualization of hatching success was created based on Unweighted UniFrac data from the core metric output.

Results

The ratios of common bacterial taxa are similar between species regardless of their hatching success rates. The Unweighted and Weighted UniFrac distance PCoA plots on hatching success show that high hatching success and low hatching success nests all bundle together as similar with no clear separation [12] [13] [14]. Beta diversity significance also shows that high hatching success nests are just as similar to one another as they are to nests with mid to low hatching success [12] [13]. These results are similar to those found in the original study [1].

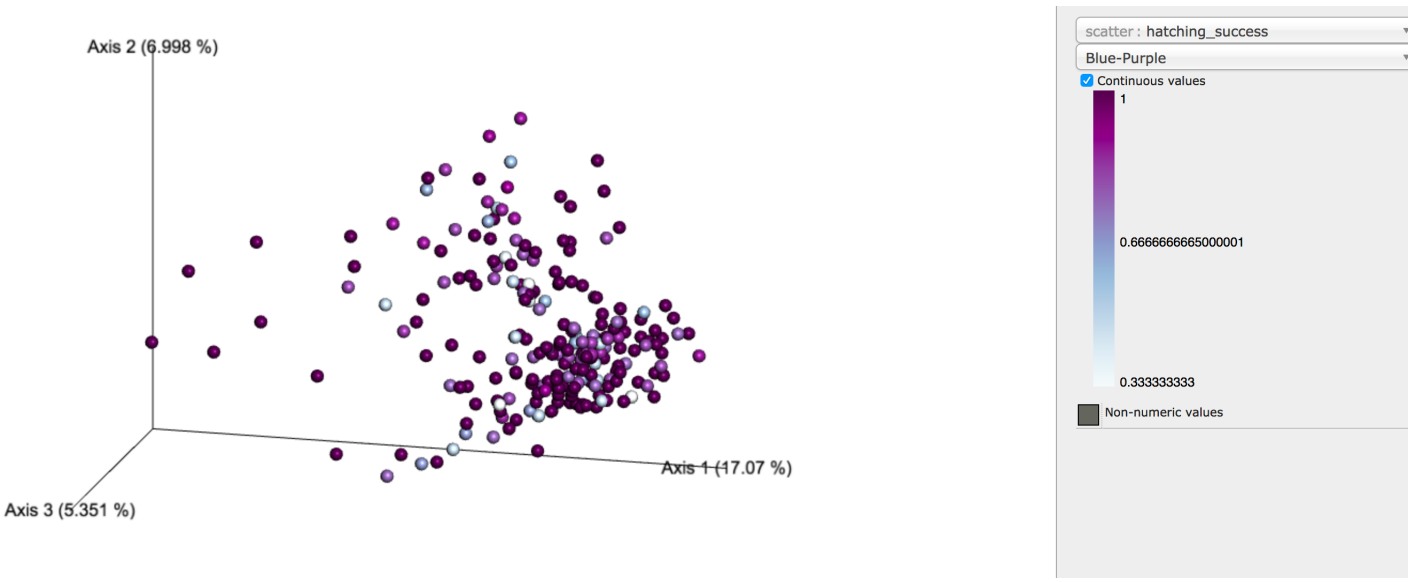


Figure 2 Unweighted UniFrac PCoA. High Hatching Success is dark purple with a gradient shift to blue for low hatching success nests. [12] [13].

