Laboratory contamination over time during low-biomass sample analysis

Authors: Laura S. Weyrich[1,2], Andrew G. Farrer[1], Raphael Eisenhofer[1,2], Luis A. Arriola[1], Jennifer Young[1T], Caitlin A. Selway[1], Matilda Handsley-Davis[1,2], Christina Adler[3], James Breen[1*], and Alan Cooper[1,2]

Affiliation:

[1]Australian Centre for Ancient DNA, University of Adelaide, Adelaide, Australia

[2] ARC Centre of Excellence for Australian Biodiversity and Heritage, University of Adelaide, Adelaide, Australia

[3]Faculty of Dentistry, University of Sydney, Sydney, Australia

[T]**Current address:** Advanced DNA, Identification, and Forensic Facility, University of Adelaide, Adelaide, Australia

***Current address:** South Australian Health and Medical Research Institute (SAHMRI), Adelaide, Australia

Abstract:

Bacteria are not only ubiquitous on earth but can also be incredibly diverse within clean laboratories and reagents. The presence of both living and dead bacteria in laboratory environments and reagents is especially problematic when examining samples with low endogenous content (*e.g.* skin swabs, tissue biopsies, ice, water, degraded forensic samples, or ancient material), where contaminants can outnumber endogenous microorganisms within samples. The contribution of contaminants within high-throughput studies remains poorly understood because of the relatively low number of contaminant surveys. Here, we examined 144 negative control samples (extraction blank and no-template amplification controls) collected in both typical molecular laboratories and an ultraclean ancient DNA laboratory over five years to characterize long-term contaminant diversity. We additionally compared the contaminant content within a homemade silica-based extraction method, commonly used to analyse low-endogenous samples, with a widely used commercial DNA extraction kit. The contaminant taxonomic profile of the ultraclean ancient DNA laboratory was unique compared to the modern molecular biology laboratories, and changed over time according to researchers, month, and season. The commercial kit contained higher microbial diversity and several human-associated taxa in comparison to the homemade silica extraction protocol. We recommend a minimum of two strategies to reduce the impacts of laboratory contaminants within low-biomass metagenomic studies: 1) extraction blank controls should be included and sequenced with every batch of extractions and 2) the contributions of laboratory contamination should be assessed and reported in each high-throughput metagenomic study.

Main Text:

In the new era of culture-independent microbiome research, targeted amplicon or 'metabarcoding' approaches are now routinely used to amplify DNA from microbial species across the tree of life. However, these methods lack the ability to select for either specific species or to exclude contaminants [1]. Although these techniques have provided invaluable insight into otherwise cryptic microbial communities, the increased sensitivity and lack of target specificity leaves microbiota studies particularly susceptible to the effects of contamination. Such effects are widespread, as several recent studies have indicated that contaminant microbial DNA can be routinely isolated from laboratory reagents and surfaces [2-4] and that this signal has significantly impacted the interpretation and characterization of microbiota in high-throughput sequencing studies. For example, Salter *et al.* recently demonstrated that bacterial DNA present in laboratory reagents is present in both quality-filtered 16S ribosomal RNA (rRNA) gene and shotgun metagenomic datasets and significantly impacts the interpretation of results [3]. Multiple microbial contaminants have already been identified within the published 1,000 Genomes dataset and other medical genomic studies [4,5]. Despite these findings, the routine assessment of microbial background contamination is still not required, or fully reported, in microbiota studies.

While the presence of contaminant DNA is widespread, the effects are particularly problematic in low-biomass samples that contain very little endogenous DNA [6] (*e.g.* preterm infant swabs, tissue samples, such as placenta, tumour biopsies, or breast tissue, and some environmental samples, such as ice or calcite). In low-biomass samples, a small contaminant signal from laboratory reagents can easily overpower the intrinsic signal from the sample. This is similarly an issue in current

palaeomicrobiology studies that examine ancient, degraded microbiota, such as mummified human tissue, preserved faeces (coprolites), or calcified dental plaque (calculus) [6-8]. In ancient samples, the amount of endogenous DNA attributed to the original source can be extremely low (*e.g.* <0.05% of the total DNA in the sample) and is damaged, fragmented, and intermixed with longer, higher-quality modern DNA fragments from contaminant species [9]. Therefore, monitoring and understanding the contributions of contaminant DNA, especially in low-biomass or ancient samples, is critical to ensure that reported results are only based on the endogenous DNA.

Microbial contaminant DNA (*i.e.* background or exogenous DNA) is a mixture of DNA from both environmental and laboratory sources, with the former including factors such as soil at a burial site, air within the sampling facility, and microorganisms from people touching the sample, while the latter involves reagents, glassware, labware, and surfaces [7]. Environmental contamination in low-biomass samples may be difficult to control or monitor, but the laboratory contaminants can be monitored by including extraction blank (EBC) and no-template amplification (NTC) controls and assessed using bioinformatics tools (*e.g.* SourceTracker [10]). An EBC is an empty tube introduced during the extraction steps to collect DNA from the laboratory environment and the reagents throughout processing [11]. Similarly, a NTC is simply an amplification reaction that lacks the addition of DNA from biological samples. These controls should be amplified and sequenced along with other samples and are critical steps to identify and exclude contaminant taxa from downstream analyses, reducing noise and ensuring any results are based solely on endogenous DNA [12]. Despite this, there are surprisingly few published resources describing contaminant taxa found in extraction blank or no-template controls [3,13,14].

In this study, we used 16S rRNA metabarcoding to characterise the contaminant diversity in 144 EBCs and NTCs using laboratory techniques specifically designed for low-biomass material. We also explored differences in microbial contamination within two different types of laboratory facilities: a state-of-the-art, purpose-built ancient DNA clean laboratory over the course of five years, and three typical modern molecular biology laboratories over one year. Lastly, we investigated differences between a common commercial DNA extraction kit and a homemade DNA extraction method typically applied in the ancient DNA field. Overall, this study is designed to assess contaminant profiles over time and identify more potential contaminant sequences in both high- and low-biomass research.

Materials and Methods

Sample collection

Four different types of sample were used: ancient dental calculus (calcified dental plaque), modern dental calculus, EBCs, and NTCs. Dental calculus samples were obtained from ancient and modern humans as described by Adler *et al.* [11]. A single EBC was included in each batch of extractions by treating an empty tube as if it was a biological sample throughout the DNA extraction and library preparation process. Similarly, NTC samples were created during the 16S rRNA library amplification stage by processing tubes without adding any known template DNA. Both EBCs and NTCs were subsequently included through to DNA sequencing a ratio of one control sample for every ten biological samples.

Description of laboratory facilities

DNA extraction occurred in two different types of laboratory facilities: a purpose-built, ultra-clean ancient DNA laboratory (ancient lab) and three typical modern molecular biology laboratories (modern labs). The ancient lab is physically remote from the university campus in a building with no other molecular biology laboratories and contains a HEPA-filtered, positive pressure air system to remove DNA and bacteria from external sources. The HEPA filter function is checked annually and changed every ten years. The surface and floors within the laboratory are cleaned weekly with a 5% bleach (NaClO) solution and are illuminated with ceiling mounted UV lights for 30 minutes each night. UV light bulbs are changed annually. Users entering the ancient lab are required to have showered, wear freshly laundered clothing, avoid the university campus prior to entry, and cannot bring personal equipment (*e.g.* phones, writing equipment, and bags) into the facility.

Standard personal laboratory wear includes disposable full-body suits, surgical facemasks, plastic see-through visors, and three layers of gloves to allow frequent changing without skin exposure (including one inner elbow-length pair of surgical gloves). All liquid reagents within the ancient lab are certified DNA-free, and the outer surface of all plastic ware and reagent bottles are decontaminated prior to entering the laboratory (cleaned with 5% bleach and treated with UV (2x, 40W, 254nm UV tubes at a distance of 10cm for 10 minutes) within a UV oven (Ultra Violet Products). All DNA extractions and amplification preparations are performed in a room separate to sample preparation and are completed in still-air cabinets that are cleaned with bleach and UV treated for 30 minutes (3x, 15w, 253.7nm tube lamps; AURA PCR) prior to beginning any work. In addition, ancient samples from different sources (*e.g.* soil, plants, and other animals) are processed in separate, dedicated rooms to minimise cross-contamination. In contrast, the modern laboratories are located over 2 km away from the ancient lab at the University of Adelaide (n=2) and the University of Sydney (n=1). All three modern labs are typical of most molecular biology laboratories and are not routinely decontaminated and contain users that routinely use latex gloves but are not required to wear body suits or masks. DNA extracted within the modern labs comes from a wide range of sources (*e.g.* humans, mammals, and environmental samples), although microbiome extractions were only performed on days when no other material was being extracted. In all facilities, DNA was extracted and prepared for amplification in still-air cabinets that are cleaned before and after each use with 5% bleach.

DNA extractions

Several specialized DNA extraction protocols have been developed within ancient DNA studies to remove environmental contamination and enhance the recovery of the endogenous DNA. The extraction method selected for this study has previously been described for work on ancient dental calculus [12]. Each ancient sample was first decontaminated using a published protocol [11], while modern samples were not decontaminated. The decontamination procedure included exposure to UV radiation for 15 minutes on each side of the sample, submersion of the sample in 5% bleach for 5 minutes, followed by submersion in 90% ethanol for 3 minutes to remove any residual bleach, and 5 minutes of drying. Decontaminated ancient calculus was then wrapped in aluminium foil and pulverized into power with a steel hammer and placed into a sterile 2mL tube. The EBCs were empty tubes exposed to air for 30 seconds in the same room during sample decontamination and were included in the extraction process as if they contained a sample.

