environmental microbiology

Environmental Microbiology (2019) **21**(7), 2440–2468 doi:10.1111/1462-2920.14632

ANCHOR: a 16S rRNA gene amplicon pipeline for microbial analysis of multiple environmental samples

Emmanuel Gonzalez,[1,2] Frederic E. Pitre[3,4] and Nicholas J. B. Brereton [3*]

[1] *Canadian Centre for Computational Genomics, McGill University and Genome Quebec Innovation Centre, Montréal, QC, H3A 0G1, Canada.*

[2] *Department of Human Genetics, McGill University, Montreal, H3A 1B1, Canada.*

[3] *Institut de Recherche en Biologie Végétale, University of Montreal, Montreal, QC, H1X 2B2, Canada.*

[4] *Montreal Botanical Garden, Montreal, QC, H1X 2B2, Canada.*

Summary

Analysis of 16S ribosomal RNA (rRNA) gene amplification data for microbial barcoding can be inaccurate across complex environmental samples. A method, ANCHOR, is presented and designed for improved species-level microbial identification using paired-end sequences directly, multiple high-complexity samples and multiple reference databases. A standard operating procedure (SOP) is reported alongside benchmarking against artificial, single sample and replicated mock data sets. The method is then directly tested using a real-world data set from surface swabs of the International Space Station (ISS). Simple mock community analysis identified 100% of the expected species and 99% of expected gene copy variants (100% identical). A replicated mock community revealed similar or better numbers of expected species than MetaAmp, DADA2, Mothur and QIIME1. Analysis of the ISS microbiome identified 714 putative unique species/strains and differential abundance analysis distinguished significant differences between the Destiny module (U.S. laboratory) and Harmony module (sleeping quarters). Harmony was remarkably dominated by human gastrointestinal tract bacteria, similar to enclosed environments on earth; however, Destiny module bacteria also derived from nonhuman microbiome carriers present on the ISS, the laboratory's research animals. ANCHOR can help substantially improve sequence resolution of 16S rRNA gene amplification data within biologically replicated environmental experiments and integrated multidatabase annotation enhances interpretation of complex, nonreference microbiomes.

Received 15 November, 2018; revised 7 April, 2019; accepted 15 April, 2019. *For correspondence. E-mail nicholas.brereton@umontreal.ca; Tel. +15143436111; ext. 82096

Background

Over the last 50 years, the 16S ribosomal RNA (rRNA) gene has been one of the most commonly used molecular barcodes for profiling bacteria present within complex microbial communities. Initially, short oligo (20 nt) catalogues (fingerprints) were produced through RNase T_1 digestion (Fox *et al*., 1977) before increasingly more advanced sequencing technologies and bioinformatics approaches allowed for more resolved phylogenetic relationships to be inferred from directly from sequences (Olsen *et al*., 1986; Pace *et al*., 1986; Muyzer *et al*., 1993; Schloss *et al*., 2009; Caporaso *et al*., 2010). The utility of barcoding technology relies on the very highly conserved function of 16S rRNA leading to sequence regions of hyperconservation within the 16S rRNA gene. Primers can be designed to target this conserved region and amplify proximal hypervariable sequence regions (an amplicon) not under functional constraint as a potentially unique barcode of life.

Woese *et al*. (Woese *et al*., 1983) first described how secondary structure of 16S rRNA can vary between species (Rehakova *et al*., 2014; Ziesemer *et al*., 2015), leading to diversity in hypervariable regions so readily exploited as barcodes but also resulting in a lack of *universally* conserved sequence regions (Martinez-Porchas *et al*., 2017). Despite this, while universal primers do not exist (over 27,000 papers contain the terms '16S rRNA' AND 'universal primers'), there are a substantial number of commonly used primer pairs that will likely amplify 16S rRNA gene regions in over 90% of known and well-characterized bacterial species (Klindworth *et al*., 2013).

One of the difficulties in identifying species using 16S barcoding is that intragenomic variation is often present (variation between gene copies within a genome). Pei et al (Pei *et al*., 2010) investigated 822 bacterial genomes (copy numbers varied between 1 and 15) and found very high sequence variation within species in some cases, such as 21.8% sequence diversity in *Borrelia afzelii* K78 (a likely pseudogene) or 11.5% diversity in *Caldanaerobacter subterraneus* subsp. tengcongensis MB4(Acinas *et al*., 2004). In most cases, however, there is little variation in secondary structure of 16S rRNA between different gene copies, resulting in the majority varying by less than 1% in sequence similarity and the exceptions to this usually retaining secondary structure (<1% diversity)(Pei *et al*., 2010). This intragenomic functional constraint was most severely illustrated in *Thermoanaerobacter tengcongensis*, where 16S rRNA gene copies *rrsB* and *rrsC* vary by 6.70% but secondary structure varies by only 0.52% (*as predicted by free energy minimization). Such high variation between gene copies could be a considerable challenge; between the 8485 bacterial genomes gathered within rrnDB database (12.8.18) (Klappenbach *et al*., 2001), the average intragenomic gene copy number is 4.7, with three copies being the most frequent. The maximum known 16S rRNA gene copy number in rrnDB is currently *Aneurinibacillus soli* CB4 and *Brevibacillus formosus* NF2, each with 17 copies, as well as *Clostridium beijerinckii*, which has 16 copies (part of Kozich's mock community investigated below).

Understanding the nature of biological variation in this molecule and recognizing the potential challenges associated with unknown biology can serve to increase the power of 16S rRNA technology. Kou *et al*. (Kou *et al*., 2018) demonstrated the biological power of this in studying the effect of metal pollution on soil when identifying putative cross-domain functional niche replacement of a nitrate-oxidizing archaea by the metal tolerant nitrospirae bacteria *Nitrospira moscoviensis.* Recognition of the variable utility of barcoding technology can be used to identify when single species resolution is not possible using a specific amplicon, enabling recognition of when species *can* be confidently identified.

A method designed deliberately for high complexity systems and the retention of maximal information in each step of sequence processing is presented, ANCHOR. The approach borrows heavily from RNAseq techniques with classical biological experimental design in mind, in particular a focus on identifying bacteria species and utility for hypothesis query using replicated samples (Weiss *et al*., 2017; Gonzalez *et al*., 2018).

Experimental procedures

ANCHOR method

Data sets used for benchmarking. Two artificial data sets [Even and Staggered (Kopylova *et al*., 2016)], two mock communities [Kozich's mock (Kozich *et al*., 2013) and Kleiner's mock (Kleiner *et al*., 2017)] and a real-world data set [ISS data set (Lang *et al*., 2017)] have been investigated using ANCHOR (see supplementary file 1 – data set specifics, for more information). The increasing data set complexity is used to assess the challenges of real-world systems and test the method's potential for biological discovery. The ISS data set [surface swabs were taken on May 9, 2014 (Lang *et al*., 2017)] was deliberately selected as technically nonideal and biologically complex: sampling had no replicated biological comparison design. A design was applied *a posteriori*, predicated on sampling location with unbalanced replication (destiny module = 10n while harmony module = 4n). Procedural specifics for each data set are included in supplementary file 1; these include threshold testing for high-count sequence identification, high-count sequence annotation and low-count sequence capture steps, a primer wild card step (optional for when degenerate primers are used), parameters used in comparative methods, chimera flagging and differential abundance analysis.

Preprocessing. Raw paired-end reads from Illumina MiSeq can be used directly as a starting point for the ANCHOR pipeline (Fig. 5). Trimming the sequences, controlling for high-quality reads and removing primers constitute alternative starting points. Whenever possible, the primers were left within the read sequences in the data set presented here. Retention of primer sequences is recommended, even when degenerate primers are used to allow for exploration of PCR bias and to ensure no amplicons are annotated as species that could not be amplified. This is also in line with the intention of ANCHOR to alter amplified sequences as little as possible and allow for observation unexpected biology.

Contig assembly. Amplicons are assembled using fast read aligners such as Mothur (Schloss *et al*., 2009), FLASH (Magoč and Salzberg, 2011), PEAR (Zhang *et al*., 2013), USEARCH (Edgar, 2010) and PANDAseq (Bartram *et al*., 2011). Fast-read aligners provide assembled contigs (potential amplicons) with diverse lengths and qualities. Users can choose to discard low-quality contigs containing a high percentage of mismatches or ambiguous bases (Ns), or limit contigs to a targeted amplicon length. If faithful sample representation is of concern, it is important to allow for unexpected amplicon length, as any target region of the 16S rRNA gene (and rRNA molecule) has the potential to vary between species [intervening sequences >10 nts are common (Pei *et al*., 2010)]. As an example, Kleiner's mock community contained three relatively abundant amplicon lengths: 465, 453 and 440 nt (±2) (Supplementary file 1 figure).

High-count sequence identification. The assembled amplicons can be dereplicated (reduce sequence pool to unique sequences) to speed up processing time using tools such as Mothur (Schloss *et al*., 2009) (used here), CD-Hit (Li and Godzik, 2006), USEARCH (Edgar, 2010) or FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). Dereplication provides a count for each unique sequence. As an example, the 12 samples in Kleiner's mock community generate unique sequences with counts ranging from 1 (represented in a single sample) to up to 5005 (represented across all samples). High-count sequences are used as a confident basis, or anchor, for analysis. This confidence derives from the idea that high-count sequences will more likely represent accurate sequences from microbes. High-count sequences are selected through a count threshold decided by the user, based on the biological system or hypothesis under investigation. Two recommended default options are provided for selection based on the number of biological replicates used and with relevance to the biological question being posed: a *minimum difference* or a *high confidence* threshold. A *minimum difference* threshold is based on one count in each biological replicate of a single factor or condition, counted across all samples of an experiment; the factor or condition with the fewest biological replicates representing the minimum requirement for observing a difference (ANCHOR requires counts in at least three biological replicates). A *high confidence* threshold is based on three counts in each biological replicate of a single factor, counted across all samples of an experiment. The choice of high count threshold is paramount to the analysis and revolves around the dilemma that: (i) a low threshold will likely inflate OTU numbers but sequences from low abundance microbes will be retained (type II error protection) and (ii) a high threshold will restrain OTU numbers but sequences from low abundance microbes are more likely to be discarded (type I error protection). The *minimum difference* threshold stemmed *from the use of a minimum significantly* different characteristics calculation to drive clustering, suggested by Gyllenberg [(Gyllenberg, 1963); reported in Sneath (1964), Sokal, (1965) and Lapage *et al*. (1973)].