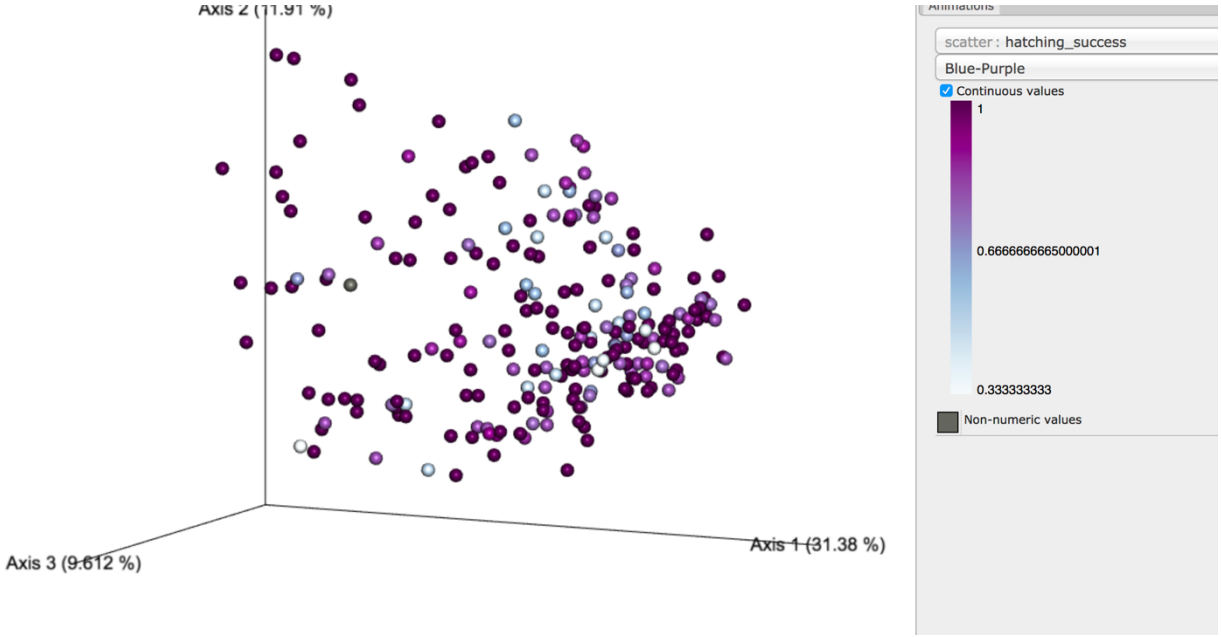


Figure 3 Weighted UniFrac PCoA. High Hatching Success is dark purple with a gradient shift to blue for low hatching success nests. [12] [13].

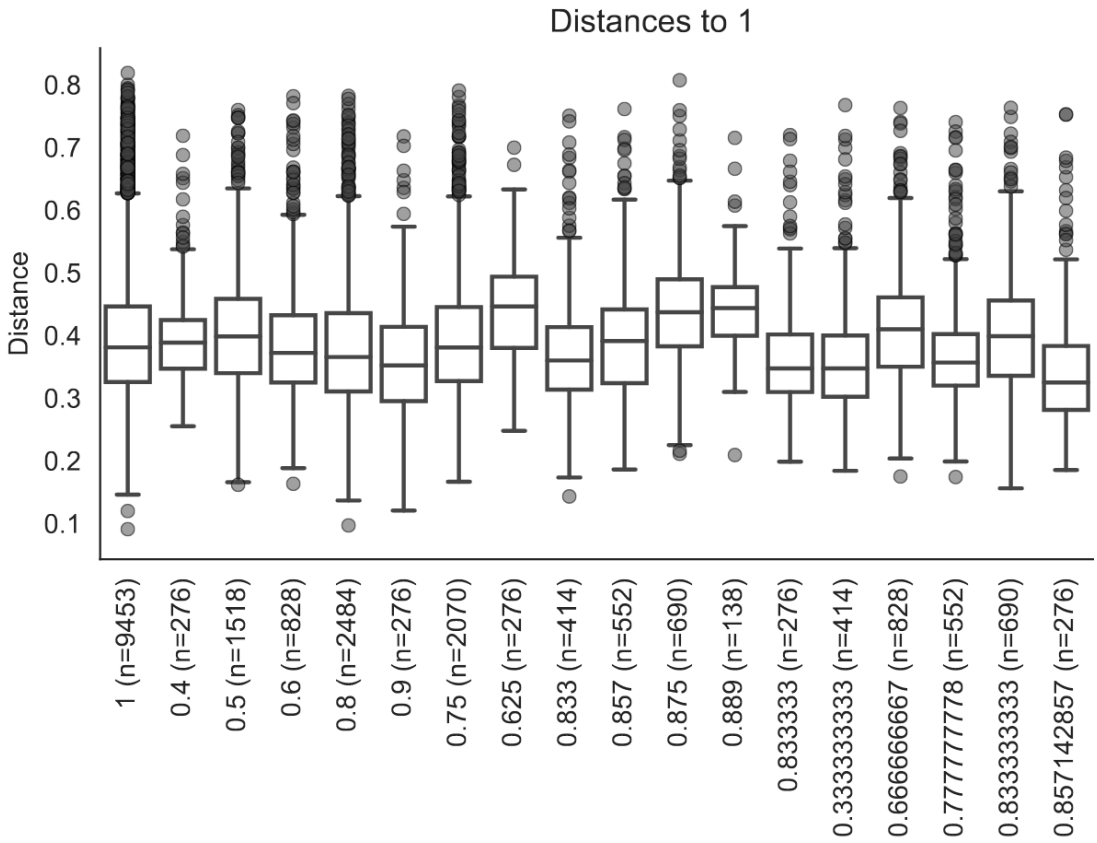


Figure 4 Box and Whisker plot of beta diversity from Unweighted UniFrac. [12] [13].

