Non-Bacterial Components of the Primate Microbiome:
Against a Bacterial Bias

Turned in on
October 18, 2019
by
Daniel G. Leonard
for the class
ANTH 438 Primate Life History Evolution
at
University of Illinois at Urbana-Champaign
taught by
Dr. Kathryn B. Clancy

1 Introduction

Since the discovery that the world in which we live is entirely filled with organisms too small to see with the eye, biologists have drawn a line between the microorganism and the multicellular organism. However, the search for the evolutionary apex of cellular organisms has blurred this line, from the theory of lateral gene flow (**Woe98**) to the discovery that archaea and eukarya share more in common with each other than either shares with bacteria (**Woe77**). In researching the microbiome – the physical commensal relationship between animals and their microbial inhabitants – it is important to consider the full diversity of microbial taxa. This paper seeks to investigate biases in microbiome research toward specific microbial taxa and seek out how the less-studied varieties of microbial inhabitants may have effects unaccounted for.

Despite a growth of research interest in archaea since its identification by **Woe77**, little research has been conducted into their relationships to other organisms, such as humans. As the domain was originally understood as a clade of extremophiles due to their original isolation from geothermal vents (**DeL01**), it would be expected that archaea would be a minor part of more familiar environments; however, it is now known that members of archaea exist with great frequency in non-extreme environments, such as plankton, lakes, and soils (**DeL01**). Regardless, research into the composition of human microbiota retains a strong bias toward bacteria (**Llo16**; **The12**). Comparative primatology has often failed to expand microbiota research beyond bacteria, frequently excluding unknown or infrequent phyla from results of DNA extraction (**Gar19**; **Mal18**; **Ork191**; **Ork19**). At its extreme, the term "microbe" has been used synonymously with the term "bacteria" despite the far more precise definition of the latter (**Mal18**).

2 Analysis of previous research

In order to assay the relative biases of microbiome research it would be necessary to compare the references to microbial phyla in published literature to the prevalence of such microbes in microbiomes. As much research into the human microbiome is undertaken from a medical perspective, NCBI PubMed contains a wealth of full-text and abstracts in its catalog, which is easily indexed for specific terms and phrases. Thus, NCBI PubMed's article archive was chosen as a sample for language analysis.

2.1 Methods

A search for the exact phrase "human microbiome" was conducted on NCBI PubMed in October 2019, fetching 1,779 results, which were downloaded as an XML file. Using R version 3.6.1 (**RCol19**) on Microsoft

Windows 10 18362.418, these results were indexed with `pubmed.mineR` version 1.0.16 (**Ran19**) and queried for metadata from the NCBI PubTator service (**Wei19**). PubTator analyzes the text of PubMed abstracts and full-text articles to determine the names of species and taxonomic tanks mentioned in such articles. Excluding mentions to *Homo sapiens*, the taxonomic ranks were looked up in the NCBI Taxonomy Database (**Nat19**) using `ncbitax2lin` (**Xue19**) to convert the database to CSV.

References to taxonomic ranks more specific than phylum were simplified to phyla, and less specific references were recorded as "*domain* (nonspecific)." In the case of viruses, which have separate ranking system, the NCBI value "no rank – 0" was used in place of phylum. The number of references to each phylum was totaled and categorized by domain. All R code executed is presented in

2.2 Results

Of the 1,779 articles indexed, 1,730 mentioned a taxonomic rank or species. This totaled 5,354 individual mentions, of which 2,419 were to species or ranks other than humans. As visible in the majority of references, with the phyla Proteobacteria and Firmicutes second only to nonspecific references to bacteria as a domain. Eukarya were second to bacteria in total number of references; however, the most frequently discussed phylum within the domain was Chordata, which likely stems from discussion of vertebrate microbiomes such as Mus musculus rather than indicating the presence of chordates in the human microbiome. Viruses were referenced moderately frequently, with nonspecific virus references making up the greatest proportion of the domain's coverage, followed by bacteriophages. Finally, archaea were referenced least frequently of all, with nine total references amounting to less than any viral phylum.