Annotation selection. Once high-count sequences are selected, they can be annotated against databases relevant to the user's preference or experimental design. Default recommended annotation uses four sequence repositories with strict BLASTn criteria (>99% identity and coverage) providing each amplicon with up to four pools of labels from: NCBI-curated bacterial and Archaea RefSeq, NCBI nr/nt, SILVA and Ribosomal Database Project (RDP). Annotation selection against the four databases is based on *de novo* metatranscriptomics strategy (Brereton *et al*., 2016; Gonzalez *et al*., 2018), where all potential annotation is retained to allow for informed annotation selection and downstream interpretation. BLASTn returns with a query identity and coverage <99% are discarded and the highest identity/coverage scores are selected per query. When the highest identity/coverage for a given high-count sequence is shared amongst different blast returns from a database, all are retained as equally 'good' annotation and designated as *ambiguous hits* (borrowed from the idea of secondary annotation in metatranscriptomics [Brereton *et al*., 2016; Gonzalez *et al*., 2018)]. Accurately reporting ambiguity is important as fragments of a specific 16S rRNA gene [or a gene's entire sequence (Vetrovsky and Baldrian, 2013)] are sometimes 100% similar between known species; the relative utility of a specific amplicon as a barcode therefore varies based on the species present in a sample as well as the technology used for sequencing and data processing.

A prioritization strategy for high-count sequence annotation is recommended for complex samples and is used here to annotate the ISS and both mock community data sets (Supplementary files 1, 3, 4 and 5). Annotation using the NCBI-curated Bacterial and Archaea RefSeq database at 100% identity and coverage is selected if present. If no 100% species hit is found, annotation using the best bitscore from any species >99% identity from the four repositories is selected. If no >99% species hit is found, annotation using the best bitscore from any taxon >99% identity from the four repositories is selected. Selection against 'unknown bacteria' annotation is also applied. This allows for NCBI-curated bacterial and Archaea RefSeq to be prioritized, the stringent curation criteria (see RefSeq Targeted Loci Project) leading to comparatively fewer database errors (see *L. monocytogenes* in results from Kozich's mock). NCBI nr/nt inclusion allows for annotation of nonbacterial/archaeal amplicons present in most data sets (even very highly ambiguous sequences can be biologically valuable for identifying confounding effects in downstream interpretation involving sample abundance). Selection against 'unknown bacteria' then ultimately leads to previously observed but uncharacterized sequences (known unknowns) with 100% identity to high-count sequences often being annotated using phylogenetic placement [while powerful, this is not prioritized due to a high error rate (Edgar, 2018)].

When a high-count sequence (or OTU) is best annotated by multiple hits, ambiguity is recorded in the output by an annotation label corresponding to the lowest common taxonomic level and the suffix '_MS' (for multiple species). BLASTn returns rejected due to databases prioritization are also made available for downstream data exploration. Presenting multiple species hits is an essential step for identifying annotation that *does not* present ambiguity, thus allowing for more confident species calls (identifying when single species-level annotation is sensible).

A limitation of this stringent (>99% identity) annotation strategy is that less well-characterized bacteria, such as most members of the TM7 group/Candidatus Saccharibacteria (Hugenholtz *et al*., 2001), will very often be annotated as unknown due to a lack of knowledge and associated databases entries from which to draw comparison against ANCHOR OTUs. Sequences are presented (OTU table) alongside OTUs annotated as unknown to allow for lower similarity BLASTn or reannotation as new species of bacteria are discovered and characterized. Conversely, high similarity annotation derived from uncurated sequence repositories can contain extensive errors and, while extremely valuable for biological discovery, need to be carefully reviewed on a record by record basis with independent consensus and peer-review of entries in-mind.

While reviewing annotation can be time-consuming in complex systems, thorough data analysis is the best way to maximize biological findings given inconsistencies across databases.

Low-count sequence capture. Sequences rejected as high-count sequences can account for a nonnegligible proportion of a given data set (e.g., 62.7% in Kleiner's mock with a high confidence count threshold of 12). Based on an assumption that rejected sequences from low abundance species are more likely to be distant (dissimilar) from high-count sequences than rejected sequences originating from technical errors, low-count sequences are binned to high-count sequences in a second BLASTn (query: low-count sequences; subject: high-count sequences), the distinction becoming progressively more important as sample complexity increases. A reduced low-count binning threshold of 98% identity/coverage is recommended and was selected for the presented data sets (note that the coverage threshold is applied on both queries and subjects). No new high-count sequences permitted to be formed during this process (owing to the theory that the majority of captured sequences should not derive from low-abundance species). Low-count sequences with <98% identity/coverage to a high-count sequence are fully discarded. The proportion of discarded data can vary across experiments: 3.9% of the initial amplicons were discarded in Kozich's mock data set, 19.9% in Kleiner's mock data set and 18.3% in the ISS data set.

Accession ID collapsing. The previous steps provide a count matrix for all high-count sequences as well as annotation. Attribution of a low-count sequence to a high-count sequence can be imprecise if high-count sequences are highly similar. To this regard, an *accession ID collapsing* step groups sequences with the same database accession ID into OTUs. While relying on database integrity, this step has an advantage over collapsing high-count sequences based on common taxonomy by separating different sequences assigned to a common taxonomic label (collapsing sequences have to share >99% identity to a common accession ID, so are <2% dissimilar). For example, at a high-count threshold of 12, 158 contigs share a total of 34 different accession IDs leading to 24 different annotation labels in Kleiner's mock. Another advantage is to create a count profile closer to a completely *de novo* approach but which also takes advantage of valuable, known, biology. It should be noted that the collapsing to a shared annotation accession does not, of course, reduce the negative impact of taxonomic mislabelling based on database errors. However, homologous accessions deposited online are unlikely to be artificial amplicons and so represent a useful fixed point of confidence to group highly similar sequences (<1% difference). Accession ID collapsing is employed throughout with the benchmarking data sets and is a strongly suggested option for multiple potentially complex biological samples. Alternative options are *de novo* (count matrix is based on high-count sequences alone) and *taxonomic annotation collapsing* (all sequences with a common taxon label are collapsed together into OTUs). The *de novo* option has all the advantages of being a sequence database-independent method, although it can separate sequences whose difference may only be attributed to small technical variations. This is problematic in high-complexity samples where low-count sequences can derive from either technical error or low sequencing depth (relatively low abundance species). Although reducing the number of final OTUs, *taxonomic annotation collapsing* has the disadvantage of relying very heavily on database input and integrity. For example, two distant sequences (low percentage similarity) can be collapsed together into a same label, thus obscuring important information and potentially confounding postprocessing interpretation (e.g., summating contradictory responses to a condition). This bias is directly linked to database integrity and should improve as the quality of records improves over time.

Terminology. An OTU suffix of 'Multiple Species' (_MS) is used to highlight when multiple species are equally likely annotation based on 16S rRNA gene amplicon sequence similarity (*see* Annotation selection; or MG, MF, etc. for multiple genera or family level annotation etc.). Presenting each possible species is preferred over moving up a taxonomic level to genus as, for example, other species within a genus can often be confidently discounted, which can be biologically informative. The term OTU has been criticized due to the bioinformatics step of 97% sequence clustering previously used in Qiime1 (Caporaso *et al*., 2010; Nguyen *et al*., 2016; Callahan *et al*., 2017; Edgar, 2017). More recent approaches have tried to convey effective increases in 16S rRNA gene amplification technical resolution with 'sub-OTU' (Janssen *et al*., 2018; Knight *et al*., 2018; Kou *et al*., 2018), 'ZOTU' (Edgar, 2016) and 'ESV'/'ASV'

(Callahan *et al*., 2016; Callahan *et al*., 2017). The term ZOTU is not used here due to accession collapsing and low-count capturing steps of the analysis, which are highly effective in producing biologically useful data from complex samples, but the term could be used to describe high-count sequences, representing *de novo* zero-difference OTUs that could potentially represent sequencing errors or gene copy-specific sequences. The terms ESV and ASV are not used here as they would presume accurate variant construction; comprehensively producing the *exact* biologically accurate sequences is not currently technically feasible to our knowledge without some errors in complex data, although such high confidence and resolution is certainly desired. Here, the term OTU is used as valuable in terms of interpretation of complex biological data (the focus of ANCHOR is upon maximizing biological discovery) and not as related to a 97% clustering threshold. Even though sequence resolution is extremely high in ANCHOR; the authors found returning to the term OTU affords a simple means to express and discuss potential biological discoveries while also considering technical errors and differing levels of sequence conservation (varying functional constraint). This use is in accordance with the early use of the term OTU as well as the practical considerations under discussion prior to its conception (Sneath, 1957; Sneath, 1964). Interpretation of a real-world data set (*see* ISS) illustrates some of this biological value, in particular valuable given the occurrences of ambiguous annotation.

Output files. The main output files are:

i. detailed OTU table consisting of all high-count sequences, detailed BLASTn output, taxonomic assignment (including ambiguous assignments) and count (including captured low-count sequences);
ii. count matrix;
iii. taxonomy table;
iv. ambiguous hits table;
v. secondary annotation table.

Several other files are produced (graphs, statistics, high-count mapping and low-count mapping) although are of less utility for downstream analyses such as diversity (alpha/beta) or differential abundance analysis (Anders *et al*., 2013; Love *et al*., 2014; Love *et al*., 2015). All scripts are provided at https://github.com/gonzalezem/ANCHOR.

Results and discussion

ANCHOR benchmarking: artificial data sets

Even data set. The recommended parameters (high-count sequence identification ≥3 counts, sequence annotation ≥99% and low-count sequence capture ≥98% parameters) used 99.9% of the total initial sequences with an average count per operational taxonomic unit (OTU) of 100.2 (Supplementary File 2). A high-count threshold of 3 and annotation using 99% identity and coverage led to a number of OTUs similar to expected: 99.7% of the expected species are observed (1073/1076). Reducing the high-count threshold to 2 overestimated (inflated) the number of different OTUs (1840/1076), whereas increasing the high-count threshold to 4 and higher underestimated different sequences (1064/1076). Lowering annotation similarity increased the number of the high-count sequences collapsing: 1082 high-count sequences are collapsed into 1056 OTUs with annotation at 90%, 1062 at 95%, 1068 at 97% and 1070 at 98% identity and coverage. The influence of the low-count sequence capture is minor here where 90%, 95%, 97% and 98% identity and coverage to high-count sequences all resulted in an increase in the initial reads used from 95.4% to 99.9%.

Staggered data set. The default recommended parameters (high-count sequence identification ≥3 counts, sequence annotation ≥99% and low-count sequence capture ≥98% parameters) captured 99.6% of the expected species in the Staggered data set, using 99.9% of the total initial reads with an OTU count of 99.8 on average (Supplementary File 2) agreeing with the projected count of 100. The random count distribution had little impact upon OTUs characterization as similar general observations about the parameters are found between the two artificial data sets. The default recommended parameters producing 1089 high-count sequences that collapsed into 1072 OTUs corresponding to 99.6% of the expected sequences. The only variation from the Even data set came, as expected, from the count distribution that varied greatly between OTUs (species abundance levels were randomly distributed amongst sequences in this data set).

