UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

NOVEL TECHNIQUES FOR THE DESCRIPTION AND INTERPRETATION OF MICROBIAL COMMUNITIES IN THE MODERN HUMAN GUT AND ANCIENT HUMAN ORAL MICROBIOME.

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

ALLISON E. MANN

Norman, Oklahoma

2018

NOVEL TECHNIQUES FOR THE DESCRIPTION AND INTERPRETATION OF MICROBIAL COMMUNITIES IN THE MODERN HUMAN GUT AND ANCIENT HUMAN ORAL MICROBIOME

A DISSERTATION APPROVED FOR THE
DEPARTMENT OF ANTHROPOLOGY

BY

Dr. Cecil M. Lewis Jr., Chair

Dr. Christina Warinner

Dr. Paul Spicer

Dr. Patrick Livingood

Dr. Paul Lawson

Dr. Krithivasan Sankaranarayanan

Dr. Lawrence Weider

Acknowledgments

I would like to extend my sincere gratitude to the many people who have assisted me throughout my doctoral program. In particular, this dissertation would not have been possible without the continuing advice and council of my chair, Dr. Cecil M. Lewis, Jr. who has been a constant source of encouragement and inspiration. I am also indebted to Dr. Christina Warinner and Dr. Krithivasan Sankaranarayanan for their guidance. Additionally, I am extremely grateful to my committee members Dr. Paul Spicer, Dr. Patrick Livingood, Dr. Paul Lawson, and Dr. Lawrence Weider, all of whom provided valuable feedback and suggestions to this work. I would be remiss to forget the invaluable support of my fellow lab mates including Nisha, Rita, Justin, Dave, Tanvi, and Christine who were consistent sources of advice and willing to respond to frantic late–night text messages. I would also like to thank my co–authors on chapter four of this dissertation and in particular Susanna Sabin. Finally, I would like to thank my parents, Gretchen and David Mann, my sister Juliet, and Jared who has been my most patient supporter throughout this process.

Contents

List of Tables

List of Figures

Abstract

Studies of human–associated microbial communities are increasingly integrated into biological anthropology, allowing us to explore the role of microorganisms in aspects of human evolution, health, and disease. Despite technological advances in the genomic characterization of microbial ecosystems in both modern populations and in the past, specific challenges remain in documenting the presence of taxa, functional potential, and interaction of members of these dynamic ecosystems. In this dissertation, three studies designed to address some of the challenges of metabarcoding and metagenomic studies of the modern human gut and ancient oral microbiome are presented. In chapter two, the eukaryotic component of the modern human gut is assessed with a metabarcoding approach using the internal transcribed spacer region. In chapter three, rare bacterial taxa are characterized using a joined serial filtration and metabarcoding approach. Finally, in chapter four, the preservation of ancient DNA in archaeological dental calculus is discussed with particular attention to both the benefits of this substrate as a reliable source of ancient DNA as well as the potential challenges associated with DNA preservation in archaeological materials.

Dissertation Keywords: Gut microbiome, ancient DNA, dental calculus, microeukaryotes, metabarcoding, metagenomics, *Vampirovibrio*, *Blastocystis*.

Chapter 1

Introduction

Studies of human–associated microbial communities—collectively known as the human microbiome—using metagenomic (the sequencing of whole genomes) and metabarcoding (profiling a community of organisms via targeted amplicon sequencing) techniques have illuminated the role of microbes in the maintenance of human health [90, 186, 24, 103, 59], the evolution of the human immune system [82, 14, 104], as well as the potential use of microbes as agents of global public health [109, 206, 119]. Studies of microbiome communities in modern human populations in non–industrial societies [131, 165, 34], and microbiomes sourced from archaeological remains [182, 2, 194] document the evolution of these microecosystems and the hosts in which they inhabit. While innovations in technologies designed to generate and analyze large biomolecular datasets have provided new ways to characterize microbiome communities, understanding the activity and functional potential of microbial organisms in these complex environments remains challenging. Consider, for example, that while members of all high level microbial lineages—both living (bacteria, archaea, eukaryota) and non–living

(viruses)—are active in microbiome habitats, nearly all microbiome studies to date focus exclusively on bacteria and, to a lesser extent, archaea, excluding other ecologically important taxa. The reason for this omission is primarily due to the relative technical difficulty of characterizing certain microbes from genomic data, a paucity of available comparative databases, and the absence of many of these taxa in industrialized groups, which were the focus of early microbiome studies. The exclusion of non–bacterial microbes from studies of human–associated microbial ecosystems is predicted to have consequences for the interpretation of the function and composition of the total ecosystem [9, 30]. As an example, consider that while micro– and macroeukaryotic organisms in the human gut are far outnumbered by bacterial cells, they may serve as critical sources of microbial predation and resource competition [71, 115, 102], not unlike large apex predators in macroecosystems [115]. Likewise, viruses, and in particular bacteriophages are abundant in the human gut and may make up much of the genomic "dark matter" [113] of which cannot be assigned to any known organism [50]. Therefore, the interpretation of variation in bacterial and archaeal community composition is incomplete lacking data from the full ecosystem, including viruses and eukaryotes.

Though metagenomic and metabarcoding approaches to the study of bacterial communities are well established in published literature, they too are prone to specific technical biases and challenges. Metabarcoding approaches to profiling bacterial communities in microbial ecosystems typically involve the high–throughput sequencing of one or a series of hyper–variable regions of the 16S rRNA gene. The advantages to this approach include

its near universality among bacteria and archaea, allowing for the design of primers that cover a wide range of phylogenetically divergent microbes, as well as its relatively deep taxonomic discrimination to the genus or species level [44]. Moreover, as the 16S rRNA gene has been the gold standard for bacterial systematics since its establishment as a phylogenetic marker by Carl Woese and George Fox in the late 1970s [207], an abundance of data for comparison is readily available in public genomic databases. Despite this, it is well known that primers for variable regions of the 16S rRNA gene are not truly universal across all possible bacterial groups and the choice of primer set will impact the taxonomic resolution and inclusion of particular groups in the resulting data [214]. Moreover, 16S rRNA copy number variation among bacteria and archaea skews community representation estimates [114]. As the number of copies of the 16S rRNA gene is inconsistent at high taxonomic levels [189], correction for this bias is a difficult and ongoing challenge in microbiome studies using a metabarcoding approach [114]. Even with the full 16S rRNA gene, differentiation of bacterial organisms at low taxonomic levels may not be feasible. As taxonomic categories (e.g., phylum, genus, species, etc) are somewhat arbitrarily defined, the definition of standardized threshold criteria for determining the taxonomic status of an organism is complicated. Consider, for example, that although clustering reads at 97% sequence identity is often cited to be an approximation of a “species” level percent identity cut off for the 16S rRNA gene in many investigations of the human microbiome [101], different levels of clustering identity have been proposed [53, 129]. Studies of the 16S rRNA gene of bacterial and archaeal

organisms have suggested that the optimal sequence similarity to differentiate microbes at the species level is around 98.7% if the full gene is available [125, 209]. Even if the full 16S rRNA gene is available, however, many closely related, yet functionally distinct organisms cluster together at 97% sequence identity, masking potentially important taxa or increasing the risk of false positives. In a study of full length reference sequences of the 16S rRNA gene from clinically important bacteria it was found that many are similar to potentially benign environmental taxa—meeting or exceeding the 97% identity threshold which may lead to erroneous conclusions in microbiome studies [197]. Thus, the determination of microbial species is better defined by multiple criteria and not the targeting of a single gene (or single gene region) alone. While target–independent metagenomic studies are unhindered by the primer biases inherent to metabarcoding approaches, metagenomic techniques are not without similar challenges. Because metagenomic sequencing is untargeted and instead relies on the shredding of whole genomes after which sequencing adapters are ligated, the probability of any single read deriving from a particular organism is heavily dependent on genome size, with large genome organisms contributing more to the total genomic library than those with smaller genomes [17]. While this is not an issue with the metagenomic sequencing of a single organism, genome size among currently known bacteria is variable, ranging from approximately 140 kbp to 14 Mb [76] and in communities of diverse microorganisms, extreme variation in genome size is expected to skew community estimations in the same way that variation in copy number of the 16S rRNA gene will with metabarcoding studies.

Finally, while modern microbiome studies of diverse extant populations provide insight into the evolution of human associated microbial communities, research involving preserved microbiome remains from archaeological contexts augments these studies by directly sampling past populations. Ancient DNA (aDNA) generated from archaeological materials is subject to its own suite of challenges that are characteristic of the natural taphonomic processes associated with highly degraded DNA. These include processes that degrade the sugar–phosphate backbone and fragment the DNA, spontaneous chemical alterations to specific bases, and a higher risk of exogenous contamination [81, 162, 39, 148, 77]. As authentic aDNA tends to be highly fragmented, amplicon based studies targeting hyper–variable regions of the 16S rRNA gene are not recommended for community reconstruction [214] and instead, metagenomic techniques to characterize ancient microbiomes are typically used. Though the ability to generate and verify aDNA datasets has improved, still much is needed to learn regarding the circumstances under which aDNA preserves and whether there are particular organisms or genomic structures that are more or less amenable to preservation, producing systematic taxonomic biases in data generated from archaeological sources.

This dissertation comprises three studies that use novel methodological and analytical techniques designed to advance microbiome studies of both modern and ancient populations and address some of the concerns raised above. In particular, the goal of this dissertation is to outline experimental methods to improve the taxonomic and community resolution of the modern human gut and ancient human oral microbiome. The first study: “Mi-

croeukaryotic and dietary survey of the gut by internal transcribed spacer metabarcoding" documents the microeukaryotic component of the gut microbiome in three human and two animal populations representative of diverse geographic, subsistence, and behavioral contexts by metabarcoding sequencing of the internal transcribed spacer (ITS) region. The second study of this dissertation: "Enrichment of non–dominant bacterial taxa in human fecal samples through serial filtration" demonstrates the impact of cell and genome size on bacterial community composition and introduces a potentially viable method for targeting specific bacteria through the serial filtration of human fecal samples and 16S rRNA metabarcoding. Finally, the third study of this dissertation: "Differential preservation of endogenous human and microbial DNA in dental calculus and dentin" examines the prospect of ancient dental calculus as a reliable source of metagenomic data for characterization of the ancient human oral microbiome, as well as potential taxonomic biases that may be due to natural taphonomic processes or induced during the handling, preparation, or sequencing of archaeological microbiome materials.

Chapter 2

Microeukaryotic and dietary survey of the gut by internal transcribed spacer metabarcoding

2.1 Abstract

Public health initiatives, advancements in sewage and water treatment systems, and major overhauls to subsistence patterns in many parts of the world have dramatically changed our relationships with micro– and macroeukaryotic organisms that spend most or part of their life cycle in the human gut. The consequence of this change on human health and the gut microbiome is hypothesized to have had dramatic impacts on diseases common in industrial societies. To understand the effects of this shifting relationship, documenting the presence of microeukaryotic taxa in geographically distant human populations with diverse lifestyles is of paramount importance. By profiling the microeukaryotic component of the human gut in industrialized and non–industrialized populations, questions regarding the ancestral state of the gut microbiome and potential interventions can be addressed. In this study, hunter–gatherers and rural agriculturalist populations living in the first and last regions of the world to be populated by *Homo sapiens* as well as indi-

viduals living in urban–industrial settings are surveyed for microeukaryotic diversity and inferred diet via internal transcribed spacer (ITS) metabarcoding. Findings indicate that non–industrial populations are more similar in terms of microeukaryotic diversity to each other than individuals living in industrial locales who are characterized by a stark depression of microeukaryotic diversity. In addition, the detection of the common protist genus *Blastocystis* in all human groups documents the global rise of single–celled microeukaryotic organisms in the human gut.

2.2 Introduction

An estimated 300 species of helminth (worm) and 100 single–celled microeukaryote (protist) parasites are known to infect humans, many of which are the causative agents of diseases responsible for significant morbidity and mortality worldwide [38]. Approximately 3.5 billion people are infected by at least one microeukaryotic or helminth parasite globally, and an estimated 450 million, mostly children, are ill as a result [199]. Many parasitic eukaryotes have expansive geographic scope and thus disproportionately contribute to the overall burden of parasitic infections in human populations. For example, the estimated global burden of human whipworm (*Trichuris trichiura*) is 429.6 to 508.0 million infected individuals while human hookworm (*Ascaris lumbricoides*) is estimated to currently infect 771.7 to 891.6 million individuals world wide [152]. While many eukaryotic species that inhabit the vertebrate gut are classified as parasites, others have a more controversial place in the clinical sphere and may be better classified as commensals, symbionts, or

opportunistic pathogens [64, 32, 20]. While public health initiatives and intensive de–worming programs have dramatically reduced the prevalence of soil–transmitted helminths in many parts of the world [43, 144], rising rates of single–celled microeukaryotic infections including *Giardia*, *Cryptosporidium*, *Entamoeba*, and *Blastocystis* may indicate that microeukaryotic parasites are quickly filling this vacant ecological niche [144, 183] with unknown future consequences.

Archaeological, genetic, and historical records of eukaryotic parasites documents their antiquity and evolutionary importance to human populations [11, 12, 69, 158, 61, 138, 38], yet the last 100 years of human history marks a significant shift in our relationship with these organisms. The removal of eukaryotic taxa from the human gut microbiome—defined as the totality of commensal, mutualistic, or parasitic organisms living in the gastrointestinal tract—is hypothesized to be an important component in the rise of atopic diseases including food allergies, hay fever and asthma, irritable bowel syndrome, multiple sclerosis, lupus, and other purported “diseases of civilization” [19, 15, 40, 184, 51, 208]. Anthropogenic forces including climate change, increased urbanization, the globalization of agricultural and market goods trading, ecotourism, and political unrest continue to shift the distribution and frequency of eukaryotic parasites in human and non–human populations alike [25, 16, 203, 7, 145] which speaks to the need to rapidly identify potential parasitic eukaryotes and better understand their ecological significance in the gut microbiome. Traditional methods for the detection of eukaryotes from fecal samples include morphological identification of eggs, cysts, larvae, or adult

specimens via simple light microscopy, yet the ability to reliably distinguish between closely related organisms is notoriously difficult [65, 37, 29]. Moreover, morphological classification schemes for microeukaryotic species may mask cryptic genetic diversity in otherwise visually indistinguishable organisms. For example, members of the microeukaryotic genus *Blastocystis* are remarkably simple morphologically speaking, resembling tiny soap bubbles under the microscope. This morphological simplicity conceals an exceptional genetic and clinical diversity with no fewer than 17 distinct subtypes isolated from humans, other mammals, birds, and other animals [3, 4, 20, 32]. Conversely, the species status of morphologically distinct parasites is sometimes contested [106]. Thus, morphology–independent methods for characterizing microeukaryotic taxa in mixed microbial communities may clarify the presence and role of these organisms in the evolution of the human gut microbiome.

As organisms classified as eukaryotic parasites—and especially those designated as “protists”—are not a natural phylogenetic group and instead largely a label of convenience, choosing an appropriate genetic barcode for their classification is challenging. Genomic studies of microeukaryotic diversity in human–associated microbial ecosystems often target a single variable region of the 18S rRNA gene [5, 202, 164, 141, 127] or to a lesser extent, shotgun metagenomics [47]. While the 18S rRNA gene is highly conserved across eukaryotes allowing the creation of near universal primers, it has fewer hypervariable regions than its homologue in bacteria, the 16S rRNA gene [166]. Moreover, the taxonomic resolution of the full 18S rRNA

gene is relatively poor, generally only able to resolve to the genus level in a limited number of taxa [168, 10]. An alternative well–described eukaryotic barcode, the mitochondrial cytochrome oxidase complex, is ill suited to mixed microeukaryotic systems as many lack the complex itself [201] or even mitochondria all together [166]. Metagenomic approaches are similarly problematic as eukaryotic taxa in mixed–microbial ecosystems are often swamped out by bacterial cell density, rendering their presence, and therefore influence on these communities, obfuscated.

In the current study, an alternative target for the molecular characterization of microeukaryotes in the human gastrointestinal tract, the internal transcribed spacer (ITS) region which separate the eukaryotic ribosomal RNA gene (rRNA) complex, is explored. The full eukaryotic rRNA complex consists of the small subunit (SSU) 18S rRNA gene followed by internal transcribed spacer one (ITS1), the 5.8S rRNA gene, internal transcribed spacer two (ITS2) and finally the large subunit (LSU) 28S or 26S rRNA gene. The total eukaryotic ITS region is defined as the 5.8S rRNA gene and two flanking ITS regions [166]. Following transcription by RNA polymerase I the two ITS regions are excised from the transcript to produce a mature ribosomal RNA [166]. While both ITS1 and ITS2 are noncoding regions of the genome, the ITS2 region is necessary for proper ribosome biogenesis [116], which may account for the higher conservation of ITS2 as compared to ITS1 among eukaryotes, the latter of which evolves via length expansions through the incorporation of variable repeat units [191, 130, 187], though the higher variability in ITS1 has been contested [190]. As a result, the ITS region is highly length

variant across eukaryotic organisms [166]. The benefit of using the ITS region for phylogenetic studies of eukaryotic taxa includes high conservation of the SSU, LSU, and 5.8S rRNA gene, allowing for the design of near–universal primers as well as the loose functional constraint of the ITS1 and ITS2 regions which allows for better taxonomic resolution even among closely related organisms [74, 83, 80]. Moreover, similar to the 16S rRNA gene in bacteria [114], the ITS region is multicopy allowing for the detection of microeukaryotes even if relatively few cells are present [80].

While the ITS region is recognized as an appropriate genetic barcode for fungal organisms [89, 166] and is routinely used to characterize the "mycobiome" from a variety of human and environmental microbial ecosystems [67, 48], it is rarely used to document non–fungal eukaryotic members of microbial communities. This is in part due to the paucity of full ITS region databases or references for non–fungal eukaryotes [168], especially considering the abundance of well–curated databases for other eukaryotic barcode sequences like the 18S rRNA gene [153, 73]. Nevertheless, the ITS region is argued to have better taxonomic resolution than other amplicon based detection methods for microeukaryotes [168, 166] and due to its fundamental role in the biogenesis of the mature eukaryotic ribosome, the region is conserved across all eukaryotes, linking taxa as distant as microscopic fungi and multicellular vertebrates [168]. Previous research targeting one or both ITS regions have been used to document the phylogenetic diversity of a variety of eukaryotes, revealing cryptic species that are otherwise morphologically identical [74, 56], as well as tracing the anthropogenic dispersal of

microeukaryotic parasites [203]. Due to its high conservation and relatively sharp taxonomic resolution, the utility of ITS region metabarcoding in microbiome environments merits investigation.

In the current study, ITS1 and ITS2 were sequenced from 18 individuals spanning a geographic distance that encompasses the cradle of humanity through to the last subarctic landmass colonized by human populations. Fecal samples collected from six rural agriculturalists living in the Kibale National Park in Uganda, six hunter–gatherers from the Matses tribe living in the Peruvian Amazon, and four urban–industrialized individuals living in Norman, Oklahoma, USA as well as a single bovid and a single wild Ugandan Red Colobus monkey serving as non–human controls were sequenced using a metabarcoding approach to: (1) document the utility of the ITS1 and ITS2 regions in characterizing a range of micro– and macroeukaryotic organisms in human and non–human fecal samples, (2) demonstrate the predictive value of microeukaryotic taxa for subsistence strategy, lifestyle, or environment and, (3) evaluate the applicability of ITS metabarcoding for dietary reconstruction from fecal samples. The results of this study suggest that while the ITS region as a whole may provide high resolution taxonomic surveys of both dietary and microeukaryotic organisms, the ITS1 and ITS2 region document overlapping, but different taxonomic profiles. Additionally, we find that non–industrial populations generally have a more diverse suite of microeukaryotes and have more microeukaryotic taxa in common amongst themselves than between industrialized and non–industrialized groups, the former of which has a relatively depressed microeukaryotic diversity. Finally, results from this study

demonstrate that microeukaryotes assigned to the genus *Blastocystis* are a highly genetically diverse and cosmopolitan group, detected in all three human populations and the single bovid group. The results of this study add to the growing literature of the role of microeukaryotic organisms in the human gut microbiome by providing an alternative method for their detection as well as documentation that, like studies of bacterial diversity in non–industrial human populations [131, 34, 182, 210], microeukaryotic diversity is similarly linked to lifestyle with a sharply decreased diversity of microeukaryotes in the gut of industrialized peoples which is strongly suggestive of a significant shift from the ancestral state of this human–associated micro ecosystem.

2.3 Methods

2.3.1 Samples. Human and non–human fecal samples collected from rural agriculturalists (n=6), a domesticated bovid (*Bos tarus*, n=1), and a Ugandan Red Colobus monkey (*Procolobus tephrosceles*, n=1) living in the Kibale National Park in Western Uganda as well as previously collected fecal samples from the Matses hunter–gatherer population (n=6) and urban residents of Norman Oklahoma (n=4) [131] were prepared for sequencing at the Laboratories for Molecular Anthropology and Microbiome Research (LMAMR) in Norman, Oklahoma, USA. The Matses, Norman, and Ugandan human populations were chosen to represent a diversity of lifestyle and subsistence strategies, a wide geographic range with variable environmental contexts, and vastly different exposure to microeukaryotic acquisition (Figure 2.1). Individuals from Norman have diets typical of urban–industrialized populations

with regular consumption of processed food items [131]. Diet amongst the Matses consists primarily of gathered plantains and starchy root vegetables and local protein sources including wild game and fish [131]. The residents of Kibale Park live in close proximity to livestock and wildlife [137] and consume a diet primarily composed of maize, beans, bananas, and starchy root vegetables supplemented by wild foods found in the nearby wetlands and forests [78]. The herbivorous ruminant bovid and folivorous Ugandan Red Colobus monkey stand as a contrast to the omnivorous human populations included in this study. Sample information including the geographic location, age, and sex are provided in Table 2.1.

Table 2.1: **Additional context and information for individuals contributing samples to this study**. Sample name, geographic and organism origin, age and sex for all individuals represented in this study.

Sample	**Location**	**Genus**	**Age**	**Sex**
HS2374	Uganda	*Homo*	4	Male
HS2416	Uganda	*Homo*	87	Male
HS2363	Uganda	*Homo*	40	Male
HS2380	Uganda	*Homo*	23	Female
HS2432	Uganda	*Homo*	25	Male
HS2446	Uganda	*Homo*	21	Male
SM05	Peru	*Homo*	1	Male
SM29	Peru	*Homo*	50	Female
SM01	Peru	*Homo*	30	Male
SM02	Peru	*Homo*	25	Female

Table 2.1 continued from previous page

SM31	Peru	*Homo*	30	Male
SM32	Peru	*Homo*	21	Female
NO15	USA	*Homo*	50	Female
NO16	USA	*Homo*	47	Male
NO7	USA	*Homo*	32	Female
NO20	USA	*Homo*	26	Male
BO2072	Uganda	*Bos*	Adult	Female
RC2109	Uganda	*Colobus*	Adult	Male

2.3.2 DNA extraction and ITS amplification. Before extraction, raw fecal samples were first homogenized into a slurry to ensure sample consistency. Homogenized fecal samples were extracted using the Power Viral Environmental RNA/DNA kit (Qiagen: 28000-50) including the optional bead–beating step. Extracted DNA was quantified using a Qubit fluorometer before being diluted to a 1:10 concentration. The ITS1 and ITS2 targeted PCR reactions were performed separately. The primers ITS1f (5'–GCTGCGTTCTTCATCGATGC–3') and ITS2r (5'–GCTGCGTTCTTCATCGATGC–3') were used to target the ITS1 region [204] while ITS3f (5'–GCATCGATGAAGAACGCAGC–3') and ITS4r (5'–TCCTCCGCTTATTGATATGC–3') target the ITS2 region. Conditions for PCR amplification of both ITS1 and ITS2 were identical. Each reaction included 4 μL of Phusion HF buffer (Thermo Scientific), 1 μL each of the forward and

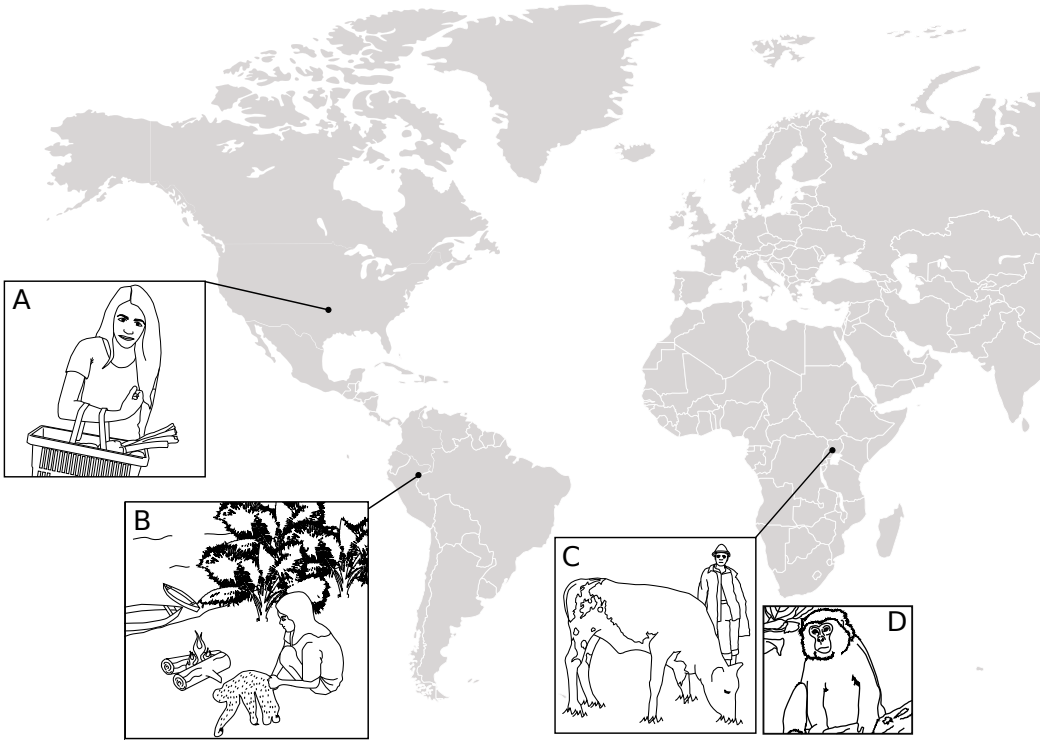

Figure 2.1: **Geographic location of human and non–human animal individuals represented in the current study**. (A) Residents of Norman, OK, USA have access to food produced through industrial means and subsist primarily on processed and pre–packaged foods. (B) The Matses population living in the Peruvian Amazon primarily forage and hunt for food in the surrounding rainforest. Illustrated is a Matses woman preparing a small sloth for a meal (Modified from [131]). (C) People living in the Kibale National Park in Uganda subsist on local agricultural products and live in close proximity to livestock (image modified from: https://www.pexels.com). (D) Living in close proximity to rural farmers in the Kibale National Park, the endangered Ugandan Red Colobus monkey is folivorous, subsisting primarily of leaves and bark of native and invasive plant species [198] (image modified from:Charlesjsharp (Own work, from Sharp Photography, sharpphotography) [CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0)], via Wikimedia Commons.

reverse primer, 0.4 μL of 10 nM dNTPs, 9.6 μL of nuclease free water, 0.2 μL Physion HS II enzyme (Thermo Scientific), and 0.8 μL of BSA (2.5 mg/mL). The PCR temperature profile included an initial amplification of 98° C for 30 seconds followed by 35 cycles of 98° C for 30 seconds, 52° C for 30 seconds, and 72° C for one minute. Amplification was completed with a final elongation step for 72° C for five minutes. PCR products were purified using a 4x bead cleanup (Sera-mag SpeedBeads) involving an initial incubation with the beads for 15 minutes and two subsequent washes with 150 μL of 80% ethanol. After drying, the beads were resuspended in 20 μL of EB buffer and incubated for 15 minutes before the cleaned product was separated from the beads and placed into a new tube. Cleaned PCR products were then prepared for Illumina sequencing.

2.3.3 Illumina library preparation and sequencing. Separate ITS1 and ITS2 libraries were constructed from cleaned PCR products using the Kappa Hyper Prep Kit (Kapa Biosystems: KR0961). During end repair, samples were diluted to a concentration between 50 and 100 ng of DNA in a 12.5 μL volume. To each sample, 1.75 μL of end repair buffer and 0.75 μL of the end repair enzyme mix were added for a final reaction volume of 15 μL. Thermocycler conditions for the end repair step were 20° C for 30 minutes followed by 65° C for 30 minutes. Next, Illumina adapters were ligated to the blunt–ended DNA in a reaction of 1.375 μL of 10 μM diluted adapters, 7.5 μL of the ligation buffer, 2.5 μL of the ligation enzyme, and 1.125 μL of nuclease free water per sample tube. Samples were incubated at 20° C for one hour

after which they were immediately placed in a -20° C freezer overnight. The indexing PCR was performed in triplicate using the Phusion Hot Start II kit. For each library, 4 μL of Phusion Buffer (5X), 1 μL of BSA, 0.4 μL dNTPs, 0.3 μL of Phusion HS II enzyme, 6.4 μL nuclease free water, and 2 μL each of unique i5 and i7 primers were added. After pooling, the indexed libraries were purified using a 1.8X bead cleanup after which all samples were run on a Fragment Analyzer to determine smear concentration before diluting to 10 nM. Samples were then pooled into two separate equimolar libraries. The first library pool was run through Pippen Prep (Sage Science) for size selection at a target fragment length of 200 to 750 bp. To evaluate the impact of ITS length variation, the second library was run through Pippen Prep at a target fragment length of 200 to 900 bp. In both cases, both ITS1 and ITS2 libraries were pooled pre–size selection. Finally, samples were sequenced on an Illumina MiSeq using a 2x250 paired–end chemistry.

2.3.4 Computational methods.

Denovo database construction

After demultiplexing, reads were split into ITS1 and ITS2 amplicons and assessed for correct directionality using a custom python script (See Appendix A). Reads were then merged and quality filtered using AdapterRemoval 2.0 with a minimum quality score of 30, a minimum alignment length of 10 bp, and a minimum insert size length of 25 bp. Representative OTUs for both ITS1 and ITS2 datasets were generated via denovo clustering of all sample

reads at 97% identity using USEARCH [52] after first sorting by length and dereplication (ITS1 n=1,679,647; ITS2 n=5,488,309). A total of 99.1% of all ITS1 reads (n=1,664,881) were successfully clustered into 1,269 OTUs and 98.5% of all ITS2 reads (n=5407737) were successfully clustered into 1,474 OTUs. The taxonomy of the resulting representative sequences were predicted using BLAST against the full NCBI database with a minimum percent identity of 80%, a minimum evalue of 1e-10, and a maximum of 1,000 hits the output of which was imported into MEGAN to predict the lowest common ancestor for each OTU. Of all OTUs generated for each dataset using the methods described above, only 30.5% of the ITS1 and 52.4% were given a eukaryotic taxonomic assignment. Those OTUs that were not assigned a eukaryotic taxonomy were primarily bacterial in origin (ITS1: 14.9%, ITS2: 7.1%) or could not be assigned to any taxonomic group (ITS1: 54.8%, ITS2: 40.5%). Any OTUs that were assigned to a non–eukaryotic node in the NCBI tree (i.e., bacteria, archaea, unassigned) were removed from the database before downstream processing.

Sample taxonomic profile generation

Representative taxonomic profiles were generated for each sample by clustering merged and quality filtered data (ITS1 $\bar{x}$ read depth: 89,178 $\pm$87,377.6; ITS2 $\bar{x}$ read depth: 148,130.3 $\pm$117,844.0) A.2 against the eukaryotic–only denovo database with a minimum percent identity set to 97% using the closed reference OTU picking protocol implemented in QIIME [27] using the USEARCH clustering algorithm [52]. Of all reads generated for each ITS region,

51% of ITS1 reads were assigned to an OTU while 62% of ITS2 reads were assigned to an OTU. Resulting taxonomic profiles were normalized to account for sequencing depth and number of representative eukaryotic taxa in each sample by rarefying to the lowest non–industrial read count in each respective dataset (ITS1: 850, ITS2: 840). A comparison of ITS1 and ITS2 profiles for each sample was performed by Procrustes analysis [70] to assess intersample taxonomic differences.

Analysis of *Blastocystis* OTUs

Finally, a neighbor–joining tree of all OTUs assigned to *Blastocystis* as well as all published *Blastocystis* ITS regions uploaded to the NCBI Nucleotide database was constructed by first aligning all sequences using MAFFT [94] after which the alignment was manually checked using the Geneious [95] software package. Pairwise distance between all *Blastocystis* OTUs was calculated using the Geneious [95] software. A cladogram of all *Blastocystis* aligned OTUs was visualized using iTOL version 3 [107].

2.4 Results

2.4.1 Sequencing Results. An average of 97 thousand raw reads were generated for all samples using the ITS1 marker while an average of 372 thousand raw reads were generated for ITS2. The rate of paired end merging and quality retention of reads for both the ITS1 and ITS2 datasets is generally high with the majority of samples retaining 70% or more of all raw reads (Supplementary Table A.2). One notable exception to this is sample HS2446

which had a retention rate of 58% for the ITS1 dataset and 68% for the ITS2 dataset. This low rate for the ITS1 dataset of HS2446 is attributed to the high number of reads that were truncated before merging (31,350 reads) due to poor quality scores at one or both ends of the paired reads [112]. For the ITS2 data generated from the HS2446 individual, the relatively low merge rate appears to be a combination of high levels of truncated reads (2,398 reads), as well as discarded singleton and paired reads (1,047 reads). Reads that could not be merged were primarily bacterial in origin or could not be assigned to a taxonomic node using the described methods. For the ITS1 dataset, 57.7% of all unmerged reads were unable to be assigned taxonomy and 33.7% were assigned to a bacterial taxonomic group with the majority (21.4%) assigned to the bacterial taxon *Bifidobacterium bifidum*. For the ITS2 dataset, 65.1% of all unmerged reads were unable to be assigned taxonomy and 13.3% were assigned to a bacterial taxonomic group with the highest observed species, *Lactobacillus ruminis* contributing 3.6% of all unmerged reads. Overall, the majority of ITS amplicons for each sample were available for use in downstream analyses.