3 The Role of Non-Bacterial Organisms in the Microbiome

3.1 Eukarya

Yeasts have been found to colonize infant guts gradually, with an isolation rate of 13% at 28-46 days old (**Ben84**) and 50% as soon as four months of age (**Ell75**). Notably, this rate of colonization was consistent between breast-fed and formula-fed infants, in contrast to the colonization rates of bacterial phyla which differ significantly between the two feeding mechanisms. From a life-history perspective, this presents a novel question of how yeast colonize the infant gut and whether the mechanism by which breastmilk contains microbiota associated with the mother's gut microbiome (**Mar13**) is also capable of providing yeast, whose eukaryotic cells differ in size by orders of magnitude from prokaryotic cells.

However, research conducted on yeast inhabitants of animals focuses primarily on their pathological role. The genus *Candida* is well known for its role in infectious disease in humans, particularly in the immunocompromised (**Kou11**). However, species of yeast, including many *Candida* spp., have frequently been identified on the human body, including the vagina, gut, and skin, without any associated mycosis (**Man10**). Even for species known to be infectious agents, between 12% (**Cho86**) and 80% (**Soe07**) of asymptomatic healthy women are vaginal carriers. In place of viewing yeast through the lens of infection, it should be considered as a part of the ecosystem that makes up the human body. How it reaches and colonizes infants, and why its prevalence makes up far less of the population than do bacterial commensals, merits further study.

3.2 Archaea

In 1966 it was discovered that humans both respire and flatulate methane, which is not known to be biologically produced by any bacteria or eukaryote. Following this observation, methanogenic prokaryotes, from what would later be called the archaea, were identified in human feces (**Not68**). The archaeal domain remains the only taxon with members known to be biological producers of methane (**Tha06**), and as such is unique in its residence within the human gut. Despite widespread colonization by a diverse array of archaeal species, the ecological niche which they occupy within the human body and whether they are transferred vertically or environmentally remains unknown (**Dri11**). It has been theorized that many more lineages of archaea remain undiscovered, as the DNA analysis methods used to identify bacterial genomes are incompatible with archaeal cell walls (**Hor15**).

3.3 Viruses

Despite not fitting many definitions of biological life, viruses are the most diverse form of biological material on the planet. They are known to outnumber bacteria by a factor of nearly ten (**Ore97**) and are a major driver of evolution in prokaryotes. Although humans are susceptible to a wide variety of viruses, the viral community in the microbiome, so fast it has itself been called a "virome" (**Wyl12**), is dominated primarily by bacteriophages whose sequences are novel (**Dut14**). The phages identifiable by genetic sequence are primarily those known to attack bacterial phyla associated with the microbiome, yet appear to be vastly more diverse than the virome of the ocean (**Wal14**). Unlike the theories of bacterial colonization, how such a vast and dynamic virome is able to establish itself by adulthood remains unknown. The effect of age, and whether milk carries a virome of similar diversity would be intriguing avenues for further study.

4 Limitations

Utilizing both life-history and ecological frameworks for approaching non-bacterial components of the microbiome, it is clear that the roles such taxa play within our bodies is underexplored. While this paper sought to touch on these issues, it is limited by both methodology and its brevity. The use of NCBI PubMed for language analysis suffers from a bias toward medical research, and so microbiome research will often consider the roles of known pathological agents rather than in-depth discussions of all possible inhabitants. Eukarya is also a domain of organisms highly diverse in physical makeup and metabolic strategy, while only yeasts were discussed here. The roles of other unicellular parasites (for instance, *Toxoplasma gondii*) and of helminthic worms, which each form diverse monophyletic groups, can also be explored via the lens of the microbiome's life-history. More research is needed in this area to investigate how the diverse community of microbes within humans evolved and what role it plays even in the absence of pathology.