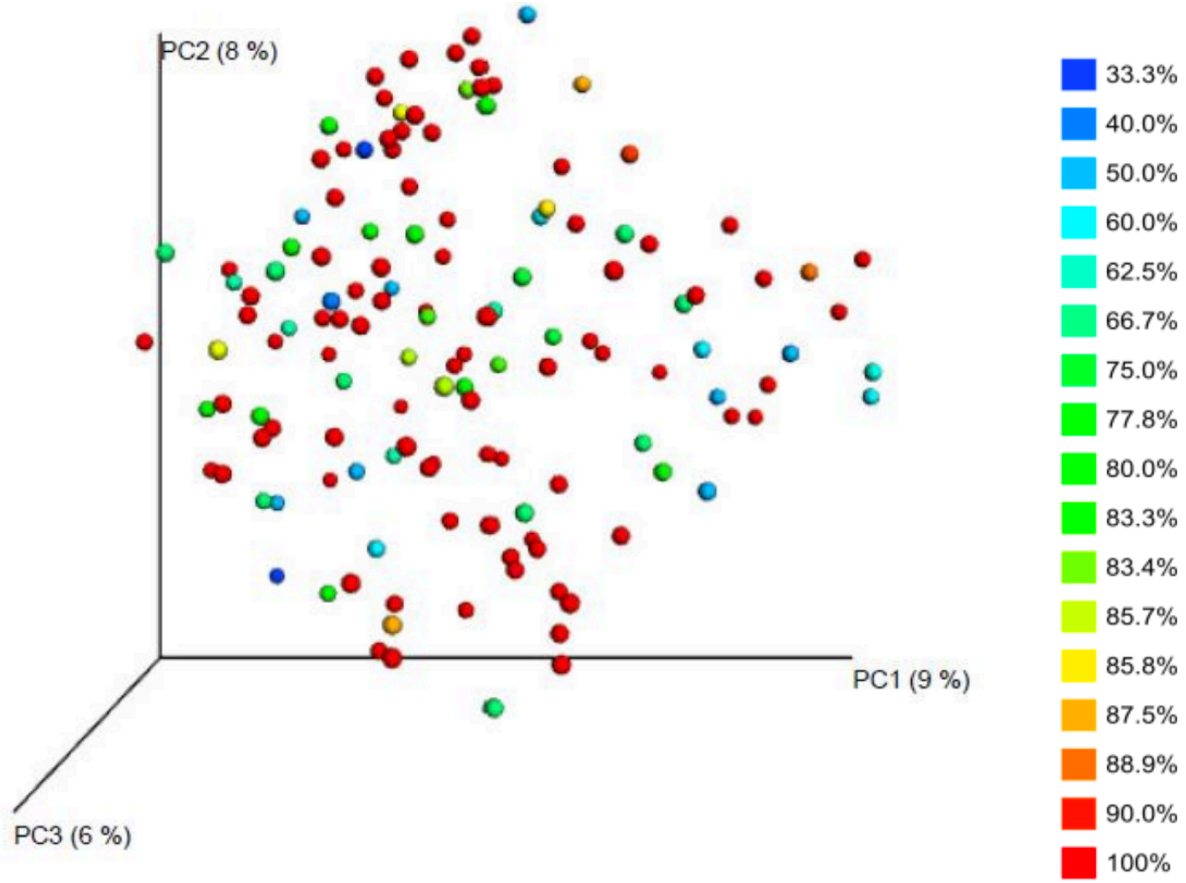


Figure 5 Unweighted UniFrac PCoA plot colored by Hatching Success from [1] [12] [13].