Following decontamination, DNA was extracted using the QG-based method previously described for the extraction of ancient microbiome material [12] (referred to as 'QG'). All reagents for the QG extraction method were prepared in a 'sample-free' room in the ancient DNA facility, and all reagents were aliquoted immediately upon opening and frozen until further use to avoid cross contamination. Where possible, certified 'DNA-free' reagents and lab ware were purchased (*e.g.* water and plastic tubes). All other reagents were opened solely within a sterilized hood within the ancient DNA facility. All chemicals were prepared for the extraction with previously unopened DNA and RNA-free certified water (Ultrapure water; Invitrogen). Briefly, 1.8 mL of 0.5 M ethylenediaminetetraacetic acid (EDTA; Life Tech), 100 µL of 10% sodium dodecyl sulphate (SDS; Life Tech), and 20 µL of 20 mg/mL proteinase K (proK; Life Tech) were added to each sample, and the mixture

was rotated at 55°C overnight to decalcify the sample. Released DNA was then purified by adding silica (silicon dioxide; Sigma Aldrich) and 3 mL of binding buffer (*e.g.* QG buffer; Qiagen; modified to contain 5.0M GuSCN; 18.1mM Tris-HCl; 25mM NaCl; 1.3% Triton X-100) [15]. The silica was pelleted, washed twice in 80% ethanol, dried, and resuspended in 100 µL of TLE buffer (10mM Tris, 1mM EDTA, pH 8) twice to elute the DNA, which was then stored at -20°C until amplification. All chemicals were prepared for the extraction with previously unopened DNA and RNA-free certified water (Ultrapure water; Invitrogen). For QG extractions performed in the modern laboratories, unopened aliquots of DNA extraction reagents were transported to the modern laboratory, and the modern samples were extracted following the ancient DNA approach described above.

In contrast to ancient DNA extractions, many modern microbiome studies decrease cost and time by using commercial DNA extraction kits to isolate DNA. In order to compare the nature and extent of contaminant DNA in the ancient method to a typical commercial microbiome DNA extraction kit, we analysed an additional set of EBCs created during extractions using a PowerBiofilm® DNA Isolation Kit (MOBIO) from concurrent oral microbiome research conducted in the same modern labs (referred to as 'kit' EBCs).

Library Preparation

To minimise additional variables, a simple 16S rRNA amplicon sequencing approach was used in this study to compare the different sample types. Briefly, the V4 region of the bacterial 16S rRNA encoding gene was targeted for amplification using degenerate Illumina fusion primers, as previously described [1]: forward primer 515F (AATGATACGGCGACCACCGAGA TCTACACTATGGTAATTGTGTGCCA

GCMGCCGCGGTAA) and barcoded reverse primer 806R (CAAGCAGAAGA CGGCATACGAGATnnnnnnnnnnnnAGTCAGTCAGCCGGACTACHVGGGTW TCTAAT) [1]. The string of n's in the reverse primer refers to the unique 12 bp barcode used for each sample. Primers were resuspended in TLE buffer within the ancient facility and distributed to the modern laboratory. In both facilities, all PCR amplification reactions were prepared using ultraclean reagents with strict ancient DNA protocols [9]. Each PCR reaction contained 17.25 µL DNA-free water (Ultrapure water; Invitrogen), 2.5 µL 10X reaction buffer (20 mM Tris-HCl, 10 mM $(NH_4)_2SO_4$, 10 mM KCl, 2 mM $MgSO_4$, 0.1% Triton® X-100, pH 8.8@25°C; ThermoPol Buffer; New England Biolabs;), 0.25 uL Taq polymerase (Platinum Taq DNA Polymerase High Fidelity; Thermo Fisher Scientific), 1.0 µL $MgCl_2$ (Thermo Fisher Scientific), 1.0 µL of each primer at 10 uM (IDT), and 2.0 µL of genomic DNA; each reaction was performed in triplicate. 16S rRNA amplification occurred under the following conditions: 95°C for 5 min; 37 cycles of 95°C for 0.5 min, 55°C for 0.5 min, 75°C for 1 min; and 75°C for 10 min. NTC reactions were also included in triplicate. PCR products were quantified (QuBit; Thermo Fisher Scientific) and pooled in batches of 30 samples at equal nanomolar concentrations prior to purification (Ampure; New England Biolabs). Each pool of purified PCR products was quantified (TapeStation; Agilent) before being combined into a single library. All amplicons were sequenced using the Illumina MiSeq 2x150 bp (300 cycle) kit.

Bioinformatics Analysis

After sequencing, fastq files for the forward and reverse reads were created using the Illumina CASAVA pipeline (version 1.8.2). Overlapping forward and reverse reads were joined (based on a maximum of 5% nucleotide difference over a

minimum 5bp overlap) using BBmerge (sourceforge.net/projects/bbmap/). Only successfully merged sequences were used in downstream analyses. The resulting fastq file was then imported into QIIME (MacQIIME v1.8.0), a bioinformatics pipeline-based software for the analysis of metagenomic data [16]. All further analysis of the amplicon datasets was conducted within the QIIME package. Libraries were demultiplexed using a Phred base quality threshold of less than or equal to 20, with no errors allowed in the barcodes. Operational taxonomic units (OTUs) were determined by clustering sequences at 97% similarity using UClust [17], and representative sequences (*i.e.* cluster seed) were selected for each cluster. By default, clusters with fewer than five sequences were eliminated from the analysis to reduce noise and spurious findings. Lastly, 16S rRNA gene sequences were given taxonomic assignments using the Greengenes 13_8 database if the sequence was at least 80% similar [18,19]. Taxonomic diversity measurements (alpha- and beta-diversity) and statistical analyses were performed and visualized in QIIME. Samples were rarefied to a minimum of 150 sequences (Figure 2) and a maximum of 1,000 sequences for diversity analyses, as many controls contained low sequence counts. Statistical differences between groups were identified using a PERMANOVA test for beta diversity (adonis), nonparametric t-test for alpha diversity (Monte Carlo), or Kruskal-Wallis and G-tests for detection of specific taxa associated with different treatments.

Results

Low bacterial diversity is routinely obtained from laboratory extraction controls.

The EBCs and NTCs were sequenced alongside the ancient and modern biological samples; all sample types were pooled together at equimolar concentrations. Despite the equimolar pooling, we routinely obtained fewer reads from control samples (EBCs and NTCs) compared to the dental calculus samples, likely due to poor amplification of control samples, the quantification of poor DNA libraries, and clean-up strategy employed. Compared to the ancient and modern calculus samples, 6.4-fold fewer reads on average were obtained from EBCs, and 7.6-fold fewer were obtained from NTCs (Figure 1A). As well as containing fewer reads overall, the control samples contained fewer taxa that could be identified than the biological samples. In the ancient laboratory, 719 total OTUs were observed in ancient biological samples (calculus), while only 415 were identified in the EBCs and 228 in NTCs (Figure 1B). In the modern laboratories, 286 total OTUs were described in the modern calculus samples, versus 208 in the EBCs and 102 in the NTCs. The OTU diversity that appears within the EBCs is similar to the differences in diversity observed between modern and ancient biological specimens, potentially reflecting minor cross contamination during DNA extraction. Across different extraction methods, the EBCs for the commercial extraction kit contained 261 OTUs, around 25% more than the in-house method conducted in the modern laboratory. Overall, the laboratory controls were largely dominated by a single phylum, Proteobacteria (Figure 2), and alpha diversity was significantly lower than in the biological samples extracted within the same laboratory (Monte Carlo; $p=<0.0001$ and $T=>11.0$ in all comparisons between any group of controls and all biological samples). While the

diversity within laboratory controls was considerably lower than the biological samples, these results demonstrate the extent of background microbial contamination even within an ultra-clean laboratory with 'DNA-free' reagents, and clearly highlight the need to routinely monitor and report background contamination within all research facilities.

Extraction blank controls detect >50% more contaminant taxa than no-template controls

Many studies, including some in palaeomicrobiological research, have simply reported failed EBC and NTC amplification reactions (often via simple visual comparison on an agarose gel) as a means to determine that their samples are free from contamination [21,22]. This approach is clearly inadequate, and importantly, also fails to appreciate the extent of contamination introduced during the extraction process, even though this issue is well described in the literature [14,23,24]. In our comparisons, EBCs were taxonomically far more diverse than NTCs (Figure 1B) and contained more microbial genera (415 versus 228 genera in the ancient lab, and 208 versus 102 genera in the modern labs). This pattern suggests that if just NTCs were used to monitor the presence of laboratory contamination, at least 53% of the total laboratory contamination may go undetected. These results highlight the need for the standard reporting of both EBCs and NTCs in both modern and ancient metagenomics research.