Appendix A

```
# R script to analyze PubMed abstracts
# Author: Dan Leonard

# Import pubmed analysis library
library(pubmed.mineR)
# Import file-reading library
library(readr)

# Load XML search result file from PubMed into list
abstracts <- xmlreadabs("pubmed_result.xml")

# Load NCBI lineage data into list
lineages <- read_csv("lineages-2017-03-17.csv",
    col_types = cols(
        .default = col_character(),
        tax_id = col_integer()
    )
)

# Run PMIDs through PubTator
pubtators <- lapply(abstracts@PMID, pubtator_function)
# Remove [" No Data "] from PubTator results
pubtators <- pubtators [! pubtators %in% list(" No Data ")]

# Get list of vectors of species names
# Species names are located in column 5
species <- sapply(pubtators, "[[", 5)
# Remove nulls
species [sapply(species, is.null)] <- NULL
# Flatten
species <- unlist(species)

# Remove species names, leave NCBI numerical ID
species <- sapply(species, function(x) sapply(strsplit(x, ">"), "[[", 2))
# Convert to numeric form
species <- as.numeric(species)
# Remove references to Homo sapiens (NCBI ID 9606)
species.nohuman <- species [! species %in% 9606]

# Get domain names
species.nohuman.domains <-
    lineages$superkingdom[match(species.nohuman, lineages$tax_id)]
# Get phyla names
species.nohuman.phyla <-
    lineages$phylum[match(species.nohuman, lineages$tax_id)]

# Create data frame
phylogeny <- data.frame(
    Domain=
        c(
            lineages$superkingdom[
                match(species.nohuman, lineages$tax_id)
```

```
            ]
        ),
    Phylum=
        c(
            lineages$phylum[
                match(species.nohuman, lineages$tax_id)
            ]
        ),
    Norank=
        c(
            lineages$`no rank`[
                match(species.nohuman, lineages$tax_id)
            ]
        ),
    stringsAsFactors=FALSE
)
# Add additional column "Name"
phylogeny [ , c("Name")] <- NA
# Remove unhelpful term "cellular organisms"
phylogeny$Norank[phylogeny$Norank == "cellular organisms"] <- NA
# Use virus types as name if present
phylogeny$Name <- phylogeny$Norank
# Use phylum as name if present
phylogeny$Name[is.na(phylogeny$Name)] <-
    as.character(phylogeny$Phylum[is.na(phylogeny$Name)])
# Use "<domain> (nonspecific)" as name if previous two not present
phylogeny$Name[is.na(phylogeny$Name)] <-
    paste(
        as.character(phylogeny$Domain[is.na(phylogeny$Name)]),
        "(nonspecific)",
        sep=" "
    )
# Remove phylum and virus type columns
phylogeny <-
    data.frame(
        Domain=phylogeny$Domain,
        Name=phylogeny$Name,
        stringsAsFactors=FALSE
    )
# Use table() to count occurrences
phylogeny <- as.data.frame(table(phylogeny))
# Remove extraneous values
phylogeny <- subset(phylogeny, Freq != 0)
# Sort
phylogeny <- phylogeny[order(phylogeny$Domain, -phylogeny$Freq),]

# Create dictionary for looking up colors
colors <-
    data.frame(
        Colors=c("Red", "Green", "Yellow", "Blue"),
        Domains=c("Archaea", "Bacteria", "Eukaryota", "Viruses"),
        stringsAsFactors=FALSE
    )
```

```
# Create list of colors matching domains
cols <-
    colors$Colors[
        match(phylogeny$Domain, colors$Domains)
    ]

# Set margins
par(mar=c(7,4,4,1)+0.1)
# Print phylum plot
plot <- barplot(
    height = phylogeny$Freq,
    ylab = "log References",
    main = "References to specific phyla in PubMed search for \"Human Microbiome\"",
    col = cols,
    names.arg = phylogeny$Name,
    log = "y",
    xaxt = "n"
)
# Add X-axis labels
text(
    plot,
    0.95,
    labels = phylogeny$Name,
    srt = 45,
    adj = c(1.1,1.1),
    xpd = TRUE,
    cex = 0.8
)
# Add legend for domain colors
legend(
    "topright",
    plot,
    legend = colors$Domains,
    fill = colors$Colors
)
```