2.4.2 Taxonomic profile differences between ITS1 and ITS2. A total of 387 unique OTUs could be taxonomically identified after clustering at 97% in the ITS1 dataset while a total of 701 could be identified in the ITS2 dataset.

Of those that could be assigned a taxonomic identification, 205 ITS1 OTUs were assigned to the genus level or below (52.97%) while only 213 (30.39%) could be assigned to the genus level or below in the ITS2 dataset.

Although many taxonomic groups are represented in both ITS1 and ITS2 representative sets, 47 genus or below level taxonomic groups were only found in the ITS1 data including important microeukaryotic and dietary sequences (*Blastocystis*, *Brassica*, *Sesamum*, *Zea*). Similarly, 46 low–level taxonomic groups were only detected in the ITS2 dataset including *Schistosoma*, *Tetratrichomonas*, and *Spinachia*) (Supplementary Table A.1).

After filtering out non–microeukaryotic OTUs (e.g., plants, animals), a Procrustes analysis of paired ITS1 and ITS2 data in each samples illustrates different taxonomic profiles generated using either approach within the same sample (Figure 2.2). The most extreme difference between datasets generated from the same sample is detected in samples NO15 and SM02 both of which have relatively lower microeukaryotic OTU assignments than other samples included in this dataset. Compounding the observed differences between the ITS1 and ITS2 data in these two samples, both are dominated by *Blastocystis* in the ITS1 data. This implies that differences between datasets will be more acute if information is already sparse in one or both ITS regions. Interestingly, the dominant taxonomic group detected in the ITS2 dataset generated for sample SM02 is *Chalara*, a fungal endophyte of trees. While members of *Chalara* have been found in the gut of wood eating beetles [213], is not a typical fungus found in the gut of mammals and therefore it is unlikely that the presence of this taxa in one of the Matses samples is a natural occurrence. Instead, as the OTU assigned to this taxonomy only is 93% similar to its closest match (NCBI accession: JN604461.1), it is likely this represents a unknown fungal species not currently characterized in the NCBI database.

No host DNA was detected in the merged or unmerged data generated in either ITS datasets. As the length of the ITS1 and ITS2 regions in mammals are generally longer than most microeukaryotic taxa [35] (*Homo sapiens* ITS1: 1069 bp ITS2: 1166 (NCBI accession: KY962518); *Bos taurus* ITS1: 1071 bp ITS2: 1037 bp (NCBI accession: DQ222453; *Procolobus tephroscelles* No data available), these results are consistent with the size selection protocol used prior to sequencing and either were not amplified during the initial PCR or were removed during size trimming after pooling for sequencing. By effectively removing the host from consideration, more sequencing space is therefore available for microeukaryotic taxa.

2.4.3 Dietary results. Amplicons of potential dietary sources were detected in both the ITS1 and ITS2 datasets (Table 2.2). Consistent with expectations based on food availability, dietary amplicons retrieved from the Norman samples include a variety of plant derived food sources common in industrialized societies including cilantro (*Corandrium sativum*), tomato (*Solanum lycopersicum*), spinach (*Spinacia oleracea*), the *Apiaceae* family which includes celery, parsley, cilantro, carrots and other vegetables and herbs, members of the *Brassica* genus which include cabbage and its relatives, *Cucumis* which includes cucumber, and the high level eukaryotic taxonomic group *Vaccinieae* which may be attributed to the consumption of berries including cranberry and blueberry. Inferred dietary reads from the Matses includes *Salmonidae*, members of which include various bony fishes and *Myrteae* which may be associated with fruit consumption. Poten-

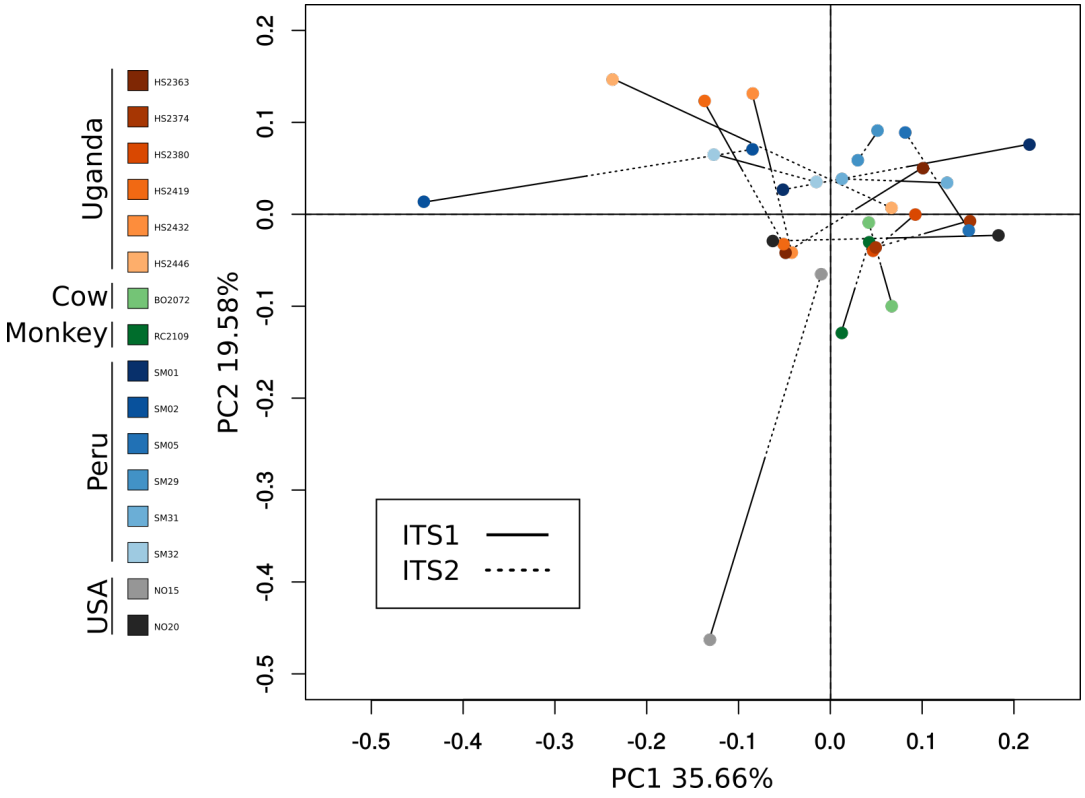

Figure 2.2: **Procrustes plot illustrating differences between ITS1 and ITS2 datasets.** Each pair of dots connected by a line represents the microeukaryotic profile of a single sample via either the ITS1 or ITS2 dataset. While many samples have similar taxonomic profiles, the microeukaryotic diversity summarized by the ITS1 and ITS2 region is clearly different. The most extreme differences in ITS1 and ITS2 taxonomic profiles are found in samples SM02 and NO15, both of which have relatively low numbers of microeukaryotic OTU assignments. Therefore, differences between ITS1 and ITS2 datasets are intensified when information is sparse in one or both internal transcribed spacer regions.

tial dietary reads from Uganda include those for peanuts (*Arachis hypogaea*) and other legumes and tubers (*Ipomoea*, *Phaseolus*), sesame (*Sesamum indicum*), grains (*Sorghum*, *Eleusine*), and wild tomato (*Solanum pennellii*). The pseudocereal genus *Amaranthus* is found in both the Matses and

Uganda groups, as is *Musa* which includes bananas and plantains. Interestingly, corn (*Zea mays*) is present in all three human populations at either the genus or species level reflecting the global reach of corn as foodstuff. Among the non–human animals included in this study, grasses, sedges, (*Cynodon*, *Cyperaceae*, *Paspalum scrobiculatum*), and other plants (*Desmodium*/Desmodieae, *Dichondra repens*, *Hydrocotyle*, etc) predominate the dietary reads found in the bovid. Dietary reads from the *Procolobus tephrosceles* individual reflects its folivorous feeding behaviors and includes species of trees (*Ficus*, *Macaranga*, *Mimusops*, *Theobroma grandiflorum*), other plant taxa (Lauraceae, Moraceae, etc), and an insect species (*Lepidocyrtus koreanus*). Interestingly, while some dietary OTUs are detected in either the ITS1 or ITS2 datasets, many are found in both, reinforcing their presence in our samples.

Table 2.2: **Dietary OTUs observed in each sample group.** Inferred dietary OTUs observed in ITS1 and ITS2 datasets.

Taxon	Resolution	Locus	Source
Apiaceae	Family	Both	Norman
Brassica	Genus	ITS1	Norman
Coriandrum sativum	Species	Both	Norman
Cucumis	Genus	ITS1	Norman
Euphorbia	Genus	ITS2	Norman
Rubus	Genus	Both	Norman
Solanum lycopersicum	Species	ITS1	Norman
Spinacia oleracea	Species	ITS2	Norman
Trifolieae	Sub Family	ITS1	Norman
Vaccinieae	Tribe	Both	Norman
Cecropia	Genus	ITS2	Matses
Cecropia peltata	Species	ITS1	Matses
Maclura	Genus	ITS2	Matses

Table 2.2 continued from previous page

Myrteae	Tribe	ITS1	Matses
Salmonidae	Family	ITS2	Matses
Arachis	Genus	Both	Uganda
Arachis hypogaea	Species	ITS1	Uganda
Eleusine	Genus	Both	Uganda
Ipomoea	Genus	ITS1	Uganda
Phaseolus	Genus	IGS1	Uganda
Sesamum indicum	Species	IGS1	Uganda
Solanum	Genus	IGS1	Uganda
Solanum pennellii	Species	ITS2	Uganda
Sorghum	Genus	Both	Uganda
Amaranthus	Genus	Both	Uganda & Matses
Musa	Genus	Both	Uganda & Matses
Poaceae	Family	ITS2	Norman & Matses
Zea mays	Species	ITS1	Norman & Matses

Table 2.2 continued from previous page

Zea	Genus	ITS1	All groups
Cynodon	Genus	ITS1	Bovid
Cypereae	Tribe	ITS1	Bovid
Desmodieae	Tribe	ITS2	Bovid
Desmodium	Genus	Both	Bovid
Dichondra repens	Species	ITS1	Bovid
Hydrocotyle	Genus	ITS1	Bovid
Paspalum	Genus	ITS1	Bovid
Paspalum scrobiculatum	Species	ITS2	Bovid
Ficus	Genus	Both	Colobus
Lauraceae	Family	ITS1	Colobus
Lepidocyrtus koreanus	Species	ITS2	Colobus
Macaranga	Genus	Both	Colobus
Mimusops	Genus	ITS2	Colobus
Moraceae	Family	ITS1	Colobus

Table 2.2 continued from previous page

Rauvolfioideae	Sub Family	ITS1	Colobus
Theobroma grandiflorum	Species	ITS2	Colobus
Urera	Genus	ITS1	Colobus
Vanguerieae	Tribe	ITS2	Colobus

2.4.4 Microeukaryotic results. Microeukaryotic diversity, defined as the number of unique OTUs attributed to a sample, is highest amongst the Ugandan Red Colobus (RC2109) and Bovid (BO2072) samples in both the ITS1 and ITS2 datasets (Figure 2.3). Of the human samples in this study, the Matses and Ugandan samples share more taxa amongst each other than do the Matses or Ugandan with Norman. As expected with an industrialized food production source and water purification systems, the Norman samples have the lowest total microeukaryotic diversity of all human samples. For Norman sample NO7 and NO16 the number of ITS1 and ITS2 OTUs that could be defined through closed reference clustering did not pass rarefaction limitations, respectively, and were therefore unable to classify. Importantly, the ITS1 dataset is driven by the presence of *Blastocystis* which was not detected in our ITS2 data. *Blastocystis* was found in all human groups using the ITS1 marker, irrespective of diet or environment.

The predominant type of microeukaryotes in all samples is fungi. Other microeukaryotes found in human derived samples include likely non–pathogenic single–celled eukaryotic taxa including *Blastocystis hominis* and *Entamoeba dispar* as well as eukaryotic parasites including *Schistosoma mansoni*. While no parasites were detected in the *Procolobus tephrosceles*, the Bovid sample was positive in the ITS2 dataset for a variety of high level taxonomic groups that include potentially parasitic species including *Digenea* a diverse class of flatworms in the *Platyhelminthes* phylum, and single–celled parasites (Parabasalia, *Simplicimonas sp.*, *Tetratrichomonas sp.*, *Blastocystis*). While members of the *Parabasalid* group include symbiotic mi-

croeukaryotes in some insect groups, *Tritrichomonas foetus*, a member of the *Parabasalid* group, is a sexually transmitted disease in cattle that can cause spontaneous abortion [118] the control of which is of significant interest to the global cattle industry [33]. Population based parasite results are found in Table 2.3.

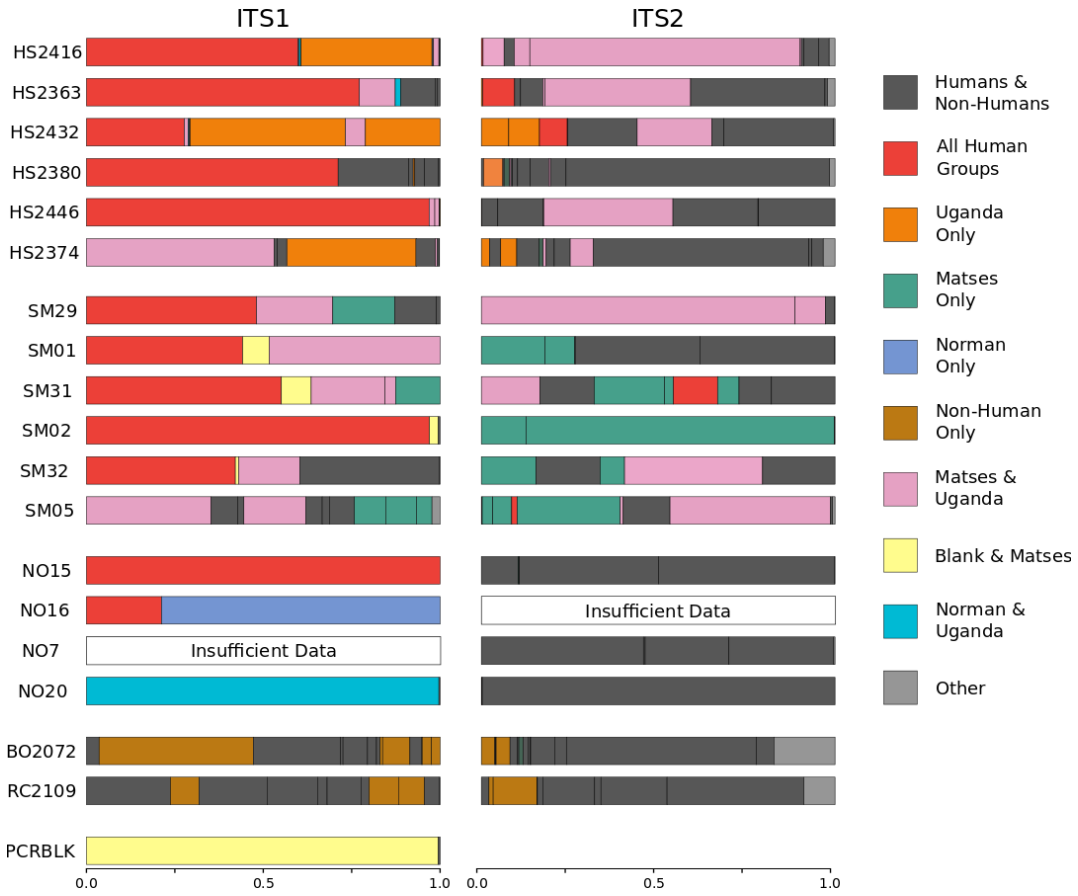

Figure 2.3: **Microeukaryotic taxonomic profiles for all samples colored by specificity to different groups.** Taxonomic profile barcharts colored by OTU specificity to each group.

Table 2.3: **Observed parasite and commensal microeukaryotic organisms.** Inferred microeukaryotic parasites and potential commensal organisms detected in each study population. Most detected organisms in human groups are suspected commensal organisms.

Taxon	Level	Locus	Source
Entamoeba	Genus	ITS1	Uganda & Matses
Entamoeba dispar	Species	ITS2	Uganda
Schistosoma mansoni	Species	ITS2	Uganda
Blastocystis hominis	Species	ITS1	Uganda, Matses, Norman & Bovid
Digenea	Class	ITS2	Bovid
Parabasalia	Phylum	ITS2	Bovid
Simplicimonas	Genus	ITS2	Bovid
Simplicimonas sp. GABC1	Species	ITS2	Bovid
Tetratrichomonas sp. IdnP1	Species	ITS2	Bovid

2.4.5 Blastocystis. Consistent with previous studies of the microeukaryotic diversity of the gut [180], *Blastocystis hominis* is the predominant non–fungal eukaryote detected in the gut using the ITS1 dataset. Through phylogenetic analysis, *Blastocystis hominis* subtypes one through three are all present in the human samples while an unknown subtype appears to be found in the bovid sample (Figure 2.4 and Figure 2.5). While only detected

using the ITS1 dataset, OTUs assigned to *Blastocystis* are highly diverse, reflecting the acute genetic diversity known to reside in this genus. Importantly, *Blastocystis* is detected in all populations with subtypes one through three found in the Uganda and Matses samples while two Norman samples (NO16 and NO15) were found positive for subtype three and potentially subtype one or three, respectively.

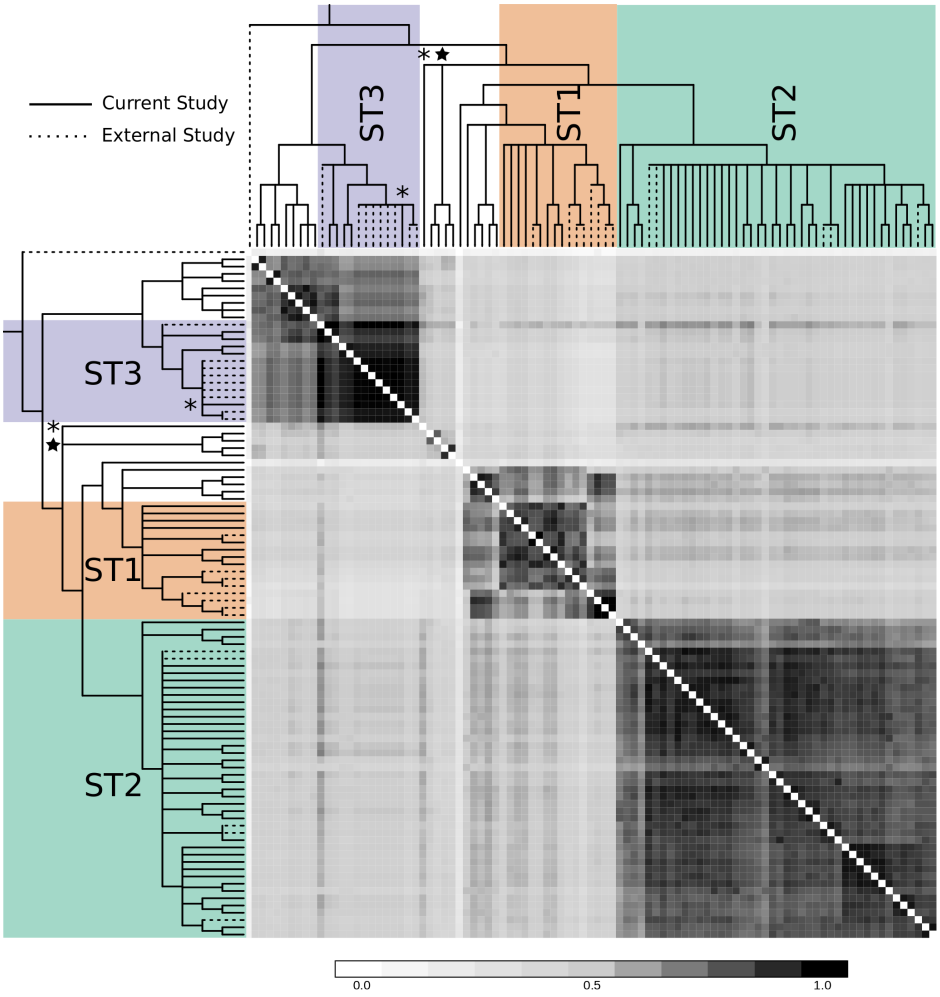

Figure 2.4: **Pairwise distance and cladogram of all *Blastocystis* ITS1 regions from this dataset and previously published *Blastocystis* ITS regions.** Cladogram and pairwise identity matrix of *Blastocystis* OTUs identified in the ITS1 dataset clustered with previously published *Blastocystis* OTUs. Three subtypes of *Blastocystis* are present in the human groups represented in the current study with all three found in the Ugandan and Matses samples while the two Norman samples test positive for *Blastocystis* OTUs that are suspected subtype one varieties. The two *Blastocystis* OTUs detected in the Norman samples are indicated by a ⋇ symbol. In addition to the human *Blastocystis* lineages, a suspected bovid strain of the microeukaryote was detected in the single bovid sample and is indicated by the ★ symbol.

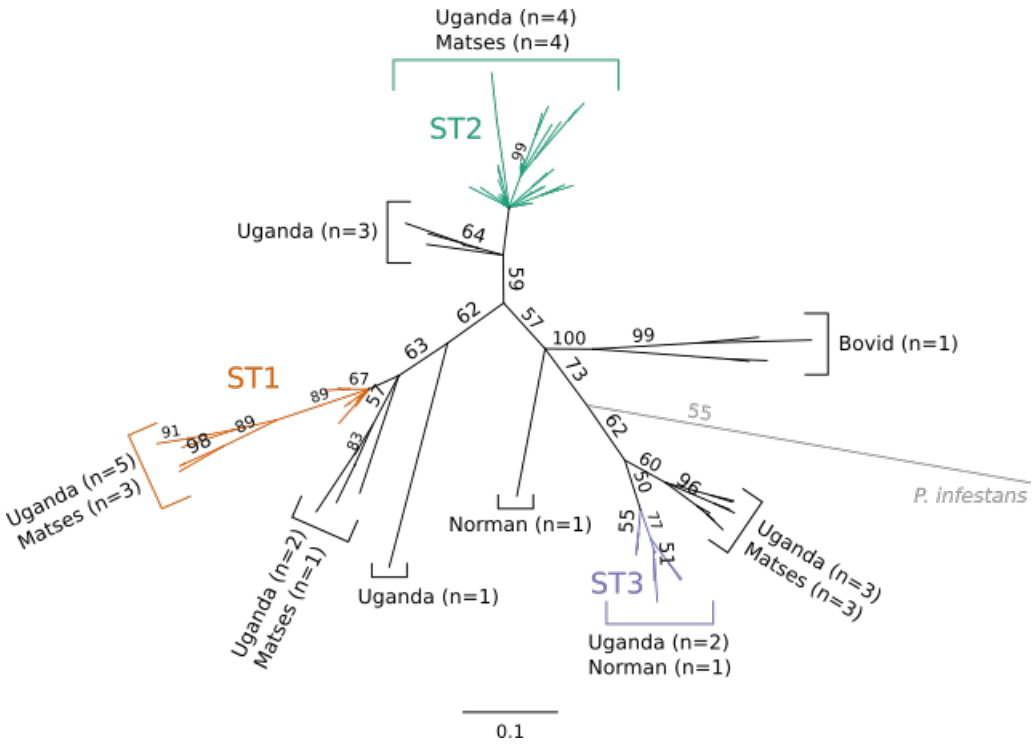

Figure 2.5: **Neighbor–joining tree of all *Blastocystis* OTUs generated in the current study and previously published *Blastocystis* ITS1 regions.** All three inferred subtypes of *Blastocystis* are found in the Matses and Ugandan human samples while only subtype three is found in the Norman samples. In addition to those *Blastocystis* OTUs found in the human samples, a separate lineage of *Blastocystis* is detected in the bovid sample and may represent a non–human lineage of the organism. While basal branches of the tree are relatively robust, bootstrap values of the terminal branches are more variable, indicating that the ability of the ITS1 region to resolve subtype variation is limited.

2.5 Discussion and Conclusions

2.5.1 Benefits and limitations of the internal transcribed spacer region for eukaryotic surveys of the gut.

The results of this study demonstrate the utility of targeting both the ITS1 and ITS2 region for a more complete

microeukaryotic and dietary survey of fecal samples. As the ITS1 and ITS2 region characterize different taxonomic groups at different levels of taxonomic resolution, they are best used in tandem if the full ITS region is not a reasonable target. For example, while 37 genus level OTUs were found in both the ITS1 and ITS2 datasets, an additional 28 genera were only detected using the ITS1 region and 36 were only detected in the ITS2 data. Though amplification of the ITS region is the currently accepted barcode for studies of fungal diversity in environmental or microbiome studies [166], the current study confirms its applicability to other microeukaryotic and dietary taxa. Currently, a limiting factor for this type of targeted metabarcoding survey is the paucity of data available for non–fungal eukaryotic organisms. A 2005 study of ITS2 regions in eukaryotes found that while more than 40% of fungal taxa in the NCBI database had at least one representative region listed, less than 4% of metazoa had similar annotations [168]. Those that do have representative ITS sequences are biased towards certain groups, for example flowering plants, which may represent a historical basis for the characterization of ITS among certain non–fungal groups [168]. While databases for the ITS1 and ITS2 [88, 169, 100] regions are available, no current database exists for the full ITS region for non–fungal organisms. Because the ITS region is characterized by large insertions and deletions, a full database of eukaryotic organisms may be useful for the design of research programs that target or omit particular taxa. For example, in the current study, the taxa of interest, microeukaryotes, have an expected shorter ITS length than the host mammal species [35]. Therefore, limiting the elongation time during PCR preparation

effectively removes the host DNA from analysis, increasing the sequencing space for microeukaryotes of interest.

2.5.2 Dietary DNA reflects industrial and non–industrial subsistence strategies. Reads from suspected dietary sources are consistent with the known subsistence patterns of the populations included in this study. Dietary sources detected in the Norman individuals includes plant sources typical of industrialized societies including tomatoes (*Solanum lycopersicum*), spinach (*Spinacia oleracea*), members of the cabbage family (*Brassica*), and other common fruits, vegetables, and herbs that can be easily purchased at grocery stores within the United States. Fewer dietary sources were detected in the Matses, but a diet rich in fish, plantains or bananas, and corn has previously been documented in this population [131] which is confirmed in this study. Other previously documented sources of food for the Matses including sloth, species of new world monkey, and a variety of reptiles and birds [131], are not represented in publicly available ITS databases. Because of this, the paucity of dietary information for this population is perhaps unsurprising. Dietary data generated in this study for rural agriculturalists living in Uganda are consistent with the known diet of this area which is rich in legumes and grains [78].

Like dietary data generated for the human derived fecal samples, the implied diet of the single bovid and Ugandan Red Colobus monkey are within expectations of these two species. In particular, the Ugandan Red Colobus monkey included reads assigned to the *Ficus* genus and *Moraceae* family, both of which may represent one of the primary food sources of Red Colobus

living in the Kibale National park, *Ficus natalensis* [198]. The provenience of other potential dietary OTUs is less clear. For example, the plant genus *Euphorbia* includes approximately 2,000 known species some of which are used in traditional medicine [172, 128] as well as some applications in the pharmaceutical or industrial food sciences field [173, 172] but the origin of this taxa in the Norman population is unknown.

2.5.3 Microeukaryotic diversity and parasitic infection rates higher in non–industrialized groups. Microeukaryotic diversity in the Matses and Ugandan individuals are higher than that found in Norman which is comparatively depressed. Both phylogenetic diversity (PD) and the number of observed OTUs as measures of α diversity in the ITS1 dataset are significantly higher (Mann–Whitney U Test, ITS1: p = 0.02) in the Matses (ITS1: PD: $\bar{x}$5.1 $\pm$ 1.2; observed OTUs: $\bar{x}$9.3 $\pm$ 5.0) and Ugandan individuals (ITS1: PD: $\bar{x}$7.3 $\pm$ 0.9; observed OTUs: $\bar{x}$24.7 $\pm$ 12.0) as compared to the Norman individuals (PD: $\bar{x}$2.1 $\pm$ 0.8; observed OTUs: $\bar{x}$1.7 $\pm$ 0.6). While both measures of α diversity are significantly higher in the Ugandan individuals (PD: $\bar{x}$4.5 $\pm$ 0.9; observed OTUS: $\bar{x}$18.3 $\pm$ 7.4) as compared to the Norman population (PD: $\bar{x}$2.3 $\pm$ 0.5; observed OTUs: $\bar{x}$7.0 $\pm$ 2.6) in the ITS2 dataset (Mann–Whitney U Test, PD: p = 0.02; observed OTUs: p = 0.3), comparisons of the Matses individuals (PD: $\bar{x}$2.6 $\pm$ 1.1; observed OTUs: $\bar{x}$9.0 $\pm$ 7.6) and Norman yielded no significant results. Importantly, the Matses and Ugandan individuals share more microeukaryotic OTUs between them than either share with the Norman samples, despite the geographic distance that

separate these two populations. This suggests patterns of microeukaryotic diversity in human populations are strongly influenced by lifestyle and not by environment. Consistent with previous studies, the most commonly detected non–fungal microeukaryotic organisms found in this dataset include the commensal or opportunistic pathogens *Blastocystis hominis* and *Entamoeba dispar* [180, 127, 164, 54, 202]. One individual from the Ugandan population, however, tested positive for the blood fluke *Schistosoma mansoni*, a pathogenic species responsible for schistosomiasis. Schistosomiasis is one of the most widespread parasitic diseases in sub–Saharan Africa, only surpassed by malaria [1]. Currently, an estimated 732 million people are vulnerable to the disease according to the WHO and in 2008 more than half of all schistosomiasis cases were recorded in African nations [1].

2.5.4 Blastocystis is a cosmopolitan parasite found in all human groups. *Blastocystis* is a genus of single–celled, requisite anaerobes found in the *Stramenopile* phylum and is closely related to diatoms and algae [20]. Visually similar to soap or water bubbles under the microscope, *Blastocystis* cells are immotile and are apparently the only genus of the *Stramenopile* phylum to regularly infect humans [20]. Globally, *Blastocystis* is reported to be the predominant non–fungal eukaryotic taxon recovered from human fecal samples and is usually asymptomatic [75]. Dissemination of *Blastocystis* is via the fecal–oral route [156], though it may be zoonotically transmitted as well [3]. While morphologically ambiguous, members of the *Blastocystis* genus are highly genetically diverse with no fewer than 17 documented sub-

types currently described [3, 32]. Subtypes one through nine are regularly found in humans with some geographic structuring in terms of relative frequency of *Blastocystis* subtype distributions [8]. In addition to humans, *Blastocystis* species have been characterized in non–human mammals, birds, reptiles, amphibians, and insects [32] with no single subtype of *Blastocystis* being specific to humans [20]. The degree of genetic diversity within this genus is surprising with some suggestion that it should be taxonomically reevaluated [3]. Consider, for example, that those *Blastocystis* isolates named *Blastocystis hominis* for their initial isolation source are as genetically diverse as all known and sequenced members of the *Cryptosporidium* genus [20].

From this dataset, three of the 17 known subtypes of *Blastocystis* were detected. While other OTUs of *Blastocystis* generated in this study are suggestive of different strains of each subtype, insufficient phylogenetic resolution is available to definitively define separate clades of the organism (Figure 2.5). Surprisingly, *Blastocystis* OTUs were detected in all three human groups and a separate clade of *Blastocystis* was detected in the bovid sample. While the risk of being a carrier for *Blastocystis* is elevated in geographic regions with poor sanitation and higher risk of tainted food or water sources, direct evidence of the mode of transmission for *Blastocystis* is inconclusive [4]. For example, while estimated rates of human infections with *Blastocystis* are as high as 45% in Columbia [156], in Japan they are predicted to be as low as 0.5% [164]. While the predicted rate of *Blastocystis* in human populations living in the United States is relatively low (between 11 and 23% [6]), it is

not uncommon for multiple subtypes of the genus to be found within a single family unit [163]. In the current study, two otherwise healthy individuals from the Norman cohort (NO16 and NO15) tested positive for *Blastocystis*, both of which falls phylogenetically close to other published *Blastocystis* subtype 3 ITS1 sequences (Figure 2.5, Figure 2.4). All three subtypes are found in the Matses and Ugandan samples with many individuals exhibiting co–infections with distinct strains of *Blastocystis*. All three subtypes detected in this study have been linked to chronic gastrointestinal disorders including irritable bowel syndrome (IBS) either alone or as coinfections as well as in asymptomatic infections [46, 20, 179].

2.5.5 Conclusions. The results from this study document the utility of the internal transcribed region for characterizing microeukaryotic and dietary diversity of both human and non–human gut microbiomes. Comparative analysis of the presence or absence of specific eukaryotic taxa in human populations with diverse subsistence strategies, lifestyles, and environments documents that, like bacterial communities that live in the gut, the microeukaryotic diversity of the human gut is strongly linked to lifestyle and not the external environment. Specifically, the loss of microeukaryotic diversity in individuals living in sanitized urban–industrialized regions of the world is suggestive of a major shift in our relationship with these microorganisms. Despite this shift, there are some microeukaryotes including *Blastocystis* that are present in all human groups. Whether the global distribution of *Blastocystis* is in response to the removal of soil–transmitted helminths [144] or if the organism repre-

sents an ancestral heirloom microeukaryote in the human gut is beyond the scope of this study. Understanding the natural distribution of microeukaryotes in diverse human populations is the first step in understanding the ecological importance and function of these diverse organisms in complex communities of bacteria, archaea, viruses, and eukaryotes.