We examined the impact of overall laboratory contamination on ancient samples by bioinformatically filtering (removing) all contaminant OTUs from ancient dental calculus samples. For the ancient samples prepared with the specialised facility, an average 92.5% of the sequence reads were contaminants, but importantly,

293 only accounted for 28% of the genera identified within these samples. This indicates
294 that endogenous signal can be identified even in low-endogenous samples once
295 contaminant taxa are removed.
296
297 **Extraction blank and no-template controls reflect laboratory environment**
298 Previous studies have detected differences in the contaminants present in
299 different laboratory facilities [3]. In our study, the laboratory environments explained
300 more of the taxonomic diversity observed in the EBCs and NTCs than the extraction
301 or amplification methods used to generate them (Figure 3). For example,
302 Proteobacteria dominated the EBCs and NTCs from the ancient laboratory, while
303 Firmicutes were more dominant in EBC and NTC controls from the modern
304 laboratories. In fact, different types of controls (*i.e.* EBC or NTC) from the same
305 laboratory clustered with others of the same sample type in a Principle Coordinates
306 Analysis (PCoA) of unweighted UniFrac values ($p=<0.001$, $R^2=0.083$; Figure 3A),
307 despite large variation and significant differences in each lab (Figure 1B). Despite the
308 sample type (*e.g.* EBC or NTC) driving the majority of the signal, taxa distinguishing
309 each laboratory could also be detected, with specific *Paenibacillus* taxa only found in
310 the modern laboratories, while the ancient laboratory contained both bacterial
311 (*Comamonas*, *Pseudomonas*, *Acinetobacter*, *Enterobacter*) and archaeal
312 (*Methanobrevibacter*) taxa that were not observed in the modern labs. In addition,
313 several bacterial taxa were identified in both lab types, but were significantly
314 increased in one location. The ancient laboratory contained significantly higher levels
315 of certain *Acinetobacter*, *Comamonas*, and *Pseudomonas* taxa compared to the
316 modern laboratories (Kruskal-Wallis; Bonferroni-corrected $p=<0.05$), while
317 Erythrobacteraceae and *Staphylococcus* taxa were increased in abundance in the

318 modern laboratories. With the exception of the *Staphylococcus* taxa, each of these
319 taxa had been previously identified in laboratory reagents [3]. This suggests that some
320 contaminant taxa are relatively universal across laboratories and are therefore either
321 introduced in the manufacturing of laboratory reagents and labware or have a
322 fundamental niche in low-nutrient, laboratory environments.

323 We next examined the genera that were likely to be in the reagents
324 themselves, rather than the laboratories, by looking for shared taxa within the EBCs
325 generated during extractions in both the ancient lab and modern labs. Of the 69
326 dominant genera (*i.e.* observed at >0.1%), 17 were present in the reagents used in the
327 in-house QG DNA extraction process used in both types of facility. These taxa
328 included *Cloacibacterium*, *Flavobacterium*, *Paenibacillus*, *Novosphingobium*,
329 *Sphingomonas*, *Limnohabitans*, *Tepidomonas*, *Cupriavidus*, *Ralstonia*, *Acinetobacter*,
330 *Enhydrobacter*, *Pseudomonas*, and *Stenotrophomonas*, and four unidentified genera
331 within Comamonadaceae, Erythrobacteraceae, Enterobacteriaceae, and
332 Pseudomonadaceae (Table 1). Within the ancient laboratory EBCs, the 26 most
333 dominant genera included *Acinetobacter* (39%), followed by three genera within the
334 Comamonadaceae family (totalling 11.3%), *Pseudomonas* (8%), *Novosphingobium*
335 (1.5%), *Ralstonia* (1%), *Cloacibacterium* (1%), and others (Table 1). In the EBCs
336 from the modern laboratories, *Paenibacillus* was the most prevalent of the 43
337 dominant genera (46%), while two Erythrobacteraceae (16.5%), Comamonadaceae
338 (6.1%), *Cloacibacterium* (3.9%), *Corynebacterium* (2.5%), *Enterococccus* (2.5%),
339 *Staphylococcus* (2.2%), *Enhydrobacter* (1.8%), Microbacteriaceae (1.7%), a
340 Pseudomonadaceae (1.4%), *Ralstonia* (1.3%), and *N09* (1.2%) taxa were the next
341 most prevalent within the reagents (Table 1). Although the same extraction method
342 and reagents were used, only three of the dominant taxa (*i.e.* identified at >1%

prevalence) were the same within both laboratories (Comamonadaceae, *Cloacibacterium*, Pseudomonadaceae), highlighting the heterogeneity of taxa identified with EBCs. While many of these taxa have been previously identified as laboratory contaminants, the diversity within the modern laboratories also includes some human-associated taxa that have been cultured from the oral cavity, gut, and skin (*e.g. Corynebacterium*, *Enterococcus*, and *Staphylococcus*, respectively). This suggests that the additional precautionary measures used within the ancient laboratory help reduce the introduction of human-associated microorganisms in metagenomic data sets.

DNA extraction kits contain microbiota indicative of the human mouth

We compared the diversity of taxa present within EBCs from the widely used ancient DNA extraction method and the commercial PowerBiofilm® DNA Isolation Kit, used in the same modern laboratory. While the latter kit has been shown to have the lowest bacterial background contamination of standard microbiome kits [3], microbial diversity within the kit EBCs was significantly higher than the in-house QG method (Figure 1B), suggesting that kit-based DNA extractions are more prone to background contamination. On a PCoA plot constructed using unweighted UniFrac distances, the kit EBCs clustered away from the QG EBCs and NTCs, including those processed in the same laboratory (adonis; $p=<0.001$, $R^2=0.04$; Figure 4A), demonstrating that a unique microbial community profile originates from the kit. This profile was not solely dominated by Firmicutes, like the other control samples from the modern lab, but contained taxa from several unique phyla (Acidobacteria, Gemmatimonadetes, and Verrucomicrobia). These unique phyla included 15 distinct taxa that were also not observed in the extractions using the ancient DNA extraction

method, including *Alicyclobacillus* (n=9), *Halomonas*, *Pseudonocardia*, *Vogesella*, *Allobaculum* (n=2), and *Akkermansia* taxa (Kruskal-Wallis; p=<0.05; Table 2). Several of these taxa are known to be resistant to sterilization treatments, including pasteurization [25]. In addition, several OTUs were more likely to be found in higher abundances in the kit EBCs than any other control samples (G-test; p=<0.05) and include specific *Bradyrhizobiaceae*, *Neisseria*, *Corynebacterium*, *Fusobacterium*, *Streptococcus*, *Micrococcus*, and *Halomonas* taxa. While *Bradyrhizobium* and *Micrococcus* have previously been identified as laboratory contaminants [3,4], the remaining taxa are commonly found in the human mouth. Concerningly, many of these human oral taxa have been previously reported from low-biomass samples, such as placenta and tumor tissue, which were examined without EBCs [22,26]. This suggests that DNA extraction kits used in modern molecular biology laboratories may be contributing unique microbial signals in addition to those generated within the laboratory environment.

Contaminant taxa change over time

Much of the variation identified in this study is laboratory-specific, so in order to test how seasonal changes, different researchers, or time might alter the microbial diversity observed in controls, we assessed the EBC and NTC records from the ancient lab facility over five years (2012-2016). Bacterial community structure in the ancient lab was linked to the researcher (adonis; p=0.001,R^2=0.073), the extraction year (adonis; p=<0.01,R^2=0.022), the extraction month (adonis; p=<0.001,R^2=0.044; Figure 4B), and wet / dry seasons (adonis; p=0.001,R^2=0.081). However, each of these signals was less significant and drove less variation within the data set when compared to the differences observed between laboratory facilities or between

393 extraction methods. Very few specific taxa were significantly associated with
394 temporal variation, although linked changes in overall diversity were observed. 32
395 OTUs were associated with the month in which the extraction was performed and
396 were largely present during dry months (Oct-January; dominated by
397 *Comamonadaceae* (2), *Bradyrhizobiaceae* (11), and *Gemmatimonadetes* (2) taxa;
398 Kruskal-Wallis; Bonferroni corrected p=<0.05), while only two OTUs
399 (*Thermobispora* and *Actinomycetales* taxa) were linked to wet seasons. Interestingly,
400 five OTUs (*Leptotrichia*, *Comamonadaceae* (3), and *Burkholderia*) were also
401 associated with the lab researcher (Kruskal-Wallis; Bonferroni corrected p=<0.05).
402 While we cannot rule out the confounding nature of these variables (*e.g.* links
403 between different researchers being more active in the laboratories at different times),
404 these observations suggest that contaminant taxa change over time and need to be
405 continually monitored, even in the cleanest molecular facilities.

Discussion

Overview

While several studies have now reported on contaminant DNA within laboratory reagents, the systematic inclusion of extraction blank controls has not yet been widely embraced in metagenomic research. Several studies on human microbiota have been criticised for their lack of careful controls [14,24,27], as the unfounded results of such studies have potentially serious repercussions and have hindered scientific progress. A similar phenomenon occurred with the new field of ancient DNA in the early 1990s, when research teams, reviewers, and editors failed to adequately test for contamination [28–30], leading to many spurious results. This seriously undermined the credibility of ancient DNA research [23] and resulted in the formation of a robust set of guidelines [9]. Here, we surveyed the largest collection of extraction blank and no-template amplification negative control samples to date (n=144) with the goal of better describing contaminant DNA in microbiome studies to avoid pitfalls similar to those observed in the ancient DNA field.

Contaminant diversity remains underestimated

We identified 861 contaminant taxa over five years within a single ultra-clean laboratory facility. Before this publication, the largest collection of contaminant taxa was published by Salter *et al.* and included 93 contaminant genera [3]. Within our study, we found 71 of the taxa identified by Salter *et al.* across all labs and methodologies. However, only 29.5% of the Salter *et al.* taxa (21 of their 71 taxa) were identified as dominant taxa within our study across all methods and labs. This indicates that laboratory microbial contamination is not yet well described and is likely to be unique across different laboratories, protocols, seasons, and researchers.

Of the 21 taxa shared across studies, four genera (*Ralstonia*, *Acinetobacter*, *Pseudomonas*, and *Stenotrophomonas*) have now been routinely identified in at least four of the six publications that examine laboratory contamination [2–4,13,31,32]. All of these taxa are classified as Proteobacteria, as are 55% of the dominant contaminant taxa (38/69) identified within our study and 63% (34/92) within the Salter *et al.* study. While contamination is highly diverse, this finding indicates that Proteobacteria appear to be the most widespread source of laboratory contamination. Proteobacteria encompasses several families of bacteria that are known to be UV and oxidation resistant.