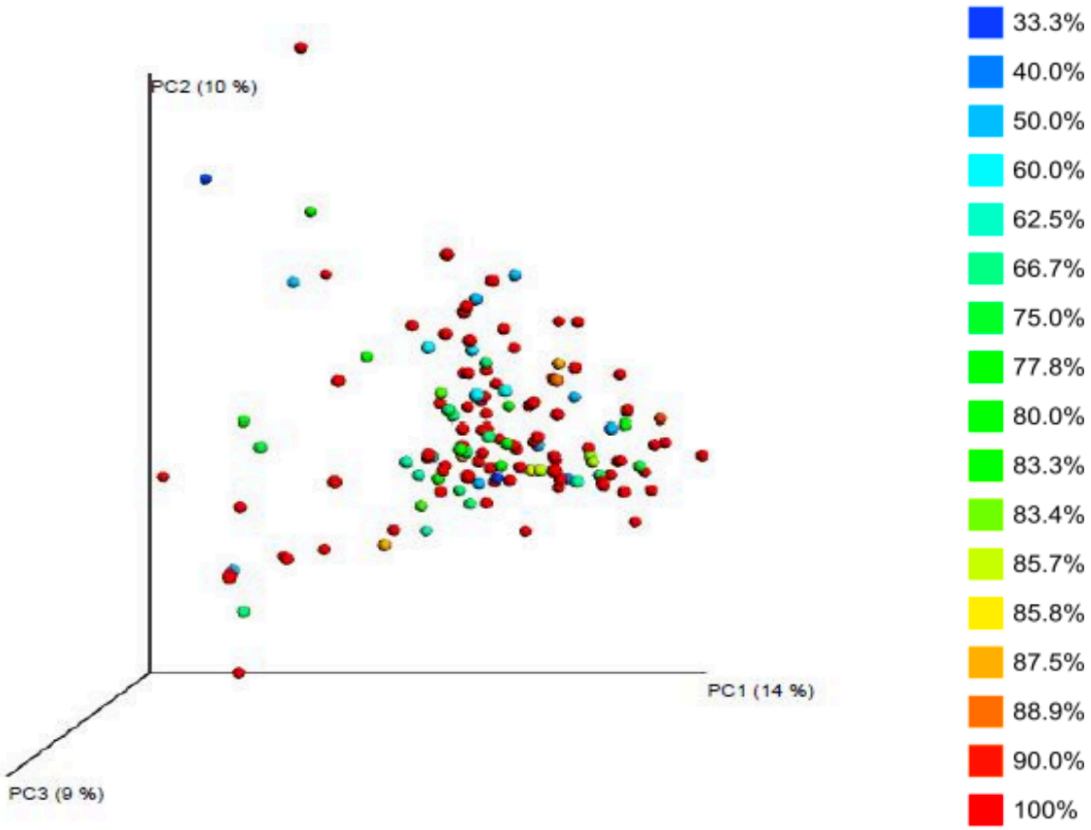


Figure 6: Weighted UniFrac PCoA plot colored by Hatching Success from [1] [12] [13].

The most common bacterial taxa found on European eggshells include Proteobacteria, Actinobacteria, Firmicutes, and Bacteroidetes.