Chapter 3

Enrichment of non–dominant bacterial taxa in human fecal samples through serial filtration

3.1 Abstract

The human gut is a complex ecological system primarily inhabited by diverse bacterial organisms. While current molecular techniques including metagenomic and metabarcoding sequencing are effective tools for characterizing bacterial community composition, these methods are confounded by issues of bacterial genome size and copy number variation in common marker genes among different bacterial groups (e.g., 16S rRNA gene). These variations may mask organisms that are present in low abundance or those with comparatively small genomes and therefore bias interpretations of whole community structure. In this study, fecal samples representing two divergent human groups were processed through a serial filtration protocol to enrich for underrepresented taxa in the human gut. Results from this experiment indicate that serial filtration of fecal samples and 16S rRNA metabarcoding may be an appropriate screening tool to identify samples for which further investigations including whole genome reconstruction and culturing of non–dominant taxa

may be accomplished.

3.2 Introduction

Bacteria are the most diverse group of organisms on earth, profoundly eclipsing the diversity of *Archaea* or *Eukaryota* [85]. Testament to this diversity, Bacteria can be found in every natural ecological niche, from marine sediments 11,000 meters deep at the bottom of the Mariana Trench [68], to 75° C alkaline hot springs in Yellow Stone National Park [42], to the harsh acidity of the human stomach [192]. The diversity of Bacteria further manifests in an astounding variety of cell size, structure, and form. The largest known bacterium belongs to the group *Thiomargarita*; initially misclassified as a protist, a single *Thiomargarita namibiensis* cell is 0.1 to 0.3 mm in diameter—approximately the same size as a *Drosophila* eye [108] or a single period on this printed page. Comparing a single cell of *Thiomargarita namibiensis* to one of the smallest known bacterial taxa, *Candidatus actinomarinidae* at 0.013μm [66] is like a grain of rice next to the Great Pyramid of Giza. The genome size of large bacteria also tends to be quite substantial, with the largest bacterial groups often exhibiting a high degree of polyploidy [108], while small bacteria—and especially those that live as symbionts of other microorganisms—tend to have a radically reduced genome size due to their ability to assimilate the molecular mechanisms or metabolic resources of their host [123]. For example, many members of the bacterial endosymbiont genera *Rickettsia*, *Mycoplasma*, and *Buchnera* have eliminated genes involved in energy metabolism required for free–living bacteria [126]. Consider, for

example, that while free–living bacteria have a predicted proteome typically between 1,500 to 6,000 proteins, bacteria that rely on some aspect of a host cell for their survival or reproduction may produce as few as 500 to 1,000 proteins [126].

This variation in genome size is a challenge for the characterization of bacteria in mixed microbial ecosystems as organisms with larger genomes will have an artificially inflated representation in metagenomic surveys [17], masking the presence of taxa with smaller genome sizes and presumably smaller cell sizes. Micro–filtration of environmental microbial communities including soils [150, 149] and water sources [193] demonstrate the utility of bacterial cell–size fractionation of mixed microbial communities for the enrichment or detection of non–dominant taxa, yet to date no such analysis of the human gut microbiome has been performed. The human gut microbiome is a rich microbial ecosystem chiefly inhabited by bacterial organisms from divergent lineages representing 12 phyla [86]. As a rich source of nutrients, the gastrointestinal tract of certain vertebrates and invertebrates have been found to be host to large bacterial species [108], which may indicate that smaller cells with reduced genome size are underrepresented in molecular characterizations of the gut microbiome. As bacteria that are classified as epibionts (those that live on the surface of another organism) or endosymbionts (those that live within another host cell or organism) are difficult to culture as compared to free–living bacteria due to their specialized host–dependent life–cycles, the detection and functional characterization of these reduced genome organisms [123] using molecular techniques is highly desir-

able.

The purpose of the current study is to partition the bacterial community of fecal samples collected from different dietary ecologies through serial filtration steps to enrich for underrepresented bacterial taxa. Results from this study suggest that serial filtration is an effective method for decreasing the overall representation of dominant bacterial species and thereby increasing the overall proportion of non–dominant under–characterized bacterial organisms. Individuals from the non–industrial population in this study exhibited higher than expected proportions of operational taxonomic units (OTUs) from the phylum *Cyanobacteria* which were assigned to the order *Vampirovibrio*. These OTUs likely represent members of the candidate phylum or class *Melainabacteria* [176, 45], a known group of gut symbionts that are to date described only by their presence in metagenome datasets [45]. Therefore, 16S rRNA metabarcoding sequencing of serial filtered fecal samples may be a cost–effective method of screening appropriate candidate samples for further investigations including full genome reconstruction and culture isolation of novel strains.

3.3 Methods

3.3.1 Samples. A total of six human fecal samples from hunter–gatherers originating in the Peruvian Amazon (n=4) and residents of Norman, Oklahoma, USA living a typical urban industrialized lifestyle (n=2) [131] were chosen to maximize the expected taxonomic diversity and overall community composition of bacterial taxa living in the gut of study participants. To test

the impact of large eukaryotic parasite cells on the initial filtration tests, two individuals from the hunter–gatherer group that had previously tested positive for microeukaryotic infection via microscopy were included. Descriptions of individuals from which samples were collected in the current study can be found in Table 3.1.

Table 3.1: **Sample metadata.** Sample ID, geographic location, age, and sex for all samples included in the current study.

Sample ID	Location	Age	Sex	Parasite Status
SM01	Peru	30	M	Negative
SM29	Peru	50	F	Negative
SM31	Peru	30	M	Positive
SM02	Peru	25	F	Positive
NO7	USA	32	F	Unknown
NO15	USA	50	F	Unknown

3.3.2 Fecal cell size filtration. Sterile cell strainers at filter mesh sizes 200 μm, 60 μm, 20 μm, and 5 μm (pluriStrainer: 43-50200-03, 43-50060-03, 43–50020–03, 43–50005–03) were selected to filter fecal slurry samples. Beginning with the 200 μm filter, approximately 1.5 mL of homogenized fecal slurry was added to the filter placed on top of a UV sterilized 50 mL falcon tube. Loaded filters were covered tightly with parafilm before centrifuging for two minutes at 4,000 rpm. If filters appeared wet or remaining liquid was

observed post centrifugation, samples were centrifuged for an additional two minutes at 4,000 rpm. Material that did not pass through the filter was lifted by washing the filter with 1 mL of nuclease free water and aspirating up and down with a 1200 μl pipette tip. All liquid from the upper portion of the filter was then removed and placed in a clean 1.5 mL Eppindorf tube. Any sample or water that passed through the filter into the 50 mL falcon tube was first vortexed briefly before being passed through the next filtration stage. These steps were repeated for all filter sizes in descending order from 200 μm to 5 μm. Parallel to sample filtration a single filter negative control was processed using nuclease free water in place of fecal material. Finally, flow through from the 5 μm filter was removed and placed into a clean 5 mL tube. All samples were concentrated to approximately 200 μL of liquid sample in an Eppendorf Vacufuge set to 30° C. In addition to fecal samples and the single filter negative control, approximately 1.5 μL pure cultured *Escherichia coli* suspended in Tryptic Soy Broth (Sigma–Aldrich 43592-800ML) was processed in an identical manner to serve as a positive control. Quantitative PCR results targeting the V4 region of the 16S rRNA gene for the *E. coli* control can be found in Supplementary Table B.1.

3.3.3 DNA extraction and quantification. All samples were extracted using a DNAeasy Power Soil Extraction Kit (Qiagen) following the manufacturer's protocol after an initial heat lysis step for 10 minutes at 60° C. The full volume of all samples (approximately 200 μL) was added to bead beating tubes for extraction. Extractions were then quantified

for total DNA yield using a Qubit fluorometric assay. Proportion of eukaryotic and bacterial DNA in each sample was estimated via quantitative PCR assay using primers specific to the eukaryotic internal transcribed spacer one region (ITS1f: 5'–TCCGTAGGTGAACCTGCGG–3'; ITS2r: 5'–GCTGCGTTCTTCATCGATGC–3') [204] and the bacterial 16S rRNA gene, V4 region (F515: 5'-CACGGTCGKCGGCGCCATT-3'; R806: 5'-GGACTACHVGGGTWTCTAAT-3') [26] the results of which can be found in Supplementary Figure B.2.

3.3.4 16S rRNA amplicon library preparation and sequencing. Before sequencing, samples were grouped based on their quantitation cycle (Cq) value as determined by qPCR and diluted so that all grouped samples would amplify at the same number of cycles. Sample dilutions were then amplified using unique Illumina specific barcoded 16S rRNA V4 primers. Each PCR reaction consisted of 4 μL of Phusion HF buffer (Thermo Scientific), 1 μL of the Illumina forward primer (10 μM), 2.0 μL of 10 nM dNTPs, 5.0 μL of nuclease free water, 0.2 μL Phusion HS II enzyme (Thermo Scientific), and 0.8 μL of BSA (2.5 mg/mL). Illumina barcoded reverse primers (2.5 μM) were added separately to each reaction tube at a volume of 4.0 μl. 3.0 μl of DNA was added to each reaction tube before placing in the thermocycler. The cycling conditions for the 16S rRNA library build includes an initial amplification of 98° C for 30 seconds followed by 27 to 29 cycles of 98° C for 15 seconds, 54° C for 30 seconds, and 72° C for 30 seconds. Amplification was completed with a final elongation step for 72° C for five minutes. PCR for each

sample was done in triplicate and checked for the V4 expected amplicon size via gel electrophoresis. After confirming all PCR reactions were successful, samples were pooled and run on a 1% agarose gel for 150 minutes at 70 volts to ensure adequate separation of the 100bp ladder and sample amplicon size. An equal volume (5 μL) of each sample was pooled together and 1 μL of each blank and standard was added to the pool. Pooled samples were run through PippenPrep to select read sizes from 300 to 450 bp to remove any dimer or residual high–molecular weight DNA prior to sequencing on an Illumina MiSeq using a 2 x 250 chemistry.

3.3.5 Computational methods. Samples were demultiplexed and converted from BCL to fastq files using the Illumina bcl2fastq conversion software version 1.8.4. An average of 32,324 ± 8,544.1 reads were generated across all samples (median = 32,017). Reads were merged using Pear (version 0.9.8) [212] with an average merge rate of 99.2% (range = 98.6% to 99.7%). Merged reads were then quality filtered using Sickle version 1.33 [93] with a quality filter of 30 and minimum read length of 100 bp. All reads were clustered into OTUs at a 97% percent identity threshold using the Usearch version 10.0.240 denovo pipeline [52]. Taxonomy for each OTU was assigned using the `assign_taxonomy.py` script as implemented in QIIME version 1.9 [27] using the EzBioCloud 16S ribosomal RNA gene database as a reference [211]. Quality filtered reads were then mapped onto the denovo reference dataset using the `-otutab` option in Usearch version 10.0.240 [52]. A total of 97.3% of all quality filtered reads mapped to OTUs in the denovo dataset.

A QIIME formatted otu map was then generated from the resulting Usearch OTU table and taxonomy assignments using a custom python script (see Appendix B) which was then made into a hdf5 formatted biom table using the QIIME script `make_otu_table.py` [27]. Before downstream analysis, the biom table was first rarefied to 8,000 sequences to maximize the inclusion of reads from all true samples (Supplementary Figure B.1). Taxonomic summaries for each sample were generated in R [155] using the ggplot2 library [205]. Maximum likelihood trees were generated using RAxML version 8.2.11 [178].

3.4 Results

3.4.1 Bacterial and eukaryotic cell abundance at sequential filter levels. To estimate the proportion of eukaryotic and bacterial cells captured at each filter level, a quantitative PCR (qPCR) assay was performed targeting the universal V4 hypervariable region of the 16S rRNA gene and the eukaryotic targeted internal transcribed spacer one (ITS1) region of the eukaryotic ribosomal RNA complex. Relative cell abundance is approximated from the quantitation cycle (Cq) number in which fluorescence is detected from each sample. A low Cq value indicates a higher copy number of the targeted gene region while a high Cq value indicates a low copy number. Expectations of the experimental design were that all filter levels would have low copy number for the V4 region of the 16S rRNA gene with a higher detected bacterial presence in the flow through. Conversely, a higher copy number of the eukaryotic ITS1 targeted region was expected at the higher filter levels but not

Table 3.2: **Estimate of bacterial and eukaryotic abundance by quantitative PCR.** Quantitative PCR results indicate an overall low frequency of eukaryotic cell density at any pore size with the highest estimated frequency of ITS1 copy number at the 200 μm level. In contrast, the estimated cell density of bacteria as quantified by V4 copy number remains relatively consistent across all pore sizes.

Filter Level	V4	ITS1
200 μm	14.6 $\pm$ 0.6	26.4 $\pm$ 1.8
60 μm	16.7 $\pm$ 2.3	29.4 $\pm$ 2.3
20 μm	15.8 $\pm$ 1.9	28.7 $\pm$ 2.4
5 μm	15.5 $\pm$ 1.7	29.1 $\pm$ 2.1
Flow Through	16.8 $\pm$ 1.6	30.5 $\pm$ 1.3

in the flow through. Cq values for negative controls (blanks) including those collected during extraction, PCR, and filter negatives are indicative of a contamination controlled experiment, with an average V4 Cq value across all negative controls of 32.4 $\pm$ 0.2. No amplification of any negative controls was detected in the ITS1 assay. Cq values as measured by ITS1 amplification across all samples are high (Supplemental Figure B.2), indicating a lower than expected eukaryotic cell capture at any filter level. The highest average Cq value for the ITS1 region was detected at the 200 μm filter (26.4 $\pm$ 1.8) while the highest average Cq value could be found in the sample flow–through (30.5 $\pm$ 1.3), consistent with expectation that large eukaryotic cells will be trapped at higher pore sizes allowing bacterial cells to flow through. Unexpectedly, a considerable amount of bacteria, as measured by low Cq value, are captured at the higher pore size filters and the flow through is on average the most diluted source of bacterial cells (Table 3.2).

As the diet of individuals in this study are highly fibrous and a fair amount

of organic material is trapped at higher pore size filters, it is possible that bacterial cells aggregated with this organic material accounts for the relatively consistency of bacterial cell mass as estimated by qPCR. Alternatively, larger conglomerations of bacterial cells may prevent individual cells from passing through the filters even at high pore size. Interestingly, while bacterial cell density as estimated by qPCR fluorometry indicates a relatively consistent distribution of bacteria across each filtration membrane level, the median phylogenetic diversity and number of observed OTUs decrease with each subsequent filter size, only to increase slightly in the flow through (Figure 3.1), demonstrating the impact of filtration on bacterial alpha diversity.

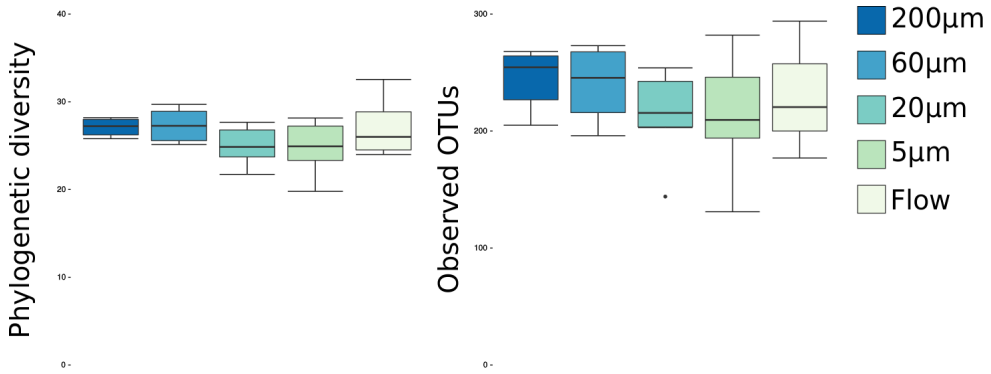

Figure 3.1: **Measures of alpha diversity decrease with decreasing pore sizes.** Alpha diversity as measured by the observed phylogenetic diversity and number of unique OTUs in each filter level among all samples. The highest phylogenetic diversity and number of OTUs is detected at the 200 µm and 60 µm filtration level, which sharply decreases at the 20 µm and 5 µm level, only to recover slightly in the flow through.

3.4.2 Sample specific taxonomic shifts at small pore sizes. An increased frequency of otherwise non–dominant bacterial phyla including

Cyanobacteria and *Verrucomicrobia* are detected within certain samples at small pore sizes with the highest increase evident in the flow through (Figure 3.2a). For example, in sample SM02 the proportion of *Cyanobacteria* recovered from the unfiltered sample is 3.7% of the total bacterial community. This proportion of *Cyanobacteria* increases nearly ten fold to 34.4% in the flow through of the same sample (Figure 3.2b). Similarly, the phylum *Verrucomicrobia* increases in the unfiltered to flow through of sample NO15 and SM29 from 29.3% to 44.7% and 0.5% to 9.5% respectively.

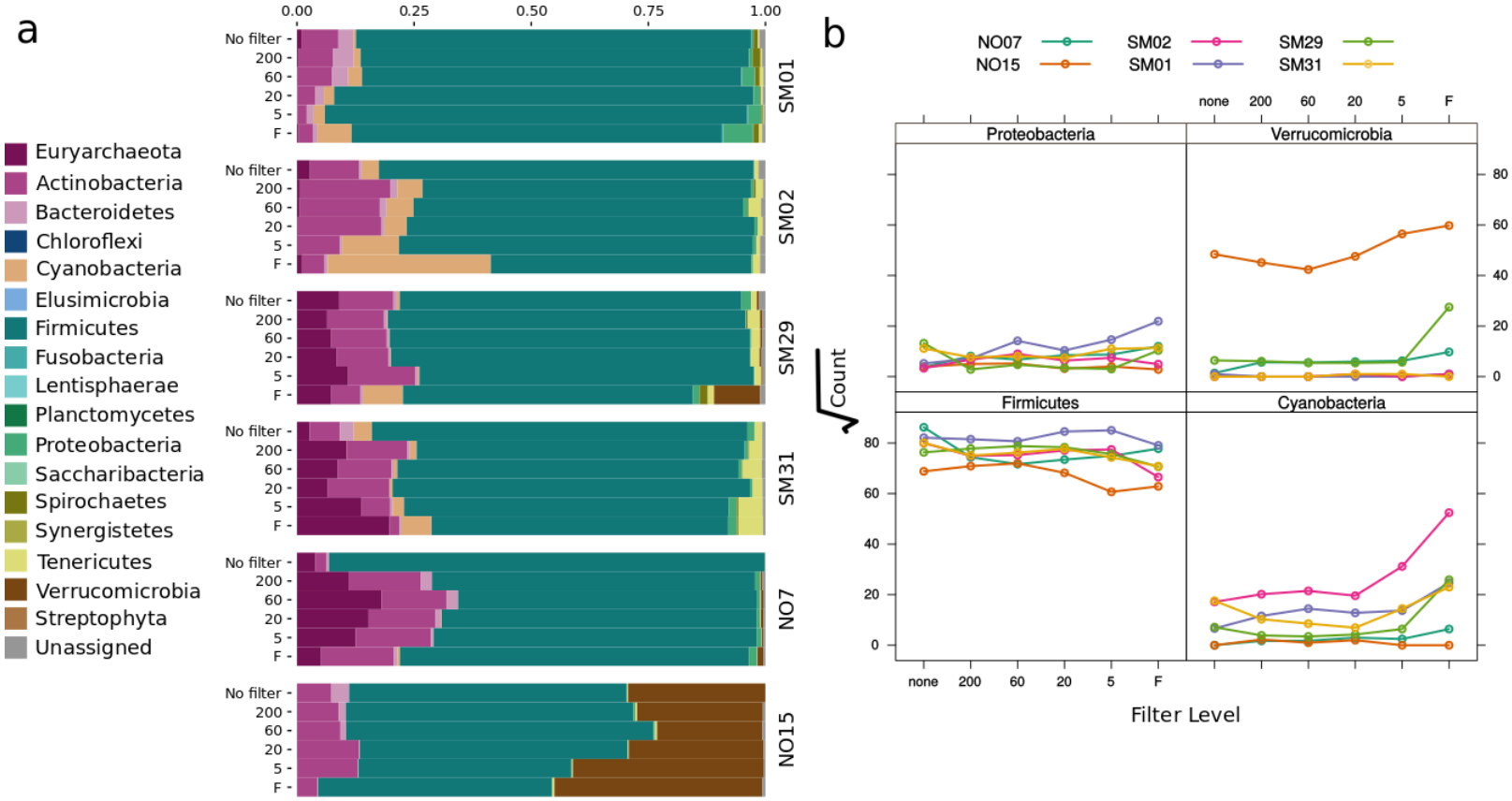

Figure 3.2: **Observed phylum–level taxonomic shifts related to filtration level.** (a) Sample–specific phylum–level taxonomic shifts in unfiltered, 200 μm, 60 μm, 20 μm, 5 μm filtered, and flow through samples. (b) Changes in proportion of selected phyla in all samples over each filtration level. While the proportion of most phyla is relatively consistent across larger pore size filters, frequencies of select phyla are increased in the smaller size filter and flow through.

Shifts in the frequency of particular OTUs was calculated among all OTUs detected in at least two samples at a minimum proportion of 0.5% of the entire sample the results of which can be found in Figure 3.3. The count of each filtration level is normalized by the maximum observed count in the within–sample series of filter levels so that $Y_{norm} = \frac{Y - X_{min}}{X_{max} - X_{min}}$, resulting in a normalized score from one to zero. Results of this analysis support the increased proportion of OTUs assigned to the *Cyanobacteria* phylum among the Peruvian samples with more than one OTU driving this pattern, suggesting it may be due to an intrinsic characteristic of these bacteria (e.g., small cell size). Conversely, other OTUs, particularly in the *Firmicutes* phylum, are found at higher frequencies in the non–filtered samples and decrease in the filtered samples. The decreased frequency of this and other dominant phyla in the filtered samples likely allows for the increased sensitivity in the ability to detect non–dominant taxa.

3.4.3 Source of *Cyanobacteria* in fecal samples. OTUs assigned to the *Cyanobacteria* phylum were assigned to various levels of taxonomic resolution to the under characterized bacterial order *Vampirovibrio* which includes a variety of bacteria isolated from environmental sources as well as the mammalian gut (Figure 3.4, Figure 3.5). To determine whether potential *Vampirovibrio* OTUs detected in this study are environmental contaminants, or potentially incorporated into the gut through consumption of water, all *Vampirovibrio* sequences in the EzBioCloud database [211] and OTUs defined in this study were clustered at 97% sequence similarity and built into a Maxi-

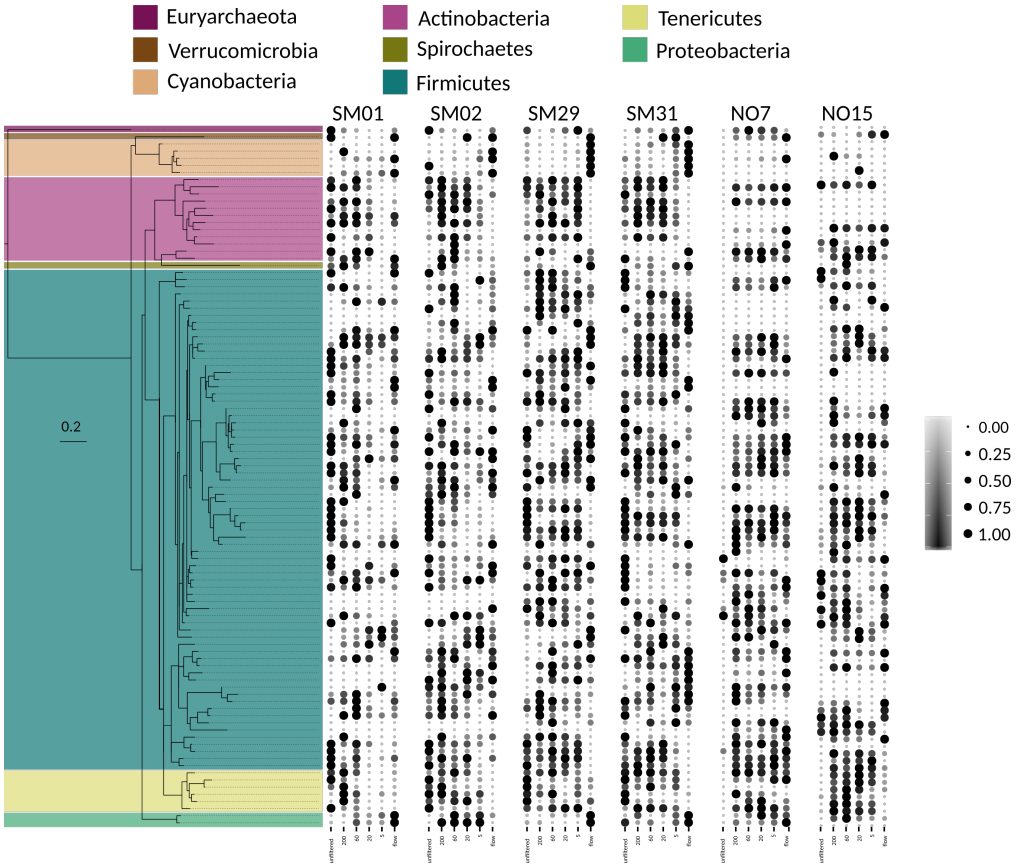

Figure 3.3: **Bubble chart of OTU–level taxonomic shifts in each sample at different filter levels.** Phylogenetic tree of high–frequency OTUs present across samples colored by phylum with corresponding change in frequency in each sample. Bubble size and intensity indicates the filtration level in which the highest and lowest proportion of the OTU is found.

mum Likelihood tree using RAxML [178]. Additionally, a pair–wise distance matrix was generated to detect clusters of similar 16S rRNA V4 sequences among taxa classified as Vampirovibrio (Figure 3.4). Of the 12 OTUs defined as *Vampirovibrio* generated from this study, four are more than 97% different in terms of sequence identity than any other *Vampirovibrio* in the EzBioCloud database [211]. Of those that cluster with the published *Vampirovibrio* OTUs,

isolation sources of clustered published sequences include the human gut microbiome but also other mammals sources including pig, snub–nose monkey, bighorn sheep, sheep, hamadryas baboon, rat, and flying fox feces as well as bovine rumen samples (NCBI Accession: HF996393, AB506276, GU303703, GQ451200, EU474510, AB494937, GQ451255, FJ879341, EU474538, EU464248, FR888536, HQ716357, EU466334, EU469690).

3.5 Discussion and Conclusions

3.5.1 Limitations to fecal filtering methods. Filtration methods described in this study are markedly different from those described using filters to select different bacterial cell sizes in environmental sources [149, 150, 193] in that the high viscosity, presence of undigested fibrous material, and sample volume limitations of fecal samples necessitates a higher filtration pore size than those that may have a more precise separation of bacterial cells at sub–micron levels. Despite this, preliminary testing of this filtration method using a pure *Escherichia coli* culture illustrates that despite the relatively large pore size of even the highest filtration levels, individual or clusters of bacterial cells are easily trapped on the filter itself (Supplementary Table B.1). Because of this, future filtration methods may require additional filtration steps at lower pore sizes as well as a more efficient washing of the filter membrane to enhance the specificity of the filtration process. However, while the presence of bacterial cells at each filtration level is expected based on our positive control, results from this study indicate nevertheless that at small pore sizes non–dominant taxa may become enriched, allowing for the targeted enrich-

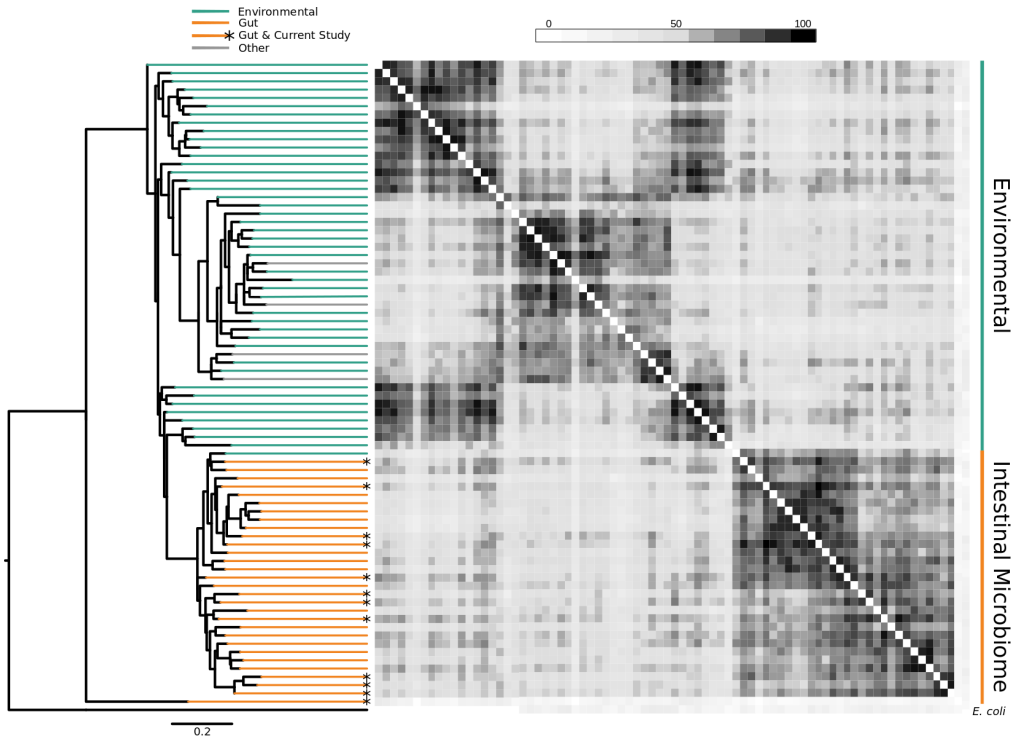

Figure 3.4: **Position of OTUs assigned to the *Vampirovibrio* order in this study relative to other published *Vampirovibrio* sequences.** Maximum likelihood tree of 97% identity clustered OTUs generated from all *Vampirovibrio* reference sequences found in the EzBioCloud database [211] as well as OTUs from the current study designated as *Vampirovibrio*. Lines from tree tips for each OTU are colored to represent "environmental", "gut", and "other" isolation sources for each OTU as defined by the NCBI genbank entry. Heatmap represents the pairwise similarity of OTUs to each other. All OTUs in this study cluster with sequences isolated from mammalian gut sources.

ment of small cell sized bacterial organisms.

3.5.2 Enrichment of non–photosynthetic *Cyanobacteria*: Implications for future research.

The phylum *Cyanobacteria* is arguably the most important bacterial groups in the genesis of multicellular life on earth. Early photosynthetic *Cyanobacteria* are thought to be major contributors to the fixation

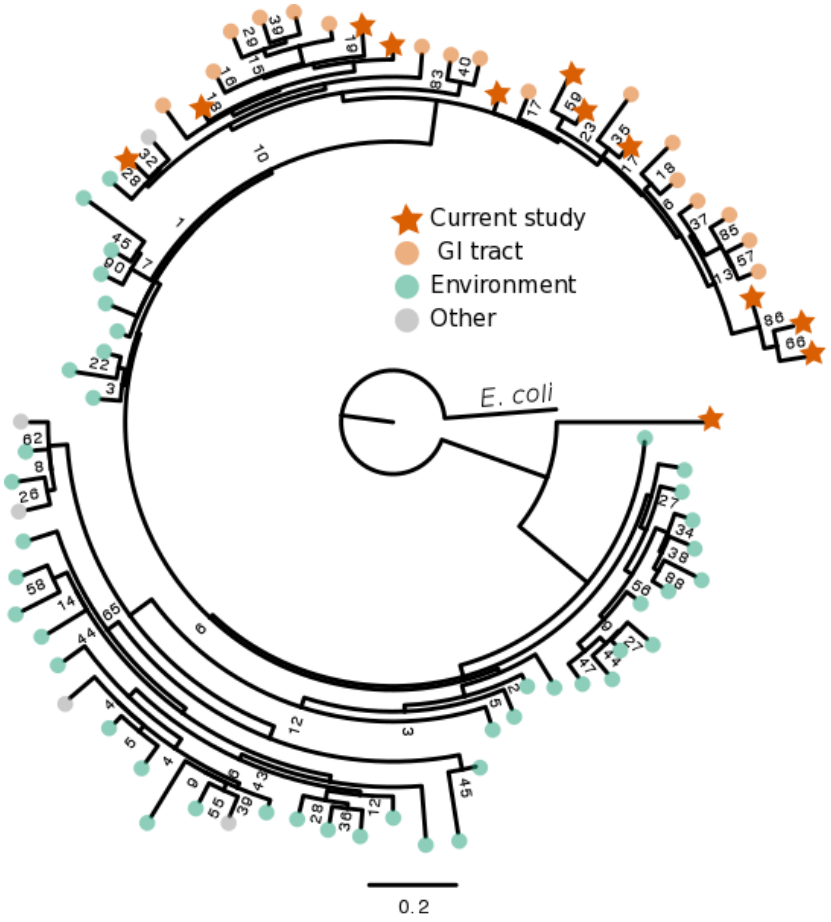

Figure 3.5: **Maximum likelihood tree of environmental and gastrointestinal *Vampirovibrio*.** Alternative maximum likelihood tree of 97% identity clustered OTUs generated from all *Vampirovibrio* reference sequences found in the EzBioCloud database [211] as well as OTUs from the current study designated as *Vampirovibrio*. While bootstrap values across the tree are low, clustering of OTUs by isolation source indicates the phylogeny represents two distinct groups.

of oxygen in the Earth's atmosphere [23] starting approximately 3.6 billion years ago [28] eventually leading to what is known as the "Great Oxidation Event" [177], the vanguard of multicellular life on Earth. Additionally, an early endosymbiotic *Cyanobacteria* eventually lost the ability to live outside of its

host cell and became the chloroplast organelle in plants [157], transmitting photosynthesis to multicellular life. Despite the significance of this microbial phylum, little is known about the early phylogenetic history of *Cyanobacterium* [45]. Recently, however, a new group of bacteria, Melainabacteria has been described as either a sister phylum of *Cyanobacterium* [45] or class within the *Cyanobacterium* phylum [175] which may shed some light into the early evolutionary history of these organisms. Members of *Melainabacteria* are non–photosynthetic and appear to be prevalent in both the environment and gastrointestinal tract of a variety of vertebrate taxa [45, 176, 175]. A valuable finding from these filtration experiments is the enrichment of potential members of this under–described bacterial group in filtered fecal samples from the Peruvian, but not Norman, individuals in this study.