To benchmark ANCHOR resolution of species present in a sample, and whether ambiguity is driven by a clustered database (i.e., all Greengenes database comprises 97% identity sequence clusters represented by single sequences), a simple nonartificial data set is selected and examined with a more extensive annotation process, Kozich's mock community data set (Kozich *et al*., 2013).

ANCHOR benchmarking: Kozich's single sample mock community

Results and discussion. A total of 26 ANCHOR OTUs were inferred based on 95.6% of the initial reads (Table 1). The total count was 4568 at an average abundance of 176 counts per OTU. All OTUs were annotated at a taxonomic level of species with the exception of one at genus level (*Clostridium*_1), which had a very low (minimum) with count of 3 (0.06% of the total) and was flagged as a potential chimera. From 20 expected

Table 1. Kozich's mock community data set expected species information as found from OTU annotation in ANCHOR.

ANCHOR OTUs	Expected species	Tax level	Ambiguous annotation	Identity %	Total counts
Acinetobacter baumannii_1	*A. baumannii*	Species	Unique	100.0	407
Actinomyces odontolyticus_1	*A. odontolyticus*	Species	Unique	100.0	356
Bacillus MS_1	*Bacillus cereus*	Species	8 = *Bacillus anthracis, B. cereus, B. gaemokensis, B. mycoides, B. pseudomycoides, B. thuringiensis, B. toyonensis, B. wiedmannii*	100.0	377
Bacteroides vulgatus_1	*B. vulgatus*	Species	Unique	100.0	204
Bacteroides vulgatus_2	*B. vulgatus*	Species	Unique	100.0	42
Bacteroides vulgatus_3	*B. vulgatus*	Species	Unique	100.0	21
Clostridium MS_1	*C. beijerinckii*	Species	4 = *C. beijerinckii, C. diolis, C. puniceum, C. saccharoperbutylacetonicum*	100.0	277
Clostridium beijerinckii_1	*C. beijerinckii*	Species	Unique	100.0	27
Deinococcus radiodurans_1	*D. radiodurans*	Species	Unique	100.0	116
Enterobacterales MS_1	*Escherichia coli*	Species	8 = *Brenneria alni, E. coli, E. fergusonii, E. marmotae, E. vulneris, Shigella boydii, S. flexneri, S. sonnei*	100.0	198
Enterococcus MS_1	*Enterococcus faecalis*	Species	14 = *E. canintestini, E. canis, E. dispar, E. durans, E. faecalis, E. faecium, E. hirae, E. lactis, E. mundtii, E. olivae, E. ratti, E. rivorum, E. saigonensis, E. villorum*	100.0	196
Helicobacter pylori_1	*Helicobacter pylori*	Species	Unique	100.0	355
Lactobacillus MS_1	*Lactobacillus gasseri*	Species	4 = *L. gasseri, L. hominis, L. johnsonii, L. taiwanensis*	100.0	139
Listeria MS_1	*L. monocytogenes*	Species	5 = *L. innocua, L. ivanovii, L. marthii, L. seeligeri, L. welshimeri*	100.0	156
Neisseria meningitidis_1	*N. meningitidis*	Species	Unique	100.0	303
Porphyromonas gingivalis_1	*P. gingivalis*	Species	Unique	100.0	104
Pseudomonas aeruginosa_1	*Pseudomonas aeruginosa*	Species	Unique	100.0	144
Rhodobacter MS_1	*Rhodobacter sphaeroides*	Species	3 = *Rhodobacter johrii, R. megalophilus, R. sphaeroides*	100.0	53
Staphylococcus MS_1	*Staphylococcus aureus/ S. epidermidis*	Species	13 = *S. aureus, Staphylococcus capitis, Staphylococcus caprae, S. chromogenes, S. epidermidis, S. haemolyticus, S. hominis, S. lugdunensis, S. pasteuri, S. petrasii, S. saccharolyticus, S. simiae, S. warneri*	99.605	4
Staphylococcus MS_2	*S. aureus/S. epidermidis*	Species	12 = *S. aureus, S. capitis, S. caprae, S. epidermidis, S. haemolyticus, S. hominis, S. lugdunensis, S. pasteuri, S. petrasii, S. saccharolyticus, S. simiae, S. warneri*	100.0	599
Streptococcus agalactiae_1	*S. agalactiae*	Species	Unique	100.0	218
Streptococcus MS_1	*Streptococcus pneumoniae*	Species	*S. pneumoniae, S. pseudopneumoniae*	100.0	24
Streptococcus mutans_1	*S. mutans*	Species	Unique	100.0	226

Unexpected ANCHOR OTUs	Tax level	Ambiguous annotation	Identity %	Total COUNTS
Bacillus anthracis_1	Species	Unique	100.0	4
Clostridium_1[a]	Genus	Unique	99.209	3
Staphylococcus chromogenes_1	Species	Unique	100.0	15

Ambiguity refers to annotation for a given OTU comprising multiple species with equal BLASTn scores. The parameters were a high-count threshold of 3, 99% ANCHOR annotation selection and 98% low-count sequences capture (see method). Data available in Supplementary File 3.

a. Retained for interest but flagged as a potential chimera by UCHIME during QC (difference from *C. beijerinckii* falls between 1–40 nt, which is 100% similar to *bacillus* and both staph sequences).

species, all 20 species were found in 23 different OTUs (out of a total of 26), each with 100% identity. Ten OTUs had ambiguous annotation (see Table 3), in that the utility of the amplified region (average size of 253 nt) to distinguish a *single* species from a specific list of equally likely species would not be possible without the *a priori* information of the expected species (due to conservation of the 16S rRNA gene amplified region between specific species).

The amplicon originating from *Listeria monocytogenes* EDG-e/BAA-679 was correctly annotated as potentially all of the *L. monocytogenes* group species suggested by Collins *et al*. (Collins *et al*., 1991): *Listeria innocua*, *Listeria ivanovii*, *Listeria marthii*, *Listeria seeligeri, Listeria welshimeri* (represented by the OTU label of *Listeria*_MS1); however, it was incorrectly *not* annotated as *L. monocytogenes*, being the only expected species that was not identified. Upon mining all six gene copies from each of the available, up-to-date, fully annotated type or representative strain genomes of species *L. monocytogenes* (str. NCTC 10357), *L. welshimeri* (str. SLCC5334), *L. seeligeri* (str. SLCC3954), *L. innocua* (str. Clip11262) as well as the most commonly used clinical *L. monocytogenes* strains str. EGD-e/BAA-679 (supplied for the mock here), str. EGD (distinct from EGD-e (Bécavin *et al*., 2014)), str. 10403S and the serotype 4b str. F2365, alignments show that the amplified region is 100% conserved across all 48 gene copies, suggesting the single variant *L. monocytogenes* 16S rRNA gene sequence entry in NCBIs 16S bacterial and archaeal database (NR_044823.1; str. NCTC 10357) may be inaccurate. The OTU annotated as Listeria MS_1 did indeed map perfectly (100% identity) to the amplified region conserved across all 16S rRNA gene copies in all these species/strains (including the BAA-679 genome).

All but one (109/110) of the expected gene copies could be mapped perfectly (100% identity) to OTUs (see Table 2). A single expected *Staphylococcus epidermidis* gene copy (one of five; labelled in-house *S. epidermidis* ATCC 12228 Se-*rrsE* here) was not captured using ANCHOR. The high proportion of ambiguous annotation hits could result from the choice of a small amplified V4 region length (~250 nt). The amplified 16S rRNA gene region from *Staphylococcus aureus* and *S. epidermidis* was identical except for the single variant *S. epidermidis* gene copy (not detected), making any differentiation between the two species impossible without a longer or different 16S RNA gene target region. Barring this single gene copy exception, ANCHOR precisely differentiated all gene copies that varied at the amplified region and allowed evaluation of the accuracy of count distribution within these species. The number of gene copies did not drive the count variation between different species in this mock community, as would be expected outside of synthetic data (due to varying population numbers and relative metabolic rates/regulation). However, when comparing the counts between OTUs representing different gene copies *within* a species (Table 2; Supplementary file 3), the total observed counts were strictly proportional to the number of gene copies sharing an identical amplified region. For example, the *C. beijerinckii* strain NCIMB 8052 genome contains 14 16S rRNA gene copies, each of which is unique at full length but only one of which varies in the amplified region from the others (see Cb-*rrsD*), resulting in two expected variant amplicons with an expected count ratio of 13:1. The two OTUs (Clostridium MS_1 and Clostridium beijerinckii_1) aligned perfectly (100% identity) with the two expected amplicons and had counts of 220 and 20, agreeing relatively well with the expected with a ratio of 11:1. Similarly, the *Bacteroides vulgatus* strain ATCC_8482 genome contains seven genes copies (Fig. 1A), six of which are unique at full length (Fig. 1B) but where only three variant amplicons would be expected at a count ratio of 5:1:1 (Fig. 1C). The three OTU sequences aligning (100%) to these expected amplicons (Bacteroides vulgatus_1, Bacteroides vulgatus_2 and Bacteroides vulgatus_3) had roughly similar counts of 204, 42 and 21 respectively (Fig. 1C and Table 4). This result suggests a good integration between counts inferred from the data set and reference sequences produced independently; however, these conclusions were derived by knowing the composition of the mock community *a priori* and, while suggesting promising potential from ANCHOR, differentiating high-count sequences formed due to technical error from those accurately representing gene copies would be currently be impossible using real-world uncharacterized data.

These results show the capability for sequence output from ANCHOR to accurately reflect 16S rRNA gene copies, expected from genome sequences, from a very simple data set. However, ANCHOR also found ambiguity of annotation for 10 of the 26 OTUs, where the amplified region is conserved across two or more species, making unique species-level annotation for many members of this sample difficult (impossible using this target region) if the expected species list was not available. It is important to point out that the ambiguity for some species does not undermine the species that can clearly be identified, but rather demonstrate the necessity for biologists to thoughtfully and thoroughly explore data output. Three OTUs were constructed representing unexpected microbes (two species and one genus), two of which had low counts suggesting either repeated technical error or low abundance contamination. The exception was Staphylococcus_chromogenes_1, which had a count comparable with expected species. This simple mock community proved to include complex features that ANCHOR detected and accurately reported. Amplicon ambiguity (sequence shared across multiple species) was notably a common feature of the data set as only half of the OTUs were unique to a species. While this mock community was informative in illustrating how amplicon ambiguity can affect 16S rRNA gene barcoding results, single sample analysis of a very simple community does not reflect standard biological design using real-world data.

ANCHOR benchmarking: Kleiner's replicated mock community

Results and discussion. Three types of samples were constructed *in vitro* for Kleiner's mock: equal-cell, equal-

Table 2. Kozich's mock community data set gene copies from expected species.