Analysing contaminants is critical for the successful interpretation of low-biomass samples

We identified several human oral microbiota taxa present in the commercial extraction kit, including *Fusobacterium, Streptococcus,* and *Corynebacterium* [33], while previous studies have previously identified additional human oral taxa contaminants, including *Haemophilus* and *Peptostreptococcus* [31]. Worryingly, one of these taxa in particular, *Fusobacterium*, has recently been identified both as a component of the 'placental microbiome', and as a component of breast cancer tissue, in low-biomass studies that did not consider background contamination from laboratory reagents or environments [22,26,34]. It remains unclear whether this taxon is a laboratory contaminant, or whether it can escape the oral cavity and contribute to inflammatory processes elsewhere in the body. Other non-oral taxa identified within this study as contaminants have also previously been reported as important taxa within studies that failed to use controls [35]. There is clearly a need for more detailed metagenomic studies, or the use of improved 'oligotyping' 16S rRNA gene analysis

456 methods of contaminant taxa, to better identify specific strain differences and
457 determine whether such taxa are contaminants or are actually present in the body and
458 can cause systemic disease. The lack of contaminant assessment has already
459 negatively impacted the metagenomics field [14], and it is critical that editors and
460 reviewers are aware of this issue.
461
462 *Bacterial DNA is still obtained from ultra-clean reagents in ultra-clean facilities – no*
463 *facility is contaminant free.*
464 Contaminant taxa were identified in EBCs and NTCs within five different
465 laboratory facilities, including a state-of-the-art, ultra-clean ancient DNA facility. In
466 the latter, the specialized conditions and procedures did not prevent low levels of
467 bacterial diversity, and a wide-range of contaminant taxa was still observed – with the
468 dominant taxa all known to resist disinfectant measures, including treatment with
469 aromatic or oxidative compounds (*i.e.* bleach) (*Acinetobacter* [36], *Comamonas* [37],
470 or other disinfectant compounds (*Pseudomonas* [38])). These mechanisms of
471 disinfection resistance have contributed to nosocomial infections in hospitals (*i.e.*
472 *Acinetobacter* [39]) and to contamination of cell culture reagents (*e.g. Achromobacter*
473 [40]). Of note, *Deinococcus*, a taxa that can notoriously survive UV irradiation [41],
474 *Alicyclobacillus*, known to survive pasteurization [25], and other species known to
475 degrade oxidative compounds (*e.g. Pasteurella* [42]) were not observed in the
476 specialised ancient DNA facility, but were identified within the modern laboratory.
477 While measures to reduce contamination have prevented the introduction of human-
478 associated microorganisms into the ancient lab EBCs, these numerous strategies did
479 not eliminate or completely prevent the introduction of bacterial contaminant DNA.
480 This suggests that each research facility will likely contain unique microorganisms

able to resist decontamination measures, although it is plausible that contaminant DNA could be routinely introduced into the facility from other source and represents living species found elsewhere, rather than in the actual facilities utilized in this study. Regardless, this finding reiterates that every laboratory is susceptible to bacterial DNA contamination and that researchers should consistently monitor the contamination present within their own facility as a best practice.

Non-kit approaches provide unique contaminant signals

In this study, we identified several taxa in a commonly used DNA extraction kit that were absent in the homemade ancient DNA extraction method. The ancient DNA method was developed to obtain more DNA from samples with low-endogenous DNA, and this and other similar extraction methods are now routinely applied in ancient DNA studies to examine ancient microbiota and metagenomes [11,43,44]. In this study, the ancient DNA method produced extraction blanks that had lower microbial diversity and were less likely to contain human oral taxa than extraction blanks generated using a commercial kit. This suggests that commercially available kits may contain more DNA contamination than homemade methods that source clean materials. It is likely that the assembly of kit-based reagents in a separate facility provides an additional opportunity to contaminate reagents with laboratory DNA. This also suggests that ancient DNA extraction methods and strategies could be applied in modern low-biomass studies to potentially reduce contaminants that originate from humans.

In the future, studies of low-biomass or low endogenous count routinely employ shotgun sequencing to better identify contaminant taxa, as strain-level identifications increase specificity in tracking contaminants. In many cases, the ancient DNA field

has now shifted to utilizing shotgun DNA sequencing as the gold-standard method (12). Shotgun sequencing also produces many other important molecular signals (*e.g.* signatures of ancient DNA damage), functional analysis, and strain markers to delineate which species are endogenous and which are contaminants. For example, distinct strains within a single genus could be identified as either a contaminant or an endogenous species, which would be critical for examining oral species in low-biomass tissues. In addition, damage profiles of DNA contamination could be used to distinguish fragmented, extracellular DNA within reagents versus species living within the laboratory. Current approaches aimed at eliminating contamination in shotgun sequenced metagenomes have had varied levels of success (reviewed in [3]), and new bioinformatic tools and models will undoubtedly improve our ability to identify and account for contaminant signals within metagenomic data sets (45). However, the need to routinely include EBCs and NTCs within microbiome data sets will likely always be necessary when examining low biomass samples, even when other methodologies, such as shotgun metagenomic sequencing, are applied.

Contamination assessment needs to be routinely reported as a publication requirement.

Contaminant sequences introduced during sample processing and library construction significantly contribute to signals from biological samples, especially those that are low-endogenous or low-biomass in nature. This study confirms that contaminant taxa that are unique to the extraction method and facility, are related to the material being extracted, and change over time within a single facility, although these levels of contamination can be somewhat mitigated by routine decontamination measures of the facility and potentially the reagents themselves (46). Therefore, the

presence of contaminants needs to be considered in all future studies of both human and environmental microbiota. We recommend that all researchers routinely record potential sources of contamination DNA (reagent batches or lot numbers; dates of extractions and amplifications; researchers performing such duties, *etc.*) and critically propose that researchers routinely include extraction blank controls during the extraction process to monitor the bacterial DNA introduced into their samples. Minimally, one control should be included in at least every batch of extractions and amplifications performed. Adding carrier DNA into control samples may also improve contaminant DNA detection (47). If controls were not included in existing data sets, an assessment of previously identified contaminant taxa within study datasets should also be minimally included in the published analysis. For example, researchers could report how many known contaminant taxa are present within a dataset or provide evidence to demonstrate that the removal of known contaminants does not impact the sample signal or conclusions of the paper. To facilitate this process, we have included a text file that includes a list of all the contaminant taxa observed here, as well as a separate file of only the dominant taxa. The inclusion of negative extraction blank controls should be regarded as minimal requirements for any metagenomics research and should become standard requirements of reviewers and journal editors.

Acknowledgements

We would like to acknowledge Paul Gooding at the Australian Genomic Research Facility for technical help during DNA sequencing. This research was funded by the Australian Research Council (L.S.W; A.C.).

References

1. Caporaso JG, Lauber CL, Walters WA *et al.* Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J* 2012;**6**:1621–4.

2. Tanner MA, Goebel BM, Dojka MA *et al.* Specific Ribosomal DNA Sequences from Diverse Environmental Settings Correlate with Experimental Contaminants. *Appl Environ Microbiol* 1998;**64**:3110–3.

3. Salter SJ, Cox MJ, Turek EM *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology* 2014;**12**, DOI: 10.1186/s12915-014-0087-z.

4. Laurence M, Hatzis C, Brash DE. Common Contaminants in Next-Generation Sequencing That Hinder Discovery of Low-Abundance Microbes. *PLoS One* 2014;**9**, DOI: 10.1371/journal.pone.0097876.

5. Kearney MF, Spindler J, Wiegand A *et al.* Multiple Sources of Contamination in Samples from Patients Reported to Have XMRV Infection. *PLoS One* 2012;**7**, DOI: 10.1371/journal.pone.0030889.

6. Weiss S, Amir A, Hyde ER *et al.* Tracking down the sources of experimental contamination in microbiome studies. *Genome Biol* 2014;**15**, DOI: 10.1186/s13059-014-0564-2.

7. Weyrich LS, Dobney K, Cooper A. Ancient DNA analysis of dental calculus. *Journal of Human Evolution* 2015, DOI: 10.1016/j.jhevol.2014.06.018.

8. Warinner C, Speller C, Collins MJ. A new era in palaeomicrobiology: prospects for ancient dental calculus as a long-term record of the human oral microbiome. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 2015;**370**:20130376.

9. Cooper A, Poinar HN. Ancient DNA: Do It Right or Not at All. *Science* 2000;**289**:1139–1139.

10. Knights D, Kuczynski J, Charlson ES *et al.* Bayesian community-wide culture-independent microbial source tracking. *Nat Meth* 2011;**8**:761–3.

11. Adler CJ, Dobney K, Weyrich LS *et al.* Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the Neolithic and Industrial revolutions. *Nat Genet* 2013;**45**:450–5.

12. Weyrich LS, Duchene S, Soubrier J *et al.* Neanderthal behaviour, diet, and disease inferred from ancient DNA in dental calculus. *Nature* 2017;**544**:357–61.

13. Glassing A, Dowd SE, Galandiuk S *et al.* Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. *Gut Pathogens* 2016;**8**:24.

592 14. Lauder AP, Roche AM, Sherrill-Mix S *et al.* Comparison of placenta samples
593 with contamination controls does not provide evidence for a distinct placenta
594 microbiota. *Microbiome* 2016;**4**:29.

595 15. Rohland N, Hofreiter M. Ancient DNA extraction from bones and teeth. *Nat*
596 *Protocols* 2007;**2**:1756–62.

597 16. Caporaso JG, Kuczynski J, Stombaugh J *et al.* QIIME allows analysis of high-
598 throughput community sequencing data. 2010:335–6.