Twelve distinct OTUs assigned to the order *Vampirovibrio* from this study are potentially uncultured members of the *Melainabacterium* group. The earliest members of *Vampirovibrio* were isolated from natural aquatic environments [72, 55]. An interesting aspect of these environmental *Vampirovibrio* are their predatory nature. Small, Gram–stain negative, and motile, *Vampirovibrio* cells seek out and attach to their prey via a pad of spikes that penetrate the prey cell membrane forming a T–type conjugation formation through to the prey cell cytoplasm [176]. The *Vampirovibrio* group obtained its name by the subsequent vampire–like resource acquisition by which the predator cell ingests the prey's cell contents through protease and other enzymatic activity [176]. *Melainabacterium* isolates from the gastrointestinal tract of animals are generally thought to be obligately fermentative organisms [45]

based on the partial reconstruction of five *Melainabacterium* genomes [45], yet still much is needed to clarify the phylogenetic relationship and functional potential of these organisms in environmental and host–associated microbial ecosystems. Importantly for the purposes of this study, currently described members of *Melainabacterium* meet requirements for potentially masked taxa in mixed microbial ecosystems which may be better clarified by serial filtration as they (1) are small in size as compared to other free–living bacteria [55] and, (2) have a highly reduced genome size, especially those found in the gut [45].

3.5.3 Conclusion. This study demonstrates the utility of serial filtration of fecal samples for the enrichment of otherwise undercharacterized taxa in metagenomic or metabarcoding study of the gut microbiome. As many bacterial organisms that are potentially important residents of these communities are under characterized, serial fecal filtration and 16S rRNA metabarcoding may be a cost effective screening technique to identify samples for potential whole genome metagenomic sequencing of non–dominant taxa for which whole genome assembly and functional annotation would advance understanding of the role of these taxa in microbiome habitats. In addition, these screening techniques may identify samples for which culturing of these microbes may be accomplished. The application of novel experimental techniques and the inclusion of non–industrial populations for the genomic characterization of bacterial members of the gut microbiome is one avenue to attempt to clarify the genomic “dark matter” that predominates in metagenomic

studies of the human gastrointestinal tract.

Chapter 4

Differential preservation of endogenous human and microbial DNA in dental calculus and dentin

4.1 Abstract

Dental calculus (calcified dental plaque) is prevalent in archaeological skeletal collections, and a rich source of oral microbiome and host-derived ancient biomolecules. Recently, it has been proposed that dental calculus may provide a more robust environment for DNA preservation than other skeletal remains, but this has not been systematically tested. In this study, shotgun-sequenced data from paired dental calculus and dentin samples from 48 globally distributed individuals are compared using a metagenomic approach. Overall, we find that dental calculus is a consistently richer and less contaminated source of ancient DNA than dentin. The majority of DNA in dental calculus is microbial and originates from the oral microbiome; however, a small but consistent proportion of DNA ($\bar{x}$0.08 $\pm$ 0.08%, range 0.007–0.47%) derives from the host genome. Host DNA content within dentin is variable ($\bar{x}$13.70 $\pm$ 18.62%, range 0.003–70.14%), and for a subset of dentin samples (15.21%), oral bacteria contribute $>$20% of total DNA. Human DNA in

dental calculus is highly fragmented, and is consistently shorter than both microbial DNA in dental calculus and human DNA in paired dentin samples. Finally, we find that microbial DNA fragmentation patterns are associated with guanine-cytosine (GC), content, but not aspects of cellular structure.

4.2 Introduction

Dental calculus is a mineralized form of dental plaque [204], a sequentially generated microbial biofilm [120] that entraps microbial, dietary, host, and ambient debris during spontaneous calcification events [196]. Unlike body mucosal surfaces that have continual cell turnover, teeth do not remodel. Consequently, they act as relatively stable environments for bacterial colonization during biofilm development [121], making the formation of dental calculus difficult to prevent without mechanical removal. As such, dental calculus is prevalent in the archaeological record, and due to its excellent morphological preservation, it has long been an attractive target for microscopic analysis [13, 60, 79, 49, 97]. More recently, dental calculus has been explored as a source of ancient DNA (aDNA), and it has been shown to retain an excellent record of the human oral microbiome [2, 194], as well as serve as an alternative source of endogenous host DNA [135].

Retrieving serviceable aDNA from archaeological sources, whether from skeletal tissues (bone and dentin) or from microbiome remains (dental calculus and paleofeces), poses several challenges because after death both time and environmental factors begin to compromise the molecular stability of DNA. These processes include oxidative and hydrolytic damage to in-

dividual bases, hydrolytic lesions on the sugar–phosphate backbone, DNA fragmentation due to nuclease activity, and general degradation by microorganisms involved in the decomposition process [136, 39]. As a result, aDNA accumulates predictable forms of damage characterized by DNA loss, extreme DNA fragmentation, depurination, and high–levels of terminal cytosine deamination [81, 162, 39]. In addition to damage, ancient samples can also acquire exogenous contamination that may obscure any remaining endogenous signal. Susceptibility to contamination appears to be tissue specific, with petrous bone and tooth dentin generally exhibiting the highest proportions of endogenous human DNA among archaeological skeletal and soft tissues [62, 148, 77]. Recent studies have suggested that dental calculus may be more resistant to environmental contamination than other sources of aDNA, and higher overall DNA yields have been reported from dental calculus than from any other archaeological source of aDNA [194, 135]. While these patterns are compelling, they have been reported from a small number of samples and are not controlled for variables such as temporal age, depositional context, or geographic location [194, 214, 135]. As such, the prospect of ancient dental calculus as a dependable source of well–preserved, endogenous aDNA has not yet been systematically tested.

In the present study, metagenomic sequencing data from paired dentin and dental calculus samples from 48 individuals are compared to test whether endogenous DNA exhibits a different degree of preservation in dental calculus than in other skeletal tissues from the same individual. Individuals included in this study represent seven archaeological sites spanning diverse

geographic, environmental, and temporal ranges, providing control over individual and group dynamics. Thirty–six paired samples were selected from a single medieval cemetery in Kilteasheen, Ireland in order to examine intra–site variation in preservation. Additionally, 12 paired samples from six different archaeological sites on three continents spanning a broad temporal range were selected to control for environment, burial context, time period, and individual dynamics that may impact preservation quality (Figure 4.1). We compare DNA preservation within archaeological dental calculus and dentin from the same tooth, with a specific focus on four main measures: (1) DNA abundance, (2) microbial community composition and contamination, (3) human DNA content, and (4) DNA fragmentation and damage patterns. Our findings confirm that calculus is a richer source of total DNA when compared to dentin, with a low, albeit consistent, proportion of endogenous human DNA. Microbial profiles of dental calculus suggest that it retains a robust signal of the human oral microbiome and is relatively resistant to exogenous contamination, while dentin is typically dominated by environmental microbial sources. A subset of dentin, however, contains DNA from oral bacteria (possibly deriving from postmortem colonization during decomposition processes), indicating that dentin may serve as an alternative source for DNA from individual oral microbes in some cases. With respect to DNA degradation, we find that DNA fragmentation patterns within dental calculus are associated with the genomic source of the DNA (human vs. microbial) but not with cellular structure (e.g., microbial cell wall type or presence of a surface–layer). Additionally, human DNA is consistently shorter in dental calculus than in paired dentin samples,

which may reflect differences in the manner by which human DNA is incorporated into each tissue. Finally, we observe a systematic loss of short AT–rich DNA fragments that is particularly marked in bacteria with low to medium GC content genomes.

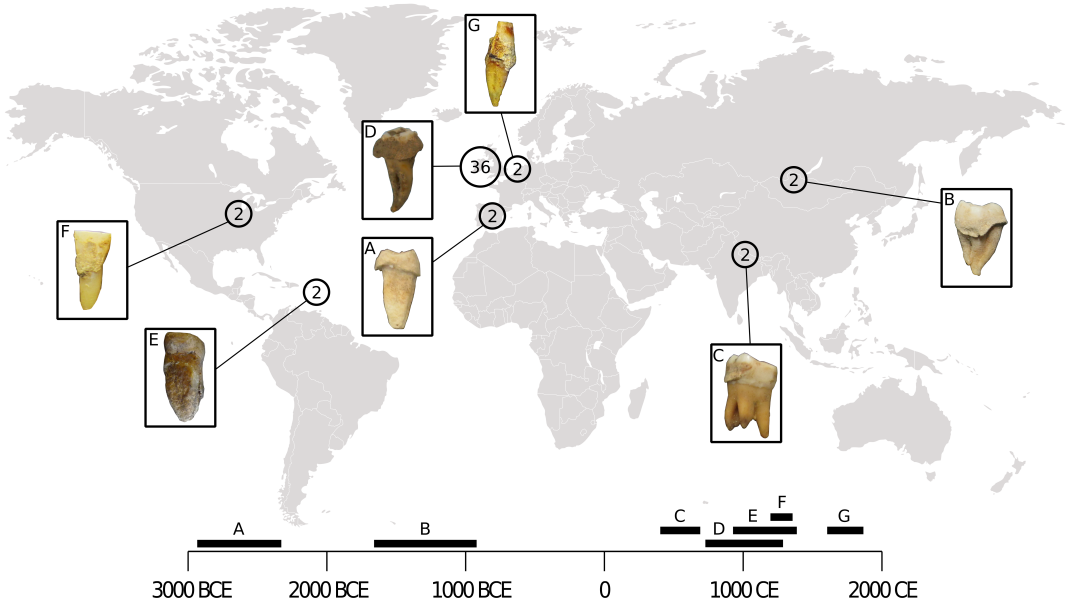

Figure 4.1: **Geographic locations and temporal periods of archaeological teeth included in this study.** (a) Camino del Molino, Spain; (b) Khovsgol, Mongolia; (c) Samdzong, Nepal; (d) Kilteasheen, Ireland; (e) Anse à la Gourde, Guadeloupe; (f) Norris Farms, Illinois, USA; (g) Middenbeemster, the Netherlands. For each site, representative teeth with dental calculus deposits are shown in boxes. The number of teeth (dentin–calculus pairs) analyzed per site is provided within the indicated circles, and corresponding letters on the time line indicate the time period represented by each site.

4.3 Methods

4.3.1 Samples. Paired dental calculus and dentin samples were obtained from seven geographically and temporally distinct sites: the 19th century site

of Middenbeemster in the Netherlands (n=2, 1611–1866 CE), the Copper Age site of Camino del Molino in Spain (n=2, 2340–2920 BCE), the Bronze Age site of Samdzong in Nepal (n=2, 400–650 CE), the Iron Age site of Hovsgol in Mongolia (n=2, 930–1650 BCE), the Late Ceramic Age site Anse à La Gourde in Guadeloupe (n=2, 975-1395 CE), the Mississippian site of Norris Farms, IL, USA (n=2, 1300 CE), and the Medieval site of Kilteasheen in Ireland (n=36, 600-1300 CE). The first six sites were selected to represent global patterns of DNA preservation across diverse environments, burial contexts, and time periods. Remains and data from this site are referred to as the global dataset. A more extensive sampling of a single site in Kilteasheen, Ireland was performed to account for regional DNA preservation between individuals with similar burial contexts across a time transect of approximately six centuries. Remains and data from this site are referred to as the regional dataset.

4.3.2 DNA Extraction.

Global Sample Set

Samples were prepared for sequencing in a dedicated ancient DNA laboratory at the Laboratories of Molecular Anthropology and Microbiome Research (LMAMR) in Norman, Oklahoma, USA. Teeth were first decontaminated with bleach, then the calculus was separated using a dental scaler. The crown was separated from the root using a Dremel rotary tool. Partitioned tooth roots and calculus were further decontaminated via exposure to

UV irradiation. DNA extraction was performed as described by Warinner et al. (2014). Approximately 10–20 mg of dental calculus and 100 mg of dentin were crushed and then immersed in 1 ml of 0.5 EDTA for 15 minutes to remove any additional surface contaminates. Dental calculus samples were demineralized in a solution of 0.45 M EDTA and 10% proteinase K (Qiagen, the Netherlands) at 55° C for 8-12 hours. Dentin samples were demineralized at room temperature. After 2 days, the EDTA supernatant was removed and refreshed with new EDTA and 50 μl of protinase K (Qiagen, the Netherlands). Dentin samples were then left to demineralize for an additional 3 days at room temperature. Prior to demineralization, all samples were centrifuged and the supernatant was used for DNA extraction using a phenol-chloroform-isoamy alcohol (25:24:1) along with three blanks. Extracted DNA was isolated using silica purification and quantified using a Qubit fluorometer.

Regional Sample Set

Samples were prepared and extracted in the paleogenetics clean room at the Institute for Archaeological Sciences, University of Tubingen (INA). The surface of the dedicated sampling hood was cleaned with HPLC water and UV irradiated by an internal light source between uses. Any calculus was removed from the surfaces of the teeth using dental scalers, which were rinsed with bleach and HPLC water, then UV irradiated for 10 minutes between uses. Large calculus samples were pulverized with a tube pestle. Teeth were then sectioned horizontally at the cementoenamel junction and dentin was drilled from the pulp chamber using a dental drill. For calculus

samples weighing over 20 mg, half the pulverized material was carried over for extraction. For dentin samples over 70 mg, aliquots of approximately 50 mg were taken for extraction. Dentin and calculus samples were extracted using a modified silica-based method according to Dabney et al. (2013). Samples were submerged in a digestion buffer with final concentrations of 0.45 M EDTA and 0.25 mg/mL proteinase K and rotated overnight at 37° C. After incubation, samples were centrifuged and the supernatant was purified using a 5 M guanidine-hydrochloride binding buffer with High Pure Viral Nucleic Acid Large Volume kits (Roche). The extracts were eluted in 100 μl of a 10 mM tris-hydrochloride, 1 mM EDTA (pH 8.0), and 0.005% tween-20 buffer (TET). One extraction blank was prepared for every ten samples, and one positive control of cave bear bone powder was processed alongside each extraction batch to ensure efficiency. The extracts were quantified using a Qubit fluorometer.

4.3.3 Illumina Library Preparation.

Global Sample Set

Approximately 100 ng of DNA was used for each Illumina shotgun library at the LMAMR, Norman, Oklahoma using NEBNext DNA Library Prep Master Set (E6076) and blunt–end modified Illumina adapters. Manufacturers instructions were followed with the exception of Nebulization. Blunt–end repair was performed using 50 μl reactions with 30 μl of DNA extract for each sample which was then incubated for 20 min at 12° C and 15 min at 37° C and

purified using Qiagen MinElute silica spin columns following the manufacturers instructions. All samples were eluted in 30 μl. Prior to end–repair, Illumina adapters were ligated in 50 μl reactions. Reactions were incubated for 15 min at 20° C and purified using Qiagen QiaQuick columns before elution in 30 μl EB. Samples were then incubated for 20 min at 37° C followed by 20 min at 80° C in a final volume of 50 μl for adapter fill–in. Libraries were amplified and dual-indexed in a 50 μl PCR reaction using 15 μl template, 25 μl of a 2x KAPA U+ master mix, 5.5 μl H2O, 1.5 μl DMSO, 1 μl BSA (20 mg/ml), and 1 μl of each forward and reverse index (10 μl μM). Thermocycling conditions were 5 min at 98° C followed by 10–12 cycles of 15 seconds at 98° C, 20 seconds at 60° C, and 20 seconds at 72° C, followed by a final elongation step for 1 minute at 72° C. Amplified libraries were then purified using Agencourt AMPure XP beads and eluted in 30 μl EB. Samples were sent for sequencing on an Illumina HiSeq 2500 using a paired-end, 2 x 100 bp, rapid-run chemistry.

Regional Sample Set

Double-stranded Illumina libraries were generated using 10 μl of extract for each sample according to an established protocol [124]. Purification of the blunt-end repair and adapter ligation steps was performed using Qiagen MinElute columns. After the adapter fill-in step, the Bst polymerase was deactivated with a 20 minute incubation at 80° C. A single library blank was used for every ten samples. The libraries were then quantified using real–time quantitative PCR (qPCR, Lightcycler 480 Roche). Each library was assigned a unique pair of indices, added to the library over 2-15 indexing reactions

per library. Libraries were indexed in 100 μl reactions using varied amounts of template and H2O based on library richness, 10 μl PfuTurbo buffer, 1 μl PfuTurbo (Agilent Technologies), 1 μl dNTP mix (25 mM), 1.5 μl BSA (10 mg/ml), and 2 μl of each indexing primer (10 μM). The reactions were purified, pooled, and eluted over MinElute columns in 50 μl TET. Efficiency of the indexing reactions was evaluated using a qPCR assay. Approximately one–third of each indexed library was amplified using 3–5 μl of template in 70 μl reactions with Herculase II Fusion DNA Polymerase (Agilent Technologies). Products for each sample were pooled and quantified using an Agilent Tape Station D1000 Screen Tape kit. Amplified sample and blank libraries were pooled into two 10 nM pools for shotgun sequencing. Samples and blanks were sequenced separately on Illumina NextSeq 500 using single–end, 75–cycle, high-output kits. Samples were sequenced to a depth of approximately 5 million reads per library, and blanks were sequenced to a depth of 100,000 to 300,000 reads. Additionally, thirteen individuals of the regional sample set were re–sequenced using paired–end 150–cycle chemistry on an Illumina NextSeq 500 so they could be included in fragment length analyses.

4.3.4 Computational Methods. Sequencing data from the global and regional sample sets were computationally processed identically. Adapters were removed, paired–end data merged, and reads quality filtered using Clip & Merge [146] with a minimum base quality of 20 and a minimum fragment length of 30. Processed reads were then taxonomically binned using MALT (version 038) [185] and the NCBI full nucleotide database

with an 85% identity threshold. Metagenomic profiles were analyzed with MEGAN (community edition, v6.11.2) [87] and screened for specific taxonomic levels for fragment length and damage pattern profiles using a RMA file format parsing script [84]. Before downstream analysis, all reads were normalized across samples to the lowest number of reads in the full sample set using the default parameters in MEGAN using the option `Use Normalized Counts` ignoring any reads that could not be assigned to a taxonomic node. Mapping to the human genome (hg19) was performed using BWA [110] as implemented in EAGER [146] with a mapping quality score of 30. A species-level taxon table was exported from the bacterial and archaeal sub-trees in MEGAN and used to generate a Bray-Curtis taxonomic distance matrix in R using the vegan library (version 2.4–1) [133]. Principal coordinates were generated using the R ape library [143] and visualized as a PCoA plot using ggplot2 [205]. Potential source contribution to samples were calculated from genus level bacterial and archaeal taxonomic frequency tables using SourceTracker version 0.9.8 [98]. All source datasets were computationally processed in a manner identical to the dental calculus and dentin samples in this study. Source accession numbers: ERR1017187, ERR1019366, ERR1022687, ERR1022692, ERR1034454, ERR1035437, ERR1035438, ERR1035441, ERR1039457, ERR1039458, ERR1043165, ERR1044071, ERR1044072, ERR1051325, SRR1631060, SRR1631061, SRR1631063, SRR1631064, SRR1633008, SRR3184100, SRR3184876, SRR3189411, SRR3189416, SRR3189418, SRS014107, SRS015650, SRS018665, SRS018975, SRS019029, SRS019129, SRS019387,

SRS023538, SRS063215, SRS077312.

Additionally, a nested classification scheme designed for this study, adapted from [160], was used to classify species level proportions of environmental sources for a subset of the samples. This scheme divides bacterial and archaeal species into likely isolation source according to a species–by–species literature survey on PubMed. Pathobionts and opportunistic pathogens are designated as such when literature on the organism consistently presented it as a health threat, though it also may be a natural inhabitant of the human microbiome or soil. Assigned read counts for the classified species were tabulated and visualized using the Krona Excel Template [134]. This approach was used in conjunction with SourceTracker to control for potential biases related to the modern samples used in the latter and together present a layered representation of the types of microorganisms typically found in ancient dental calculus and dentin. Fragment lengths and damage pattern distributions were generated from MALT [185] results using a RMA file format parsing script [84] and visualized using the ggplot2 [205] and lattice [161] libraries in R. For all fragment length and damage rate analyses only those samples that were paired end sequenced were used. Additionally, only merged reads were considered to prevent artifacts brought on by unmerged pairs, the maximum length of which is fixed. Finally, GC content versus length profiles were generated using a custom python script (See Appendix C). When possible, analyses were run in parallel using GNU parallel [181].

4.4 Results

4.4.1 DNA abundance. The total amount of DNA recovered from archaeological dental calculus is substantially higher than from dentin as measured by both fluorometric quantitation (Qubit) (Figure 4.2a) and quantitative PCR (qPCR) (Figure 4.2b). Immediately following extraction, DNA yields from dental calculus as measured by fluorometry ranged from 15.4 ng/mg to 214.4 ng/mg ($\bar{x}$77.0 ng/mg $\pm$ 51.6), while dentin samples from the same individual yielded 0.2 ng/mg to 14.3 ng/mg ($\bar{x}$6.4 ng/mg $\pm$ 6.0). These measurements are moderately correlated (Pearsons coefficient: 0.59; 95% CI: 0.43, 0.72) to the average DNA copy number estimated after library construction using qPCR. This fit accounts for 34% of the total variance using a linear regression model (Figure 4.2b). Differences between the two methods may be related to the fact that DNA library construction involves several silica-based purification steps that are known to result in substantial, but stochastic, DNA loss [96]. While DNA abundance differences are not as stark in the qPCR results, the overall amount of DNA recovered from dental calculus is consistently higher than from dentin using both metrics.

4.4.2 Microbial community composition and contamination. Overall, archaeological dental calculus and dentin contain microbial DNA from two distinct communities. To explore these differences, shotgun-sequenced reads from the 48 dentin and dental calculus pairs were taxonomically binned using MALT [185], and the resulting assignments were analyzed using MEGAN [87]. A species-level PCoA plot based on a Bray-Curtis dissimilar-

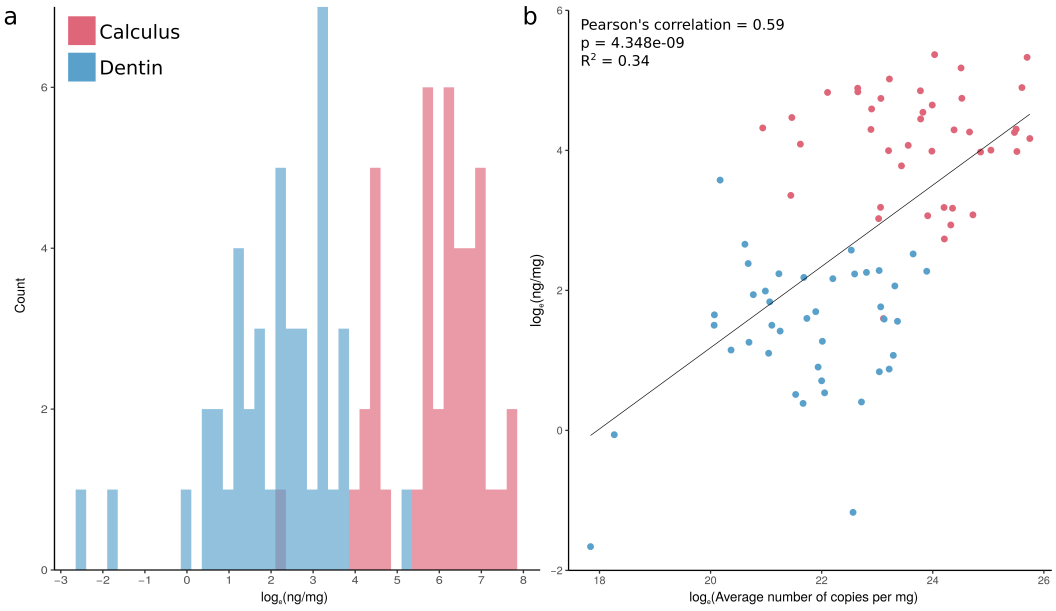

Figure 4.2: **Total DNA content of dental calculus is higher than dentin as measured by both fluorescence and quantitative PCR (qPCR) techniques.** (a) Normalized DNA yield (log transformed nanograms DNA per milligram starting material) of DNA extracts obtained from dental calculus and dentin as measured by a Qubit fluorometer using a High Sensitivity Assay (p=3.911e-11, Wilcox signed-rank paired test). (b) Linear correlation between normalized DNA yield (log(ng/mg)) and normalized DNA yield as measured by qPCR after library preparation. The average number of copies per milligram was calculated from the mean of four technical replicates. In both quantification metrics dental calculus has a higher DNA yield than dentin.

ity matrix demonstrates that dental calculus samples form a relatively tight and cohesive group that is distinct from the more diffuse distribution of microbial communities identified within dentin (Figure 4.3a). Importantly, microbial communities from each material (dental calculus or dentin) are less similar to their paired sample than they are to samples of the same material. This pattern is consistent with expectations that the microbial taxa within dental calculus represent a relatively well preserved biological community derived

from dental plaque, while dentin—being sterile in life—is expected to harbor a microbial community entirely composed of exogenous contaminant bacteria acquired through stochastic postmortem processes.

Two additional approaches were used to further characterize the nature and inferred sources of the microbial communities present within sample—types. First, microbial species—level identifications were categorized according to a nested scheme reflecting organism membership in one or more of the following source categories: environmental, uncultured environmental, human microbiome, human oral, pathobiont, and opportunistic pathogen (Figure 4.3b), which were then visualized using the Krona Excel template [134]. Category membership was determined by species presence or absence in the Human Oral Microbiome Database [31] as well as source and habitat descriptions in the 20 most recent articles in PubMed using the species name as the search keyword. Using this analysis, stark differences are observed in the inferred microbial source contributions to archaeological dental calculus and dentin, whereby dental calculus is strongly dominated by human-associated—and especially oral–associated—taxa, while dentin is primarily composed of environmental taxa.

As a separate confirmation method, SourceTracker, a Bayesian source-prediction tool [98], was used to estimate proportions of source similarity at the genus level in dental calculus and dentin based on a set of modern reference microbial communities sequenced from human dental plaque [36], human hand swabs [132], and top soil [91] (Figure 4.3c). While there are some shared similarities between dental plaque and skin, skin is typically

dominated by members of the genera *Propionibacterium* and *Staphylococcus* while the oral cavity is predominantly colonized by members of the genera *Streptococcus*. The bacterial community composition of soil is variable [57] but is typically dominated by members of the phyla *Acidobacteria* and *Proteobacteria*, distinguishing it from the microbial community found in human associated ecosystems. In agreement with the other methods presented here, archaeological dental calculus is estimated to be composed primarily of dental plaque-associated taxa, while dentin is dominated by genera associated with human skin and environmental sources (Appendix Table C.5). The highest predicted contributions of skin and soil to dental calculus samples are 7.9% and 14.0%, respectively, while the same predicted contributions for dentin are 33.5% and 89.0%, respectively. Together, these analyses suggest dental calculus is relatively robust to environmental contamination when compared to dentin.

Although most dentin samples are strongly dominated by environmental taxa, two dentin samples from the site of Norris Farms—NF47 and NF217—cluster with dental calculus samples in the PCoA (Figure 4.3a) and are estimated via SourceTracker analysis [98] to contain microbial DNA that is 56.2% and 78.4% derived from dental plaque, respectively (Appendix C.1). Unlike other teeth in this study, the Norris Farms teeth were obtained in a fragmented state. Lacking the full intact teeth, the presence of carious lesions could not be ruled out. For this reason, the Norris Farms samples were excluded from further downstream analyses. In samples for which the tooth was intact, tooth dentin is generally strongly dominated by environmental taxa; however, a sub-

set of dentin samples exhibit a slight signal of the human oral microbiome, ranging from 0.0% to 40.0%, with 7 of 46 dentin samples having a predicted oral source contribution of 20.0% or more by SourceTracker analysis.

4.4.3 Human DNA content. Although typically higher than in dental calculus, the proportion of human endogenous DNA in dentin varies substantially, ranging more than 4 orders of magnitude in this study, from 0.003% to 70.1% of all reads (Figure 4.4a). By contrast, the proportion of human DNA in dental calculus is relatively low, but consistent across all samples, differing by less than 2 orders of magnitude, from 0.007% to 0.4%. To verify these reads as authentic host DNA, and to mitigate the possibility that they represent spurious mapping to the human genome or modern contamination, a secondary verification procedure was performed whereby only those reads that met stringent mapping criteria, were assigned to the Homo sapiens node in lowest common ancestor assignment by MALT, and displayed typical ancient DNA damage profiles were included in a high confidence human dataset (see Appendix). The number of human-assigned reads following strict mapping decreased across all samples but was more severe among samples from the Kilteasheen (regional) dataset. As the strict mapping parameters used allow only one mismatch per 50 base pairs, this comparatively high loss of reads in the regional dataset likely results from the fact that these samples were sequenced using a single–end, 75 cycle sequencing strategy, rather than the paired–end 2 x 100 cycle sequencing strategy employed for the global sample set.

Of all originally designated human reads in the regional dataset, between 20.7% to 60.9% of the dental calculus reads and 0.6% to 67.1% of dentin reads pass the initial strict mapping step. Within the global dataset, between 60.1% to 88.3% of dental calculus and 74.0% to 97.9% of dentin reads pass. Reads passing strict mapping were next run through MALT with the full NCBI nucleotide database as a reference to ensure proper assignment to *Homo sapiens*. Among all dental calculus samples, the proportion of reads uniquely assigned to the *Homo sapiens* node ranges from 79.6% to 90.3%. Within all dentin samples the proportion of reads assigned to *Homo sapiens* ranges from 81.1% to 92.7%. Finally, to ensure these reads represent authentic ancient host DNA and not modern contamination, rates of terminal cytosine deamination—chemical damage signals expected of ancient DNA—were evaluated using mapDamage 2.0 [92]. While damage rates were lower after these verification steps, all samples except one dentin (KT05) present damage patterns consistent with authentic aDNA both pre– and post–verification. Among dental calculus samples, the median change in terminal damage between pre- and post-verification is a loss of 0.03% and among dentin samples the median change is a loss of 0.04% (Appendix Table C.2). Thus, although a small subset of reads may be erroneously assigned to the human genome, for most of the samples included here the majority of human assigned reads appear to be authentic ancient human DNA and not contamination or misassignments.

Next, DNA fragmentation and damage patterns of human reads in paired dentin and dental calculus samples were compared for a subset of the sam-

ples for which paired-end DNA sequence data were generated, which includes the global sample set (n=10 individuals) and a subset of the regional sample set (n=13 individuals). Overall, the median fragment lengths of total DNA recovered from both dental calculus (56-88 bp, $\bar{x}$ 72.8 bp) and dentin (53–105 bp, $\bar{x}$ 66.0 bp) are short and fall within a size range expected for archaeological samples (Appendix Table C.3). For each dental calculus sample, the median fragment length of human reads was found to be 15.5 $\pm$ 4.2 bp shorter than the overall median fragment length of DNA in each sample, which is primarily microbial in origin (Figure 4.4b). Dentin samples, however, show no pattern with respect to human DNA fragment length compared to overall DNA fragment length. Comparing human DNA in dental calculus and dentin, we find that human DNA within dental calculus is generally more fragmented than human DNA in paired dentin samples, with the median length of calculus–derived fragments being approximately 10.3 bp shorter than that of dentin-derived fragments (Wilcoxon signed–rank test, $p < 0.02$); however, this pattern is largely driven by the long human DNA fragment lengths in dentin, and further work is needed to determine if this is an artifact of sample preparation or a true biological pattern.

The relative degree of terminal cytosine deamination among human reads is variable in dental calculus and dentin pairs from the same individual (Appendix C.5). Interestingly, most dental calculus samples from the global dataset present lower initial terminal damage rates than their dentin pair. This pattern is not, however, observed in the regional dataset, thus making it unclear if this is an artifact of sample preparation or a true biological pattern.

4.4.4 Microbial DNA Fragmentation and Damage Patterns. We next investigated fragmentation and terminal cytosine damage patterns for a selection of oral microbes preserved at high abundance within dental calculus in order to determine the impact of cell wall or genomic structure on aDNA preservation. It has been previously suggested that cell wall composition may influence the preservation of microbial DNA in archaeological dental calculus and dentin [167, 2]. Fifty oral bacteria were selected and categorized into groups based on Gram stain status, the presence or absence of a surface layer (S-layer), and overall genomic GC content (Figure 4.5a and 4.5c). Additionally, a subset of 20 highly abundant bacterial taxa was analyzed to examine species-level patterns of fragmentation (Figure 4.5b) and DNA damage (Figure 4.5d).

We found no indication of a relationship between terminal cytosine damage and microbial genomic source (Figure 4.5c and 4.5d) nor a relationship between fragmentation and cell wall structure. However, we do see a small decline in average DNA fragment length in taxa with higher genomic GC content (Figure 4.5a). This pattern is also reflected among the 20–species subset of oral bacteria chosen for closer analysis, and reads assigned to *Actinomyces radicidentis* , the species with the highest GC content, have the largest displacement from the median fragment length of all selected taxa (Figure 4.5b).