Identified species with reference genomes	No. of gene copies (Variant @ full length)	Amplified Region			ANCHOR OTU (100% similarity to gene copy)	ANCHOR OTU counts
		Variant	Distribution	Gene copies		
A. baumannii ATCC 17978	6(1)	1	1	Ab-*rrsA-F*	Acinetobacter baumannii_1	407
A. odontolyticus ATCC 17982	2(1)	1	1	Ao-*rrsA*,*B*	Actinomyces odontolyticus_1	356
B. cereus ATCC 10987	12(3)	1	1	Bc-*rrsA-L*	Bacillus MS_1	377
B. vulgatus ATCC 8482	7(6)	3	5	Bv-*rrsA-D*,*G*	Bacteroides vulgatus_1	204
			1	Bv-*rrsE*	Bacteroides vulgatus_2	42
			1	Bv-*rrsF*	Bacteroides vulgatus_3	21
C. beijerinckii NCIMB_8052/ ATCC 51743	14(14)	2	13	Cb-*rrsA-C*,*E-N*	Clostridium MS_1	277
			1	Cb-*rrsD*	Clostridium beijerinckii_1	27
D. radiodurans R1	3(2)	1	1	Dr-*rrsA-C*	Deinococcus radiodurans_1	116
E. faecalis OG1RF/47077	4(2)	1	1	Ef-*rrsA-D*	Enterococcus MS_1	196
E. coli str. K12 MG1655[a]	7(6)	1	1	Ec-*rrsA-G*	Enterobacterales MS_1	198
H. pylori 26695/700392 [b]	2(2)	1	1	Hp-*rrsA*,*B*	Helicobacter pylori_1	355
L. gasseri ATCC 33323	6(1)	1	1	Lg-*rrsA-F*	Lactobacillus MS_1	139
L. monocytogenes EGD-e	6(4)	1	1	Lm-*rrsA-F*	Listeria MS_1	156
N. meningitidis MC58/BAA-335	4(1)	1	1	Nm-*rrsA-D*	Neisseria meningitidis_1	303
P. gingivalis ATCC 33277	4(1)	1	1	Pg-*rrsA-D*	Porphyromonas gingivalis_1	104
P. aeruginosa PAO/47085	4(2)	1	1	Pa-*rrsA-D*	Pseudomonas aeruginosa_1	144
R. sphaeroides 2.4.1/17023	3(2)	1	1	Rs-*rrsA-C*	Rhodobacter MS_1	53
S. aureus NCTC 8325/BAA-1718	5(5)	1	1	Sa-*rrsA-E*	Staphylococcus MS_2[c]	599[c]
S. epidermidis ATCC 12228	5(5)	2	4	Se-*rrsA-D*		
			1	Se-*rrsE*	X	-
S. agalactiae 2603V/R/BAA-611	7(1)	1	1	Stra-*rrsA-G*	Streptococcus agalactiae_1	218
S. mutans UA159/700610	5(2)	1	1	Strm-*rrsA-E*	Streptococcus mutans_1	226
S. pneumoniae TIGR4/BAA-334	4(1)	1	1	Strp-*rrsA-D*	Streptococcus MS_1	24

Full length expected gene copies from Kleiner's Mock were manually extracted from strain specific reference genomes (Supplementary File 3). The number of gene copies per genome was validated against the (very useful) University of Michigan Centre for Microbial Systems Ribosomal RNA Database (Klappenbach *et al*., 2001). Gene copies are named using *E. coli* nomenclature but are assigned a letter based on arbitrary occurrence in specific strain genome assembly to aid data navigation (these labels for specific copies should *not* be considered phylogenetically/across strains). Data available in Supplementary File 3.

a. No *E. coli* strain was provided but K12 (MG1655) had 100% similarity at the amplified region.

b. There are ambiguous nt calls in the amplified region of the *H. Pylori* 26695 assembly (none disagree with the ANCHOR OTU).

c. *S. epidermidis* (4/5) and *S. aureus* (5/5) gene copies share 100% identity for the amplified region.

protein and uneven mock communities, each with four biological replicates. A total of 159 high-count sequences were collapsed through exact shared accession IDs into 34 ANCHOR OTUs. The 34 OTUs accumulated 272,447 counts (80.2% of the assembled amplicons) with an average abundance of 8013 counts per OTU (Supplementary file 4). Of the 34 OTUs, 32 were annotated at species level, representing 97.9% of the total counts, with the exception of one OTU annotated at genus level (<2.1% total counts, Uncultured bacterium AK199) and one OTU was could not be annotated >99% identity (representing <0.1% total counts) (Table 3). Sixteen of the OTUs had ambiguous annotation where the amplicon could represent multiple species (at 100% identity). Seventeen of the 23 expected species (or strains) were identified by OTUs with 100% identity annotation. The expected OTUs accounted for 268,147 counts (97.7% of the total count) with an average count of 15,773 per species (from 47,420 counts for *Salmonella enterica* to 276 for *Desulfovibrio vulgaris*). Six OTUs were unexpected (Staphylococcus MS_1, Staphylococcus epidermidis_1, Staphylococcus epidermidis_2, Aeromicrobium fastidiosum_1, Enterobacter cloacae_1 and an unknown sequence) and represented 3.3% of the total counts.

ANCHOR did not annotate the *Paracoccus* MS_1 OTU as the expected *Paracoccus denitritificans*, but instead annotated it as *Paracoccus pantotrophus, Paracoccus bengalensis, Paracoccus ferrooxidans* or *Paracoccus versutus*, all of which share a common amplicon sequence that is 1 nt off the curated *P. denitritificans* strain 17,741. *P. pantotrophus* and *P. denitritificans* have been extensively confused in the past, with a number of *P. denitrificans* strains renamed *P. pantotrophus*, and the OTU found here is a 100% match to the *P. pantotrophus* LGM 4218 deriving from the Stanier 381 strain now recognized as mistakenly archived as *P. denitritificans* 17,741 [start with fig. 1 in Goodhew *et al.,* 1996 (Goodhew *et al*., 1996; Rainey *et al*., 1999; Kelly *et al*., 2006)]. The strain shares 100% 16S rRNA gene sequence at full length to the *P. pantotrophus type* strain GB17 (ATCC 35512; as well as WGS contigs from all four *P. pantotrophus* partial genome assemblies J40, J46, DSM1403 and DSM11073) and so is likely annotated correctly at species level by ANCHOR. The OTU Bacteria MS_1 is an example of where annotation using

Table 3. Kleiner's mock community data set expected species information as found from OTU annotation in ANCHOR.

ANCHOR OTUs	Expected species	Tax level	Ambiguous annotation	Identity %	Total counts
Agrobacterium fabrum_1	*Agrobacterium tumefaciens*	Species	Unique	100	17,289
Alteromonas MS_1	*A. macleodii*	Species	4 = *A. macleodii, A. marina, A. mediterranea, A. tagae*	100	5413
Alteromonas macleodii_1	*A. macleodii*	Species	Unique	100	1342
Bacillus MS_1	*B. subtilis*	Species	2 = *B. subtilis, B. tequilensis*	100	16,021
Bacillus MS_2	*B. subtilis*	Species	2 = *B. subtilis, B. virus*	100	2543
Bacillus MS_3	*B. subtilis*	Species	2 = *B. subtilis, B. virus*	100	2370
Bacillus subtilis_1	*B. subtilis*	Species	Unique	100	5442
Bacillus subtilis_2	*B. subtilis*	Species	Unique	99.6	268
Chromobacterium MS_1	*Chromobacterium violaceum*	Species	3 = *Chromobacterium aquaticum, C. subtsugae, C. violaceum*	100	12,685
Cupriavidus metallidurans_1	*C. metallidurans*	Species	Unique	100	34,913
Desulfovibrio vulgaris_1	*D. vulgaris*	Species	Unique	100	276
Enterobacterales MS_1	*E. coli*	Species	5 = *B. alni, E. coli, E. fergusonii, S. flexneri, S. sonnei*	100	12,066
Enterobacteriaceae MS_1	*E. coli*	Species	2 = *E. coli, Shigella dysenteriae*	99.8	131
Paracoccus MS_1	*P. pantotrophus*[a]	Species	4 = *P. bengalensis, P. ferrooxidans, P. pantotrophus, P. versutus*	100	5958
Pseudomonas MS_1	*Pseudomonas sp.*[b]	Species	5 = *Pseudomonas citronellolis, Pseudomonas delhiensis, Pseudomonas knackmussii, Pseudomonas multiresinivorans, Pseudomonas nitroreducens*	100	16,681
Pseudomonas MS_3	*P. fluorescens*	Species	2 = *Pseudomonas antarctica, P. fluorescens*	100	2114
Pseudomonas fluorescens_1	*P. fluorescens*	Species	Unique	100	10,881
Pseudomonas MS_2	*Pseudomonas pseudoalcaligenes*	Species	4 = *P. aeruginosa, P. balearica, P. pseudoalcaligenes, P. resinovorans*	100	14,714
Rhizobiaceae MS_1	*R. leguminosarum*	Species	27 = *Agrobacterium rhizogenes, A. rubi, Rhizobium sp. (x25)*	100	24,671
Rhodobacteraceae MS_1	*Uncultured bacteriumAK199*[c]	Genus[c]	3 = *Donghicola, Lutimaribacter, Oceanicola*	100	5652
Salmonella enterica_1	*S. enterica*	Species	Unique	100	40,898
Salmonella enterica_2	*S. enterica*	Species	Unique	100	6234
Salmonella enterica_3	*S. enterica*	Species	Unique	100	137
Salmonella enterica_4	*S. enterica*	Species	Unique	99.8	51
Salmonella enterica_5	*S. enterica*	Species	Unique	100	100
Staphylococcus MS_2	*S. aureus*	Species	2 = *S. aureus, S. simiae*	100	6352
Bacteria MS_1[d]	*S. maltophilia*	Species	7 = *S. succinus*[d]*, S. chelatiphaga, S. maltophilia, S. rhizophila, X. citri, X. oryzae, X. retroflexus*	100	18,638
Thermus thermophilus_1	*T. thermophilus*	Species	Unique	100	4307

Unexpected ANCHOR OTUs	Tax level	Ambiguous annotation	Identity %	Total counts
Aeromicrobium_fastidiosum_1	Species	Unique	99.8	148
Enterobacter_cloacae_1	Species	Unique	100	929
Staphylococcus MS_1	Species	3 = *S. capitis, S. caprae, S. epidermidis*	100	2126
Staphylococcus_epidermidis_1	Species	Unique	100	537
Staphylococcus_epidermidis_2	Species	Unique	99.8	443
TrueUnknown_1	N/A	Unknown	100	117

Expected species not detected

Burkholderia xenovorans; *Chlamydomonas reinhardtii*; *N. europaeae*; *N. ureae*; *Nitrososphaera viennensis*; *N. multiformis*

Ambiguity refers to annotation for a given OTU comprising multiple species with equal BLASTn scores. The parameters were a high-count threshold of 3, 99% ANCHOR annotation selection and 98% low-count sequences capture (see method). Data available in Supplementary File 4.

a. *Paracoccus dentrificans* ATCC 17741 recognized as mistakenly archived *P.* pantotrophus LGM 4218 (start with Fig. 5 Goodwin *et al*., 1996 (Goodhew *et al*., 1996; Rainey *et al*., 1999; Kelly *et al*., 2006)).

b. Mistaken for *Ps. denitrificans, nomen rejiciendum* (Bacteriology, 1982).

c. ***Uncultured bac AK199*** is not currently classified to a species or genera, and ANCHOR annotation was in the family Rhodobacteraceae as the consensus phylogenetic placement between RDP and Silva; however, the assembled ANCHOR OTU was 100% similar to the original isolate, *Uncultured bac AK199* (NCBI: JQ256816)(Lenk *et al*., 2012).

d. The *Staphylococcus succinus* (NCBI: KJ534522.1) is mistakenly annotated within the NCBI nt database; this was easily observed by both the high taxon disparity and the ambiguous annotation. This OTU would require manual curation to be relabelled correctly as Xanthomonadaceae AS (removal of the erroneous Staph hit) but highlights database integrity challenges here.