599 17. Edgar RC. Search and clustering orders of magnitude faster than BLAST.
600 2010:2460–1.

601 18. DeSantis TZ, Hugenholtz P, Larsen N *et al.* Greengenes, a Chimera-Checked 16S
602 rRNA Gene Database and Workbench Compatible with ARB. *Appl Environ*
603 *Microbiol* 2006;**72**:5069–72.

604 19. Wang Q, Garrity GM, Tiedje JM *et al.* Na?ve Bayesian Classifier for Rapid
605 Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Appl Environ*
606 *Microbiol* 2007;**73**:5261–7.

607 20. Quast C, Pruesse E, Yilmaz P *et al.* The SILVA ribosomal RNA gene database
608 project: improved data processing and web-based tools. *Nucl Acids Res*
609 2013;**41**:D590–6.

610 21. Santiago-Rodriguez TM, Fornaciari G, Luciani S *et al.* Gut Microbiome of an
611 11th Century A.D. Pre-Columbian Andean Mummy. *PLOS ONE* 2015;**10**:e0138135.

612 22. Aagaard K, Ma J, Antony KM *et al.* The Placenta Harbors a Unique Microbiome.
613 *Sci Transl Med* 2014;**6**:237ra65-237ra65.

614 23. Weyrich LS, Llamas B, Cooper A. Reply to Santiago-Rodriguez et al.: Was luxS
615 really isolated from 25- to 40-million-year-old bacteria? *FEMS Microbiol Lett*
616 2014;**353**:85–6.

617 24. Kliman HJ. Comment on "The placenta harbors a unique microbiome." *Science*
618 *Translational Medicine* 2014;**6**:254le4-254le4.

619 25. Chang S-S, Kang D-H. Alicyclobacillus spp. in the Fruit Juice Industry: History,
620 Characteristics, and Current Isolation/Detection Procedures. *Critical Reviews in*
621 *Microbiology* 2004;**30**:55–74.

622 26. Hieken TJ, Chen J, Hoskin TL *et al.* The Microbiome of Aseptically Collected
623 Human Breast Tissue in Benign and Malignant Disease. *Scientific Reports*
624 2016;**6**:30751.

625 27. Eisenhofer R, Cooper A, Weyrich LS. Reply to Santiago-Rodriguez et al: Proper
626 authentication of ancient DNA is essential. *FEMS Microbiol Ecol* 2017, DOI:
627 10.1093/femsec/fix042.

628 28. Priest FG. Age of bacteria from amber. *Science* 1995;**270**:2015; author reply
629 2016-2017.

630 29. Beckenbach AT. Age of bacteria from amber. *Science* 1995;**270**:2015-2016;
631 author reply 2016-2017.

632 30. Austin JJ, Smith AB, Fortey RA *et al.* Ancient DNA from amber inclusions: a
633 review of the evidence. 1998.

634 31. Barton HA, Taylor NM, Lubbers BR *et al.* DNA extraction from low-biomass
635 carbonate rock: an improved method with reduced contamination and the low-
636 biomass contaminant database. *J Microbiol Methods* 2006;**66**:21–31.

637 32. Grahn N, Olofsson M, Ellnebo-Svedlund K *et al.* Identification of mixed bacterial
638 DNA contamination in broad-range PCR amplification of 16S rDNA V1 and V3
639 variable regions by pyrosequencing of cloned amplicons. *FEMS Microbiol Lett*
640 2003;**219**:87–91.

641 33. Chen T, Yu W-H, Izard J *et al.* The Human Oral Microbiome Database: a web
642 accessible resource for investigating oral microbe taxonomic and genomic
643 information. *Database (Oxford)* 2010;**2010**:baq013.

644 34. Kostic AD, Gevers D, Pedamallu CS *et al.* Genomic analysis identifies
645 association of Fusobacterium with colorectal carcinoma. *Genome Res* 2012;**22**:292–8.

646 35. Mayneris-Perxachs J, Bolick DT, Leng J *et al.* Protein- and zinc-deficient diets
647 modulate the murine microbiome and metabolic phenotype. *Am J Clin Nutr*
648 2016:ajcn131797.

649 36. Ridgway HF, Olson BH. Chlorine resistance patterns of bacteria from two
650 drinking water distribution systems. *Appl Environ Microbiol* 1982;**44**:972–87.

651 37. Liu L, Zhu W, Cao Z *et al.* High correlation between genotypes and phenotypes of
652 environmental bacteria Comamonas testosteroni strains. *BMC Genomics* 2015;**16**,
653 DOI: 10.1186/s12864-015-1314-x.

654 38. Sagripanti JL, Bonifacino A. Resistance of Pseudomonas aeruginosa to liquid
655 disinfectants on contaminated surfaces before formation of biofilms. *J AOAC Int*
656 2000;**83**:1415–22.

657 39. Dent LL, Marshall DR, Pratap S *et al.* Multidrug resistant Acinetobacter
658 baumannii: a descriptive study in a city hospital. *BMC Infectious Diseases*
659 2010;**10**:196.

660 40. Gray JS, Birmingham JM, Fenton JI. Got black swimming dots in your cell
661 culture? Identification of Achromobacter as a novel cell culture contaminant.
662 *Biologicals* 2010;**38**:273–7.

663 41. Krisko A, Radman M. Biology of Extreme Radiation Resistance: The Way of
664 Deinococcus radiodurans. *Cold Spring Harb Perspect Biol* 2013;**5**:a012765.

665 42. Wackett LP, Logan MSP, Blocki FA *et al.* A mechanistic perspective on bacterial
666 metabolism of chlorinated methanes. *Biodegradation* 1992;**3**:19–36.

667 43. Willerslev E, Hansen AJ, Binladen J *et al.* Diverse plant and animal genetic
668 records from Holocene and Pleistocene sediments. *Science* 2003;**300**:791–5.

669 44. Gilbert MTP, Jenkins DL, Götherstrom A *et al.* DNA from Pre-Clovis Human
670 Coprolites in Oregon, North America. *Science* 2008;**320**:786–9.

671 45. J Lu and SL Salzberg. Removing Contaminants from Metagenomic Databases.
672 Biorxiv. *bioRxiv* 261859; doi: https://doi.org/10.1101/261859
673
674 46. A. Borst, A. Box and A. Fluit: False-positive results and contamination in nucleic
675 acid amplification assays: suggestions for a prevent and destroy strategy. *European*
676 *Journal of Clinical Microbiology and Infectious Diseases* 2004;**23**(4): 289-299.
677
678 47. Xu Z, Zhang F, Xu B, Tan J, Li S, Jin L. Improving the sensitivity of negative
679 controls in ancient DNA extractions. *Electrophoresis*. 2009;**30**:1282–5.
680
681

682 **Data Accessibility**

683 QIIME demultiplexed sequences (16SContam_seqs_forpub.fna), a phylogenetic tree

684 of representative sequences (rep_set.tre), a biom table (otu_table_clean.biom), and

685 sample metadata (SampleInformation_20180820.txt) can be accessed from

686 https://figshare.com/account/articles/7283816 (doi: 10.25909/5bdaa4431a941).

Author Contributions

LSW and AGF conceived of the study. LSW, AGF, RE, JY, CS, MHD, and CA contributed samples and completed lab work. LSW, LA, and JB completed bioinformatic analysis of the data. LSW wrote the paper, and all authors edited and contributed to the final manuscript.

692 **Figure and Table Legends**

693 **Figure 1: Lower diversity is observed in EBC and NTC samples.**

694 The number of sequenced reads from samples that were all pooled at equimolar
695 concentrations is displayed on a box and whisker plot. (B) The alpha diversity of each
696 type sample (*i.e.* the within sample diversity) was calculated using observed species
697 metric in QIIME for rarefied 16S rRNA data. Each sample was rarefied up to 10,000
698 sequences in 500 sequence intervals; the standard error at each subsampling event is
699 displayed. Calculus samples are shown in blue, while control samples (extraction
700 blank controls (EBCs) and no-template controls (NTCs)) from the ancient laboratory
701 (AL) and the modern laboratory (ML) in red and green, respectively.
702
703 **Figure 2: Microbial phyla within controls are distinct from biological samples.**
704 The proportion of different microbial phyla are shown for a wide-array of modern and
705 ancient calculus samples and controls samples (EBCs and NTCs) from both
706 laboratory facilities (modern lab (ML) and ancient lab (AL)) and two different
707 extraction methods: the method employed in ancient DNA research and a
708 commercially available DNA extraction kit (kit). Rare phyla were collapsed if the
709 represented less than 0.001% of the total phyla observed.
710
711 **Figure 3: PCoA plots of control samples highlight differences in method and**
712 **laboratory.**
713 PCoA plots of unweighted UniFrac values were plotted in QIIME to compare beta
714 diversity differences (between samples differences) in all samples (A) or in different
715 laboratories (B). The different laboratory facilities are represented by ML (modern
716 lab) and AL (ancient lab), and the two control types are represented by EBC
717 (extraction blank control) or no-template control (NTC).
718
719 **Figure 4: PCoA analysis of extraction method and seasonal variation on**
720 **contaminant communities**
721 The modern and ancient calculus samples were removed from the analysis presenting
722 in Figure 3, and a PCoA plot was constructed of only control samples to identify
723 differences between the extraction method and laboratory in control samples (A). (B)
724 UniFrac values from controls samples (EBCs and NTCs) from the ancient laboratory
725 over a five-year period (2012 – 2016) are colored on a PCoA plot according to month.

Table 1: Dominant contaminant genera are largely unique within each laboratory.