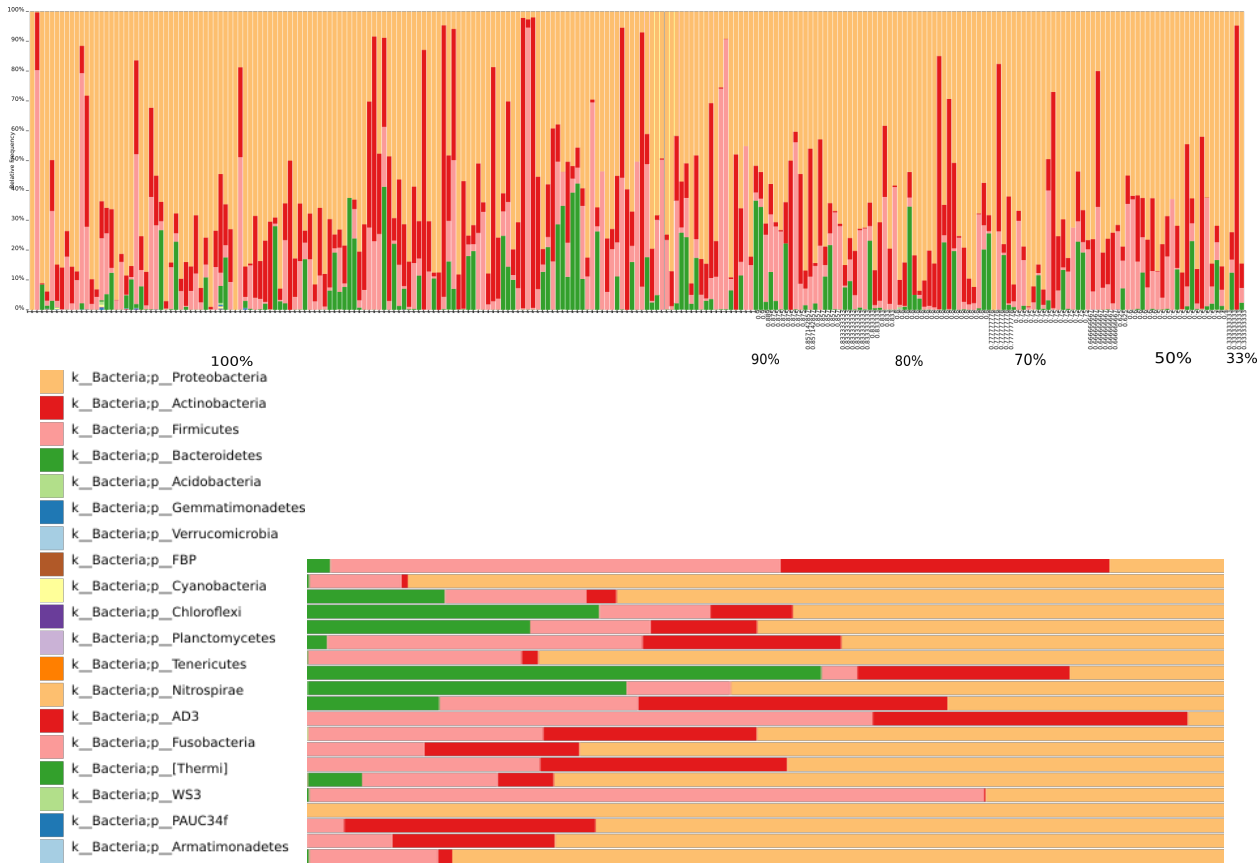


Figure 7: Taxa barplot showing hatching success by nest. The legend is bellow and to the left. The bottom right shows a subset for a particular species, the little owl Athene noctua, which had an average hatching success rate of 100.0% in the original study [1].

Discussion

The results are negative, in fact even though the original paper outlines the sequencing methods in detail, it focuses more heavily on the results of the culturing methods since those could be seen as a positive result. However, in science even negative results deserve to be considered, as focusing only on the positive is harmful for the scientific community. It removes the search for truth and knowledge instead searching for popularity and publishing power. It is worth noting that the bacterial taxa on European bird eggs do not have an influence on the clutch's survival. In addition, the detailing of sequence analysis methodology shows great value and encourages reproduction. Therefore, in this study I have worked to ensure my methods are well described, since a focus on reproduction allows a paper to benefit all of science and health, not just a particular niche or subfield.

Annotated Bibliography

1. *Peralta-Sánchez et al. "Bacterial density rather than diversity correlates with hatching success across different avian species" FEMS Microbiology Ecology, 2018, Vol. 94, Issue 3.*

The authors investigated whether bacterial density and bacterial assembly influenced the hatching success of wild birds in the Hoya de Guadix plateau of Spain. Nests were visited daily and swabs were gathered at the beginning of incubation and after hatching was complete. The length and width off each egg was measured with a caliper after swabbing. The collected swabs were used for culturing in agar plates which were used to estimate the bacterial density. Fingerprinting was performed with ARISA and 16s rRNA sequencing was performed with HiSeq Illumina which were used to study bacterial community assembly. They found from the culturing that density was negatively correlated with hatching success. However, the sequencing data did not show a strong correlation with hatching success. They noted some differences in the Unweighted and Weighted UniFrac data they they theorized could potentially mean that rare taxa had a small effect.

2. *Leinonen, Rasko et al. "The European Nucleotide Archive" Nucleic Acids Research, 2011, Vol. 39, Database issue, D28–D31*

The European Nucleotide Archive is a publicly available database of nucleotide sequences and part of the International Nucleotide Sequence Database Collaboration along with GenBank and the DNA Databank of Japan. It contains the Sequence Read Archive, the Trace Archive, and the EMBL-Bank. The EMBL-Bank and SRA both allow the public to submit data, and many improvements to how data can be submitted are covered in this paper. Data can be downloaded in several formats including xml and fastq files. Data can be searched and even downloaded in bulk with FTP. This is a very useful tool and it is fantastic that the researchers and employees at the European Bioinformatics Institute made this archive a reality.