4.4.5 GC Content Shifts. Finally, the relationship between fragment length and mean GC content was examined for five prevalent oral genera

in dental calculus samples and a single prevalent soil genus in dentin samples to further evaluate the influence of genomic structure on microbial DNA survival. Genera were chosen to maximize the range of GC content with two genera each representing low, medium, and high expected genomic GC content (Figure 4.6). Among all published genomes available in the NCBI database, members of the genus *Methanobrevibacter* range in genomic GC content from 24.2% in *M. wolinii* to 32.6% in *M. ruminantium*. The common oral methanogen *M. oralis* is expected to have a GC content of 27.9%. Members of *Fusobacterium* also have low GC content, ranging from 26.0% (F. perfoetens) to 35.1% (*F. necrophorum*). Among the medium GC content genera, the GC content of *Tannerella* ranges from 37.7% (*T. sp. CAG:118*) to 56.5% (*T. sp. oral taxon HOT-286*), and *Porphyromonas* range from 42.7% (*P. gingivicanis*) to 56.3% (*P. bennonis*). Finally, for high GC content genera, *Actinomyces* ranges from 49.6% (*A. coleocanis*) to 73.1% (*A. dentalis*), and *Streptomyces* ranges from 56.4% (*S. sp. WAC00263*) to 71.1% (*S. sp. NBRC110027*). In comparing reads assigned to these genera in our samples, we detect an increase in GC content at shorter read lengths among all chosen genera except for *Streptomyces*, where no shift was observed. Importantly, this shift is greater for genera with moderate and low expected GC content. For example, in *Methanobrevibacter* and *Fusobacterium* the length at which the mean GC content begins to substantially shift (1 z score) is 39 bp and 48 bp, respectively (Appendix C.4). This shift occurs at longer lengths for *Tannerella* and *Porphyromonas* at 59 bp and 58 bp, respectively, while *Actinomyces* does not present a substantial shift until 35 bp. No shift is ob-

served in *Streptomyces*, although this may occur in short fragments that are below the length–filtering threshold (30 bp) used for this dataset.

4.5 Discussion and Conclusions

4.5.1 Dental calculus is a richer source of genetic material than dentin.

In agreement with the findings of previous studies [195, 135, 194], overall DNA recovery from ancient dental calculus was found to be substantially higher than from dentin, and this pattern is consistent through time and across preservation contexts. This higher DNA content of archaeological dental calculus compared to dentin likely reflects biological differences between the two substrates in cellular composition and structure during life, as well as decomposition patterns after death.

Dental calculus is formed from dental plaque, a dense microbial biofilm that has been estimated to contain more than 200 million cells per milligram [174]. Approximately 70% of the dry weight of plaque consists of microbial cells [122], and a large proportion of the biofilm matrix itself is composed of extracellular bacterial DNA, which provides both structural support and protection to its microbial inhabitants [139]. Furthermore, the mineralization process that leads to dental calculus formation involves rapid inter– and intracellular crystal formation by calcium phosphates, including hydroxyapatite, which strongly bind DNA. The result is a dense crystalline structure that is relatively inert and resistant to microbial attack, enzymatic action, and non–acidic chemical alteration [195]. Although the surface of dental calculus remains porous [171], penetration of substances towards the internal layers of

the calculus matrix is restricted [200], which may account for its high DNA preservation qualities. Calculus is not an entirely closed system, however. It has been shown to disproportionately lose soluble small metabolites over time [188], suggesting that the mineralized matrix allows some degree of water movement.

In contrast to dental calculus, dental hard tissues are largely acellular, with live cells in mature teeth being limited to a layer of odontoblasts lining the pulp chamber wall, a sparse distribution of entrapped cementocytes within apical cementum, and a layer of cementoblasts around the periodontal ligament [111, 122, 105]. Most cells within teeth are instead found within the dental pulp and consist of perivascular cells, blood cells, and pulpal blood vessels [151, 122], all of which decompose readily after death through a combination of necrosis and microbial invasion [195]. Thus, while the majority of cells within dental calculus are found within a mineralized structure conducive to preservation, the majority of cells within teeth are not, which may partially explain the large differences in total DNA yield between the two substrates. However, further studies of total DNA yields from freshly extracted teeth and their component tissues are needed to fully understand these differences.

4.5.2 Dental calculus and dentin harbor distinct microbial communities. Microbial DNA obtained from dental calculus and dentin derive from distinct communities. In agreement with previous studies [195, 194], the microbial community in dental calculus is dominated by human–associated oral taxa, and DNA derived from these organisms greatly exceeds that originat-

ing from environmental sources. The consistent preservation of a strong oral microbiome signal in all 48 dental calculus samples in this study suggests that this pattern is typical for archaeological dental calculus. By contrast, microbial DNA within dentin primarily derives from environmental sources. This distinction between the two substrates is preserved across geography, burial environment, and temporal period.

The tight clustering of all calculus samples included in this study, in contrast to with the diffuse distribution of dentin samples in the PCoA (Figure 4.3a), indicates that the oral microbiome signal is relatively uniform and stable across diverse contexts, as expected for a preserved biological community. In comparison, the diffuse distribution of dentin samples reflects the diverging influences of different environmental microbes and the absence of a consistent microbial composition. Despite the presence of DNA belonging to oral microbes in some dentin samples, none of those included in this analysis join the calculus cluster, indicating they are result of stochastic preservation of particular oral microbes and do not retain a signal of a biological community.

4.5.3 Dentin is a source of oral microbial DNA. Interestingly, although most microbial DNA within dentin is environmental in origin, we find that oral bacteria contribute $>$ 20% of total DNA in approximately one third of the dentin samples (7 of 46) in this study. Notably, the teeth in this study were free of oral pathology, such as caries, and with the exception of the two excluded Norris Farms teeth, care was taken to avoid sampling the tooth surface. Thus, dental infection and incomplete calculus removal are unlikely explanations for

the presence of oral microbial DNA we observe in dentin. Our results agree with data recently reported by [147], in which human–associated microbes constituted 15% of the organisms identified in a metagenomic study of over 100 archaeological, root–derived dentin samples. These findings suggest that members of the oral microbiome may also participate in postmortem dental decomposition, although to a lesser extent than environmental microbes.

The presence of DNA from oral taxa in dentin has important implications for the study of ancient commensal microbes and their evolution. Although prevalent, dental calculus is not always present or preserved in archaeological skeletal collections. Additionally, dental calculus may be absent or found in low abundance in young individuals or for certain populations or time periods. If insufficient calculus is available for study, it may be possible to instead access aspects of the oral microbiome through tooth dentin. Although the stochastic processes of postmortem microbial growth would preclude oral microbiome community–level analyses, genetic sequencing of dentin could nevertheless provide access to the genomes of individual oral taxa for analysis.

4.5.4 Dental calculus is a source of host DNA. Although dentin generally contains a higher proportion of human DNA than dental calculus, many dental calculus samples in this study have comparable proportions of human DNA to their dentin pair, with one dental calculus sample from the Netherlands (S454) having a higher proportion of DNA assigned to the human genome than its paired dentin sample (Appendix Table C.2). This pattern is

largely driven by the high variability of human DNA preservation within dentin.

Excepting in cases of infection [167], nearly all DNA in dentin should originate from the host genome at the time of death. However, archaeological teeth typically contain much lower proportions of human DNA due to post-mortem degradation and exogenous microbial growth. By contrast, the relative proportion of human DNA is uniformly low and relatively consistent in dental calculus. When dentin is strongly degraded and the relative proportion of host DNA in dentin is very low ($<$ 0.1%), the absolute amount of human DNA within dental calculus can exceed that of dentin. In such cases, the genetic richness of dental calculus appears to compensate for its low relative proportion of human DNA. Although obtaining host DNA from dental calculus using shotgun sequencing is generally inefficient given its low relative abundance, dental calculus has been shown to be a valuable reservoir for recovering host DNA using DNA capture methods [135].

4.5.5 Human DNA in dental calculus is highly fragmented. We find human DNA from dental calculus to be consistently shorter than the total DNA from the same sample, and on average shorter than human DNA recovered from the paired dentin sample. As a microbial biofilm, dental calculus is not a human tissue and does not contain human cells. The mechanisms by which human DNA is incorporated into dental calculus are not well understood but are presumed to include passive adsorption of human DNA from oral fluids and shed mucosal cells, as well as more active incorporation through host inflammatory responses, including a kind of immune response mediated by

neutrophils known as NETosis [194]. Neutrophils are an essential cell type of the innate immune system that are recruited by macrophages during active microbial infection [170]. Particularly important in the pathogenesis of periodontal disease, neutrophils are recruited in high numbers into the gingival crevice to attack dental plaque bacteria [159, 139]. Previous research has found that most human proteins recovered from both modern and ancient dental calculus are associated with the innate immune system, and specifically with neutrophils [194].

If host immune activity is a major contributor of human DNA to dental calculus, the role of neutrophils in this activity may partially explain the higher degree of human DNA fragmentation in dental calculus than in dentin. While neutrophils and other immune cells are capable of phagocytizing individual or small aggregates of microbial cells, large pathogens or those that can thwart phagocytosis by forming biofilms stimulate the formation of neutrophilic extracelluar traps (NETs) [21, 99]. NETs are composed of decondensed chromatin that is released from the nuclear membrane and mixed with disarticulated histones and granules containing antimicrobial proteins before being ejected from the cell [170, 99, 117, 18]. The expelled chromatin traps the offending microbes while simultaneously promoting destruction of the entrapped cells [170]. Interestingly, along with the granular proteins, disarticulated histones and short fragments of DNA ($<$ 100 bp) are also potent antimicrobials, likely increasing the antibiotic effect of NETs [22, 18].

Many bacteria have been shown in vitro to stimulate NETosis, including the oral bacteria *Porphyromonas gingivalis* and *Aggregatibacter actino-*

mycetemcomitans [140]. Moreover, certain bacteria subvert NETosis by producing extracellular nucleases (DNases) which are either bound to the cell membrane or secreted from the cell [41]. These nucleases free trapped bacteria by degrading the DNA backbone of the excreted NETs [170]. This activity is particularly prevalent during periodontal disease, and a wide range of oral bacteria including *P. gingivalis*, *Tannerella forsythia*, *Fusobacterium nucleatum*, and *Parvimonas micra* are able to produce extracellular nucleases [139].

If host DNA is incorporated into dental calculus in an acellular form—either through NETosis or by another mechanism—the exposed DNA would be vulnerable to a variety of damaging processes, including both hydrolysis and extracellular nuclease activity. This may explain why human DNA within dental calculus exhibits a high level of fragmentation that is poorly correlated with that of oral microbes in the same sample [197].

4.5.6 Cell wall structure is not correlated with microbial DNA fragmentation or damage. It has been previously proposed that certain microbial cell wall attributes, such as the presence (Gram–positive) or absence (Gram–negative) of a thick peptidoglycan layer, may influence the preservation of microbial DNA and therefore contribute to biases in taxonomic analyses of archaeological dental calculus [167, 2]. However, a subsequent investigation of four dental calculus samples failed to find such a correlation [214]. In this study, we test this hypothesis in 48 dental calculus samples and find no relationship between attributes such as cell wall peptidoglycan structure or the

presence of an S–layer and DNA fragmentation or terminal cytosine damage patterns. This analysis does not preclude other aspects of cellular structure (e.g., spore formation or the presence of mycolic acids) that may impact aDNA preservation but which were not tested in this study. Our analysis of reads assigned to specific bacteria suggests that fragmentation and damage patterns may be taxonomically structured. However, the consistency and biological basis of these patterns is beyond the scope of the data presented here.

4.5.7 Loss of short AT–rich DNA fragments may contribute to taxonomic skew. Analysis of the relationship between genomic GC content, DNA fragment GC content, and DNA fragment length reveals an inverse relationship between DNA fragment length and GC content in taxa with low– and medium–GC genomes, suggesting a systematic loss of short AT–rich fragments. Short DNA fragments lack thermostability and are easily lost through denaturation. The melting temperature of short DNA fragments is primarily dependent on DNA sequence and length, in addition to environmental conditions and additional factors [142], and in general sequences with longer lengths and higher GC content have higher melting temperatures. Short DNA fragments from taxa with lower GC content genomes are expected to be more susceptible to loss through denaturation because their melting point for a given fragment length will be lower, and this may contribute to taxonomic skew. Interestingly, we found that high GC–content genera had slightly shorter median fragment lengths, which is consistent with a higher retention

rate of short DNA fragments.

Although we observe these patterns in archaeological dental calculus, it is unclear if this is an artifact introduced during sequencing preparation or a naturally occurring taphonomic process. Importantly, this effect is weaker or absent from high genomic GC content genera, which include many soil bacteria [58]. If the loss of short AT–rich fragments is primarily taphonomic and not methodological in nature, greater taxonomic skew may be expected in libraries generated using a single–stranded library preparation, which is known to retain a higher proportion of shorter DNA fragments than the double–stranded DNA library preparation method used in this study [63]. Documentation of potential taxonomic–specific biases in recovery of DNA is critical as they impact downstream interpretations of metagenomic data, affecting accurate description of these ancient microbial ecosystems.

In this study, we use metagenomic sequence data to explore patterns of preservation in microbial– and host–derived DNA in a large and diverse set of paired archaeological dentin and dental calculus samples (n=48 individuals, n=96 samples). We demonstrate that dental calculus is a rich source of well–preserved oral microbiome DNA and a consistent source of highly fragmented and low abundance human DNA. We find that cell wall structure has no significant association with microbial DNA preservation, but that all samples exhibit systematic loss of short AT–rich DNA fragments, a trend that disproportionately affects taxa with low and moderate GC content genomes. Finally, we show that approximately one third of teeth retain DNA from the oral microbiome and thus tooth dentin may serve as an alternative source of

oral bacterial DNA in the absence of preserved dental calculus deposits.

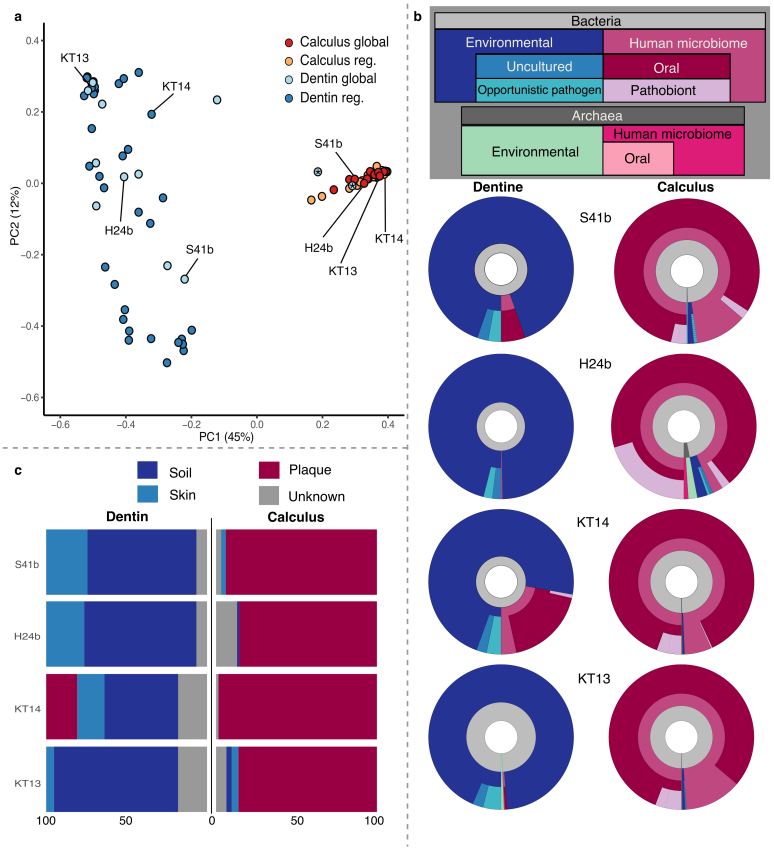

Figure 4.3: **Microbial communities represented in archaeological dental calculus and dentin are distinct.** (a) Principal Coordinates Analysis (PCoA) of Bray–Curtis distances of all bacterial and archaeal species–level assignments from dental calculus and dentin. Color indicates material type and membership in global or regional dataset. Dentin samples marked with an asterisk belong to individuals NF47 and NF217 in the global dataset, and may represent the impact of carious lesions (Appendix C.1). (b) A subset of four sample pairs were selected to further demonstrate differences in dental calculus and dentin microbial communities. Species represented in the MALT results were sorted into a layered classification scheme and the proportions of reads assigned to each taxon were used to generate Krona plots. (c) Stacked bar plots of Bayesian SourceTracker results for the four selected pairs display estimated proportions of source contribution at the genus level, using modern plaque, skin, and soil datasets as model sources. Both approaches show overwhelming abundance of environmental bacteria within dentin samples, while most microbial DNA within the paired calculus samples are native to, and most likely derive from, the human microbiome.

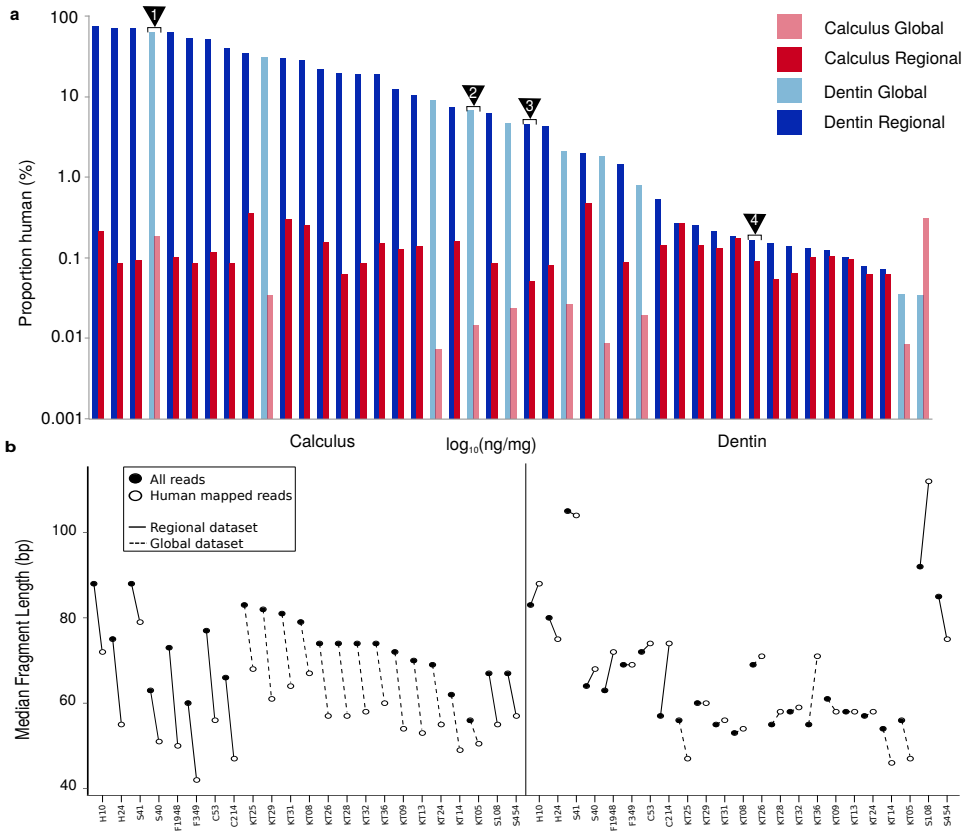

Figure 4.4: **Human DNA in dental calculus shows consistent patterns of low relative abundance and high fragmentation.** (a) Relative percentage of human DNA in all paired calculus and dentin samples calculated from de–duplicated reads mapped to the hg19 human reference genome using BWA. While the majority of dentin samples have an overall higher percentage human DNA, this value varies substantially by sample. Calculus is comparatively consistent between samples albeit on average lower than their paired dentin sample. Sample pairs corresponding to those in Figure 3 are indicated by numbered triangles: (1) S41, (2) H24b, (3) KT14, (4) KT13. (b) Median fragment length of merged reads mapping to the human genome compared to all merged reads in both dental calculus and dentin. Human assigned reads in dental calculus are shorter than expected independent of age, laboratory processing protocol, or sample preservation. Human mapped reads in dental calculus and dentin were further verified for authenticity using strict mapping parameters (Appendix Figure C.2, Appendix Table C.2).

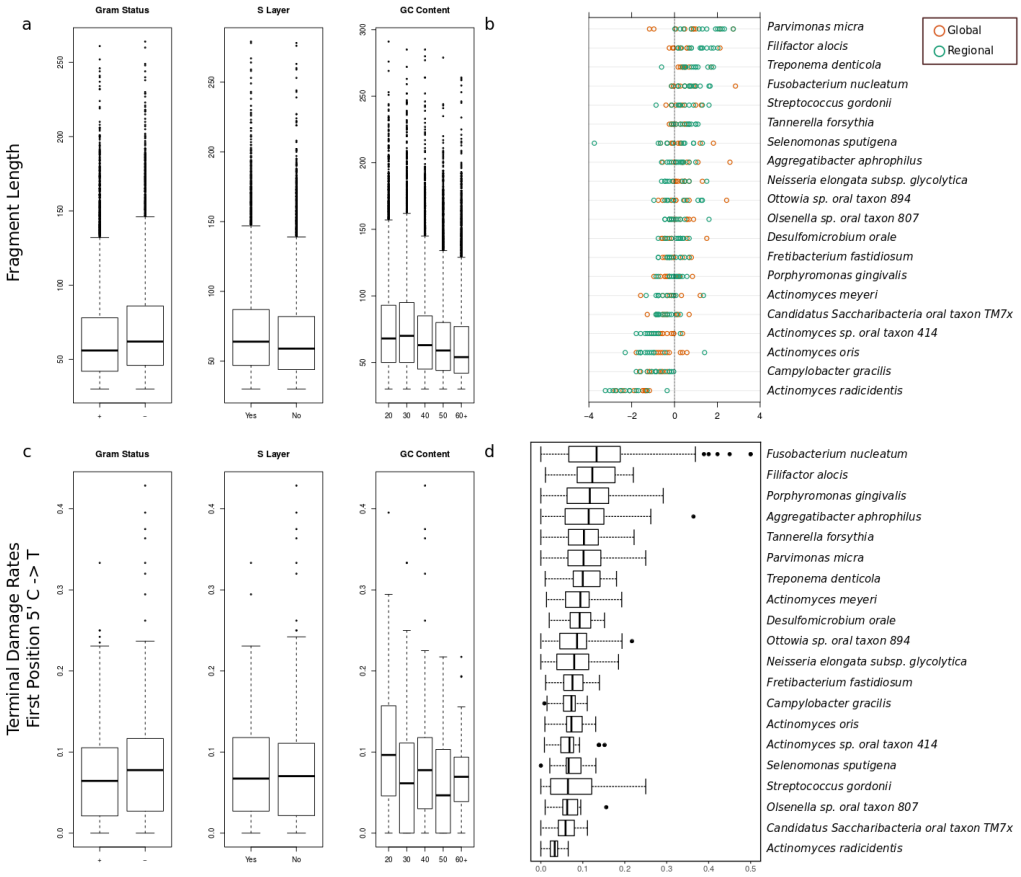

Figure 4.5: **Fragment length and damage rates among bacterial taxa within calculus.** (a) Fragment length distribution of 50 high frequency species–level bacteria among all dental calculus grouped into three metadata categories: gram status, the presence or absence of a surface layer (S–layer), and the overall genomic GC content of the organism as documented from the reference genome in the NCBI database. Input was normalized to 400 randomly chosen reads per sample to mitigate the impact of sample specific read length profiles. (b) Deviation of the median fragment length from overall sample median fragment length of a subset of 20 oral bacteria in dental calculus colored by dataset origin. (c) Terminal cytosine damage rates (C to T substitution ratio at the first position of the 5 end of the molecule) among 50 bacterial species in dental calculus grouped by gram status, presence or absence of an S–layer, and overall genomic GC content. (d) Terminal cytosine damage rates (C to T substitution ratio at the first position of the 5 end of the molecule) among 20 oral bacteria.

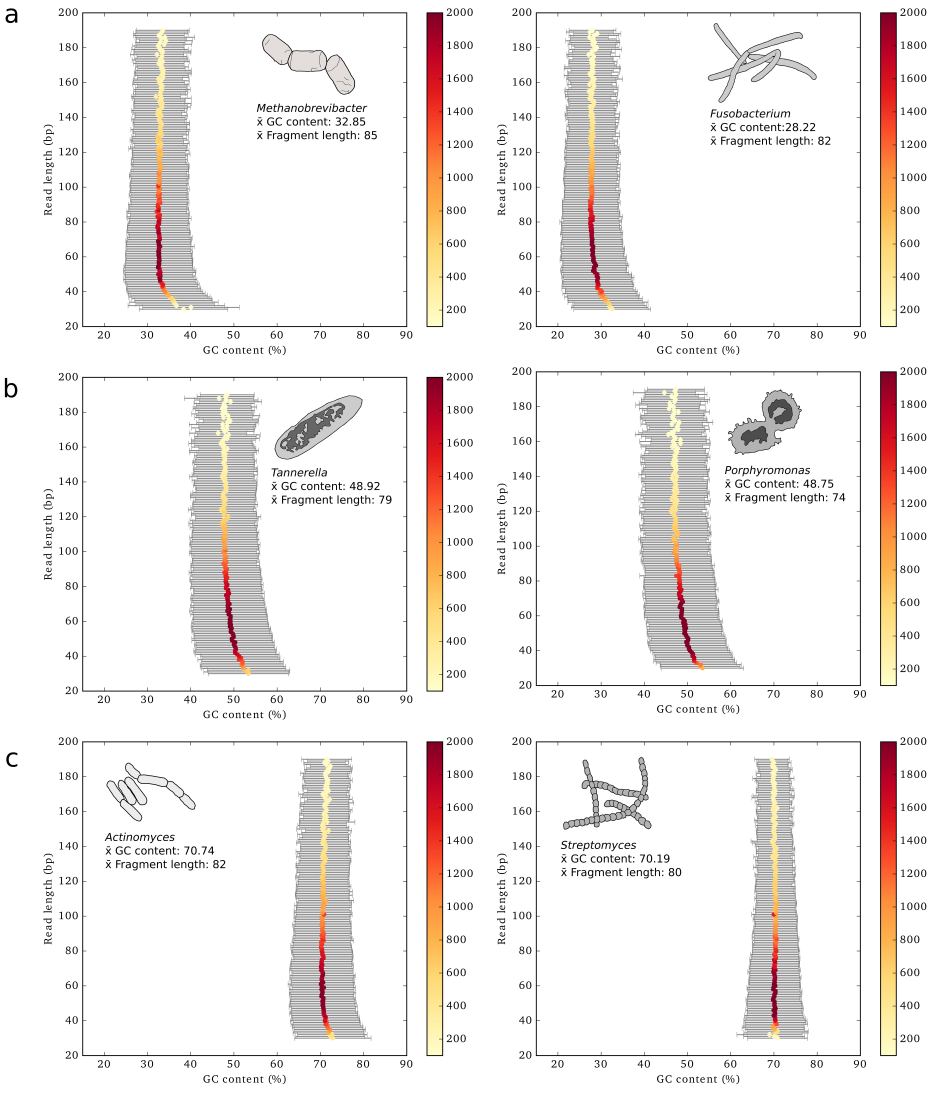

Figure 4.6: **Relationship of GC content to fragment length in five prevalent oral genera and one soil genus (*Streptomyces*).** (a) Two low expected GC content genera, *Methanobrevibacter* and *Fusobacterium* binned by read lengths wherein each dot represents the mean GC content for all reads at that length. (b) Two moderate expected GC content general, *Tannerella* and *Porphyromonas* binned by read lengths wherein each dot represents the mean GC content for all reads at that length. (c) One high expected GC content oral genus (*Actinomyces*) and one high expected GC content soil genus (*Streptomyces*) binned by read lengths wherein each dot represents the mean GC content for all reads at that length.

Chapter 5

Conclusions

The purpose of this dissertation is to outline experimental techniques designed to improve the taxonomic and community resolution of members of the human microbiome. Specifically, three studies employing novel sample processing and analytical techniques were presented to address known complications in characterizing the members of the human gut microbiome in extant populations, as well as the ancient human oral microbiome, using metabarcoding and metagenomic techniques.

In chapter two: “Microeukaryotic and dietary survey of the gut by internal transcribed spacer metabarcoding”, the eukaryotic component of the human gut microbiome was characterized in three human groups and two animal species using the internal transcribed spacer regions one (ITS1) and two (ITS2). Results of this study suggest that like bacterial diversity between industrialized and non–industrialized populations [131, 34, 165, 210], microeukaryotic diversity is higher in non–industrialized populations with more shared taxa among groups with more traditional subsistence strategies than with industrialized groups. In addition, the presence of the protist genus

Blastocystis was detected in all three human groups, regardless of geography or subsistence strategy, which suggests that this microeukaryotic taxa is widespread among human groups. Chapter three: "Enrichment of non–dominant bacterial taxa in human fecal samples through serial filtration" demonstrates the utility of serial cell–sized fractionation of fecal samples for the enrichment of relatively small bacterial taxa that are to date under characterized. In particular, the presence of putative non–photosynthetic bacterial relatives of the phylum *Cyanobacteria* in filtered samples from non–industrial populations illustrates the applicability of this method to screen for appropriate samples for further investigation of undercharacterized taxa. As members of this bacterial group have to–date only been characterized by their presence in metagenomic samples and similarity to close relatives in the environment [45, 176, 175], enrichment for these taxa may enable the further characterization of these taxa using metagenomic techniques or the selection of samples where culturing of these microbes may be possible. Finally, chapter four: "Differential preservation of endogenous human and microbial DNA in dental calculus and dentin" evaluates the preservation qualities of archaeological dental calculus and finds that while dental calculus is a dependable source of high yield ancient DNA and preserves a microbial signature consistent with a human oral microbiome, the organism source may impact DNA preservation as measured by fragmentation profiles. Taxonomically structured biases, therefore, may be intrinsic to this archaeological substrate, however, the impact of these biases on the retrieval and interpretation of endogenous host and microbial DNA requires further investigation.

Future applications of the methods described here are hoped to further clarify the interaction of under described taxa—including low frequency bacteria and eukaryotes—in the human gut microbiome with other community members, as well as better characterize ancient human microbiome sources by delineating expected taxonomic shifts by whether these are true representations of the ancient microbiome state or if they are instead the result of taphonomic or other natural or artificial preservational differences. The omission of microbial eukaryotes as well as other low frequency or otherwise under reported taxa is expected to have consequences for the interpretation of data generated for microbiome habitats [9, 115, 50]. Additionally, as the inclusion of microbiome data generated from archaeological sources including paleofeces [182] and dental calculus [194, 2, 214] are increasingly used to understand the evolutionary history and ancestral state of a variety of human–associated microbial ecosystems, a full appreciation of the potential biases stemming from the characteristic degeneration associated with ancient biomolecules including DNA is of paramount importance. Because microbiome studies have the potential to inform on not only the evolution of these diverse microbial communities, but also have the potential to inform on aspects of modern human health and disease, understanding the human microbiome as a ecological system, holistically defined, will augment our ability to predict the impact of dysbiosis and the presence or absence of specific microbes on these ecological systems.

Bibliography

[1] Abiola Fatimah Adenowo et al. Impact of human schistosomiasis in sub-saharan africa. *The Brazilian Journal of Infectious Diseases*, 19(2):196 – 205, 2015.

[2] C. J. Adler et al. Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the neolithic and industrial revolutions. *Nat Genet*, 45(4):450–5, 455e1, 2013.

[3] Mohammed A. Alfellani et al. Diversity and distribution of blastocystis sp. subtypes in non-human primates. *Parasitology*, 140(8):966971, 2013.

[4] Mohammed A. Alfellani et al. Variable geographic distribution of blastocystis subtypes and its potential implications. *Acta Tropica*, 126(1):11 – 18, 2013.

[5] Linda A. Amaral-Zettler et al. A method for studying protistan diversity using massively parallel sequencing of v9 hypervariable regions of small-subunit ribosomal rna genes. *PLOS ONE*, 4(7):1–9, 07 2009.

[6] Omar M. Amin. The epidemiology of blastocystis hominis in the united states. *Research Journal of Parasitology*, 1:1–10, 2006.

[7] Evilena Anastasiou et al. Prehistoric schistosomiasis parasite found in the middle east. *The Lancet Infectious Diseases*, 14(7):553 – 554, 2014.

[8] Lee O'Brien Andersen and Christen Rune Stensvold. Blastocystis in health and disease: Are we moving from a clinical to a public health perspective? *Journal of Clinical Microbiology*, 54:524–528, 2015.

[9] L.O. Andersen et al. Waiting for the human intestinal eukaryotome. *International Society for Microbial Ecology*, 7:1253–1255, 2013.

[10] Ian C. Anderson et al. Potential bias of fungal 18s rdna and internal transcribed spacer polymerase chain reaction primers for estimating fungal biodiversity in soil. *Environmental Microbiology*, 5(1):36–47, 2003.

[11] A. Araujo et al. Parasites adn probes for prehistoric human migrations? *Trends in Parasitology*, 3, 2008.

[12] A. Araujo and L.F. Ferreira. Paleoparasitology and the antiquity of human host–parasite relationships. *Memorias do Instituto Oswaldo Cruz*, 95, 2000.

[13] Philip L. Armitage. The extraction and identification of opal phytoliths from the teeth of ungulates. *Journal of Archaeological Science*, 2(3):187–197, 1975.

[14] Yasmine Belkaid and Timothy Hand. Role of the microbiota in immunity and inflammation. *Cell*, 157:121–141, 2015.

[15] R.G. Bell. Ige, allergies and helminth parasites: A new perspective on an old conundrum. *Immunology and Cell Biology*, 74(4):337–345, 2014.

[16] Lea Berrang-Ford et al. Conflict and human african trypanosomiasis. *Social Science & Medicine*, 72(3):398 – 407, 2011. 13th International Medical Geography Symposium.

[17] B. Beszteri et al. Average genome size: a potential source of bias in comparative metagenomics. *The ISME Journal*, 4:1075–1077, 2010.

[18] R. K. Bhongir et al. Dna-fragmentation is a source of bactericidal activity against pseudomonas aeruginosa. *Biochemical Journal*, 474, 2017.

[19] Staci D. Bilbo et al. Reconstitution of the human biome as the most reasonable solution for epidemics of allergic and autoimmune diseases. *Medical Hypotheses*, 77, 2011.

[20] Kenneth F. Boorom et al. Oh my aching gut: irritable bowel syndrome, blastocystis, and asymptomatic infection. *Parasites & Vectors*, 1(1):40, Oct 2008.