The 69 genera that dominated EBC control samples are displayed for all sample types and include the proportion identified in each sample type. Genera were identified if dominant if they were found to be above 0.01% of the total genera identified within each laboratory. Taxa highlighted in green represent genera that dominated EBCs in the ancient laboratory, while unhighlighted are those from the modern EBC samples. If the genera were identified in previous studies that examined contamination, the reference number is shown in the right hand column.

Table 2: Extraction methods contain unique taxa.

OTUs identified as statistically significant (Kruskal-Wallis Bonferroni Corrected p-value <0.05) between the two extraction methods in the modern laboratory are listed. OTUs highlighted in green were significantly within the QG method, while highlighted OTUs were significant in the kit extraction method.

Figure 1

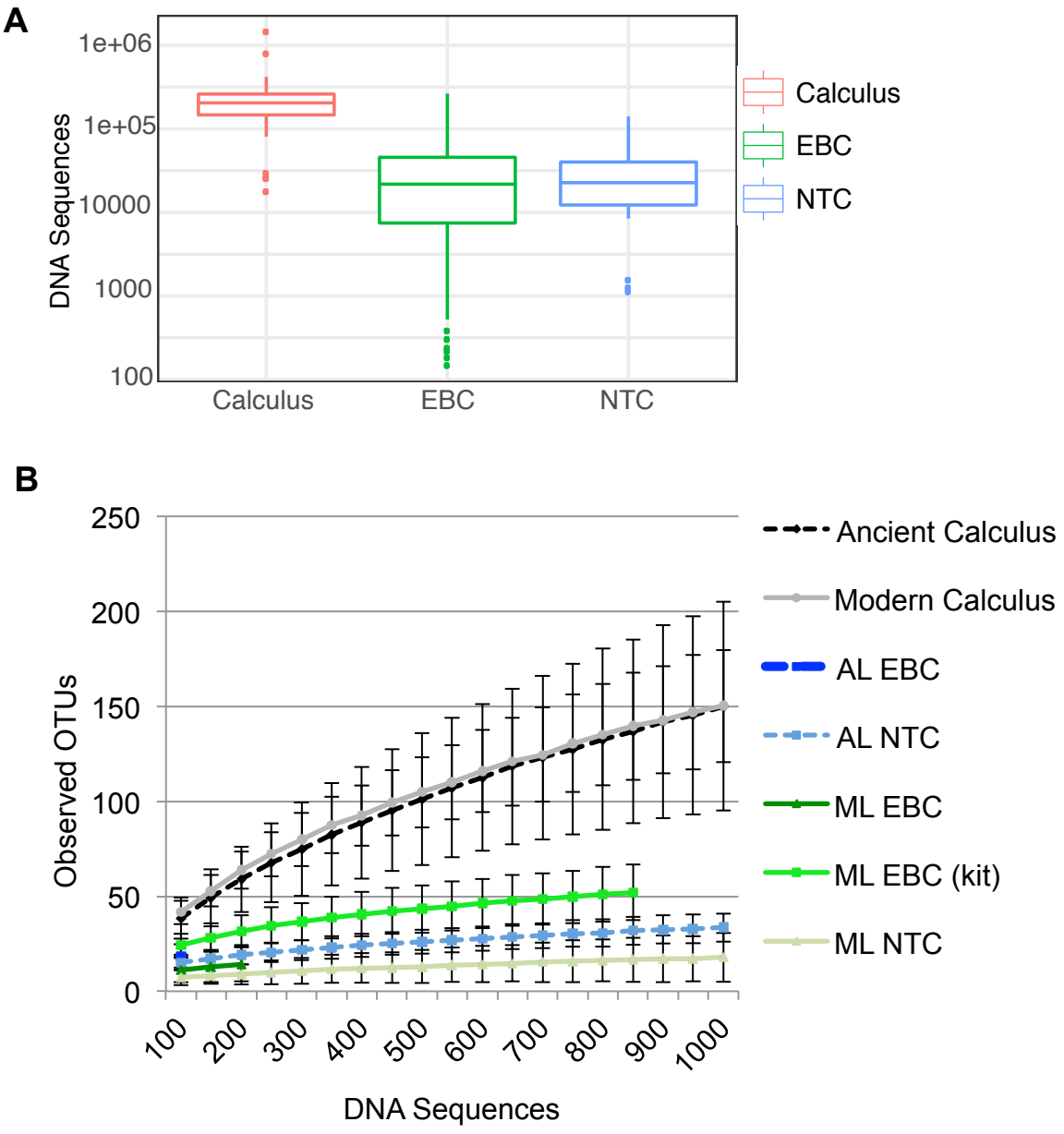

A
1e+06
1e+05
10000
1000
100
DNA Sequences
Calculus
EBC
NTC
Calculus
EBC
NTC
B
250
200
150
100
50
0
Observed OTUs
100
200
300
400
500
600
700
800
900
1000
DNA Sequences
Ancient Calculus
Modern Calculus
AL EBC
AL NTC
ML EBC
ML EBC (kit)
ML NTC

Figure 2

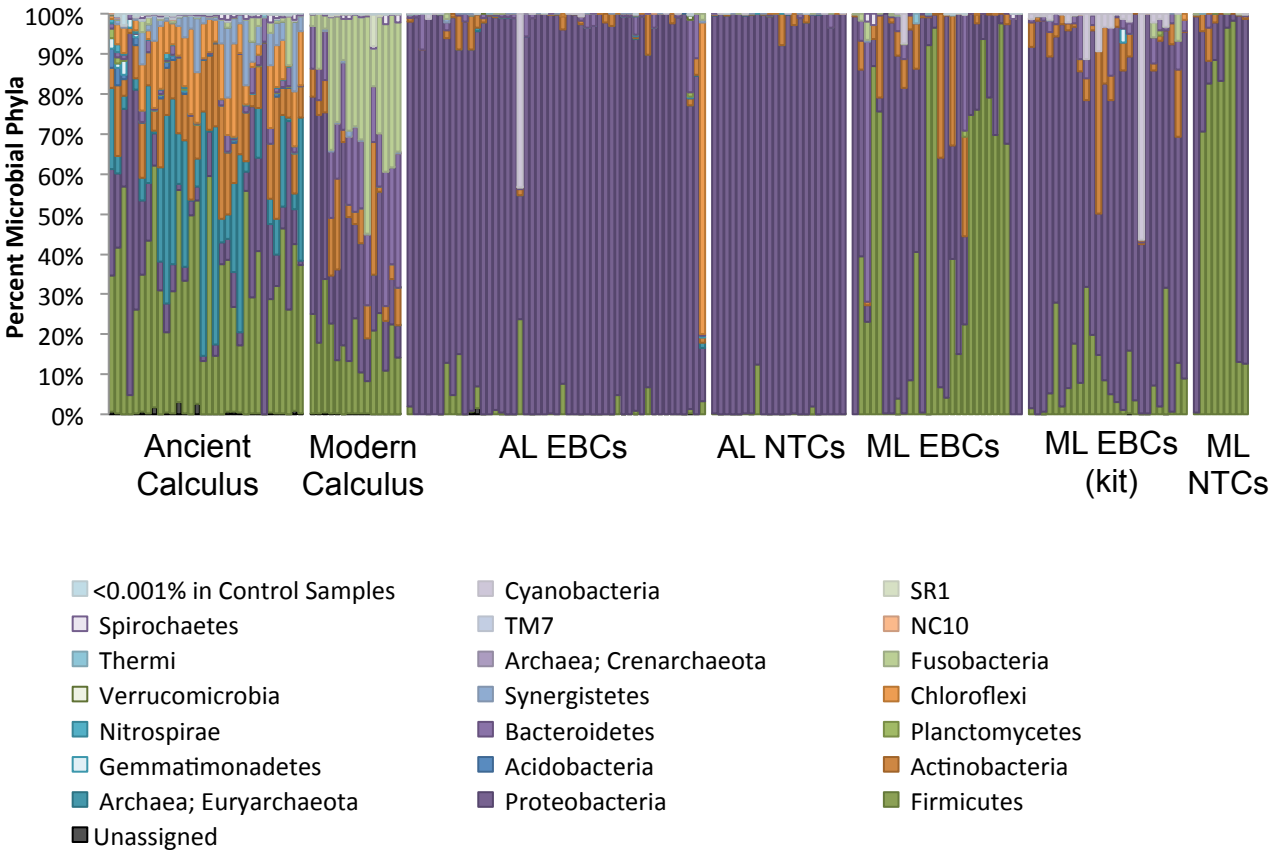
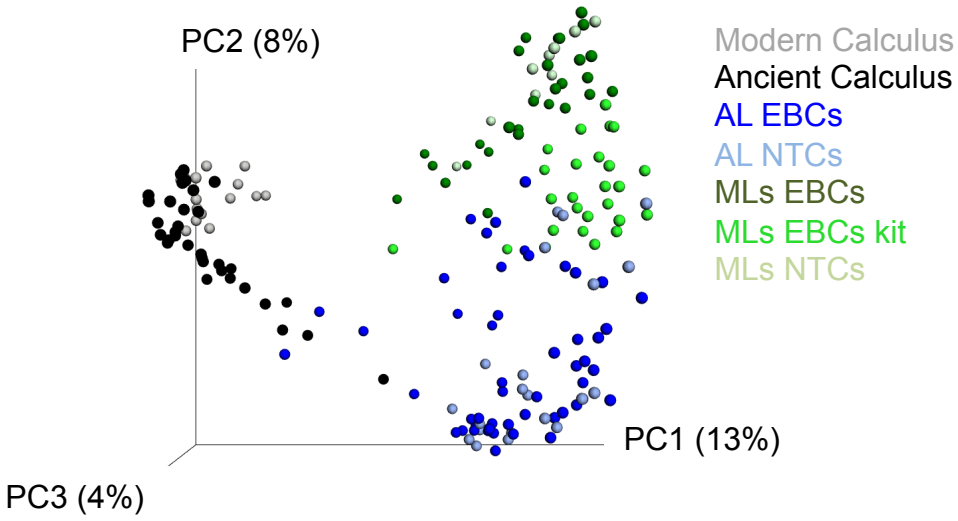