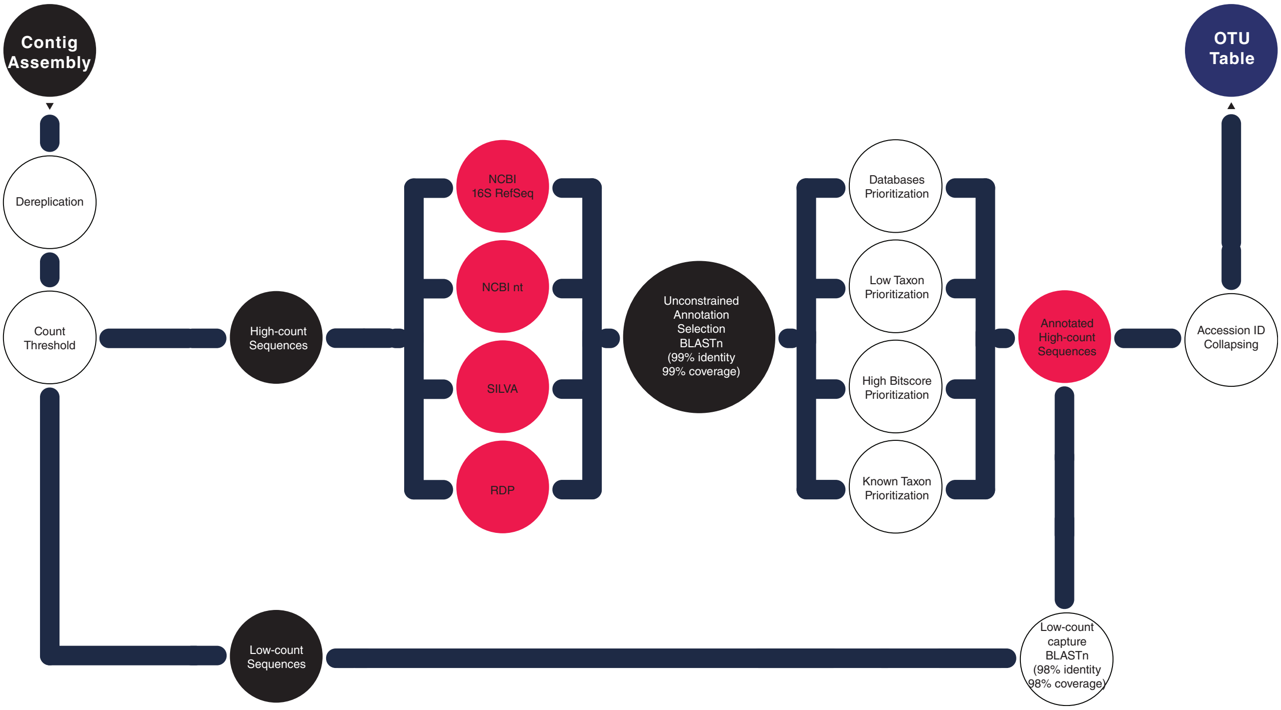
Contig Assembly
Dereplication
Count Threshold
High-count Sequences
NCBI 16S RefSeq
NCBI nt
SILVA
RDP
Unconstrained Annotation Selection BLASTn (99% identity 99% coverage)
Databases Prioritization
Low Taxon Prioritization
High Bitscore Prioritization
Known Taxon Prioritization
Annotated High-count Sequences
Accession ID Collapsing
OTU Table
Low-count Sequences
Low-count capture BLASTn (98% identity 98% coverage)

Fig. 1. ANCHOR sequence processing diagram.
Four design targets were: (1) Fastq-ready, no preprocessing required from users, (2) no sequence modification (sequence integrity retained), (3) low resource demanding, and (4) integrated exhaustive cross-database annotation. [Correction added on 18 June 2019, after first online publication: Figure 1 caption has been corrected in this version]. [Color figure can be viewed at wileyonlinelibrary.com]

the uncurated NCBI nr/nt database can cause difficulty. The correct ambiguous annotation for this OTU includes *Stenotrophomonas chelatiphaga, Stenotrophomonas maltophilia, Stenotrophomonas rhizophila, Xanthomonas citri, Xanthomonas oryzae* and *Xanthomonas retroflexus* [*S. maltophilia* was previously placed in the genus *Xanthomonas* before becoming the type species of *Stenotrophomonas* (Palleroni and Bradbury, 1993)]. As the 16S rRNA gene target region is conserved across these species (share a common sequence in the amplified region), the correct automated ANCHOR annotation should therefore be Xanthomonadaceae MS_1. However, as there is a single (likely) mistakenly annotated NCBI database entry of *Staphylococcus succinus* for this sequence (KJ534522.1; a firmicutes as opposed to proteobacteria), the lowest shared taxon is used, 'bacteria'. The benefits of using a rich but uncurated database (after a prioritized screening of a curated database) generally outweigh the drawbacks of potential database mistakes as they are easily identifiable; both the OTU and culprit database entry stand out as distinct in ANCHOR output as well as the entry itself (KJ534522.1) being over 30% dissimilar from the consensus *S. succinus* sequences, including those published after peer-review. However, the substantial impact of a single poorly annotated sequence entry highlights the need for careful user scrutiny of automated output if the meaningfulness of data is to be maximized.

- *16S rRNA gene methodology comparison.* Count distribution was very similar between all methods with the most substantial difference coming from the number of OTUs/ASVs constructed between the methods from the same data, ranging from 26–56,205 OTUs (Table 4; method data and parameters are provided in supplementary files 1 and 4). Although most of the methods were assessing OTUs (ASV for dada2) at genus level, Dada2 and Qiime1 were also capable of assessing OTUs/ASVs as species, identifying five and eight expected species respectively.

Qiime1 found 948 OTUs, a high proportion of which (68.7%) represented the 18/23 expected species present in the mock at either species or genera level. In total, 96 Qiime OTUs were annotated at species level, 744 at genus level, 49 at family level, 26 at order level, 20 at class level and 12 at phylum level. Mothur found a total of 56,205 OTUs with 38,509 annotated at genus level, 17,432 at family level, 129 at order level, 71 at class level, 63 at phylum level and 1 domain. Only two expected species were not detected using Qiime1 and 4 using Mothur, both of which detected *Nitrosomonas.* Despite 97% clustering, which has been a recent source

Table 4. Kleiner's mock community assessed using five different methods.

	Mothur	Qiime1	Dada2	MetAmp	Anchor
Number of expected species	23	23	23	23	23
Expected species (Species ID)	N/A	8	5	N/A	16
Expected species (Genera ID)	19	11	11	17	1
No. of unexpected OTUs/ASVs^{a}	17,037	297	31	8	6
Average count per OTU/ASV	5	360	4478	4864	8013
Total counts (% raw reads)	275,610 (53.6%)	340,895 (66.2%)	259,699 (50.5%)	126,459 (24.6%)	272,941 (53.0%)

Method	OTUs/ASVs annotated as genera (% total counts)					OTUs/ASVs annotated as species (% total counts)				
Expected species/taxon	Mothur	Qiime1	Dada2	MetaAmp	ANCHOR	Mothur	Qiime1	Dada2	MetaAmp	ANCHOR
A. fabrum	-	-	-	-	-	-	11 (6.7)	-	-	1 (6.3)
A. macleodii	1172 (2.3)	51 (2.2)	1 (2.1)	1 (2.5)	-	-	-	1 (0.5)	-	2 (2.5)
B. subtilis	6592 (9.5)	40 (8.0)	4 (9.8)	1 (8.8)	-	-	30 (1.5)	-	-	5 (9.8)
B. xenovorans	-	1 (<0.1)	-	-	-	-	-	-	-	-
C. reinhardtii	-	-	-	-	-	-	-	-	-	-
C. violaceum	2989 (4.6)	4 (4.9)	1 (4.6)	1 (4.4)	-	-	-	-	-	1 (4.7)
C. metallidurans	8582 (12.6)	3 (13.6)	-	1 (10.3)	-	-	-	1 (12.5)	-	1 (12.8)
D. vulgaris	103 (0.1)	2 (0.1)	-	1 (0.1)	-	-	-	1 (0.1)	-	1 (0.1)
E. coli	2681 (4.2)	203 (4.6)	1 (4.4)	1 (4.5)	-	-	2 (<0.1)	-	-	2 (4.5)
N. europaeae	5 (<0.1)	2 (<0.1)	-	-	-	-	-	-	-	-
N. ureae			-	-	-	-	-	-	-	-
N. viennensis	-	-	-	-	-	-	-	-	-	-
N. multiformis	7 (<0.1)	1 (<0.1)	-	1 (<0.1)	-	-	-	1 (<0.1)	-	-
P. pantotrophus	872 (2.0)	23 (2.3)	1 (2.1)	1 (2.4)	-	-	1 (<0.1)	-	-	1 (2.2)
Pseudomonas sp.	1 (16.5)	110 (11.6)	4 (16.5)	3 (15.4)	-	-	-	-	-	1 (6.1)
P. fluorescens					-	-	-	-	-	2 (4.8)
P. pseudoalcaligenes					-	-	1 (<0.1)	-	-	1 (5.4)
R. leguminosarum	5504 (14.2)	43 (2.0)	2 (15.5)	2 (19.3)	-	-	1 (<0.1)	-	-	1 (9.1)
Uncultured bac AK199^{a}	2695^{a} (4.3)	8^{a} (2.3)	1^{a}(2.1)	1^{a} (2.6)	1^{a} (2.1)	-	-	-	-	-
S. enterica	943 (13.3)	29 (3.9)	2 (17.4)	1 (17.7)	-	-	6 (13.6)	-	-	5 (17.4)
S. aureus	2397 (3.5)	113 (3.7)	4 (3.5)	1 (3.6)	-	-	-	-	-	1 (2.3)
S. maltophilia	4444 (6.9)	17 (7.3)	1 (7.1)	1 (7.3)	-	-	4 (<0.1)	-	-	1 (6.8)
T. thermophilus	1059 (1.5)	1 (1.7)	-	1 (1.0)	-	-	-	1 (1.4)	-	1 (1.6)
Total OTUs/ASVs in expected sp.	39,168	651	22	18	1	0	56	5	0	27

Kleiner's mock community is composed of 12 samples: 3 conditions (types) × 4 sample replicates. Only amplicons within the length range of 436–467 nt were selected to allow for comparisons across methods. Method-specific parameters used (defaults where possible) and resulting data are available in Supplementary File 4.