[21] N. Branzk et al. Neutrophils sense microbe size and selectively release neutrophil extracellular traps in response to large pathogens. *Nature Immunology*, 15, 2014.

[22] V. Brinkmann and A. Zychlinsky. Neutrophil extracellular traps: Is immunity the second function of chromatin? *The Journal of Cell Biology*, 198, 2012.

[23] Jochen J. Brocks et al. Archean molecular fossils and the early rise of eukaryotes. *Science*, 285(5430):1033–1036, 1999.

[24] R.M. Brotman. Vaginal microbiome and sexually transmitted infections: An epidemiologic perspective. *Journal of Clinical Investigation*, 121:4610–4617, 2011.

[25] Cyril Caminade et al. Impact of climate change on global malaria distribution. *Proceedings of the National Academy of Sciences*, 111(9):3286–3291, 2014.

[26] J. Gregory Caporaso et al. Global patterns of 16s rrna diversity at a depth of millions of sequences per sample. *Proc Natl Acad Sci USA*, 108:4516–4522, 2011.

[27] J Gregory Caporaso et al. Qiime allows analysis of high-throughput community sequencing data. *Nature Methods*, 7:335–336, 2011.

[28] Tanai Cardona. Early archean origin of heterodimeric photosystem i. *Heliyon*, 4(3):e00548, 2018.

[29] David A. Caron et al. Defining dna-based operational taxonomic units for microbial-eukaryote ecology. *Applied and Environmental Microbiology*, 75:5797–5808, 2009.

[30] David A. Caron et al. Protists are microbes too: a perspective. *The ISME Journal*, 3:4–12, 2009.

[31] T. Chen et al. The human oral microbiome database: a web accessible resource for investigating oral microbe taxonomic and genomic information. *Database 2010*, 2010.

[32] Amandine Cian et al. Molecular epidemiology of blastocystis sp. in various animal groups from two french zoos and evaluation of potential zoonotic risk. *PLOS ONE*, 12(1):1–29, 01 2017.

[33] B.L. Clark et al. The effect of tritrichomonas foetus infection on calving rates in beef cattle. *Australian Veterinary Journal*, 60(3):71–74, 1983.

[34] Jose C. Clemente et al. The microbiome of uncontacted amerindians. *Science Advances*, 1(3), 2015.

[35] Annette W. Coleman. Analysis of mammalian rdna internal transcribed spacers. *PLOS ONE*, 8(11), 11 2013.

[36] HMP Consortium. A framework for human microbiome research. *Nature*, 486, 2012.

[37] Nathalie M.L. Cote et al. A new high-throughput approach to genotype ancient human gastrointestinal parasites. *PLOS ONE*, 11(1):1–18, 01 2016.

[38] F.E. Cox. History of human parasitic diseases. *Infectious Disease Clinics of North America*, 18, 2004.

[39] Jesse Dabney et al. Ancient dna damage. *Cold Spring Harbor perspectives in biology*, 5(7):a012567, 2013.

[40] D.Z. Dagoye et al. Wheezing, allergy, and parasite infection in children in urban and rural ethiopia. *Am. J. Respir. Crit. Care Med.*, 167, 2003.

[41] G. Dang et al. Characterization of rv0888, a novel extracellular nuclease from mycobacterium tuberculosis. *Scientific Reports*, 6, 2016.

[42] Kara Bowen De Leon et al. Archael and bacterial communities in three alkaline hot springs in heart lake geyser basin, yellowstone national park. *Frontiers in Microbiology*, 4:330, 2013.

[43] Nilanthi R. de Silva et al. Soil-transmitted helminth infections: updating the global picture. *Trends in Parasitology*, 19(12):547 – 551, 2003.

[44] Les Dethlefsen et al. An ecological and evolutionary perspective on human-microbe mutualism and disease. *Nature*, 449, 2007.

[45] Sara C. Di Rienzi et al. The human gut and groundwater harbor non-photosynthetic bacteria belonging to a new candidate phylum sibling to cyanobacteria. *eLIFE*, 2:e01102, 2013.

[46] F. Dogruman-Al et al. A possible link between subtype 2 and asymptomatic infections of blastocystis hominis. *Parasitology Research*, 103:685–689, 2008.

[47] Paul D. Donovan et al. Identification of fungi in shotgun metagenomics datasets. *PLOS ONE*, 13(2):1–16, 02 2018.

[48] Tiina Drell et al. Characterization of the vaginal micro- and mycobiome in asymptomatic reproductive-age estonian women. *PLOS ONE*, 8(1):1–11, 01 2013.

[49] J.V. Dudgeon and M. Tromp. Diet, geography and drinking water in polynesia: Microfossil research from archaeological human dental calculus, rapa nui (easter island). *International Journal of Osteoarchaeology*, 24(5):634–648, 2014.

[50] Bas E. Dutilh et al. A highly abundant bacteriophage discovered in the unknonwn sequences of human faecal metagenomes. *Nature Communications*, 5(4498), 2014.

[51] S. Boyd Eaton et al. Prehistoric schistosomiasis parasite found in the middle east. *The Lancet Infectious Diseases*, 14(7):553 – 554, 2014.

[52] Robert C. Edgar. Search and clustering orders of magnitude faster than blast. *Bioinformatics*, 26(19):2460–2461, 2010.

[53] Robert C. Edgar. Updating the 97 *Bioinformatics*, page bty113, 2018.

[54] Dima El Safadi et al. Children of senegal river basin show the highest prevalence of blastocystissp. ever observed worldwide. *BMC Infectious Diseases*, 14(1):164, Mar 2014.

[55] I. Esteve et al. Electron microscope study of the interaction of epibiontic bacteria with chromatium minus in natural habitats. *Microbial Ecology*, 9:57–64, 1983.

[56] R.S.J. Felleisen. Comparative sequence analysis of 5.8s rrna genes and internal transcribed spacer (its) regions of trichomonadid protozoa. *Parasitology*, 115:111–119, 1997.

[57] Noah Fierer and Robert B. Jackson. The diversity and biogeography of soil bacterial communities. *Proceedings of the National Academy of Sciences*, 103(3):626–631, 2006.

[58] K. U. Foerstner et al. Environments shape the nucleotide composition of genomes. *EMBO Reports*, 6, 2005.

[59] J.A. Foster and K.A. Neufeld. Gut-brain axis: How the microbiome influences anxiety and depression. *Trends in Neuroscience*, 36:305–312, 2013.

[60] C. Lalueza Fox et al. Dietary information through the examination of plant phytoliths on the enamel surface of human dentition. *Journal of Archaeological Science*, 21(1):29–34, 1994.

[61] G. F. Fry and J. G. Moore. Enteriobius vermicularis: 10,000 year old human infection. *Science*, 166, 1969.

[62] C. Gamba et al. Genome flux and stasis in a five millennium transect of european prehistory. *Nature Communications*, 5, 2014.

[63] Marie-Theres Gansauge et al. Single-stranded dna library preparation from highly degraded dna using t4 dna ligase. *Nucleic Acids Research*, 45(10):e79–e79, 2017.

[64] Simonetta Gatti et al. Amebic infections due to the entamoeba histolytica-entamoeba dispar complex: a study of the incidence in a remote rural area of ecuador. *The American Journal of Tropical Medicine and Hygiene*, 67(1):123–127, 2002.

[65] Jay R. Georgi and Charles E. McCulloch. Diagnostic morphometry: Identification of helminth eggs by discriminant analysis of morphometric data. *Proc. Helminthol. Soc. Wash.*, 56:44–57, 1989.

[66] Rohit Ghai et al. Metagenomics uncovers a new group of low gc and ultra-small marine actinobacteria. *Scientific Reports*, 3:2471, 2013.

[67] Mahmoud A. Ghannoum et al. Characterization of the oral fungal microbiome (mycobiome) in healthy individuals. *PLOS Pathogens*, 6(1):1–8, 01 2010.

[68] Ronnie N. Glud et al. High rates of microbial carbon turnover in sediments in the deepest oceanic trench on earth. *Nature Geoscience*, 6:284–288, 2013.

[69] M. L. C. Goncalves et al. Human intestinal parasites in the past: New findings and a review. *Memorias Do Instituto Oswaldo Cruz*, 98, 2003.

[70] J. C. Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, Mar 1975.

[71] Andrea L. Graham. Ecological rules governing helminth–microparasite coinfection. *Proceedings of the National Academy of Sciences*, 105(2):566–570, 2008.

[72] B.V. Gromov and K.A. Mamkaeva. New genus of bacteria, vampirovibrio, parasitizing chlorella and previously assigned to the genus bdellovibrio. *Mikrobiologiia*, 49:165–167, 1980.

[73] Laure Guillou et al. The protist ribosomal reference database (pr2): a catalog of unicellular eukaryote small sub-unit rrna sequences with curated taxonomy. *Nucleic Acids Research*, 41, 2013.

[74] B. J. Hackett et al. Ribosomal dna internal transcribed spacer (its2) sequences differentiate anopheles funestus and an. rivulorum, and uncover a cryptic taxon. *Insect Molecular Biology*, 9(4):369–374, 2000.

[75] I. Hamad et al. Molecular detection of eukaryotes in a single human stool sample from senegal. *PloS ONE*, 7, 2012.

[76] Kui Han et al. Extraordinary expansion of a sorangium cellulosum genome from an alkaline milieu. *Scientific Reports*, 3, 2013.

[77] H. B. Hansen et al. Comparing ancient dna preservation in petrous bone and tooth cementum. *PLOS ONE*, 12, 2017.

[78] Joel Hartter. Attitudes of rural communities toward wetlands and forest fragments around kibale national park, uganda. *Human Dimensions of Wildlife*, 14(6):433–447, 2009.

[79] A.G. Henry et al. Microfossils in calculus demonstrate consumption of plants and cooked foods in neanderthal diets. *Proceedings of the National Academy of Sciences*, 108(2), 2011.

[80] Travis Henry et al. Identification of aspergillus species using internal transcribed spacer regions 1 and 2. *Journal of Clinical Microbiology*, 38:1510–1515, 2000.

[81] Michael Hofreiter et al. Dna sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient dna. *Nucleic acids research*, 29(23):4793–4799, 2001.

[82] L.V. Hooper et al. Interactions between the microbiota and the immune system. *Science*, 8:1268–1273, 2012.

[83] Herve Hoste et al. Differences in the second internal transcribed spacer (ribosomal dna) between five species of trichostrongylus (nematoda: Trichostrongylidae). *International Journal for Parasitology*, 25(1):75 – 80, 1995.

[84] R. Hubler et al. Amps: A pipeline for screening archaelogical remains for pathogen dna, 2017.

[85] Laura A. Hug et al. A new view of the tree of life. *Nature Microbiology*, 1, 2016.

[86] Perrine Hugon et al. A comprehensive repertoire of prokaryotic species identified in human beings. *The Lancet Infectious Diseases*, 2015.

[87] Daniel H. Huson et al. Megan community edition - interactive exploration and analysis of large-scale microbiome sequencing data. *PLOS Computational Biology*, 12(6):e1004957, 2016.

[88] Laszlo Irinyi et al. International society of human and animal mycology (isham)-its reference dna barcoding databasethe quality controlled standard tool for routine identification of human and animal pathogenic fungi. *Medical Mycology*, 53(4):313–337, 2015.

[89] P. C. Iwen et al. Utilization of the internal transcribed spacer regions as molecular targets to detect and identify human fungal pathogens. *Medical Mycology*, 40(1):87–109, 2002.

[90] Priscilla A. Johanesen et al. Disruption of the gut microbiome: Clostridium difficile infection and the threat of antibiotic resistance. *Genes*, 6:1347–1360, 2015.

[91] E. R. Johnston et al. Metagenomics reveals pervasive bacterial populations and reduced community diversity across the alaska tundra ecosystem. *Frontiers in Microbiology*, 7, 2016.

[92] H. Jonsson et al. mapdamage2.0: fast approximate bayesian estimates of ancient dna damage parameters. *Bioinformatics*, 29(13):1682–4, 2013.

[93] N.A. Joshi and J.N. Fass. Sickle: a sliding-window, adaptive, quality-based trimming tool for fastq files (version 1.33), 2011.

[94] Kazutaka Katoh and Daron M. Standley. Mafft multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, 30:772–780, 2013.

[95] Matthew Kearse et al. Geneious basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28(12):1647–1649, 2012.

[96] Brian M. Kemp et al. How much dna is lost? measuring dna loss of short-tandem-repeat length fragments targeted by the powerplex 16 system using the qiagen minelute purification kit. *Human Biology*, 86(4):1–18, 2014.

[97] D.J. King et al. Corn, beer, and marine resources at casa grandes, mexico. *Journal of Archaeological Science: Reports*, 16, 2017.

[98] Dan Knights et al. Bayesian community-wide culture-independent microbial source tracking. *Nat Meth*, 8(9):761–763, 2011.

[99] M. V. Kockritz-Blickwede et al. Interaction of bacterial exotoxins with neutrophil extracellular traps: Impact for the infected host. *Frontiers in Microbiology*, 7, 2016.

[100] Urmas Koljalg et al. Towards a unified paradigm for sequence-based identification of fungi. *Molecular Ecology*, 22(21):5271–5277, 2013.

[101] Konstantinos T. Konstantinidis and James M. Tiedje. Genomic insights that advance the species definition for prokaryotes. *Proceedings of the National Academy of Sciences*, 102(7):2567–2572, 2005.

[102] Isabelle Laforest-Lapointe and Marie-Claire Arrieta. Microbial eukaryotes: a missing link in gut microbiome studies. *mSystems*, 3(2), 2018.

[103] J.G. LeBlanc et al. Bacteria as vitamin suppliers to their host: A gut microbiota perspective. *Current Opinion in Biotechnology*, 24:1–9, 2012.

[104] Yun Kyung Lee and Sarkis K. Mazmanian. Has the microbiota played a critical role in the evolution of the adaptive immune system? *Science*, 330:1768–1773, 2010.

[105] P. Lekic et al. Phenotypic comparison of periodontal ligament cells in vivo and in vitro. *Journal of Periodontal Research*, 36:71–79, 2001.

[106] Daniela Leles et al. Are ascaris lumbricoides and ascaris suum a single species? *Parasites & Vectors*, 5(1):42, Feb 2012.

[107] Ivica Letunic and Peer Bork. Interactive tree of life (itol) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Research*, 44:W242–W245, 2016.

[108] Petra Anne Levin and Esther R. Angert. Small but mighty: cell size and bacteria. *Cold Spring Harb Perspect Biol*, 7:a019216, 2015.

[109] Ruth E. Ley et al. Microbial ecology: Human gut microbes associated with obesity. *Nature*, 444:1022–1023, 2006.

[110] H. Li and R. Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25:1754–60, 2009.

[111] A. Linde and M. Goldberg. Dentinogenesis. *Critical Reviews in Oral Biology and Medicine*, 4(5):679–728, 1993.

[112] S. Lindgreen. Adapterremoval: easy cleaning of next-generation sequencing reads. *BMC Res Notes*, 2, 2012.

[113] Corie Lok. Mining the microbial dark matter. *Nature*, 522(7556):270, 2015.

[114] Stilianos Louca et al. Correcting for 16s rrna gene copy numbers in microbiome surveys remains an unsolved problem. *Microbiome*, 6(1):41, Feb 2018.

[115] Julius Lukes et al. Are human intestinal eukaryotes beneficial or commensals? *PLOS Pathogens*, 11(8):1–6, 08 2015.

[116] Jeffrey C. Mai and Annette W. Coleman. The internal transcribed spacer 2 exhibits a common secondary structure in green algae and flowering plants. *Journal of Molecular Evolution*, 44:258–271, 1997.

[117] P. Majewski et al. Inhibitors of serine proteases in regulating the production and function of neutrophil extracellular traps. *Frontiers in Immunology*, 7, 2016.

[118] Shehre-Banoo Malik et al. Phylogeny of parasitic parabasalia and free-living relatives inferred from conventional markers vs. rpb1, a single-copy gene. *PLOS ONE*, 6(6):1–14, 06 2011.

[119] Clarisse A. Marotz and Amir Zarrinpar. Treating obesity and metabolic syndrome with fecal microbiota transplantation. *Yale Journal of Biology and Medicine*, 89:383–388, 2016.

[120] Philip D. Marsh. Dental plaque as a biofilm and microbial community - implications for health and disease. *BMC Oral Health*, 6(1), 2006.

[121] Philip D. Marsh and D. Bradshaw. Dental plaque as a biofilm. *Journal of Industrial Microbiology*, 15, 1995.

[122] Grayson W. Marshall et al. The dentin substrate: structure and properties related to bonding. *Journal of Dentistry*, 25(6):441–458, 1997.

[123] John P. McCutcheon and Nancy A. Moran. Extreme genome reduction in symbiotic bacteria. *Nature Reviews Microbiology*, 10:13–26, 2012.

[124] M. Meyer and M. Kircher. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *old Spring Harbor Protoc*, 6, 2010.

[125] Kim Mincheol et al. Towards a taxonomic coherence between average nucleotide identity and 16s rrna gene sequence similarity for species determination of prokaryotes. *International Journal of Systematic and Evolutionary Microbiology*, 64:346–351, 2014.

[126] Nancy A. Moran. Microbial minimalism: genome reduction in bacterial pathogens. *Cell*, 108:583–586, 2002.

[127] Young-Do Nam et al. Bacterial, archaeal, and eukaryal diversity in the intestines of korean people. *The Journal of Microbiology*, 46(5):491–501, Oct 2008.

[128] D. Natarajan et al. Anti-bacterial activity of euphorbia fusiformisa rare medicinal herb. *Journal of Ethnopharmacology*, 102(1):123 – 126, 2005.

[129] Nam-Phuong Nguyen et al. A perspective on 16s rrna operational taxonomic unit clustering using sequence similarity. *NPJ Biofilms and Microbiomes*, 2(16004), 2016.

[130] Matthew J. Nolan and Thomas H. Cribb. The use and implications of ribosomal dna sequencing for the discrimination of digenean species. *Advances in Parasitology*, 60:101 – 163, 2005.

[131] Alexandra J. Obregon-Tito et al. Subsistence strategies in traditional societies distinguish gut microbiomes. *Nature Communications*, page 6505, 2015.

[132] J. Oh et al. Temporal stability of the human skin microbiome. *Cell*, 165, 2016.

[133] Jari Oksanen et al. *vegan: Community Ecology Package*, 2017. R package version 2.4-5.

[134] Brian D. Ondov et al. Interactive metagenomic visualization in a web browser. *BMC Bioinformatics*, 12:385, 2011.

[135] Andrew T. Ozga et al. Successful enrichment and recovery of whole mitochondrial genomes from ancient human dental calculus. *American journal of physical anthropology*, 160(2):220–228, 2016.

[136] S. Paabo et al. Genetic analyses from ancient dna. *Annu Rev Genet*, 38:645–79, 2004.

[137] Sarah B. Paige et al. Beyond bushmeat: Animal contact, injury, and zoonotic disease risk in western uganda. *EcoHealth*, 11(4):534 – 543, 2014.

[138] Niloofar Paknazhad et al. Paleoparasitological evidence of pinworm (enterobius vermicularis) infection in a female adolescent residing in ancient tehran (iran) 7000 years ago. *Parasites & Vectors*, 9(1):33, Jan 2016.

[139] L. Palmer et al. Extracellular deoxyribonuclease production by periodontal bacteria. *Journal of Periodontal Research*, 47, 2012.

[140] L. Palmer et al. Influence of complement on neutrophil extracellular trap release induced by bacteria. *Journal of Periodontal Research*, 51, 2016.

[141] Prashant K. Pandey et al. Molecular typing of fecal eukaryotic microbiota of human infants and their respective mothers. *Journal of Biosciences*, 37(2):221 – 226, 2012.

[142] Alejandro Panjkovich et al. dnamate: a consensus melting temperature prediction server for short dna sequences. *Nucleic Acids Research*, 33, 2005.

[143] E. Paradis et al. Ape: analyses of phylogenetics and evolution in r language. *Bioinformatics*, 20:289–290, 2004.

[144] O. Partida-Rodriguez et al. Human intestinal microbiota: Interaction between parasites and the host immune response. *Archives of Medical Research*, 2017.

[145] Jonathan A. Patz et al. Effects of environmental change on emerging parasitic diseases. *International Journal for Parasitology*, 30(12):1395 – 1405, 2000. Thematic Issue: Emerging Parasite Zoonoses.

[146] Alexander Peltzer et al. Eager: efficient ancient genome reconstruction. *Genome Biology*, 17(1):60, 2016.

[147] A. Philips et al. Comprehensive analysis of microorganisms accompanying human archaeological remains. *GigaScience*, 6, 2017.

[148] R. Pinhasi et al. Optimal ancient dna yields from the inner ear part of the human petrous bone. *PLOS ONE*, 10, 2015.

[149] L.M. Polyanskaya et al. Assessment of the number, biomass, and cell size of bacteria in different soils using the ”cascade” filtration method. *Eurasian Soil Science*, 48(3):288–293, 2015.

[150] Maria C. Portillo et al. Cell size distributions of soil bacterial and archaeal taxa. *Applied and Environmental Microbiology*, 79(24):7610–7617, 2013.

[151] D. Vincent Provenza. The blood vascular supply of the dental pulp with emphasis on capillary circulation. *Circulation Research*, 6(2):213, 1958.

[152] Rachel L. Pullan et al. Global numbers of infection and disease burden of soil transmitted helminth infections in 2010. *Parasites & Vectors*, 7(1):37, Jan 2014.

[153] Christian Quast et al. The silva ribosomal rna gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, 41, 2013.

[154] A. Quinlan. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26, 2010.

[155] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.

[156] Juan David Ramirez et al. Blastocystis subtypes detected in humans and animals from colombia. *Infection, Genetics and Evolution*, 22:223 – 228, 2014.

[157] John A. Raven and John F. Allen. Genomics and chloroplast evolution: what did cyanobacteria do for plants? *Genome Biology*, 4(3):209, 2003.

[158] M.A. Ruffer. Note on the presence of bilharzia haematobia in egyptian mummies of the twentieth dynasty [1250-1000 b.c.]. *British Medical Journal*, 1, 1910.

[159] M. I. Ryder. Comparison of neutrophil functions in aggressive and chronic periodontitis. *Periodontology*, 53, 2010.

[160] Susanna Sabin. Ancient dna analysis of dental remains from kilteasheen: a case study in metagenomics and an exploration of dental calculus. Master's thesis, Eberhard-Karls-Universitat Tubingen, 2016.

[161] D. Sarkar. *Lattice: Multivariate Data Visualization with R*. Springer, 2008.

[162] Susanna Sawyer et al. Temporal patterns of nucleotide misincorporations and dna fragmentation in ancient dna. *PloS one*, 7(3):e34131, 2012.

[163] Pauline D. Scanlan et al. Prevalence and genetic diversity of blastocystis in family units living in the united states. *Infection, Genetics and Evolution*, 45:95 – 97, 2016.

[164] P.D. Scanlan and J.R. Marchesi. Micro-eukaryotic diversity of the human distal gut microbiota: Qualitative assessment using culture-dependent and -independent analysis of faeces. *ISME*, 2:1183 – 1193, 2008.

[165] Stephanie L. Schnorr et al. Gut microbiome of the hadza hunter-gatherers. *Nature Communications*, page 3654, 2014.

[166] Conrad L. Schoch et al. Nuclear ribosomal internal transcribed spacer (its) region as a universal dna barcode marker for fungi. *Proceedings of the National Academy of Sciences*, 109(16):6241–6246, 2012.

[167] Verena J. Schuenemann et al. Genome-wide comparison of medieval and modern mycobacterium leprae. *Science*, 341(6142):179, 2013.

[168] J. Schultz et al. A common core of secondary structure of the internal transcribed spacer 2 (its2) throughout the eukaryota. *RNA*, 11, 2005.

[169] Jorg Schultz et al. The internal transcribed spacer 2 databasea web server for (not only) low level phylogenetic analyses. *Nucleic Acids Research*, 34(Supplement 2):W704–W707, 2006.

[170] A. Seper et al. Vibrio cholerae evades neutrophil extracellular traps by the activity of two extracellular nucleases. *PLOS Pathogens*, 9, 2013.

[171] M. Shirato et al. Observations of the surface of dental calculus using scanning electron microscopy. *Journal of Nihon University School of Dentistry*, 23, 1981.

[172] A. K. Singla and Kamla Pathak. Phytoconstituents of euphorbia species. *Fitoterapia*, 41(6):483–516, 1990.

[173] Andrea Michel Sobottka et al. Proteinase activity in latex of three plants of the family euphorbiaceae. *Brazilian Journal of Pharmaceutical Sciences*, 50(3), 2014.

[174] Sigmund S. Socransky et al. Dental biofilms: difficult therapeutic targets. *Periodontology 2000*, 28:2–55, 2002.

[175] Rochelle M. Soo et al. An expanded genomic representation of the phylum cyanobacterium. *Genome Biol Evol*, 6(5):1031–1045, 2014.

[176] Rochelle M. Soo et al. Back from the dead; the curious tale of the predatory cyanobacterium vampirovibrio chlorellavorus. *PeerJ*, 3:e968, 2015.

[177] M.E. Sosa Torres et al. *Sustaining Life on Planet Earth: Metalloenzymes Mastering Dioxygen and Other Chewy Gases. Metal Ions in Life Sciences*, chapter The magic of dioxygen. Springer, 2015.

[178] Alexandros Stamatakis. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, 2014.

[179] Christen Rune Stensvold et al. Blastocystis sp. subtype 4 is common in danish blastocystis-positive patients presenting with acute diarrhea. *The American Journal of Tropical Medicine and Hygiene*, 84:883–885, 2011.

[180] Kevin S. W. Tan. New insights on classification, identification, and clinical relevance of blastocystis spp. *Clinical Microbiology Reviews*, 21:639–665, 2008.

[181] O. Tange. Gnu parallel - the command-line power tool. *The USENIX Magazine*, pages 42–47, February 2011.

[182] Raul Y. Tito et al. Insights from characterizing extinct human gut microbiomes. *PloS one*, 7(12):e51146, 2012.

[183] Paul R. Torgerson et al. World health organization estimates of the global and regional disease burden of 11 foodborne parasitic diseases, 2010: A data synthesis. *PLOS Medicine*, 12(12):1–22, 12 2015.

[184] Joseph D. Turner et al. Intensity of intestinal infection with multiple worm species is related to regulatory cytokine output and immune hyporesponsiveness. *The Journal of Infectious Diseases*, 197(8):1204–1212, 2008.

[185] Ashild J. Vagene et al. Salmonella enterica genomes from victims of a major sixteenth-century epidemic in mexico. *Nature Ecology & Evolution*, 2018.

[186] Janneke H.H.M. van de Wijgert. The vaginal microbiome and sexually transmitted infections are interlinked: Consequences for treatment and prevention. *PLOS Medicine*, 14:1–4, 2017.

[187] Lynne van Herwerden et al. Intra- and interindividual variation in its1 ofparagonimus westermani(trematoda: Digenea) and related species: Implications for phylogenetic studies. *Molecular Phylogenetics and Evolution*, 12(1):67 – 73, 1999.

[188] Irina M. Velsko et al. The dental calculus metabolome in modern and historic samples. *Metabolomics*, 13(134), 2017.

[189] Tomas Vetrovsky and Petr Baldrian. The variability of the 16s rrna gene in bacterial genomes and its consequences for bacterial community analyses. *PLOS ONE*, 8(2):1–10, 02 2013.

[190] R. Vilas et al. A comparison between mitochondrial dna and the ribosomal internal transcribed regions in prospecting for cryptic species of platyhelminth parasites. *Parasitology*, 131(6):839846, 2005.

[191] J. von der Schulenburg et al. Extreme length and length variation in the first ribosomal internal transcribed spacer of ladybird beetles (coleoptera: Coccinellidae). *Molecular Biology and Evolution*, 18(4):648–660, 2001.

[192] Helge L. Waldum et al. *Helicobacter pylori* and gastric acid: an intimate and reciprocal relationship. *Therapeutic Advances in Gastroenterology*, 9(6):836–844, 2016.

[193] Yingying Wang et al. Quantification of the filterability of freshwater bacteria through 0.45, 0.22, and 0.1 m pore size filters and shape-dependent enrichment of filterable bacterial communities. *Environmental Science & Technology*, 41(20):7080–7086, 2007. PMID: 17993151.

[194] Christina Warinner et al. Pathogens and host immunity in the ancient human oral cavity. *Nature genetics*, 46(4):336–344, 2014.

[195] Christina Warinner et al. Ancient human microbiomes. *Journal of Human Evolution*, 79:125–136, 2015.

[196] Christina Warinner et al. A new era in palaeomicrobiology: prospects for ancient dental calculus as a long-term record of the human oral microbiome. *Phil. Trans. R. Soc. B*, 370(1660):20130376, 2015.

[197] Christina Warinner et al. A robust framework for microbial archaeology. *Annual Review of Genomics and Human Genetics*, 18(1):321–356, 2017.

[198] Michael D Wasserman et al. Estrogenic plant consumption predicts red colobus monkey (procolobus rufomitratus) hormonal state and behavior. *Hormones and Behavior*, 62:553–562, 2012.

[199] B.M. Watkins. Drugs for the control of parasitic diseases: Current status and development of schistosomiasis. *Trends in Parasitology*, 19, 2003.

[200] P. S. Watson et al. Penetration of fluoride into natural plaque biofilms. *Journal of Dental Research*, 84, 2005.

[201] Ivan Wawrzyniak et al. Complete circular dna in the mitochondria-like organelles of blastocystis hominis. *International Journal for Parasitology*, 38(12):1377 – 1382, 2008.

[202] Laura Wegener Parfrey et al. Communities of microbial eukaryotes in the mammalian gut within the context of environmental eukaryotic diversity. *Frontiers in Microbiology*, 5, 2014.

[203] Christopher M. Whipps et al. Myxobolus cerebralis internal transcribed spacer 1 (its-1) sequences support recent spread of the parasite to north america and within europe. *Diseases of Aquatic Organisms*, 60:105–108, 2004.

[204] Donald J. White. Processes contributing to the formation of dental calculus. *Biofouling*, 4(1-3):209–218, 1991.

[205] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York, 2009.

[206] Paul E. Wischmeyer et al. Role of the microbiome, probiotics, and 'dysbiosis therapy' in critical illness. *Current Opinion in Critical Care*, 22:347–353, 2016.

[207] Carl R. Woese and George E. Fox. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *PNAS*, 74:5088–5090, 1977.

[208] Chao Yan et al. Impact of environmental factors on the emergence, transmission and distribution of toxoplasma gondii. *Parasite Vectors*, 9:137, 2016.

[209] Pablo Yarza et al. Uniting the classification of cultured and uncultured bacteria and archaea using 16s rrna gene sequences. *Nature Reviews: Microbiology*, 12:635–645, 2014.

[210] Tanya Yatsunenko et al. Human gut microbiome viewed across age and geography. *Nature*, 486, 2012.

[211] S.H. Yoon et al. Introducing ezbiocloud: a taxonomically united database of 16s rrna and whole genome assemblies. *Int J Syst Evol Microbiol*, 67:1613–1617, 2017.

[212] Jiajie Zhang et al. Pear: a fast and accurate illumina paired-end read merger. *Bioinformatics*, 30(5):614–620, 2014.

[213] Ning Zhang et al. Microorganisms in the gut of beetles: evidence from molecular cloning. *Journal of Invertebrate Pathology*, 84:226–233, 2012.

[214] K. A. Ziesemer et al. Intrinsic challenges in ancient microbiome reconstruction using 16s rrna gene amplification. *Scientific Reports*, 5, 2015.