3. *Collins, Francis and Tabak, Lawrence. NIH plans to enhance reproducibility. Nature, 2014, vol. 505, pp. 612-613.*

Scientific research, and biomedical research especially, are going through a crises and more and more published papers are impossible to reproduce. The NIH believes this is not due to any purposeful misconduct but due to a variety of problems in the community such as tenure rewards, focus on high impact journals, hiding information from competition, poor study design, and more. Since reproducibility is such a core part of science, the community must improve. NIH

lays out its current plans for improvement in this paper. Improvements to mandatory training, new reviewer checklists, and a dedicated "scientific premise" reviewer are currently being piloted. NIH is also working to setup a data repository where data can be shared even if it is not associated with a published paper. NIH also encourages journals to be more open to negative data and correction papers.

4. *Ravel, Jacques and Wommack, K Eric. All hail reproducibility in microbiome research. Microbiome, 2014, 2:8*

The Editor in Chief and another editor of the journal Microbiome present their stance on reproducibility of research. They believe that the journals themselves are the key to ensuring published papers have data sets available. They provide a detailed list of useful tools and sites that researches can use including SRA, dbGAP, FigShare, Github, and iPython Notebooks, as well recommending a previous paper as an example of correct datasharing. While the named tools are well explained, it might have been more helpful if the authors had included them in a table or sidebar. With this method researchers looking for new tools could quickly reference the list, and more options could be included without bogging down the narrative. The paper ends with the declaration that Microbiome intends to ask authors to provide accessible data, but does not detail if this will be an actual requirement or simply an option that researches can easily decline.

5. *Gonzalez, Antonio et al. "Qiita: rapid, web-enabled microbiome meta-analysis" Nature Methods, 2018, Vol. 15, pp 796-798*

Qiita is a web based platform that not only stores sequence data, but stores it in the rawest form possible along with all the steps and parameters used to get from raw data to final analysis. The goal of the platform is to allow users that are not bioinformaticians to perform metanalyses on data using pipelines like QIIME2. Qiita automatically deposits data to the ENA as well. By accepting only the rawest form of data, Qiita is ensuring the data can be reused with new analytic technologies as they are created. Studies with high quality metadata are labeled as Gold studies, in order to encourage users to aim for high quality in their work as well. The lead author on the work is a programmer, so we can have confidence the backend technology is sound and well supported. Qiita is a very useful platform for new researchers and veterans alike.

6. *Boylen, Evan et al. "QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science" PeerJ Preprints 6:e27295v2*

QIIME 2 is the successor to QIIME, moving the microbial analytic pipeline platform forward and allowing new technology and opportunity. To allow for growth in technologies, QIIME2 is based around plugins. Various implementations of denoising, taxanomic alignment, and more are available as plugins and can be updated or replaced easily as new algorithms are developed. In

addition, QIIME2 provides interactive visualizations with provenance tracking. QIIME2 also works to be accessible to users from various backgrounds and is free and open source. Its design aimed for the future, long list of high profile authors, and most importantly the success of its predecessor QIMME gives confidence that QIIME2 will be a fantastic resource for the field.

7. Caporaso, Gregory et al. "QIIME allows analysis of high-throughput community sequencing data" Nature Methods, 2010, Vol. 7, Issue 5. Pp335-336

QIIME, quantitative insights into microbial ecology, is a tool that allows users to work through an entire analysis pipeline. As new sequencing technology increases the size of datasets, the authors found they could not find tools that offered a library for demultiplexing, a library for taxonomy assignment, and a set of analytic tools for continuing with the data. Using PyCogent they created the QIIME platform. It is built to be modular so that functions can be added over time. They then tested the tool on a new twin study to verify it. Written by a group of experts in the field, we know QIIME was a useful tool until its successor QIIME2 was created.

8. Amir, Amnon et al. "Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns" mSystems,2017, Vol. 2, Issue 2, e00191-16;

DeBlur is a tool for denoising datasets, or removing error sequences. To do this it uses a sub-Operational-Taxanomic-Unit approach, similar to other popular algorithms DADA2 and UNOISE2. Unlike those tools DeBlur can be more easily parallelized because it operates on every sample independently. The algorithm first sorts the sequences by abundance, then subtracts the predicted error number from neighbors based on Hamming distance. If a sequence's abundance drops to 0, it is removed as likely an error.