- = Not detected.

a. High taxon OTUs (phylum, class, order, family). ***Uncultured bac AK199*** is not currently classified to a species or genera, ANCHOR annotation was in the family Rhodobacteraceae as the consensus phylogenetic placement between RDP and Silva; however, the assembled ANCHOR OTU was 100% similar to the original isolate, *Uncultured bac AK199* (NCBI: JQ256816)(Lenk *et al*., 2012). Rhodobacteraceae OTUs/ASVs from other methods are also presented as potentially representing *Uncultured bac AK199.*

of discussion (Nguyen *et al*., 2016; Edgar, 2017), Qiime1 detected the most expected species (at genera level annotation) and, despite inflating OTUs, inferred similar count distributions to other methodologies. Qiime1 was also the only method alongside ANCHOR, which could identify *Agrobacterium fabrum* (the difficulty across methods being distinction from *Rhizobium* at genus level). While Mothur also achieved high detection of expected species (at genus level) as well as a count distribution broadly common to all the methods, the number of OTUs were extremely inflated, with 17,037 OTUs annotated at a high level of taxonomy or as unexpected taxa, including 52 unexpected genera (which would make biological interpretation of these data challenging). Read retention, which is an important consideration for accurately representing sample biology, was the highest out of all the investigated methods in Qiime1 and Mothur, at 66.2% and 53.6% of raw read counts respectively.

Dada2 found 58 ASVs, 7 of which were annotated at species level, 26 at genus level, 2 at family level, 1 at order level, 1 at class level and 21 at domain level. Sixteen of the 23 expected species were represented by ASVs, 27 in total (5 at species and 22 at genus level representing 11 expected genera). As this mock community was developed using MetaAmp version 1, it was appropriate to use as a comparison however, version 2 is now available and may further improve upon these results (Dong *et al*., 2017). MetaAmp found a total of 26 OTUs (annotated as 20 Genera, 1 Family, 1 Order, 1 Class and 3 Phyla). Of these, 17 OTUs represented expected species annotated at Genus level with three distinct OTU assembled for the three *Pseudomonas* species (but annotated as the common genus). One species was detected at genera level which was not detected by ANCHOR, *Nitrosospira multiformis*, although with very low counts. Six expected species were not detected at genera level using Dada2 or MetAmp, including *A. fabrum*, which was identified by both Qiime1 and ANCHOR. Five expected species were detected and annotated at species level with accurate count distribution by Dada2 (the second highest after ANCHOR). Read retention for Dada2 and MetAmp was 50.5% and 24.6% respectively. OTUs were similar to ANCHOR in accuracy across samples in both Dada2 and MetAmp, without substantial inflation of OTUs. The high accuracy and low OTU inflation of both methods was impressive and both would therefore be a comparable alternative to ANCHOR for analyses of 16S rRNA gene amplicons across multiple samples, such as when a biological question is posed using a replicated design.

The high numbers of OTUs produced by Qiime1 and Mothur demonstrate how inflation does not have to confound results *per se*, as the overarching biology was still observable here using count distribution to distinguish accurate OTUs. However, inflation is problematic with highly complexity (nonmock) samples, where interpreting count distribution can be more challenging. Similarly, although all methods tended to recover most of the expected organisms, annotation varied substantially. Higher taxonomic annotation can be problematic owing to the implications when querying the unknown (non-mock) systems. ANCHOR captured the diversity of Kleiner's mock community at high species-level resolution for a majority of expected species, with specific examples including identification of *A. fabrum*, which was only present in ANCHOR and Qiime1, as well as *Rhizobium leguminosarum*, having correspondingly fewer counts in Anchor and Qiime than in the other methods (due to not conflating Agrobacterium and Rhizobium). ANCHOR fell short, however, of other methodologies for three expected species. *N. multiformis* was absent using ANCHOR but successfully captured by all other methods at various raw abundances: three counts in Dada2 (detected at species level), three in MetaAmp (genus), eight in Mothur (genus) and eight in Qiime1 (genus). This general low-level abundance would fall below the 12 high-count sequence threshold, preventing *N. multiformis* from being detected by ANCHOR. *Nitrosomonas europaeae* and *Nitrosomonas ureae* were also not detected by ANCHOR but were successfully detected by both Qiime1 and Mothur (at genus level), again, likely due to very low counts (Table 4).

- *Gene copy capture.* Given the multiple OTUs annotated from the same expected species within Kleiner's Mock data set, representation of gene copies was explored using ANCHOR data. When comparing the number of gene copies represented by a single OTU (i.e., those gene copies that are conserved at the amplified region) and their respective counts within a same species, the proportion was generally represented with the exception of *Alteromonas macleodii* (expected 3:2 is 1:4; Table 5). *Bacillus subtilis* (strain 168) has an expected count ratio between five expected variant amplicons of 2:5:1:1:1, and is closely represented by ANCHOR OTUs with 100% identity to the expected amplicons at a count distribution ratio of 2:6:1:1 (the expected variant amplicon for the gene copy Bs-*rrE* was not detected). *Pseudomonas fluorescens* (strain ATCC 13525) contains six gene copies that would produce only two variant amplicons from the amplified region with an expected count distribution of 5:1. These two expected amplicons are observed perfectly (100% identity) by ANCHOR OTUs with a count distribution of 10,881:2114 (5:1). Similarly, *S. enterica* (Typhimurium LT2) has seven gene copies that are expected to produce two variant amplicons with a count distribution of 6:1. Anchor OTUs represent each amplicon at 100% identity and at count distribution of 40,898:6234 (6.6:1). While the gene copy

Table 5. Kleiner's mock community data set gene copies from expected species.

Identified species with reference genomes	No. gene copies (variant @ full length)	Variant @ amplified region			OTU (100% similarity to gene copy)	OTU counts	% Celleq avg.	% Proteq avg.	% Uneven avg.
		Variant	Distribution	Gene copy labels					
A. fabrum strain C58/ATCC 33970	4 (1)	1	4	Af-*rrsA-D*	Agrobacterium fabrum_1	17,289	21.13	33.50	45.37
A. macleodii ATCC 27126	5 (3)	2	3	Am-*rrsA,B,D*	Alteromonas macleodii_1	1342	27.20	70.34	2.46
			2	Am-*rrsC,E*	Alteromonas MS_1	5413	27.90	69.94	2.16
B. subtilis 168	10 (9)	5	2	Bs-*rrsA,C*	Bacillus subtilis_1	5442	59.45	38.66	1.89
			5	Bs-*rrsB,D,F,G,I*	Bacillus MS_1	16,021	58.85	39.34	1.81
			1	Bs-*rrsH*	Bacillus MS_2	2543	58.08	40.07	1.85
			1	Bs-*rrsJ*	Bacillus MS_3	2370	56.92	41.35	1.73
			1	Bs-*rrsE*	X	X	-	-	-
C. violaceum CV026	8 (1)	1	8	Cv-*rrsA-H*	Chromobacterium MS_1	12,685	82.71	15.03	2.25
C. metallidurans CH34	4 (1)	1	4	Cm-*rrsA,B* (x2[a])	Cupriavidus metallidurans_1	34,913	9.56	23.92	66.51
D. vulgaris Hildenborough	5 (4)	2	4	Dv-*rrsA,C-E*	Desulfovibrio vulgaris_1	276	0.00	0.00	100.00
			1	Dv-*rrsB*	X		-	-	-
E. coli K12	7 (1)	7	7	Ec-*rrsA-G*	Enterobacterales MS_1	12,066	30.41	47.39	22.20
P. pantotrophus LMG4218	1[b]	1	1	Ppa-*rrsA*	Paracoccus MS_1	5958	24.32	65.36	10.32
Pseudomonas sp. ATCC 13867	5 (3)	1	5	Psp-*rrsA-E*	Pseudomonas MS_1	16,681	41.20	44.90	13.90
P. fluorescens ATCC 13525	6 (3)	2	5	Pf-*rrsA,B,D-F*	Pseudomonas fluorescens_1	10,881	28.69	41.81	29.50
			1	Pf-*rrsC*	Pseudomonas MS_3	2114	30.09	44.18	25.73
P. pseudoalcaligenes KF707	5 (3)	1	5	Pps-*rrsA-E*	Pseudomonas MS_2	14,714	56.94	40.04	3.02
R. leguminosarum bv. viciae 3841	3	1	3	Rl-*rrsA-C*	Rhizobiaceae MS_1	24,671	22.15	57.00	20.85
S. enterica typhimurium LT2	7 (5)	2	6	Se-*rrsA,C-G*	Salmonella enterica_1	40,898	26.83	34.35	38.81
			1	Se-*rrsB*	Salmonella enterica_2	6234	27.25	34.66	38.08
S. aureus ATCC 13709/NCTC10399	6 (5)	2	1	Pa1-*rrsF*	X	X	-	-	-
			5	Pa1-*rrsA-E*	Staphylococcus MS_2	6352	5.81	84.08	10.11
S. aureus ATCC 25923	6 (3)	1	6	Pa2-*rrsA-F*					
T. thermophilus HB27	2 (1)	1	1	Tt-*rrsA,B*	Thermus thermophilus_1	4307	48.57	45.48	5.94
Unexpected species									
S. epidermidis strain 14.1.R1	6 (5)	4	3	SeR1-*rrsA,D,F*	Staphylococcus MS_1	2126	9.83	87.30	2.87
			1	SeR1-*rrsB*	Staphylococcus epidermidis_1	537	9.87	87.90	2.23
			1	SeR1-*rrsC*	X	X	-	-	-
			1	SeR1-*rrsE*	X	X	-	-	-

Full length expected gene copies from Kleiner's Mock were manually extracted from strain specific reference genomes (Supplementary file 4). The number of gene copies per genome was validated against the (very useful) University of Michigan Centre for Microbial Systems Ribosomal RNA Database (Klappenbach *et al*., 2001). Gene copies are named using *E. coli* nomenclature but are assigned a letter based on arbitrary occurrence in specific strain genome assembly to aid data navigation (these labels for specific copies should *not* be considered phylogenetically/across strains). Data available in Supplementary File 4.

a. Genome and megaplasmid.

b. Only one copy mined from all four current partial *P. pantotrophus* genomes: strains J40, J46, DSM1403, DSM 11073 (100% to amplicon in each).

distribution suggests promising potential for ANCHOR to distinguish gene copies due to in simple data, such high resolution is not currently possible using real-world complex data sets (where comprehensive reference genomes are not available).