Appendix A

Microeukaryotic and dietary survey of the gut by internal transcribed spacer metabarcoding

Code

```
#!/usr/bin/python3

'''Useage: python itsPrimerCheck.py input_R1.fastq input_R2.fastq
↪  sampleName'''

import sys
import gzip
from Bio import SeqIO
from itertools import zip_longest

primer_its1f = "TCCGTAGGTGAACCTGCGG"
primer_its2r = "GCTGCGTTCTTCATCGATGC"
primer_its3f = "GCATCGATGAAGAACGCAGC"
primer_its4r = "TCCTCCGCTTATTGATATGC"

r1_its1 = []
r1_its2 = []
```

```
r2_its1 = []
r2_its2 = []

readiter_R1 = SeqIO.parse(open(sys.argv[1]), "fastq")
readiter_R2 = SeqIO.parse(open(sys.argv[2]), "fastq")

for rec1, rec2 in zip_longest(readiter_R1, readiter_R2):
        #check for normal configuration
        if primer_its1f in rec1.seq[0:24] and primer_its2r in
↪   rec2.seq[0:24]:
                r1_its1.append(rec1)
                r2_its1.append(rec2)
        if primer_its3f in rec1.seq[0:24] and primer_its4r in
↪   rec2.seq[0:24]:
                r1_its2.append(rec1)
                r2_its2.append(rec2)
        #check for opposite configuration
        if primer_its2r in rec1.seq[0:24] and primer_its1f in
↪   rec2.seq[0:24]:
                r1_its1.append(rec2)
                r2_its1.append(rec1)
        if primer_its4r in rec1.seq[0:24] and primer_its3f in
↪   rec2.seq[0:24]:
                r1_its2.append(rec2)
                r2_its2.append(rec1)
        #check for one of the seqs not having primer seq, keep both
        if primer_its1f in rec1.seq[0:24] and primer_its2r not in
↪   rec2.seq[0:24]:
                r1_its1.append(rec1)
```

```
                r2_its1.append(rec2)
        if primer_its1f in rec2.seq[0:24] and primer_its2r not in
↪  rec1.seq[0:24]:
                r1_its1.append(rec2)
                r2_its1.append(rec1)
        if primer_its3f in rec1.seq[0:24] and primer_its4r not in
↪  rec2.seq[0:24]:
                r1_its2.append(rec1)
                r2_its2.append(rec2)
        if primer_its3f in rec2.seq[0:24] and primer_its4r not in
↪  rec1.seq[0:24]:
                r1_its2.append(rec2)
                r2_its2.append(rec1)

with open('%s_ITS1_R1.fastq' % sys.argv[3], "w") as outITS1_R1:
        SeqIO.write(r1_its1, outITS1_R1, "fastq")
outITS1_R1.close()

with open("%s_ITS1_R2.fastq" % sys.argv[3], "w") as outITS1_R2:
        SeqIO.write(r2_its1, outITS1_R2, "fastq")
outITS1_R2.close()

with open("%s_ITS2_R1.fastq" % sys.argv[3], "w") as outITS2_R1:
        SeqIO.write(r1_its2, outITS2_R1, "fastq")
outITS2_R1.close()

with open("%s_ITS2_R2.fastq" % sys.argv[3], "w") as outITS2_R2:
        SeqIO.write(r2_its2, outITS2_R2, "fastq")
outITS2_R2.close()
```

Table A.1: **Genera exclusively detected by ITS1 or ITS2**

ITS1		**ITS2**	
Genera	*Type*	*Genera*	*Type*
Acremonium	Fungi	Anaeromyces	Fungi
Blastocystis	Heterokont	Ascobolus	Fungi
Brassica	Plant	Auricularia	Fungi
Ceratobasidium	Fungi	Bensingtonia	Fungi
Cercophora	Fungi	Brettanomyces	Fungi
Cladosporium	Fungi	Caecomyces	Fungi
Clavaria	Fungi	Chalara	Fungi
Clavispora	Fungi	Chloroidium	Algae
Cucumis	Plant	Coprinopsis	Fungi
Cynodon	Plant	Cryptococcus	Fungi
Cyrenella	Fungi	Cyphellophora	Fungi
Devriesia	Fungi	Desmococcus	Algae
Dichondra	Plant	Euphorbia	Plant
Dicyma	Fungi	Fusariella	Fungi
Edenia	Fungi	Lepidocyrtus	Animal
Entomophthora	Fungi	Lepiota	Fungi
Fusarium	Fungi	Leptogium	Fungi
Glomus	Fungi	Maclura	Plant
Heveochlorella	Algae	Mimusops	Plant
Hydrocotyle	Plant	Neurospora	Fungi
Ipomoea	Plant	Nigrospora	Fungi

Table A.1 continued from previous page

Lachancea	Fungi	Oontomyces	Fungi
Microbotryum	Fungi	Orpinomyces	Fungi
Microdochium	Fungi	Panama	Fungi
Panaeolus	Fungi	Phallus	Fungi
Parasympodiella	Fungi	Phoma	Fungi
Penidiella	Fungi	Pichia	Fungi
Phaeococcomyces	Fungi	Piromyces	Fungi
Phaseolus	Plant	Podospora	Fungi
Phlebia	Fungi	Psathyrella	Fungi
Preussia	Fungi	Pseudoacremonium	Fungi
Protomyces	Fungi	Pseudozyma	Fungi
Pseudocercosporella	Fungi	Pyrenochaetopsis	Fungi
Rauvolfioideae	Plant	Rhizopus	Fungi
Readeriella	Fungi	Roussoella	Fungi
Rhizophlyctis	Fungi	Sakaguchia	Fungi
Saccharomycopsis	Fungi	Schistosoma	Trematode
Septoria	Fungi	Schizosaccharomyces	Fungi
Sesamum	Plant	Simplicimonas	Parabasalid
Stagonospora	Fungi	Sphenophorus	Animal
Stichococcus	Algae	Spinacia	Plant
Tetracladium	Fungi	Sterigmatomyces	Fungi
Tilletia	Fungi	Tetraplosphaeria	Fungi
Trifolieae	Plant	Tetratrichomonas	Parabasalid

Table A.1 continued from previous page

Tulasnella	Fungi	Theobroma	Plant
Urera	Plant	Vanguerieae	Plant
Zea	Plant	–	–

Table A.2: **Sequencing results**. The number of raw reads generated for each ITS target region, the number of those reads that pass quality filtering and merging steps, and the percentage of raw reads that were useable for downstream analysis.

Sample	Raw ITS1	Raw ITS2	Merged ITS1	Merged ITS2	Proportion Merged ITS1	Proportion Merged ITS2
HS2374	46,985	134,711	33,164	99,034	0.71	0.74
HS2416	47,433	266,009	39,990	195,437	0.84	0.73
HS2363	6,880	7,738	5,190	5,700	0.75	0.74
HS2380	1,275	38,916	1,132	33,456	0.89	0.86
HS2432	5,572	48,181	4,901	41,356	0.88	0.86
HS2446	5,381	187,306	3,103	12,8028	0.58	0.68
SM05	86,869	459,960	84,268	447,912	0.97	0.97
SM29	103,736	272,961	100,366	258,119	0.97	0.95
SM01	288,580	192,036	275,909	184,591	0.96	0.96
SM02	283,089	191,180	250,605	187,681	0.89	0.98
SM31	110,603	255,617	108,282	249,198	0.98	0.97

Table A.2 continued from previous page

SM32	191,162	117,135	177,101	114,212	0.93	0.98
NO7	191,162	117,135	177,101	114,212	0.93	0.98
NO15	110,603	255,617	108,282	249,198	0.98	0.97
NO16	6,880	7,738	5,190	5,700	0.75	0.74
NO20	1,275	38,916	1,132	33,456	0.89	0.86
BO2072	156,080	82,643	142,468	79,503	0.91	0.96
RC2109	94,102	51,610	87,030	49,597	0.92	0.96

Appendix B

Enrichment of non–dominant bacterial taxa in human fecal samples through serial filtration

Code

```
#!/usr/bin/python

import pandas as pd

otumap = pd.read_csv("map_otu97.txt", sep="\t")
grouped = otumap.groupby("OTU")["READ"].apply(lambda x: "%s" %
↪  ','.join(x))

with open("otuMap.txt", "w") as outfile:
        grouped.to_csv(outfile, sep="\t")
```

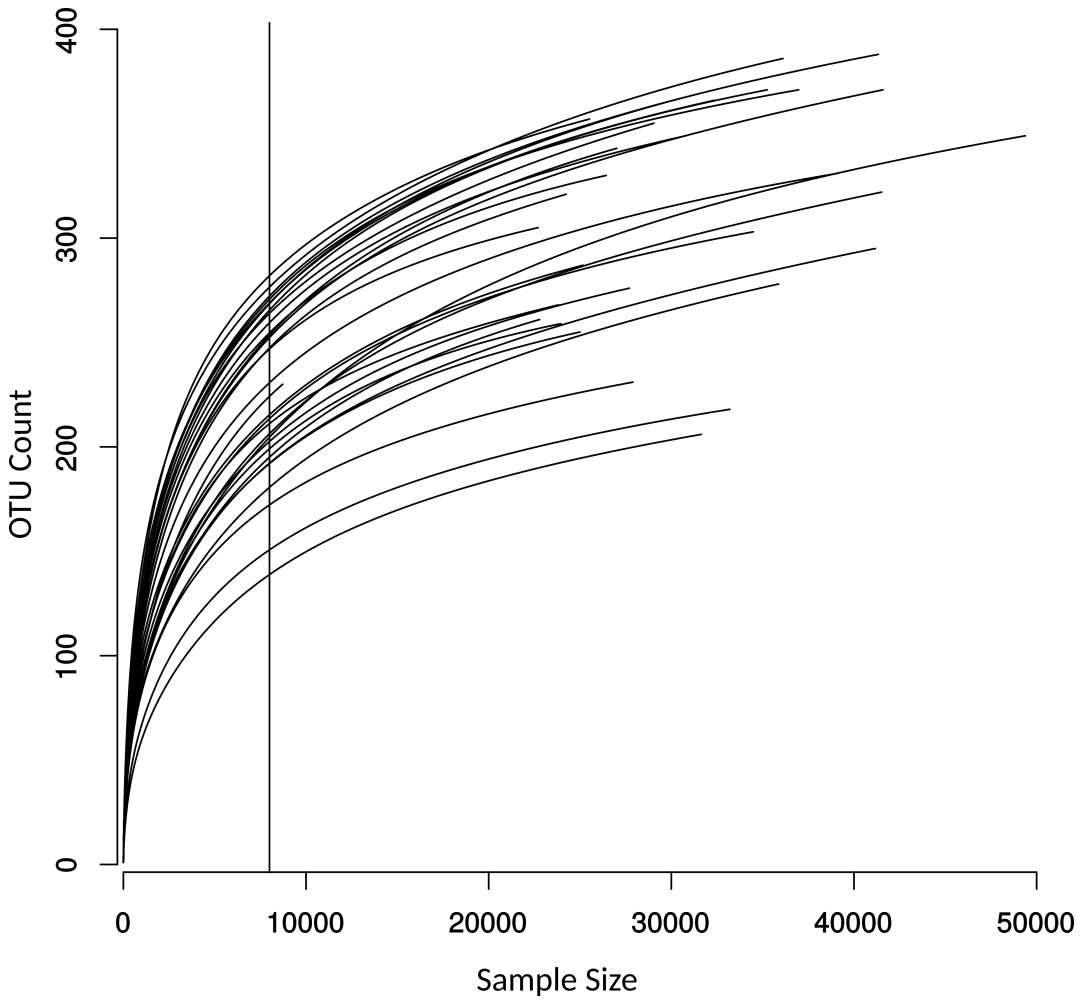

Figure B.1: **Rarefaction analysis at 8,000 read depth.** Each line represents a single true sample, vertical line represents a 8,000 rarefaction depth.

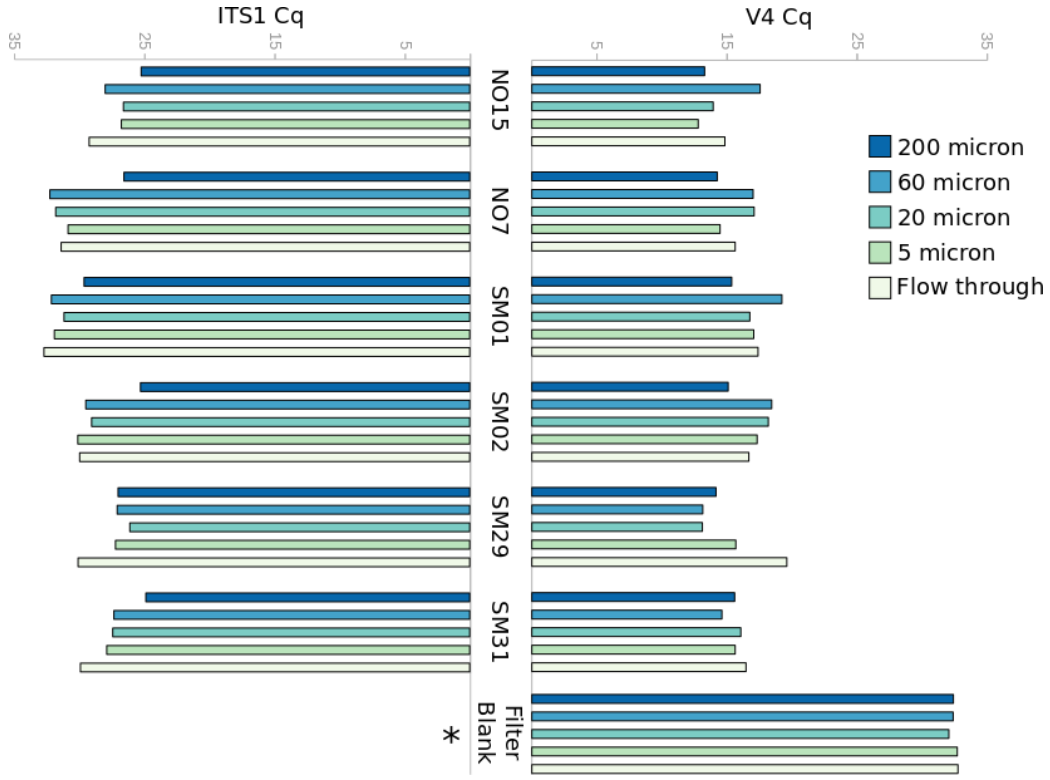

Figure B.2: **Cq value for each filter level in all samples as measured by V4 and ITS1 targeted qPCR.** A lower Cq value for either the ITS1 or V4 region indicates an earlier amplification and thus a higher frequency of the targeted taxonomic group. * No ITS regions were amplified in the filter blank.

Table B.1: **Cq value changes over filter levels with *Escherichia coli* standard**. Cq value is lowest at 200μm but remains relatively consistent over 60μm and 20μm level filters indicating that although these filter sizes are larger than an average bacterium, individual cells are trapped at higher than expected filter levels.

Filter Level	V4
200μm	19.97
60μm	14.70
20μm	14.65
5μm	17.77
Flow Through	17.56

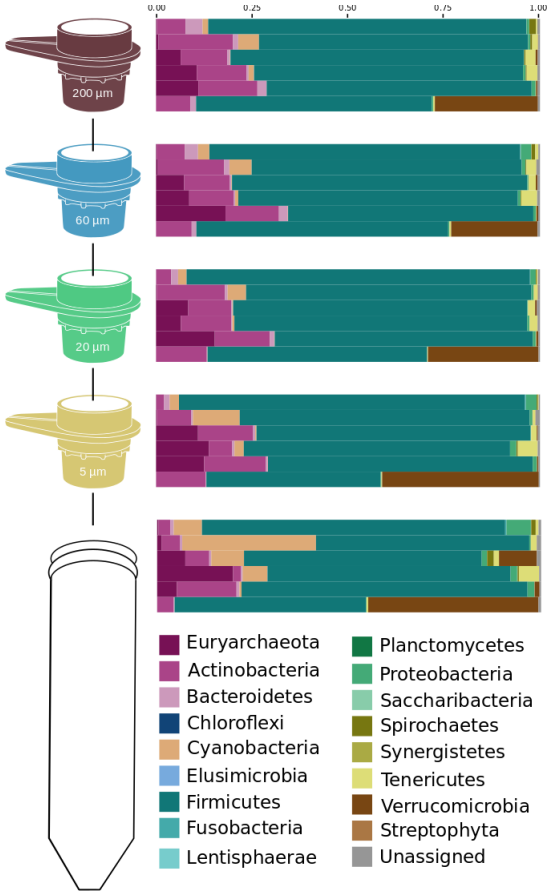

Figure B.3: **Phylum level taxonomic distribution of each filter level.** Illustration of the filtering process with phylum–level taxonomic barcharts for each sample at the corresponding filter level. Phyla remain consistent across each filter level for an individual sample except for *Cyanobacteria*, *Verrucomicrobia*, and to a lesser extent, *Tenericutes*, which are enriched at lower filter levels.

Appendix C

Differential preservation of endogenous human and microbial DNA in dental calculus and dentin

Code

```
  #!/usr/bin/python

'''Usage: python gcLenCorPlots.py -i <input fasta or fastq> [-m
↪  <method> -r <range for heatmap> -t <trim maximum length> -s
↪  <normalize to number> -ec <error bar color>]'''

import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import argparse
import scipy

from Bio import SeqIO
from Bio.SeqUtils import GC
from scipy.stats import mannwhitneyu
from scipy.signal import resample
```

```
parser = argparse.ArgumentParser()
parser.add_argument('-i', '--input', help='either a fastq or fasta
 ↪  file, must end with .fasta, .fna, .fa, .fastq, or .fq')
parser.add_argument('-m', '--method', help='options: mean, median',
 ↪  default='mean')
parser.add_argument('-ec', '--errorbarColor', help='desired error
 ↪  bar color in hex color, default is grey', default='grey')
parser.add_argument('-r', '--range', help='Range setting for the
 ↪  color bar. Accepted arguments: num, perc, max. Num colors by
 ↪  the absolute number of reads ranging from 100 to 2k, perc
 ↪  colors by percentage of total, max colors based on minimum and
 ↪  maximum read counts', default='max')
parser.add_argument('-t', '--trim', help='Maximum length trim,
 ↪  numeric')
parser.add_argument('-s', '--shuffle', help='Randomly shuffle
 ↪  results to a specific number')
args = parser.parse_args()

fastaEnds = ('.fasta', '.fna', '.fa')
fastqEnds = ('.fastq', '.fq')

if args.input.endswith(fastaEnds):
        gcContent = [GC(rec.seq) for rec in SeqIO.parse(args.input,
 ↪  "fasta")]
        lens = [len(rec) for rec in SeqIO.parse(args.input,
 ↪  "fasta")]
elif args.input.endswith(fastqEnds):
        gcContent = [GC(rec.seq) for rec in SeqIO.parse(args.input,
 ↪  "fastq")]
```

```
        lens = [len(rec) for rec in SeqIO.parse(args.input,
↪  "fastq")]
else:
        print("File extension not recognized, see help file")
print("Number of reads: %i" % len(gcContent))
df = pd.DataFrame({'length': lens, 'gcContent': gcContent})

#optional trimming/shuffle options
if args.trim is not None:
        df = df.drop(df[df.length >
↪  int(args.trim)].index).reset_index()
        print("Length trimmed to maximum %i" % int(args.trim))
        print("Number of trimmed reads: %i" % len(df.gcContent))

if args.shuffle is not None:
        df = df.sample(int(args.shuffle))
        print("Number of reads normalized to %i" %
↪  int(args.shuffle))

#stats
print("Mean GC content: %.2f" % np.mean(df['gcContent']))
print("Median GC content: %.2f" % np.median(df['gcContent']))
print("Mean fragment length: %i" % np.mean(df['length']))
print("Median fragment length: %i" % np.median(df['length']))
print("Fragment length range: %i : %i" % (min(df['length']),
↪  max(df['length'])))
print("GC content range: %.2f : %.2f" % (min(df['gcContent']),
↪  max(df['gcContent'])))
```

```
#grouped data
dfGrouped = df.groupby(by='length').agg(['count', 'mean', 'median',
↪  'std']).reset_index()
dfGrouped['perc'] = dfGrouped['gcContent',
↪  'count']/dfGrouped['gcContent', 'count'].sum()

#print out data
with open('%s_data_out.txt' % args.input, 'w') as outfile:
        dfGrouped.to_csv(outfile, sep="\t", index=False)

#calculate significance of each grouping, compared to overall
↪  distribution
dfLensGroup = df.groupby(by='length')
overall = df['gcContent']
lines = []

#limit options
plt.xlim(15, 90)
plt.ylim(20, 200)
plt.suptitle(args.input + "\n" + "n= " + str(dfGrouped['gcContent',
↪  'count'].sum()))
plt.xlabel('GC content (%)')
plt.ylabel('Read length (bp)')

#plot error bar
cm = plt.cm.get_cmap('YlOrRd')
plt.errorbar(dfGrouped['gcContent', args.method],
↪  dfGrouped['length', ''], xerr=dfGrouped['gcContent', 'std'],
↪  linestyle="None", marker="None", color=args.errorbarColor)
```

```
#color range options
if args.range == 'num':
        #plot by number of reads, range from 1k to 10k
        plt.scatter(dfGrouped['gcContent', args.method],
 ↪  dfGrouped['length', ''], c=list(dfGrouped['gcContent',
 ↪  'count']), cmap=cm, vmin=100, vmax=2000, marker='o',
 ↪  edgecolors='None', s=25, zorder=2)
elif args.range == 'perc':
        #plot by percentage instead
        plt.scatter(dfGrouped['gcContent', args.method],
 ↪  dfGrouped['length', ''], c=list(dfGrouped['perc']), cmap=cm,
 ↪  vmin=0.0, vmax=1.0, marker='o', edgecolors='None', s=25,
 ↪  zorder=2)
elif args.range == 'max':
        #plot by min to max count
        plt.scatter(dfGrouped['gcContent', args.method],
 ↪  dfGrouped['length', ''], c=list(dfGrouped['gcContent',
 ↪  'count']), cmap=cm, vmin=min(dfGrouped['gcContent', 'count']),
 ↪  vmax=max(dfGrouped['gcContent', 'count']), marker='o',
 ↪  edgecolors='None', s=25, zorder=2)

plt.colorbar()
plt.draw()
plt.savefig('%s_plot.pdf' % args.input)
```

Human Read Validation Reads mapped to the hg19 human reference genome using sensitive BWA [110] mapping parameters (-n 0.01, -l 1000,

-q 30) were de–replicated using DeDup as implemented in EAGER version 1.92 [146]. Next, de–replicated bam files were converted to fastq format using bedtools (bedtools bamtofastq) [154] and where necessary split into forward, reverse, and merged reads. Three base pairs were trimmed from either side of all merged reads using the FASTX–Toolkit (http://hannonlab.cshl.edu). Forward reads were trimmed of three base pairs exclusively on the 5 end of the read while reverse pair reads were trimmed for three base pairs on the 3 end. Trimmed reads were then remapped to the hg19 human reference genome using BWA with a higher mismatch penalty and quality mapping threshold so that approximately one mismatch would be allowed per 50 bases (-n 0.2, -l 1000, -q 37). The reads passing this quality threshold filter have a higher level of confidence of their proper assignment to the human genome. To further test the validity of these reads, they were then run through a lowest–common–ancestor algorithm via MALT using the full NCBI NT database with a percent identity threshold of 90%. Those reads that were assigned to the *Homo sapiens* node are recorded in Appendix Table C.2. Finally, to test whether these high confidence human reads are in fact ancient and not the result of background contamination, the original reads pre-damage trimming were pulled from the original fastq files, mapped to the hg19 human reference genome using BWA (-n 0.01, -l 1000, -q 30) and then assessed for terminal cytosine deamination patterns using mapDamage version 2.0 [92]. For all calculus samples and most dentin samples, the percent endogenous and damage patterns both pre and post strict map filtering are comparable, confirming the observed pattern of a low but consistent human aDNA retrieval

from calculus and a variable yield from dentin (Appendix C.2).

Bacterial Fragment Analyses Reads that mapped to the species node and all higher resolution taxonomic nodes were extracted from the MALT results for all bacteria of interest. To limit the impact of erroneous mapping only those reads with damage at the terminus of the read were considered. For all fragment length analyses, only merged reads were analyzed.

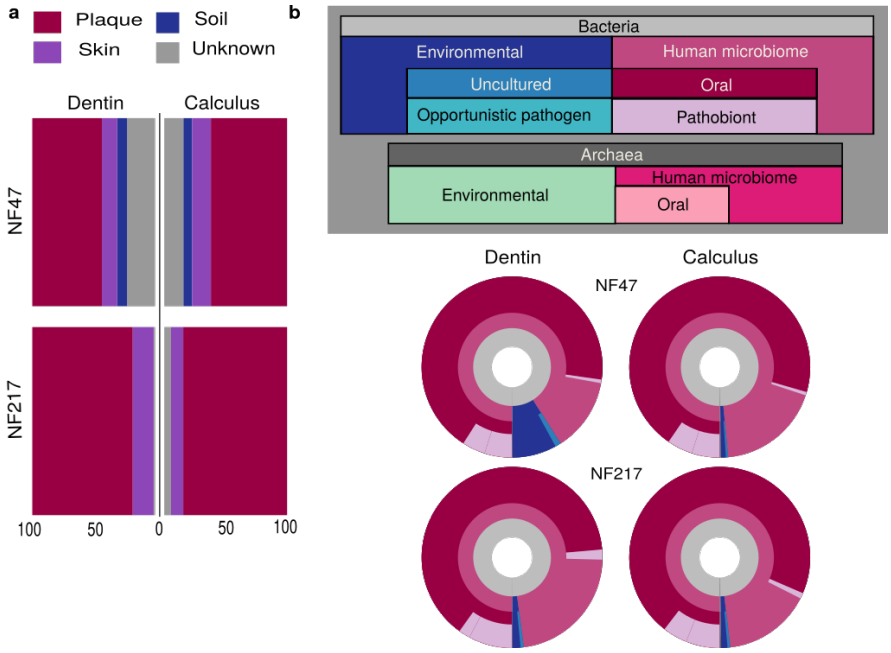

Figure C.1: **Likely signal of carious lesions on two dentin samples** (a) Bar chart displaying Bayesian SourceTracker results of estimated genetic contributions from human oral and environmental microbial sources. (c) Donut plots constructed from nested classification of species-level MALT results. The microbial communities in the Norris Farms dentin and calculus samples have a high oral contribution, as demonstrated by both the SourceTracker bar plots and the MALT classification donut plots.

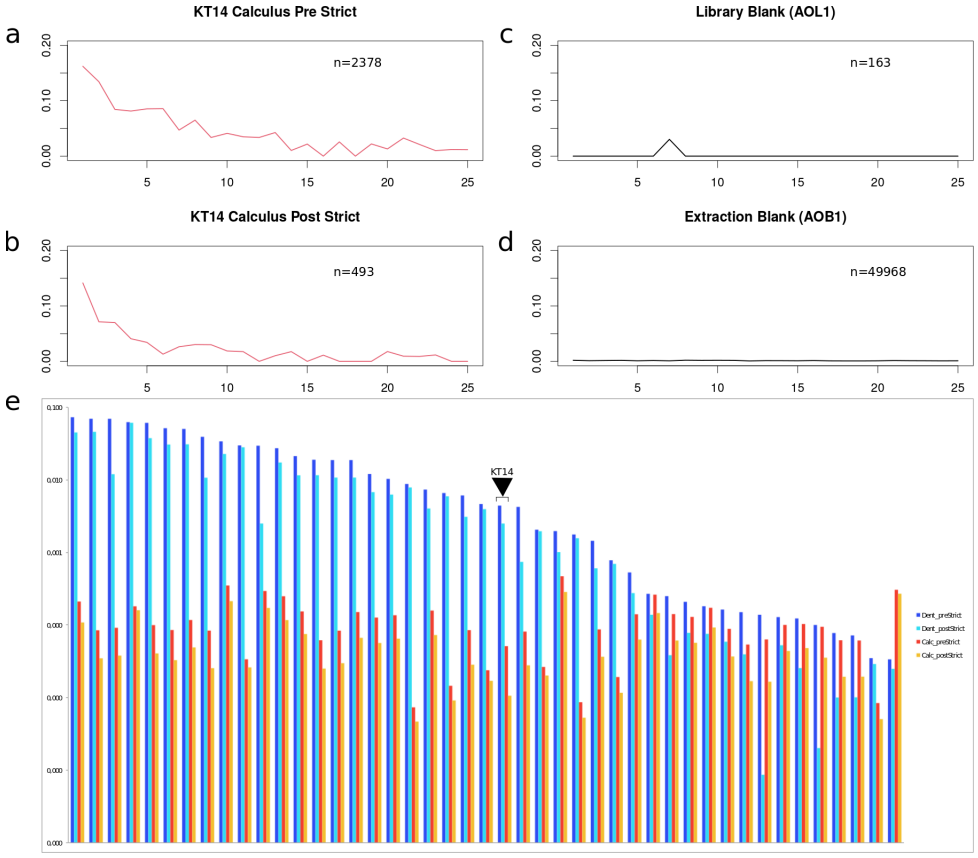

Figure C.2: **Validation of ancient human DNA authenticity.** (a) Terminal cytosine deamination rates of reads mapped to the human genome from a single calculus sample (KT14) before extra human validation steps. (b) Damage rate of human reads from calculus sample KT14 after human validation. While the damage rate of human reads post-validation drops, a clear damage signal, consistent with authentic ancient DNA is still observed. (c-d) Terminal cytosine deamination of a library build and extraction blank. Neither blanks have an observed damage signal, consistent with modern DNA. (e) Proportion human endogenous content for all paired samples both before and after strict mapping (see Supplementary Methods). While in all cases the proportion of human endogenous content drops after strict mapping, the effect is minor for most samples. Major drops are detected in certain dentin samples.

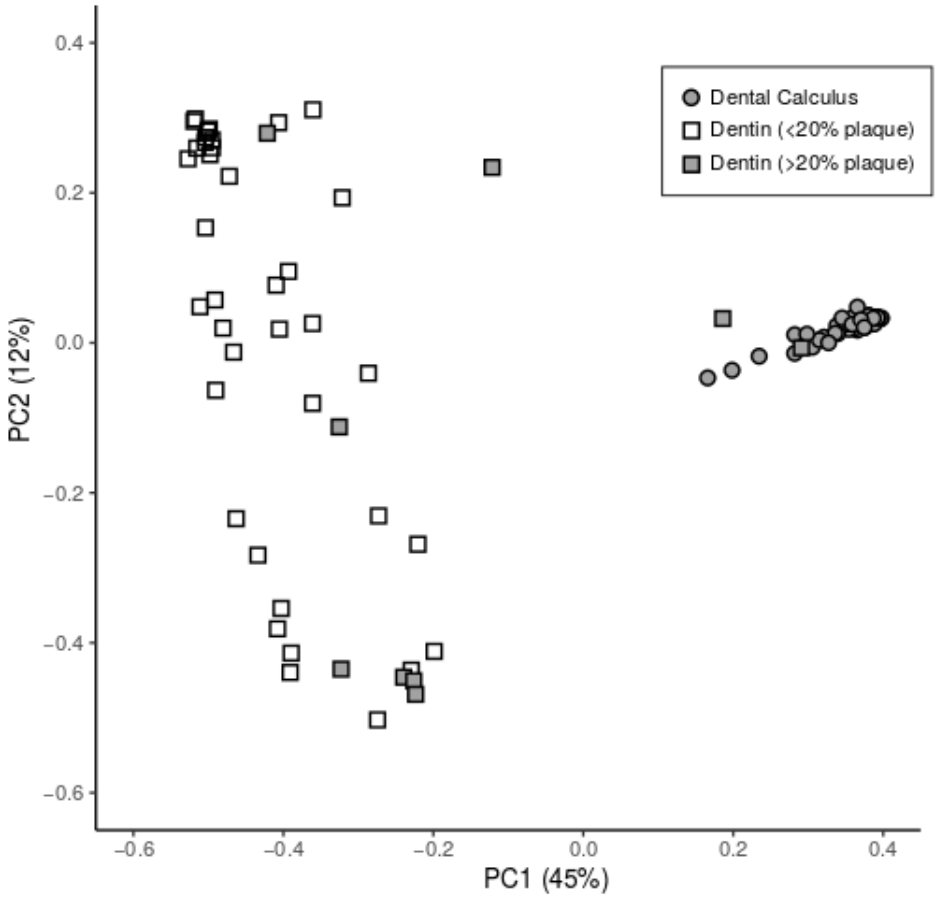

Figure C.3: **Principal Coordinates Analysis (PCoA) of Bray-Curtis distances of all bacterial and archaeal species-level assignments from dental calculus and dentin.** Dental calculus is represented as circles and dentin is represented as squares. Color indicates estimated contribution of oral taxa. Black symbols are those samples that have a predicted proportion of oral contribution of 20% or more, illustrating that some dentin samples have an unexpectedly high oral signature, though most do not cluster with the dental calculus samples, indicative of a non-biological community.

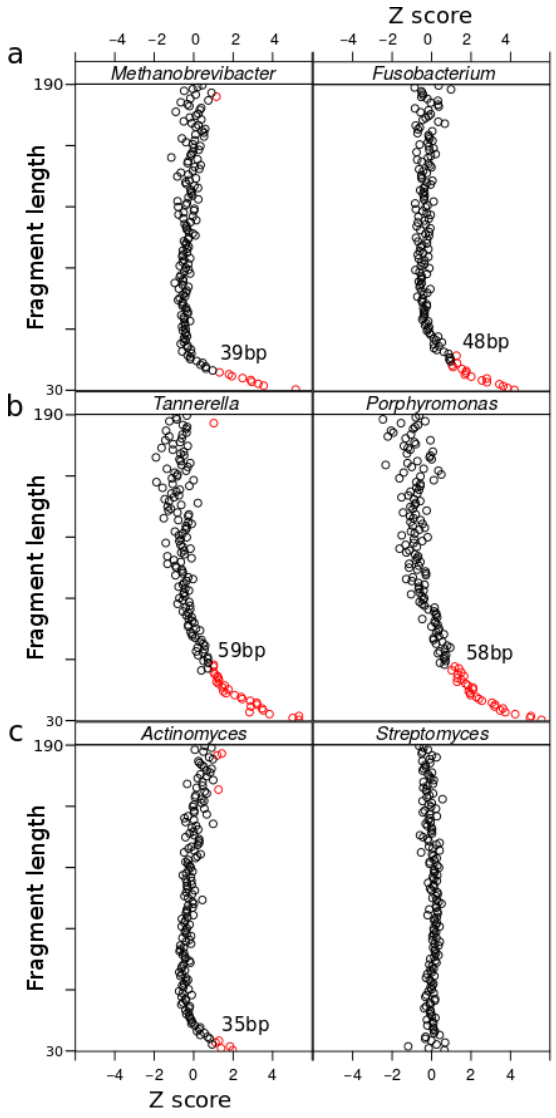

Figure C.4: **Fragment length deviation from expected mean GC content for selected bacterial genera.** Bacterial genera are organized by expected genomic GC content wherein (a) are low genomic GC taxa, (b) are moderate genomic GC taxa and, (c) are high genomic GC taxa. Each point represents a single length bins mean deviation from the overall mean of all reads mapped to the genus. Red points are those length bins that are one or more Z scores deviated from the mean GC content. For each genus the read length bin at which a major deviation can be seen (1 z score) is noted on the graph. For low or medium genomic GC content genera, this length threshold occurs at a higher fragment length than those with high genomic GC content.