9. Callahan, Benjamin et al. "DADA2: High-resolution sample inference from Illumina amplicon data" Nature Methods, 2016, Vol. 13, pp 581-583

A successor to DADA, Divisive Amplicon Denoising Algorithm, DADA2 is a tool for correcting errors in Illimuna amplicon data. Many researchers are handling the issues of amplicon errors with OTU clustering, however; the authors feel important real variations may be getting lost. DADA2 was tested against several similar methods such as Mothur, and the authors found that it the most accurate.

10. *DeSantis, et al. "Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB" Applied and Environmental Microbiology. Jul 2006, Vol. 72, Issue 7, pp 5069-5072*

At the time of the paper there was a major issue with public sequence databases having a large number of chimeric sequences. When using PCR methods the sequence fragments must be aligned and rejoined to make longer sequences, but if done improperly a sequence made from unrelated contigs can be created, a chimera. Once the chimera is out there, others may try to align data to the sequence of this nonexistent creature. To resolve this issue, the authors have created a new database that checks data for chimeras, while also providing standard alignment, and taxonomic classification. The chimera checking algorithm is based on Bellerophon, but with modifications that make it significantly faster to run and provide a divergence ratio. DeSantis and Hugenholtz have worked on many tools in the field. DeSantic worked on NAST and PyNAST and has six publications with more than 1000 citations. Hugenholtz worked on the original Bellerophon that was adapted as part of this project as well as CheckM, NAST, and CRISPR. He has a dozen publications with over 1000 citations each. Clearly these authors are good tool makers in the field.

11. *Rideout, Jai Ram et al. "Subsampled open-reference clustering creates consistent, comprehensive OUT definitions and scales to billions of sequences" PeerJ. 2014, eCollection e545*

Several familiar names from the field including Rideout, Knight, and Caporaso contributed to this paper on a technique for OTU clustering. Closed reference OTU clustering is fast and aligns sequences to a reference database. However, any novel sequences that do not align are thrown away. The alternative is De Novo clustering, which clusters sequences by comparing them to each other instead of references. But this method is very slow because the comparisons can not be easily done in parallel, so it is not reasonable to use this on large data sets. One strategy to get the best of both options is open-reference picking, where first Closed Reference OTU picking is done, and the novel sequences that remain are then de novo clustered. However, in sets with a large number of novel sequences this can still be too slow. The authors have suggested a new strategy to speed this up. After the initial closed reference run, a subsample of the remaining sequences is de novo clustered. Then the remaining sequences are close reference clustered using the new OTUs from the subsample. This can be repeated
as necessary, allowing for the benefits of open reference OTU picking with faster results. The authors ran several comparison test and found their results align well with existing methods, with improved runtimes in datasets with many novel OTUs. Since these authors are some of the common writers in the field we have good reason to trust their results.

12. *Lozupone, Catherine et al. UniFrac: A Phylogenetic Method for Comparing Microbial Communities. Applied and Environmental Microbiology,* Dec 2005, Vol. 71, Issue 12, pp. 8228-8235

This work introduces a new tool and metric for measuring and comparing diversity in microbial communities called UniFrac. The metric, called unique fraction metric, measures phylogenetic distance. This will better determine if environments are similar or distinct than just comparing the OTUs alone, as different OTUs will still share many branches in similar communities but have more evolutionary diversity in others. To display their new tool and metric, a comparison was made between cultured and uncultured samples of sea water, sea ice, and sediment. Using UniFrac they determined that where different OTUs live was less effected by geography than by the environment type. This could not have been determined without a measurement made for comparing large numbers of sequences from extremely different geographic locations and environments, perfectly demonstrating an example use case for their new tool. Knight was a professor at University of Colorado at Boulder when Lozupone was a pre-doctoral student. She has since continued to work in the field and now has her own faculty position1.