Interestingly, when considering if the OTUs did represent variant amplicons deriving from different gene copies, a useful clue was the original design of the Kleiner experiment using three growth conditions or types: cell equal, protein equal or uneven (Kleiner *et al*., 2017). While not consistent between species, the average counts per condition were strictly uniform compared across OTUs within the same expected species without exception. For example, ANCHOR OTUs Salmonella enterica_1 and 2, corresponding to genes Se-*rrsA*,*C-G* and Se-*rrsB*, respectively, had relative count distributions of 26.83% and 27.25% in equal cell samples, 34.35% and 34.66% counts in equal protein samples and 38.81% and 38.08% in uneven samples (Table 5; Supplementary file 4). Three out of the six unexpected OTU were annotated as *S. epidermidis* (Staphylococcus epidermidis_1, Staphylococcus epidermidis_2 and Staphylococcus_MS). Upon detailed investigation, these OTUs may represent an uncharacterised *S. epidermidis* species present in all samples as all were similar to gene copies identified in the partially assembled *S. epidermidis* genome NIHLM040 with Staphylococcus epidermidis_2 corresponding to 16S rRNA gene in contig NZ_AKGR01000041.1 and Staphylococcus epidermidis_1 and *Staphylococcus*_MS corresponding to the 16S rRNA gene in contig NZ_AKGR01000002.1. The most compelling indicator of this association of the three OTUs (beyond the sequence and annotation similarity) is that each had the very precise common abundance ratio shared between the three type conditions of growth in Kleiner samples (uneven:cell even: protein even = 1:31:3.5) suggesting a common organism of origin when compared to expected OTUs. ANCHOR is designed for multisample and replicated data sets and less towards single sample analysis; this is a deliberate compromise to prevent false positives from being detected and to create count matrix as a sound base for downstream biologically focused analysis (e.g., differential abundance calculations) where low abundance and sparse species have reduced value.

Real-world data testing: International Space Station data set

Results and discussion • *Total sampled environment.* ANCHOR is designed with utility for nonideal, uncharacterised biology in mind and, in particular, to provide flexibility to complex data sets and complement high uncertainty metatranscriptomics (Gonzalez *et al*., 2015; Brereton *et al*., 2016; Gonzalez *et al*., 2018). While benchmarking against synthetic or simple (mock) communities is essential, an equally important test of the technology is its utility to contribute to unknown biology in complex real-world systems. As such, ANCHOR was used to analyze surface swab data from the International Space Station (ISS). While challenging within unknown systems, the results are briefly interpreted in an attempt to establish whether they are biologically coherent and, if so, whether they can build upon the previous findings reported by Lang *et al*. (Lang *et al*., 2017) and deepen our knowledge of this unique environment. A total of 1,132,141 amplicons were assembled from the ISS samples, 553,762 of which were unique. Of these, 6833 high-count sequences were identified using a count threshold of 12 and which represented 78.7% of total amplicons after low count sequence capture. These high-count sequences collapsed into 3455 OTUs, which could be annotated at various taxonomic levels: 11 were annotated at phylum level, 58 at class, 85 at order, 284 at family, 842 at genus level and 1087 as species (Supplementary file 5). A total of 988 OTUs could not be annotated as >99% similar to anything previously reported in the queried databases (designated as TrueUnknowns to differentiate them from database entries labelled as unknown bacteria). These unknown OTUs sequences are inflated compared with annotated OTUs, as they are not collapsed based on shared annotation, and can be easily explored for biological utility (many are >98% similarity to known species) but are not automatically reported as high confidence hits using ANCHOR. Of the 1087 species level hits, comprising 74.5% of the captured sequence counts, 373 were ambiguous in that the specific amplified sequence was common to multiple known species, averaging eight but ranging to as high as 263 species (Streptophyta MS_3), leaving 714 OTUs with the potential utility to identify a single species with confidence.

Overall, the bacterial community represented 87% of the OTU sequence counts, with Eukaryotes making up 7%, Archaea 0.2% and unknown sequences ~6% (Fig. 2). At phyla level, Bacteria were dominated by 38% Firmicutes sequences, 22% Proteobacteria (11% α, 39% β, 1% δ, 46% γ and 2% ε), 19% Actinobacteria, 14% Bacteroidetes, 2% Fusobacteria and 2% Verrucomicrobia (remainder as others, Supplementary file 5). Eukaryotes were made up of 79% Chordata, all of which were derived from human mitochondrial OTUs with the exception of *Coturnix japonica*_1 (Quail mitochondrial 12S rRNA), with 8% Plantae and the remainder dominated by fungi, stramenopiles and cryptophyta (mitochondrial and chloroplast). The majority of plant OTUs were highly ambiguous (due to extensive chloroplast 16S rRNA sequence conservation) with the most abundant being Streptophyta MS_3 (common to 263 species), although

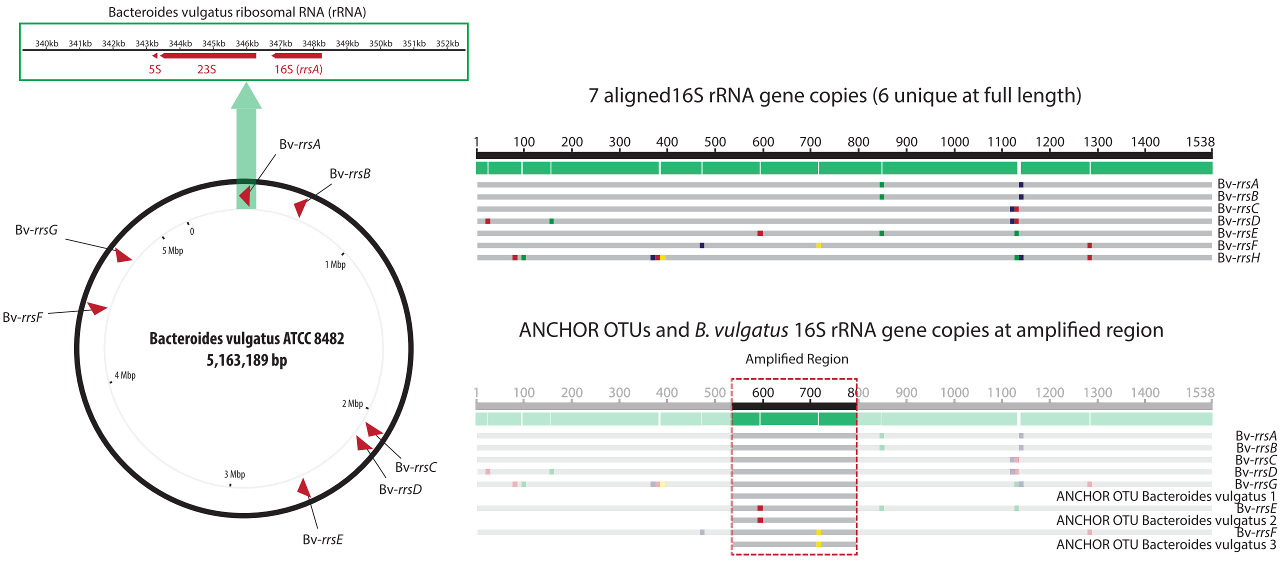
Bacteroides vulgatus ribosomal RNA (rRNA)
340kb 341kb 342kb 343kb 344kb 345kb 346kb 347kb 348kb 349kb 350kb 351kb 352kb
5S
23S
16S (rrsA)
Bv-rrsA
Bv-rrsB
Bv-rrsG
Bv-rrsF
Bv-rrsC
Bv-rrsD
Bv-rrsE
0
5 Mbp
1 Mbp
4 Mbp
2 Mbp
3 Mbp
Bacteroides vulgatus ATCC 8482
5,163,189 bp
7 aligned16S rRNA gene copies (6 unique at full length)
1 100 200 300 400 500 600 700 800 900 1000 1100 1200 1300 1400 1538
Bv-rrsA
Bv-rrsB
Bv-rrsC
Bv-rrsD
Bv-rrsE
Bv-rrsF
Bv-rrsH
ANCHOR OTUs and B. vulgatus 16S rRNA gene copies at amplified region
Amplified Region
1 100 200 300 400 500 600 700 800 900 1000 1100 1200 1300 1400 1538
Bv-rrsA
Bv-rrsB
Bv-rrsC
Bv-rrsD
Bv-rrsG
ANCHOR OTU Bacteroides vulgatus 1
Bv-rrsE
ANCHOR OTU Bacteroides vulgatus 2
Bv-rrsF
ANCHOR OTU Bacteroides vulgatus 3

Fig. 2. ANCHOR OTU and gene copy alignment for B. vulgatus ATCC 8482 in Kozich's Mock community.
The B. vulgatus ATCC 8482 genome (GCA_000012825.1 ASM1282v1) was downloaded from NCBI and explored using Geneious 7.1.9 (https://www.geneious.com). All sequences are provided in Supplementary File 4. All seven expected 16S rRNA gene copies of B. vulgatus ATCC 8482 are illustrated at full length (Bv-rrsA-H) with the three corresponding ANCHOR OTUs (amplicons) highlighted. [Correction added on 18 June 2019, after first online publication: Figure 2 caption has been corrected in this version]. [Color figure can be viewed at wileyonlinelibrary.com]

less ambiguous sequences, *Daucus* MS_1 (including *D. carota*; carrot), *Pisum sativum* (peas), *Malus* MS_1 (*Malus domestica* or *Malus_hupehensis*, Apple) and Rosales MS_1 (*Cannabis sativa* or *Ziziphus jujube*) were also identified as present (100% identity).

Results generated by ANCHOR generally agreed with the original study performed by Lang et al (Lang *et al*., 2017), in which the predominant genera observed (within the most abundant orders) were *Corynebacterium, Staphylococcus*, *Streptococcus, Finegoldia, Pseudomonas, Neisseria, Fusobacterium, Haemophilus, Akkermansia, Capnocytophaga, Selenomonas, Sphingomonas, Methylobacterium* and *Campylobacter.* Each of these genera included highly abundant OTUs, which could be annotated at species level when the data set was analyzed using ANCHOR, such as *Finegoldia magna, Haemophilus parainfluenzae* and *Akkermansia muciniphila*; although some ANCHOR OTUs were highly ambiguous, such as the most abundant OTU Staphylococcus MS_3, where the sequence was conserved at 100% identity to nine species [Table 6; Supplementary file 5; *Staphylococcus* was also the most abundant genera in Lang *et al*. (Lang *et al*., 2017)]. The additional resolution of ANCHOR analysis also yielded original bacterial species, such as *Lawsonella clevelandensis* (second most abundant OTU at 4.3% of all amplicons), as well as archaeal species, such as *Methanobrevibacter smithii*, *Methanosphaera stadtmaniae* and *Nitrosopumilus maritimus.* By again comparing results generated by ANCHOR to the analysis of this data by Lang (Lang *et al*., 2017) using Qiime1, it is possible to decipher that the *Corynebacterium* (genus) reported as dominating the ISS samples was actually constructed from: *L. clevelandensis* and *Corynebacterium tuberculostearicum*, two highly abundant species representing 4.3% and 2.2% of the sequences, and annotated by ANCHOR without ambiguity. The two major Qiime1 *Corynebacterium* sequences (OTU:495067 and OTU:1012948) corresponded to 4.25% and 2.11% of reads, respectively, and were indeed most similar to *L. clevelandensis* and *C. tuberculostearicum* (but at only 98% instead of 100% BLASTn identity NCBI nr/nt due to sequence modification).