Percent Microbial Phyla
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Ancient Calculus
Modern Calculus
AL EBCs
AL NTCs
ML EBCs
ML EBCs (kit)
ML NTCs
<0.001% in Control Samples
Spirochaetes
Thermi
Verrucomicrobia
Nitrospirae
Gemmatimonadetes
Archaea; Euryarchaeota
Unassigned
Cyanobacteria
TM7
Archaea; Crenarchaeota
Synergistetes
Bacteroidetes
Acidobacteria
Proteobacteria
SR1
NC10
Fusobacteria
Chloroflexi
Planctomycetes
Actinobacteria
Firmicutes

Figure 3

A


PC2 (8%)
PC1 (13%)
PC3 (4%)
Modern Calculus
Ancient Calculus
AL EBCs
AL NTCs
MLs EBCs
MLs EBCs kit
MLs NTCs

B

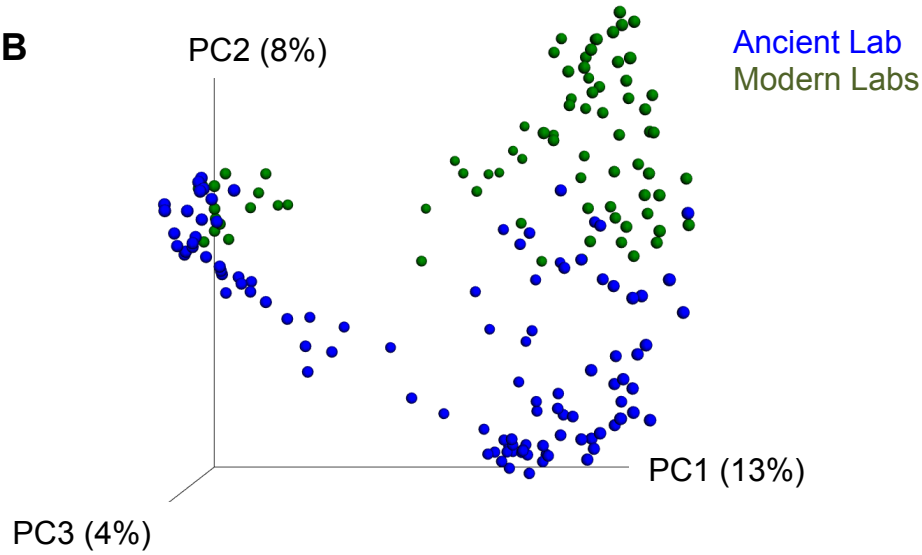
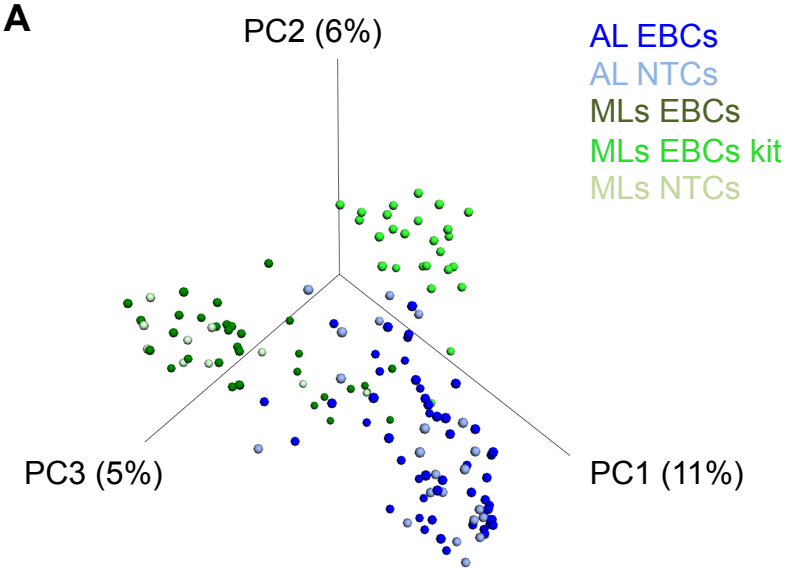

PC2 (8%)
PC1 (13%)
PC3 (4%)
Ancient Lab
Modern Labs

Figure 4

A


PC2 (6%)
AL EBCs
AL NTCs
MLs EBCs
MLs EBCs kit
MLs NTCs
PC3 (5%)
PC1 (11%)

B

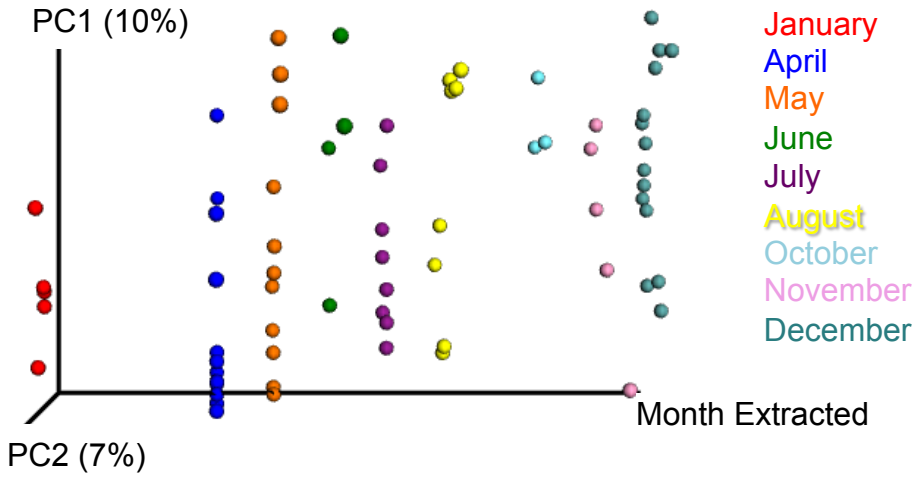

PC1 (10%)
January
April
May
June
July
August
October
November
December
Month Extracted
PC2 (7%)