13. *Lozupone, Catherine et al. Quantitative and Qualitative β Diversity Measure Lead to Different Insights into Factors That Structure Microbial Communities. Applied and Environmental Microbiology,* Feb 2007, Vol. 73, Issue 5, pp 1576-1585

The creators of UniFrac return along with other researchers to introduce a new metric added to the UniFrac tool. The added metric, weighted UniFrac, modifies the original unweighted UniFrac to factor in quantitative measures of β diversity, while leaving the original metric available to measure qualitative diversity. The authors reproduced existing studies to demonstrate both the validity of the new weighted UniFrac measure and possible use cases for both weighted and unweighted UniFrac metrics. The first study reproduced was of geothermal vents from Yellowstone with data from a 2007 paper by Mathur et al. Weighted UniFrac agreed with the original paper's quantitative metric that the temperature was not as important of a factor for diversity as the minerals available. The unweighted version, measuring only the phylogenetic difference in OTUs without considering their frequency, favored temperature as the more important factor. The second study reproduced shows another case where unweighted and weighted UniFrac show unique insights; examining the composition of microbiota in the guts of healthy and overweight mice with data from Ley et al (2005). Weighted UniFrac clustered the sequences by the mice obesity phenotypes, however unweighted UniFrac clustered them by their hereditary relationships. Quantitative and qualitative measures are both useful and having both makes UniFrac an even more powerful tool.

14. Rognes, Torbjorn et al. "VSEARCH: a versatile open source tool for metagenomics" PeerJ, 2016, 4:e2584 eCollection 2016.

VSEARCH was created as an alternative to USEARCH. USEARCH was a useful and popular tool but was closed source and was not entirely free to use. VSEARCH allows users to detect chimeras, search, cluster by similarity, dereplicate, and more. It is free and open source, making it a superior tool for academics and anyone interested in more open science and software.

15. *Price, Morgan et al. "FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments"*

FastTree 1 was a scalable way to quickly make phylogenic trees. It was a minimum evolution method, which is very fast, but not as accurate as maximum likelihood methods. FastTree 2 first uses a minimum evolution algorithm, then uses maximum likelihood nearest neighbor interchanges to improve the tree. FastTree 2 is an "an approximately-maximum-likelihood method." It is sill less accurate than true maximum likelihood approaches, but remains accurate enough for many uses with impressive speed that can be used on large datasets.

16. *Katoh, Kazutaka and Standley, Daron. "ArticleFast TrackMAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability" Molecular Biology and Evolution, 2013, Vol 30, Issue 4, pp 772-780*

The original MAFFT program from 2002 is a multiple sequence alignment tool. Version 7 is a large enough update to warrant a new paper. New features have been added such as the ability to add unaligned sequences to an existing alignment. Each new feature is described here with examples. It is great to see authors continuing to support and improve their work so many years later.

17. *Bokulich, Nikolas et al. "Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing" Nature Methods, 2012, Vol. 10, Issue 1, pp 57-59*

The authors created a strategy for improving the quality of Illumina amplicon data. As sequencing becomes faster, easier, and cheaper more data is being produced, and more data with errors is being produced. These paper outlines the tests used to produce the quality guidelines used by the Earth Microbiome Project.

Annotation references and additional data type references

- Google Scholar https://scholar.google.com/
- Editorial Board. Microbiome. https://microbiomejournal.biomedcentral.com/about/editorial-board
- Antonio González Peña. About my research. https://sites.google.com/site/antgonza/
- Catherine Lozupone PhD "School of Medicine Biomedical Informatics and Personalized Medicine" University of Colorado Anshutz Campus http://www.ucdenver.edu/academics/colleges/medicalschool/departments/medicine/BIPM/Faculty/Pages/Catherine-Lozupone,-PhD.aspx
- McDonald, Daniel et al. "The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome." Gigascience. 2012, Vol 1, Issue 1, p 7
- McKinney, Wes. "Data Structures for Statistical Computing in Python" PROC. OF THE 9th PYTHON IN SCIENCE CONF, 2010

- Chang, Qin et al. “Variance adjusted weighted UniFrac: a powerful beta diversity measure for comparing communities based on phylogeny” BMC Bioinformatics, 2011, Vol. 12, pp118
- Chen, Jun et al. “Associating microbiome composition with environmental covariates using generalized UniFrac distances” Bioinformatics, 2012, Vol. 28 Issue 16, pp 2106–2113.
- McDonald, Daniel et al. “Striped UniFrac: enabling microbiome analysis at unprecedented scale” Nature Methods, 2018, Volume 15, pp 847–848
- Weiss, Sophie “Normalization and microbial differential abundance strategies depend upon data characteristics” Microbiome, 2017, Vol. 5, Issue 27
- Lane, DJ. “Nucleic Acid Techniques in Bacterial Systematics 16S/23S sequencing”, 1991 pp115-175