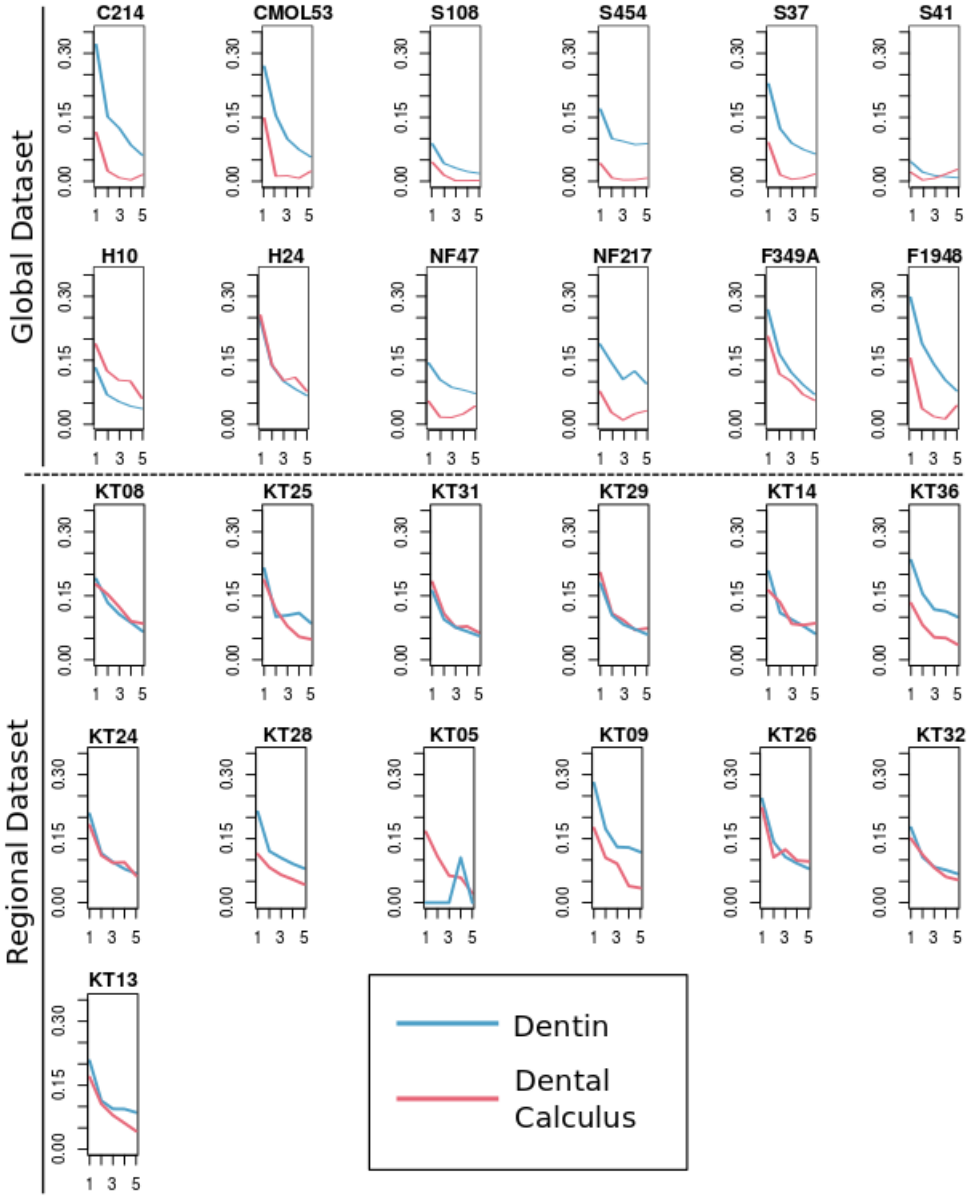

Figure C.5: **Differences in damage patterns among paired dentin and dental calculus is sample-specific.** Cytosine damage patterns for a subset of paired dentin (blue) and dental calculus (red) samples. While most dental calculus samples have a lower intial deamination rate than their dentin pair in the global dataset, this pattern is not consistently observed in the regional dataset, possibly the result of differences in laboratory preparation.

Table C.1: Sample metadata and sequencing statistics.

Country	Site	ID	Type	RawReads	QualFilter	AvLen	AvCopiesPerUl	ng/mg
Spain	Camino_del_Molino	C214	calculus	16562372	8688438	74.38	NA	NA
Spain	Camino_del_Molino	cmol53	calculus	44932086	24214108	83.26	26072816294.12	53.98
Guadeloupe	Anse_a_la_Gourde	F1948	calculus	18595784	9632823	78.97	115312571929.83	70.52
Guadeloupe	Anse_a_la_Gourde	F349A	calculus	22746512	11427674	66.73	NA	NA
Mongolia	Hovsgol	H10b	calculus	27888566	15530167	93.45	51369376363.64	70.97
Mongolia	Hovsgol	H24b	calculus	23362848	12049868	80.08	24180983122.36	21.46
Ireland	Kilteasheen	KT01	calculus	4911010	4755715	58.63	3984942405.06	124.9
Ireland	Kilteasheen	KT01	calculus	5702580	5501834	63.85	10122933333.33	2.31
Ireland	Kilteasheen	KT02	calculus	4020488	3821086	53.19	151993633136.1	64.67
Ireland	Kilteasheen	KT03	calculus	3981083	3816316	57.98	144750000000	206.25
Ireland	Kilteasheen	KT04	calculus	4553768	4374085	55.57	118946233082.71	73.89
Ireland	Kilteasheen	KT05	calculus	4667356	4379057	51.39	32640000000	15.4
Ireland	Kilteasheen	KT06	calculus	5886164	5624929	56	38862670807.45	73.18
Ireland	Kilteasheen	KT07	calculus	5282583	5078760	56.38	21297238709.68	85.52
Ireland	Kilteasheen	KT08	calculus	4600075	4433489	61.13	21200200947.87	127.88

Table C.1 continued from previous page

Ireland	Kilteasheen	KT09	calculus	4707428	4501061	54.52	120493242553.19	53.71
Ireland	Kilteasheen	KT10	calculus	4859064	4617930	54.5	36647068965.52	18.81
Ireland	Kilteasheen	KT11	calculus	4542147	4316220	52.39	32461496815.29	24.15
Ireland	Kilteasheen	KT12	calculus	3807537	3638561	55.77	54585858490.57	21.74
Ireland	Kilteasheen	KT13	calculus	4178637	3979576	53.88	10902068965.52	4.94
Ireland	Kilteasheen	KT14	calculus	4825668	4646770	51.19	2056833333.33	28.7
Ireland	Kilteasheen	KT15	calculus	7618554	7301967	57.32	8705424150.94	73.61
Ireland	Kilteasheen	KT16	calculus	4005185	3818784	52.76	37812869767.44	23.89
Ireland	Kilteasheen	KT17	calculus	3659112	3495140	55.17	15081808609.27	43.77
Ireland	Kilteasheen	KT18	calculus	5737309	5494880	58.08	62621085937.5	53.33
Ireland	Kilteasheen	KT19	calculus	7054607	6799826	58.94	26244025000	104.35
Ireland	Kilteasheen	KT20	calculus	7403326	7082748	56.54	8806688294.93	98.64
Ireland	Kilteasheen	KT21	calculus	3936817	3772624	55.07	75454518260.87	54.76
Ireland	Kilteasheen	KT22	calculus	4605691	4410357	55.99	9995000000	20.6
Ireland	Kilteasheen	KT23	calculus	6802200	6461176	55.11	10344500000	24.2
Ireland	Kilteasheen	KT24	calculus	6478560	6196160	55.12	27390875921.91	214.42
Ireland	Kilteasheen	KT25	calculus	5428976	5219375	62.63	6854951458.33	132.73

Table C.1 continued from previous page

Ireland	Kilteasheen	KT26	calculus	5775371	5561060	57.25	1241075602.09	75.2
Ireland	Kilteasheen	KT27	calculus	4142968	3996320	61.21	12133675446.43	151.35
Ireland	Kilteasheen	KT28	calculus	5201864	5012681	56.36	2442698461.54	59.69
Ireland	Kilteasheen	KT29	calculus	6061552	5836482	59.3	44075813854.75	177.23
Ireland	Kilteasheen	KT30	calculus	4897216	4726141	58.52	2095413615.89	87.28
Ireland	Kilteasheen	KT31	calculus	9355963	9011366	60.37	22181513924.05	94.05
Ireland	Kilteasheen	KT32	calculus	5405125	5200487	57.31	17022138672.77	58.66
Ireland	Kilteasheen	KT33	calculus	6653276	6429041	59.74	44707812345.68	114.89
Ireland	Kilteasheen	KT34	calculus	5923944	5697691	57.99	11950376899.7	54.38
Ireland	Kilteasheen	KT35	calculus	4302295	4134490	59.7	10409058947.37	114.66
Ireland	Kilteasheen	KT36	calculus	5469103	5224291	58.13	6872992977.1	125.96
United_States	Norris_Farms	NF217	calculus	11561738	6141258	NA	NA	NA
United_States	Norris_Farms	NF47	calculus	11945018	6363413	NA	NA	NA
Netherlands	Middenbeemster	S108	calculus	13923962	7124982	75.52	NA	NA
Nepal	Samdzong	S37	calculus	17293694	8837686	72.37	NA	NA
Nepal	Samdzong	S41b	calculus	19199306	11089749	94.44	131895158469.95	133.9
Netherlands	Middenbeemster	S454	calculus	14907918	7803710	76.72	NA	NA

Table C.1 continued from previous page

Spain	Camino_del_Molino	C214	dentin	31732062	16090474	63.48	NA	NA
Spain	Camino_del_Molino	cmol53	dentin	131971688	74127692	79.99	85658577.98	0.94
Guadeloupe	Anse_a_la_Gourde	F1948	dentin	61151808	31757589	70.18	2243590145.99	1.67
Guadeloupe	Anse_a_la_Gourde	F349A	dentin	30558738	16566928	76.66	NA	NA
Mongolia	Hovsgol	H10b	dentin	61149106	34834067	90.85	6307049446.9	0.31
Mongolia	Hovsgol	H24b	dentin	66607316	37629597	86.79	701265495.87	3.15
Ireland	Kilteasheen	KT01	dentin	5979713	5771580	58.9	NA	NA
Ireland	Kilteasheen	KT02	dentin	4153951	4029632	59.13	519357717.39	5.21
Ireland	Kilteasheen	KT03	dentin	3670693	3553557	54.4	968759712.23	3.52
Ireland	Kilteasheen	KT04	dentin	4671204	4517813	56.4	1300396313.36	7.32
Ireland	Kilteasheen	KT05	dentin	3944620	3820087	49.95	575959448	35.71
Ireland	Kilteasheen	KT06	dentin	5674209	5542331	54.42	1379696067.42	3.01
Ireland	Kilteasheen	KT07	dentin	4078202	3939964	62.54	1696910139.86	4.13
Ireland	Kilteasheen	KT08	dentin	4830306	4656617	54.55	3361880620.16	2.47
Ireland	Kilteasheen	KT09	dentin	6612577	6429918	57.55	1047288385.83	6.95
Ireland	Kilteasheen	KT10	dentin	3238307	3122323	52.68	517163149.61	4.49
Ireland	Kilteasheen	KT11	dentin	4874064	4714739	44.99	2599155263.16	8.89

Table C.1 continued from previous page

Ireland	Kilteasheen	KT12	dentin	4888032	4749761	58.59	900163350.79	14.28
Ireland	Kilteasheen	KT13	dentin	5947000	5738015	57.26	954896397.06	10.84
Ireland	Kilteasheen	KT14	dentin	6173417	5905330	49.11	2565364077.67	1.47
Ireland	Kilteasheen	KT15	dentin	4507142	4349804	NF	6107454580.15	13.13
Ireland	Kilteasheen	KT16	dentin	4937769	4780420	58.4	4382678250	8.73
Ireland	Kilteasheen	KT17	dentin	6213480	6063041	55.68	3581822222.22	2.03
Ireland	Kilteasheen	KT18	dentin	4729610	4590927	54.48	1406499152.54	6.27
Ireland	Kilteasheen	KT20	dentin	4631119	4485429	62.3	3779428846.15	1.71
Ireland	Kilteasheen	KT21	dentin	4808710	4648332	57.95	2746567241.38	4.95
Ireland	Kilteasheen	KT22	dentin	3762440	3656881	53.44	1659647683.4	9.37
Ireland	Kilteasheen	KT23	dentin	4507686	4374541	52.15	1461548031.5	4.49
Ireland	Kilteasheen	KT24	dentin	5748119	5600847	57.76	12080000000	2.4
Ireland	Kilteasheen	KT25	dentin	4429591	4272549	49.69	23760699152.54	9.71
Ireland	Kilteasheen	KT26	dentin	5034345	4872126	NF	7336801242.24	1.5
Ireland	Kilteasheen	KT27	dentin	5703823	5519751	57.09	14034270833.33	4.75
Ireland	Kilteasheen	KT28	dentin	5387360	5235938	57.47	13018172185.43	2.92
Ireland	Kilteasheen	KT29	dentin	6729054	6545793	NF	8025619533.53	9.55

Table C.1 continued from previous page

Ireland	Kilteasheen	KT30	dentin	5027319	4878335	59.09	13381414110.43	7.87
Ireland	Kilteasheen	KT31	dentin	5168663	5017943	56.02	11038830645.16	4.9
Ireland	Kilteasheen	KT32	dentin	7313342	7119676	57.79	6487446864.69	9.34
Ireland	Kilteasheen	KT33	dentin	6549519	6372460	61.6	10088003257.33	9.81
Ireland	Kilteasheen	KT34	dentin	4745793	4552565	NF	3630243750	3.57
Ireland	Kilteasheen	KT35	dentin	5431255	5227859	54.09	18603957547.17	12.43
Ireland	Kilteasheen	KT36	dentin	5978692	5771519	63.28	10374055172.41	5.84
United_States	Norris_Farms	NF217	dentin	11248574	5803813	NA	NA	NA
United_States	Norris_Farms	NF47	dentin	15232364	7730991	NA	NA	NA
Netherlands	Middenbeemster	S108	dentin	54239912	32185296	97.76	NA	NA
Nepal	Samdzong	S37	dentin	103351972	55597338	72.37	NA	NA
Nepal	Samdzong	S41b	dentin	32332152	20055715	109.9	55875367.65	0.19
Netherlands	Middenbeemster	S454	dentin	43246346	25547745	93.08	NA	NA

Table C.2: Human verification statistics

ID	Type	Reads Pre	Reads Post	Endogenous Pre	Endogenous Post	Human Node Post	Damage Pre	Damage Post	Damage Untrim
C214	calculus	730	439	0.01	0.01	358	0.12	0.00	0.15
cmol53	calculus	1779	1130	0.01	0.00	923	0.15	0.02	0.21
F1948	calculus	832	509	0.01	0.01	420	0.15	0.00	0.21
F349A	calculus	2194	1330	0.02	0.01	1058	0.20	0.03	0.25
H10b	calculus	5238	4053	0.03	0.03	3661	0.18	0.06	0.18
H24b	calculus	1754	1100	0.01	0.01	948	0.25	0.09	0.29
KT01	calculus	8568	3800	0.16	0.07	3374	0.20	0.04	0.17
KT01	calculus	11844	5557	0.25	0.12	4919	0.22	0.04	0.17
KT02	calculus	2057	644	0.05	0.02	571	0.17	0.03	0.14
KT03	calculus	2356	958	0.06	0.03	846	0.17	0.02	0.14
KT04	calculus	6594	2916	0.15	0.07	2539	0.24	0.06	0.18
KT05	calculus	2771	724	0.06	0.02	634	0.17	0.01	0.17
KT06	calculus	19701	11998	0.35	0.21	10407	0.13	0.03	0.10

Table C.2 continued from previous page

KT07	calculus	6573	2893	0.13	0.06	2536	0.21	0.03	0.19
KT08	calculus	6238	2792	0.14	0.06	2455	0.19	0.04	0.15
KT09	calculus	6930	3381	0.15	0.08	2953	0.19	0.02	0.14
KT10	calculus	2828	897	0.06	0.02	788	0.03	0.02	0.01
KT11	calculus	4087	1530	0.09	0.04	1324	0.22	0.02	0.18
KT12	calculus	2246	705	0.06	0.02	616	0.08	0.01	0.06
KT13	calculus	3522	1467	0.09	0.04	1277	0.15	0.02	0.11
KT14	calculus	2378	493	0.05	0.01	424	0.17	0.04	0.14
KT15	calculus	21482	12568	0.29	0.17	11059	0.15	0.02	0.12
KT16	calculus	3247	1085	0.09	0.03	963	0.21	0.02	0.14
KT17	calculus	2985	1146	0.09	0.03	995	0.20	0.04	0.16
KT18	calculus	14420	8035	0.26	0.15	7100	0.17	0.03	0.14
KT19	calculus	19852	10900	0.29	0.16	9578	0.15	0.04	0.12
KT20	calculus	12241	6553	0.17	0.09	5772	0.14	0.03	0.12
KT21	calculus	4776	2134	0.13	0.06	1894	0.21	0.04	0.18
KT22	calculus	6243	2697	0.14	0.06	2381	0.25	0.04	0.20
KT23	calculus	5402	1924	0.08	0.03	1665	0.19	0.05	0.15

Table C.2 continued from previous page

KT24	calculus	7264	3043	0.12	0.05	2716	0.16	0.03	0.12
KT25	calculus	5239	2293	0.10	0.04	2057	0.18	0.03	0.16
KT26	calculus	4656	1414	0.08	0.03	1245	0.22	0.05	0.15
KT27	calculus	6325	2912	0.16	0.07	2578	0.23	0.04	0.18
KT28	calculus	6833	3253	0.14	0.06	2841	0.16	0.03	0.13
KT29	calculus	5343	2211	0.09	0.04	1969	0.19	0.03	0.15
KT30	calculus	4886	2276	0.10	0.05	2014	0.16	0.03	0.12
KT31	calculus	18973	9772	0.21	0.11	8602	0.15	0.04	0.12
KT32	calculus	5186	2114	0.10	0.04	1844	0.15	0.03	0.12
KT33	calculus	5439	2235	0.08	0.03	1981	0.18	0.05	0.13
KT34	calculus	4624	1587	0.08	0.03	1418	0.18	0.03	0.13
KT35	calculus	3602	1512	0.09	0.04	1335	0.15	0.02	0.13
KT36	calculus	24578	14924	0.47	0.29	13145	0.14	0.02	0.11
NF217	calculus	4712	3587	0.08	0.06	3028	0.07	0.02	0.08
NF47	calculus	3127	2101	0.05	0.03	1796	0.05	0.01	0.06
S108	calculus	1887	1435	0.03	0.02	1191	0.05	0.00	0.05
S37	calculus	2104	1505	0.02	0.02	1285	0.09	0.01	0.10

Table C.2 continued from previous page

S41b	calculus	20086	17727	0.18	0.16	15983	0.02	0.02	0.02
S454	calculus	23888	21051	0.31	0.27	17692	0.04	0.00	0.04
C214	dentin	5621	4674	0.03	0.03	4217	0.32	0.07	0.34
cmol53	dentin	6506624	5824828	8.78	7.86	5233216	0.27	0.06	0.27
F1948	dentin	561545	498419	1.77	1.57	448899	0.29	0.08	0.29
F349A	dentin	129335	115189	0.78	0.70	103171	0.26	0.07	0.26
H10b	dentin	10413259	9800087	29.89	28.13	9003402	0.13	0.04	0.12
H24b	dentin	2480085	2237287	6.59	5.95	2028196	0.24	0.06	0.24
KT01	dentin	1578342	1001384	27.35	17.35	880214	0.18	0.04	0.14
KT02	dentin	6060	1589	0.15	0.04	1425	0.35	0.04	0.29
KT03	dentin	672997	412331	18.94	11.60	358525	0.17	0.04	0.14
KT04	dentin	1317777	815218	29.17	18.04	717216	0.20	0.03	0.16
KT05	dentin	5288	33	0.14	0.00	27	0.00	0.00	0.00
KT06	dentin	1880891	1261317	33.94	22.76	1109933	0.14	0.03	0.12
KT07	dentin	8220	3073	0.21	0.08	2681	0.32	0.05	0.25
KT08	dentin	24697	12819	0.53	0.28	11022	0.19	0.05	0.15
KT09	dentin	1368665	744080	21.29	11.57	658020	0.28	0.07	0.22

Table C.2 continued from previous page

KT10	dentin	2246	315	0.07	0.01	278	0.26	0.05	0.22
KT11	dentin	4723	95	0.10	0.00	77	0.10	0.00	0.12
KT12	dentin	3684	476	0.08	0.01	433	0.28	0.05	0.22
KT13	dentin	9370	3376	0.16	0.06	2962	0.18	0.04	0.14
KT14	dentin	260852	147775	4.42	2.50	126063	0.21	0.03	0.15
KT15	dentin	1286855	108906	29.58	2.50	99009	0.16	0.11	0.11
KT16	dentin	291541	148005	6.10	3.10	130156	0.28	0.06	0.23
KT17	dentin	3122780	1860580	51.51	30.69	1639109	0.20	0.04	0.16
KT18	dentin	12348	6355	0.27	0.14	5588	0.20	0.04	0.15
KT20	dentin	8143	3389	0.18	0.08	2948	0.24	0.04	0.19
KT21	dentin	559996	314864	12.05	6.77	275779	0.20	0.04	0.16
KT22	dentin	9138	1403	0.25	0.04	1217	0.31	0.07	0.25
KT23	dentin	819546	471063	18.73	10.77	406680	0.20	0.03	0.16
KT24	dentin	2820962	1730557	50.37	30.90	1529031	0.21	0.04	0.17
KT25	dentin	5494	2251	0.13	0.05	1938	0.19	0.05	0.15
KT26	dentin	1910366	522444	39.21	10.72	479694	0.15	0.14	0.14
KT27	dentin	405669	222574	7.35	4.03	195869	0.23	0.05	0.18

Table C.2 continued from previous page

KT28	dentin	540135	328380	10.32	6.27	289285	0.21	0.05	0.17
KT29	dentin	4552576	783512	69.55	11.97	713991	0.18	0.12	0.12
KT30	dentin	6009	1243	0.12	0.03	1094	0.21	0.05	0.16
KT31	dentin	3662568	2262240	72.99	45.08	1971214	0.16	0.04	0.13
KT32	dentin	4345121	2675717	61.03	37.58	2371441	0.18	0.04	0.14
KT33	dentin	4437119	2929234	69.63	45.97	2609224	0.14	0.04	0.12
KT34	dentin	193388	33745	4.25	0.74	30961	0.23	0.13	0.13
KT35	dentin	75613	31630	1.45	0.61	27755	0.28	0.06	0.22
KT36	dentin	113771	58385	1.97	1.01	52137	0.23	0.05	0.18
NF217	dentin	6977	4887	0.12	0.08	4332	0.18	0.06	0.19
NF47	dentin	67885	50970	0.88	0.66	43583	0.14	0.04	0.14
S108	dentin	662104	631728	2.06	1.96	585485	0.09	0.03	0.08
S37	dentin	2587590	2199786	4.65	3.96	1960765	0.23	0.05	0.22
S41b	dentin	12528010	12261919	62.47	61.14	11369788	0.05	0.02	0.05
S454	dentin	8601	6367	0.03	0.02	5598	0.17	0.07	0.17

Table C.3: Length Statistics

ID	Type	Overall Median Length	Overall Mean Length	Human Median Length	Human Mean Length	Difference Median Length
KT05	calculus	56	64.23	50.5	58.92	5.5
KT08	calculus	79	86.61	67	73.52	12
KT09	calculus	72	82.46	54	60.09	18
KT13	calculus	70	80.48	53	59.39	17
KT14	calculus	62	70.44	49	55.37	13
KT24	calculus	69	78.24	55	60.72	14
KT25	calculus	83	91.84	68	72.33	15
KT26	calculus	74	83.47	57	63.49	17
KT28	calculus	74	83.1	57	64.24	17
KT29	calculus	82	91.08	61	68.15	21
KT31	calculus	81	91.34	64	71.22	17
KT32	calculus	74	83.6	58	64.86	16
KT36	calculus	74	83.97	60	66.16	14
C214	calculus	66	74.38	47	51.61	19
CMOL53	calculus	77	83.26	56	59.56	21
F1948	calculus	73	78.97	50	53.8	23
F349A	calculus	60	66.73	42	47.3	18
H10B	calculus	88	93.45	72	75.86	16
H24B	calculus	75	80.08	55	58.22	20

Table C.3 continued from previous page

S108	calculus	67	75.52	55	60.76	12
S37	calculus	63	72.37	51	56.66	12
S41B	calculus	88	94.44	79	85.1	9
S454	calculus	67	76.72	57	63.34	10
KT08	dentin	53	60.53	54	59.14	-1
KT25	dentin	56	64.03	47	52.76	9
KT31	dentin	55	60.21	56	60.13	-1
KT29	dentin	60	66.76	60	64.5	0
KT14	dentin	54	65.64	46	52.14	8
KT36	dentin	55	65.64	71	79.35	-16
KT24	dentin	57	63.82	58	63.68	-1
KT28	dentin	55	63.58	58	63.17	-3
KT05	dentin	56	65.11	47	49.43	9
KT09	dentin	61	70.02	58	61.85	3
KT26	dentin	69	79.12	71	78.71	-2
KT32	dentin	58	64.72	59	63.26	-1
KT13	dentin	58	67.44	58	63.92	0
C214	dentin	57	63.48	74	80.23	-17
CMOL53	dentin	72	79.99	74	80.23	-2
F1948	dentin	63	70.18	72	78.89	-9
F349A	dentin	69	76.66	69	75.46	0
H10B	dentin	83	90.85	88	94.95	-5
H24B	dentin	80	86.79	75	79.71	5

Table C.3 continued from previous page

S108	dentin	92	97.76	112	113.17	-20
S37	dentin	64	72.37	68	74.58	-4
S41B	dentin	105	109.9	104	109.12	1
S454	dentin	85	93.08	75	84.88	10

Table C.4: **Extraction and library blanks**

ID	Type	Dataset	Reads	Post Quality Filter	Human Reads	Endogenous	Damage
AOB1	blank	regional	288636	108306	49968	55.99	0.00
AOB10	blank	regional	230240	45434	1629	4.29	0.00
AOB11	blank	regional	841776	704107	1138	0.21	0.00
AOB12	blank	regional	611553	500720	1106	0.26	0.00
AOB13	blank	regional	186339	49932	1595	4.64	0.01
AOB14	blank	regional	293398	108875	4741	6.94	0.01
AOB2	blank	regional	282395	116275	756	0.75	0.02
AOB3	blank	regional	258106	44562	1391	4.27	0.00
AOB4	blank	regional	286467	115160	979	1.19	0.01
AOB5	blank	regional	289450	84146	4261	6.71	0.01
AOB6	blank	regional	244717	106687	3505	4.08	0.00
AOB7	blank	regional	307136	122101	2146	2.29	0.00

Table C.4 continued from previous page

AOB8	blank	regional	177851	64348	1693	3.13	0.00
AOB9	blank	regional	288123	65004	4807	9.01	0.00
AOL1	blank	regional	218354	27143	163	0.70	0.00
AOL10	blank	regional	192157	37163	1381	4.56	0.00
AOL11	blank	regional	163301	21101	747	3.98	0.00
AOL12	blank	regional	240609	26856	1225	5.54	0.00
AOL13	blank	regional	203469	21921	457	2.83	0.00
AOL14	blank	regional	201882	27889	596	2.89	0.02
AOL15	blank	regional	157109	24895	544	3.12	0.00
AOL16	blank	regional	255675	70966	639	1.46	0.01
AOL2	blank	regional	250738	18675	83	0.50	0.00
AOL3	blank	regional	249849	20771	67	0.40	0.00
AOL4	blank	regional	259283	16488	403	3.46	0.00
AOL5	blank	regional	238344	7847	139	2.19	0.00
AOL6	blank	regional	202055	48633	628	1.68	0.00

Table C.4 continued from previous page

AOL7	blank	regional	186756	34664	416	1.49	0.00
AOL8	blank	regional	194603	36563	614	2.19	0.00
AOL9	blank	regional	164442	25076	459	2.31	0.01
LIB_CONTROL	blank	global	956074	472124	5262	7.32	0.02
N1	blank	global	649670	305725	3828	14.04	0.01
NEG_S.G.	blank	global	453184	217466	841	16.23	0.00

Table C.5: Source contribution estimates

ID	Type	Plaque	Skin	Soil	Unknown
KT12	dentin	0	0.0912	0.6641	0.2447
KT08	dentin	0.2142	0.0722	0.5553	0.1583
KT25	dentin	7.00E-04	0.1423	0.8379	0.0191
KT01	dentin	0.1147	0.0244	0.7619	0.099
KT34	dentin	0.0286	0.1587	0.7012	0.1115
KT21	dentin	0.0436	0.1806	0.7424	0.0334
KT31	dentin	0.1948	0.0527	0.6909	0.0616
KT20	dentin	0	0.1318	0.7252	0.143
KT10	dentin	0	0.1227	0.6048	0.2725
KT03	dentin	0	0.1273	0.6921	0.1806
KT07	dentin	0	0.031	0.7957	0.1733
KT33	dentin	0.1124	0.0331	0.7968	0.0577
KT16	dentin	0.0791	0.0366	0.6285	0.2558
KT29	dentin	0.2145	0.0564	0.6031	0.126
KT14	dentin	0.1923	0.1704	0.4576	0.1797
KT30	dentin	0.0845	0.0231	0.5869	0.3055
KT36	dentin	0.0413	0.1148	0.6315	0.2124
KT06	dentin	0	0.0736	0.7895	0.1369
KT24	dentin	0.4083	0.0995	0.4308	0.0614
KT28	dentin	0.1213	0.0244	0.6485	0.2058
KT11	dentin	0	0.2235	0.5956	0.1809

Table C.5 continued from previous page

KT05	dentin	0.001	0.0836	0.751	0.1644
KT15	dentin	0.1097	0.0534	0.6919	0.145
KT27	dentin	0.0698	0.0194	0.7284	0.1824
KT02	dentin	0	0.066	0.6759	0.2581
KT09	dentin	0.0035	0.0309	0.669	0.2966
KT26	dentin	0.2037	0.059	0.6548	0.0825
KT17	dentin	0.0041	0.1653	0.6706	0.16
KT22	dentin	0.0791	0.0899	0.6659	0.1651
KT23	dentin	0.3421	0.0037	0.4877	0.1665
KT04	dentin	0.035	0.094	0.719	0.152
KT35	dentin	0.1235	0.0907	0.6831	0.1027
KT01	calculus	0.9822	0.003	0	0.0148
KT01	calculus	0.9548	0	0.0093	0.0359
KT02	calculus	0.8284	0.1192	0.0032	0.0492
KT03	calculus	0.9282	0.0646	7.00E-04	0.0065
KT04	calculus	0.8933	7.00E-04	0.0016	0.1044
KT05	calculus	0.8152	0.0654	0.0681	0.0513
KT06	calculus	0.69	1.00E-04	0.0588	0.2511
KT07	calculus	0.7382	0.0415	0.085	0.1353
KT08	calculus	0.9812	0.0031	0.0119	0.0038
KT09	calculus	0.8526	0	0	0.1474
KT18	dentin	0	0.0772	0.8878	0.035
KT10	calculus	0.8097	1.00E-04	0.0549	0.1353

Table C.5 continued from previous page

KT11	calculus	0.9542	0.0137	0.0036	0.0285
KT12	calculus	0.8557	0.028	0.0605	0.0558
KT13	calculus	0.8605	0.0428	0.0324	0.0643
KT14	calculus	0.9849	0	0	0.0151
KT15	calculus	0.7201	3.00E-04	0.049	0.2306
KT16	calculus	0.738	0.1374	0.0437	0.0809
KT17	calculus	0.8204	0	0.0716	0.108
KT18	calculus	0.752	0.001	0.064	0.183
KT19	calculus	0.8963	0.0864	0.0132	0.0041
KT32	dentin	0.2779	0.0938	0.5587	0.0696
KT20	calculus	0.8121	3.00E-04	0.0366	0.151
KT21	calculus	0.7438	0.1029	0.0444	0.1089
KT22	calculus	0.9228	0.0585	0.0068	0.0119
KT23	calculus	0.8998	0.0031	0.0312	0.0659
KT24	calculus	0.8336	0	0.0283	0.1381
KT25	calculus	0.8312	0	0.0238	0.145
KT26	calculus	0.8666	0	0.0518	0.0816
KT27	calculus	0.8876	0	0.0325	0.0799
KT28	calculus	0.8733	0	0	0.1267
KT29	calculus	0.9554	0	0.0127	0.0319
KT30	calculus	0.8129	0	0.0451	0.142
KT31	calculus	0.8956	0	0.0694	0.035
KT32	calculus	0.8907	0.0213	0.0324	0.0556

Table C.5 continued from previous page

KT33	calculus	0.8325	0.0522	0.0675	0.0478
KT34	calculus	0.8033	0.0633	0.0645	0.0689
KT35	calculus	0.8384	0	0.0302	0.1314
KT36	calculus	0.8529	0	0.0166	0.1305
KT13	dentin	0	0.0483	0.7716	0.1801
C214	calculus	0.9172	1.00E-04	0.0143	0.0684
C214	dentin	0	0.0965	0.6669	0.2366
cmol53	calculus	0.86	0	0	0.14
cmol53	dentin	2.00E-04	0.1157	0.7296	0.1545
F1948	calculus	0.9127	0.0403	0.0022	0.0448
F1948	dentin	0.3855	0.2293	0.2711	0.1141
F349A	calculus	0.7264	1.00E-04	0.0121	0.2614
F349A	dentin	0	0.1373	0.6238	0.2389
H10b	calculus	0.8174	0	0.0157	0.1669
H10b	dentin	3.00E-04	0.3351	0.5246	0.14
H24b	calculus	0.8557	0	0.0139	0.1304
H24b	dentin	0	0.2366	0.6974	0.066
NF217	calculus	0.7908	0.1447	0.0333	0.0312
NF217	dentin	0.7835	0.2059	3.00E-04	0.0103
NF47	calculus	0.5933	0.1399	0.1569	0.1099
NF47	dentin	0.5621	0.0877	0.1847	0.1655
S108	calculus	0.8572	0	0.0245	0.1183
S108	dentin	0	0.1393	0.7062	0.1545

Table C.5 continued from previous page

S37	calculus	0.8186	0.1213	0.0589	0.0012
S37	dentin	0	0.3125	0.5801	0.1074
S41b	calculus	0.9389	0.0292	3.00E-04	0.0316
S41b	dentin	0	0.2567	0.6776	0.0657
S454	calculus	0.756	0	0.1512	0.0928
S454	dentin	2.00E-04	0.1992	0.6631	0.1375