Table 1

Genera Taxonomy	AL EBC	ML EBC	ML EBC (kit)	AL NTC	ML NTC	Identified Previously (Ref)
Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Actinomycetaceae;g__Actinomyces	0.000243	0.001159	0.002897	9.89E-05	1.33E-05	
Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Actinomycetaceae;g__N09	0	0.012119	0	0	0	
Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Corynebacteriaceae;g__Corynebacterium	0.000294	0.025472	0.010478	0.001034	0.00033	3
Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Dermacoccaceae;g__Dermacoccus	0	0.005987	0	0	0.00399	
Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Microbacteriaceae;g__	0.000644	0.017225	0.000293	5.38E-05	0	
Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Micrococcaceae;g__	6.30E-05	0.001636	0.002447	2.50E-06	0.00011	
Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Micrococcaceae;g__Micrococcus	0.000229	0.002165	0.002806	2.50E-06	4.00E-05	3
Bacteria;p__Bacteroidetes;c__[Saprospirae];o__[Saprospirales];f__Chitinophagaceae;g__Sediminibacterium	0.001729	0	2.17E-06	0	0	
Bacteria;p__Bacteroidetes;c__Cytophagia;o__Cytophagales;f__Cytophagaceae;g__	0.00173	0	0.000884	1.25E-06	0	
Bacteria;p__Bacteroidetes;c__Flavobacteriia;o__Flavobacteriales;f__[Weeksellaceae];g__	6.68E-06	0.001492	0	5.01E-06	6.67E-06	
Bacteria;p__Bacteroidetes;c__Flavobacteriia;o__Flavobacteriales;f__[Weeksellaceae];g__Chryseobacterium	0.000571	0.004948	0.000128	0	0	3
Bacteria;p__Bacteroidetes;c__Flavobacteriia;o__Flavobacteriales;f__[Weeksellaceae];g__Cloacibacterium	0.010137	0.039939	0.006025	0.000769	0.00207	
Bacteria;p__Bacteroidetes;c__Flavobacteriia;o__Flavobacteriales;f__[Weeksellaceae];g__Wautersiella	0.002471	8.33E-05	0.000113	0.001456	0	
Bacteria;p__Bacteroidetes;c__Flavobacteriia;o__Flavobacteriales;f__Cryomorphaceae;g__Fluviicola	0.001045	0	0	0	0	
Bacteria;p__Bacteroidetes;c__Flavobacteriia;o__Flavobacteriales;f__Flavobacteriaceae;g__Flavobacterium	0.002652	0.004036	0.000325	0	2.00E-05	3, 4
Bacteria;p__Bacteroidetes;c__Sphingobacteriia;o__Sphingobacteriales;f__;g__	0.001722	0	6.28E-05	0	0	
Bacteria;p__Bacteroidetes;c__Sphingobacteriia;o__Sphingobacteriales;f__Sphingobacteriaceae;g__	0	0.002581	0	0	0	
Bacteria;p__Bacteroidetes;c__Sphingobacteriia;o__Sphingobacteriales;f__Sphingobacteriaceae;g__Pedobacter	0.001991	2.08E-06	0	0	6.67E-06	3
Bacteria;p__Cyanobacteria;c__Chloroplast;o__Streptophyta;f__;g__	0.009702	0.000779	0.051116	2.50E-06	6.67E-06	
Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__[Thermicanaceae];g__Thermicanus	0	0.002846	0	0	0	
Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__Bacillaceae;g__Bacillus	0.000154	0.009569	0.001742	2.75E-05	6.67E-06	3
Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__Bacillaceae;g__Geobacillus	0	0.002712	0.000165	0	0	
Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__Paenibacillaceae;g__Paenibacillus	0.001262	0.465748	0.000121	2.13E-05	0.82364	3
Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__Staphylococcaceae;g__Staphylococcus	0.000884	0.022356	0.002642	0.001866	0.00876	
Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Aerococcaceae;g__	0.000368	0.006397	0.001008	1.25E-06	0	
Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Enterococcaceae;g__Enterococcus	0	0.024966	2.82E-05	0	8.01E-05	
Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Leuconostocaceae;g__Leuconostoc	0.001056	0.000633	0	0	0	
Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Lactococcus	0.003898	0	0.003962	0	0	
Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus	0.000944	0.004338	0.022634	4.01E-05	0.00388	3
Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Clostridiaceae;g__Clostridium	0.000102	0.001049	0.002926	0.000879	0.00102	
Bacteria;p__Planctomycetes;c__Planctomycetia;o__Pirellulales;f__Pirellulaceae;g__	0.000177	0.002577	0	0	6.67E-06	
Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__;g__	0.00118	0.000518	0.000938	0.001164	6.67E-06	
Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Bradyrhizobiaceae;g__	0.001811	4.16E-06	0.016817	0.001168	0	
Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Bradyrhizobiaceae;g__Bradyrhizobium	0.006279	0.0004	0.003966	0.013655	0.00169	3, 4
Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Hyphomicrobiaceae;g__Devosia	0.001031	0	0.006283	0	0	3
Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Methylobacteriaceae;g__Methylobacterium	0.021743	0.000162	0.001198	0.04655	0	3, 28
Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Phyllobacteriaceae;g__Mesorhizobium	0.005669	0.000552	0.000598	0	0	3
Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhodospirillales;f__Acetobacteraceae;g__	4.17E-07	0.001397	0.001554	0	0.00702	
Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rickettsiales;f__mitochondria;Other	0.001506	0	0.001207	0.000476	0	
Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Sphingomonadales;f__Erythrobacteraceae;g__	0.000717	0.085193	0.009221	0	0.0591	
Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Sphingomonadales;f__Erythrobacteraceae;Other	0	0.001492	8.67E-06	0	4.00E-05	
Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Sphingomonadales;f__Sphingomonadaceae;g__	0.001988	0.000327	0.001296	9.51E-05	6.67E-06	
Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Sphingomonadales;f__Sphingomonadaceae;g__Novosphingobium	0.015942	0.002879	0.006879	0.044996	0.00165	3
Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Sphingomonadales;f__Sphingomonadaceae;g__Sphingobium	0.002402	0	0.009915	1.88E-05	0	3
Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Sphingomonadales;f__Sphingomonadaceae;g__Sphingomonas	0.007868	0.004113	0.00788	0.017517	0.00798	3, 28, 4
Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__;f__;g__	0.000291	0.001116	0	1.25E-06	0	
Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__Alcaligenaceae;g__Achromobacter	0.001025	2.08E-06	0	5.13E-05	0	
Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__Burkholderiaceae;g__Burkholderia	0.029885	0	0.000282	0.00102	0	3, 4
Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__Comamonadaceae;g__	0.02727	0.061636	0.014919	0.004361	0.01679	
Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__Comamonadaceae;g__Comamonas	0.16639	0.000181	0.055275	0.131521	0.0001	3, 28
Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__Comamonadaceae;g__Limnohabitans	0.007023	0.001341	0.002967	0.010622	0	
Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__Comamonadaceae;g__Tepidimonas	0.032485	0.003524	0.000483	0.002675	0	
Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__Comamonadaceae;Other	0.054856	0.004001	0.025356	0.072105	0.00195	
Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__Oxalobacteraceae;g__Cupriavidus	0.002021	0.001418	0.000542	0.00034	0	3, 28
Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__Oxalobacteraceae;g__Ralstonia	0.010983	0.013709	0.015801	0.045934	0.00927	3, 28, 4, 29
Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__Oxalobacteraceae;Other	0.002536	0.00011	0.001636	0.018057	0	
Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Methylophilales;f__Methylophilaceae;g__	0.002073	0	0	0	0	
Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Rhodocyclales;f__Rhodocyclaceae;g__	0.00023	0.001971	0.002921	0.001048	0	
Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Aeromonadales;f__Aeromonadaceae;g__	7.81E-05	0.001607	0.00405	1.25E-06	0	
Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacteriales;f__Enterobacteriaceae;g__	0.01923	0.082828	0.039958	0.036438	0.00017	
Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacteriales;f__Enterobacteriaceae;Other	0.014787	1.25E-05	0.005398	0.030091	0.00438	
Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Moraxellaceae;g__Acinetobacter	0.393093	0.001653	0.264511	0.411117	0.00023	3, 2, 28
Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Moraxellaceae;g__Enhydrobacter	0.001156	0.018176	0.000503	0.002678	0.01944	3
Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Pseudomonadaceae;g__	0.00019	0.005277	0.006005	7.51E-05	6.67E-05	
Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Pseudomonadaceae;g__Pseudomonas	0.08139	0.004871	0.02539	0.082003	0.00877	3, 29, 4
Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Pseudomonadaceae;Other	0.005272	0.014047	0.000178	5.01E-06	0.00171	
Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__	0.005478	0.000377	0.001892	0.001645	0	
Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Lysobacter	0.001864	2.08E-06	0	0	0	
Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Stenotrophomonas	0.001809	0.001405	0.000761	0.000969	0.00039	3, 2, 28, 29, 4

Table 2

OTU Taxonomy	Mean Seqs/Sample Kit	QG
Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales	6.535714	0.008696
Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__Alicyclobacillaceae;g__Alicyclobacillus;s__	2.821429	0
Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__Alicyclobacillaceae;g__Alicyclobacillus;s__	7.214286	0
Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__Alicyclobacillaceae;g__Alicyclobacillus;s__	1.928571	0
Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__Alicyclobacillaceae;g__Alicyclobacillus;s__	2.071429	0
Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__Alicyclobacillaceae;g__Alicyclobacillus;s__	1.714286	0
Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Oceanospirillales;f__Halomonadaceae;g__Halomonas;s__	0.035714	0
Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Pseudonocardiaceae;g__Pseudonocardia;s__	0.035714	0
Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Moraxellaceae;g__Acinetobacter;s__	385.2857	0.026087
Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Bradyrhizobiaceae;g__;s__	93.42857	0
Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Bradyrhizobiaceae;g__;s__	91.85714	11.72174
Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__Alicyclobacillaceae;g__Alicyclobacillus;s__	0.857143	0
Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__Alicyclobacillaceae;g__Alicyclobacillus;s__	0.857143	0
Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Neisseriales;f__Neisseriaceae;g__Vogesella;s__	184.4643	0
Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Bradyrhizobiaceae	2.357143	0.008696
Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Moraxellaceae;g__Acinetobacter;s__	395.3929	0.626087
Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Sphingomonadales;f__Sphingomonadaceae;g__Sphingobium;s__	160.75	33.24348
Bacteria;p__Firmicutes;c__Erysipelotrichi;o__Erysipelotrichales;f__Erysipelotrichaceae;g__Allobaculum;s__	0.607143	0
Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__Alicyclobacillaceae;g__Alicyclobacillus;s__	0.428571	0
Bacteria;p__Verrucomicrobia;c__Verrucomicrobiae;o__Verrucomicrobiales;f__Verrucomicrobiaceae;g__Akkermansia;s__muciniphila	0.5	0
Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Moraxellaceae;g__Acinetobacter;s__	82.14286	0.73913
Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Bradyrhizobiaceae	1.964286	1.721739
Bacteria;p__Firmicutes;c__Erysipelotrichi;o__Erysipelotrichales;f__Erysipelotrichaceae;g__Allobaculum;s__	0.464286	0
Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__Alicyclobacillaceae;g__Alicyclobacillus;s__	0.392857	0
Bacteria;p__Cyanobacteria;c__Chloroplast;o__Streptophyta;f__;g__;s__	2	0.008696
Bacteria;p__Cyanobacteria;c__Chloroplast;o__Streptophyta;f__;g__;s__	26.10714	1.547826
Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales	1.964286	1.991304
Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacteriales;f__Enterobacteriaceae;g__;s__	5.392857	23.46087
Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Pseudomonadaceae;g__Pseudomonas;s__	0	0.008696
Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__Comamonadaceae;g__;s__	0	0.008696
Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Pseudomonadaceae;g__Pseudomonas;s__	0	0.008696
Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__Comamonadaceae;g__;s__	0	0.008696
Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacteriales;f__Enterobacteriaceae;g__;s__	5.571429	44.6
Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Oceanospirillales;f__Alcanivoracaceae;g__Alcanivorax;s__	0	0.017391
Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacteriales;f__Enterobacteriaceae;g__;s__	4.785714	20.73043
Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacteriales;f__Enterobacteriaceae;g__;s__	4.928571	23.37391
Bacteria;p__Proteobacteria;c__Epsilonproteobacteria;o__Campylobacterales;f__Helicobacteraceae;g__;s__	0	0.017391
Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacteriales;f__Enterobacteriaceae;g__;s__	0	0.017391
Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Oceanospirillales;f__Halomonadaceae;g__Halomonas;s__	0	0.026087
Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacteriales;f__Enterobacteriaceae;g__;s__	1.892857	4.33913
Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__Oxalobacteraceae;g__Cupriavidus;s__	0	0.034783
Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacteriales;f__Enterobacteriaceae;g__;s__	1.107143	5.921739
Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhodobacterales;f__Rhodobacteraceae;g__;s__	0	0.026087
Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Alteromonadales;f__Alteromonadaceae;g__Marinobacter;s__	0	0.034783