Species from half of the dominant genera outlined in Lang *et al*. (Lang *et al*., 2017) could be identified by ANCHOR without ambiguity, including: *Campylobacter hominis*, *Fusobacterium nucleatum*, *H. parainfluenzae*, *A. muciniphila*, *Capnocytophaga leadbetteri*, *Selenomonas artemidis*, *Sphingobium yanoikuyae*, *F. magna* and *C. tuberculostearicum* (Table 6; Supplementary file 5). Most of these species could be considered normal gastrointestinal tract (GIT) bacteria found predominantly in the intestine/faeces or oral cavity. In the intestine/faeces: *F. magna* [has been associated to infection (Rosenthal *et al*., 2012)], *C. hominis* (Lawson *et al*., 2001) and *A. muciniphila* (Derrien *et al*., 2004). In the oral cavity (buccal flora): *F. nucleatum* (commensal but with association to a broad range of diseases [Han, 2015)], *H. parainfluenzae* [a common oral cavity bacteria with the potential to be a serious multiresistant opportunistic pathogen (Kosikowska *et al*., 2016)], *C. leadbetteri* (Frandsen *et al*., 2008) and *S. artemidis* [the specific sequence is similar to that of the isolate ATCC 43528 as well as a number of poorly annotated

Table 6. A comparison of most abundant organisms found in Lang *et al*. (Lang *et al*., 2017).

ANCHOR OTU 19 most abundant species	% Total raw counts	Amplicon ambiguity
Staphylococcus MS_3	8.77	12 = *S. aureus, S. capitis, S. caprae, S. epidermidis, S. haemolyticus, S. hominis, S. lugdunensis, S. pasteuri, S. petrasii, S. saccharolyticus, S. simiae, S. warneri*
Lawsonella clevelandensis_1	4.32	Unique
Lactobacillus MS_5	3.98	4 = *L. animalis, L. apodemi, L. faecis, L. murinus*
Streptococcus MS_6	2.52	5 = *Streptococcus cristatus, S. gordonii, S. infantis, S. mitis, S. oralis*
Corynebacterium tuberculostearicum_1	2.20	Unique
Homo Sapiens_53	2.15	Unique
Homo Sapiens_40	1.52	Unique
Pseudomonas MS_4	1.39	9 = *Pseudomonas alcaliphila, P. chengduensis, P. composti, P. indoloxydans, P. mendocina, P. oleovorans, P. pseudoalcaligenes, P. sihuiensis, P. toyotomiensis*
Akkermansia muciniphila_1	0.93	Unique
Haemophilus parainfluenzae_1	0.92	Unique
Pseudomonas lini_1	0.82	Unique
Alistipes_2	0.81	Unique
Corynebacterium MS_9	0.81	3 = *C. ihumii, C. mucifaciens, C. pilbarense*
Homo Sapiens_4	0.80	Unique
Finegoldia magna_1	0.73	Unique
Corynebacterium MS_12	0.72	2 = *C. accolens, C. macginleyi*
Bacteroides fragilis_1	0.68	Unique
Acinetobacter johnsonii_1	0.65	Unique

Lang *et al.,* 2017			Equivalent ANCHOR OTU(% counts)
19 most abundant orders	% Total raw counts	Dominant genus	to dominant (Lang) OTU
Actinomycetales[a]	18.3	*Corynebacterium*	Lawsonella clevelandensis_1 (4.3%)[b]
Bacillales	14	*Staphylococcus*	Staphylococcus MS_3 (8.77%)
Bacteroidales	12.8	Unclassified Rikenellaceae	Alistipes_2 (0.81%)
Lactobacillales	11.1	*Streptococcus*	Streptococcus MS_6 (2.52%)
Clostridiales	11	*Finegoldia*	Finegoldia magna_1 (0.73%)
Pseudomonadales	6.1	*Pseudomonas*	Pseudomonas MS_4 (1.39%)
Burkholderiales	5.6	Unclassified Comamonadaceae	Comamonadaceae MS_5 (0.55%)
Neisseriales	2.3	*Neisseria*	Neisseria MS_1 (0.58%)
Fusobacteriales	2.2	*Fusobacterium*	Fusobacterium nucleatum_1 (0.26%)
Pasteurellales	1.7	*Haemophilus*	Haemophilus parainfluenzae_1 (0.92%)
Verrucomicrobiales	1.6	*Akkermansia*	Akkermansia muciniphila_1 (0.93%)
Rhizobiales[c]	1.3	*Methylobacterium*	Methylobacterium MS_2 (0.44%)
Flavobacteriales	1.1	*Capnocytophaga*	Capnocytophaga leadbetteri_1 (0.08%)
Selenomonadales	1	*Selenomonas*	Selenomonas artemidis_1 (0.07%)
Sphingomonadales	0.9	*Sphingomonas*	Sphingobium yanoikuyae_1 (0.17%)
Sphingobacteriales	0.8	Unclassified Sphingobacteriales	Bacteroidetes MS_5 (0.02%)
Enterobacteriales	0.8	Unclassified Enterobacteriaceae	Enterobacterales MS_2 (0.37%)
Rhodobacterales[c]	0.6	*Rhodobacter*	Pseudorhodobacter MS_1 (0.10%)
Campylobacterales	0.6	*Campylobacter*	Campylobacter hominis_1 (0.11%)

Equivalent ANCHOR OTUs to the stated dominant genera are provided (the dominant genus in the order did not include the most abundant species in all cases). All 3347 ANCHOR OTUs, relative abundance and annotation as well as count distribution, blast statistics, alternative database hits and sequences are provided in Supplementary file 5.

a. Corynebacterium has now been placed in the order Corynebacteriales (Corynebacteriales ord. nov. Goodfellow and Jones 2015);

b. The second most abundant Corynebacterium genus annotated OTU in Lang *et al*. (Lang *et al*., 2017) was equivalent to ANCHOR OTU *C. tuberculostearicum_1* at 100% similarity.

c. Revised from presented data in Lang *et al*. (Lang *et al*., 2017) using their raw data.

sequences with the NCBI nt database, all of which were isolated from the human oral cavity (Bisiaux-Salauze *et al*., 1990)]. The exceptions to this were *S. yanoikuyae*, *L. clevelandensis* and *C. tuberculostearicum.* Although *Sphingobium* species are most often found in soils, and particularly contaminated soils, *S. yanoikuyae* [which has polycyclic aromatic hydrocarbons degrading capability (Kou *et al*., 2018)] was actually first isolated from human clinical samples (Yabuuchi *et al*., 1990). *L. clevelandensis* was only first described in 2013 (Harrington *et al*., 2013) and has since been repeatedly associated with abscess formation (Bell *et al*., 2016; Menezes *et al*., 2018); however, very recent research suggests it is a common human (nasal) commensal (Escapa *et al*., 2018). *C. tuberculostearicum* has traditionally been termed a 'leprosy-derived' *Corynebacterium*, having been first isolated from a Lepromatous leprosy case (Brown *et al*., 1984; Feurer *et al*., 2004); the OTU sequence here was a unique 100% identity match with this (type) strain Medalle X. Recent research, isolating 18 *C. tuberculostearicum* strains from human clinical specimens (Hinić *et al*., 2012), demonstrated multiple antimicrobial resistance in most isolates (but importantly, 100% susceptibility to vancomycin) and classified 7 of the 18 isolates as being clinically relevant to surgical site infection [centers for disease control (CDC) criteria (Henriksen *et al*., 2010)].

Sequences identifying the presence of bacteria belonging to the family Legionellaceae and Neisseriaceae were identified by Ichijo *et al*. (Ichijo *et al*., 2016) as present in the ISS and were highlighted within Lang *et al*. (Lang *et al*., 2017) as a concern due to these families containing well-characterized pathogenic members (there was some confusion in the manuscript as to their presence). No OTU belonging to Legionellaceae was identified here but 34 OTUs within the family Neisseriaceae were present. Eighteen of these were uniquely or ambiguously annotated at species level, including *Neisseria subflava*, *N. sicca*, *N. cinerea*, *N. oralis*, *N. elongate*, *N. lactamica*, *N. meningitidis*, *N. perflava*, *N. macacae*, *N. flavescens*, *N. mucosa*, *N. pharynges*, *Morococcus cerebrosus*, *Kingella denitrificans*, *Eikenella corrodens* and *Kingella oralis.* The presence of species such as *N. meningitides* could be a concern; however, as the species shares common 16S rRNA sequence at the amplified region, it cannot be distinguished from the human commensal upper respiratory tract bacteria *N. subflava* and *N. lactamica* (OTUs potentially representing *N. gonorrhoeae* were not detected).

Nonbacterial 16S rRNA gene sequences are often not reported in barcoding studies due to a general loss of utility for differentiating species using the technology. ANCHOR reports all data for downstream biological analysis (where they may be discarded); a large number of nonbacterial OTUs were identified on the ISS (Supplementary file 5). The presence of Japanese quail (*Coturnix japonica*) DNA on the ISS could be expected due to research conducted at the Avian Development Facility (such as Skeletal Development in Embryonic Quail, ADF Skeletal). Japanese quail has been used extensively as a model organism in space, as far back as 1979 (Soyuz 32), due to their low space requirement as well as their potential as a sustainable source of food. Similarly, it is also not surprising to find evidence of common food such as peas, carrots and apples (*Daucus* MS_1, *Pisum sativum*_1 and *Malus* MS_1) due to the practical challenges of zero gravity ingestion. The most abundant archaea, *Methanobrevibacter smithii* and *Methanosphaera stadtmaniae*, are commensal human methanogens (Miller and Wolin, 1985; Hansen *et al*., 2011), being the predominant human archaea and the first isolated human archaea respectively. The incredibly small (0.5–0.9 μm length) *N. maritimus* is an ammonia-oxidizing and ubiquitous across marine and terrestrial environments (Walker *et al*., 2010) (but perhaps not extraterrestrial, as it was only present in a single sample take from a keyboard in the laboratory). Fungal mitochondrial 16S rRNA gene sequences were also identified, the majority being *Panicillium* and *Aspergillus* species, which could derive from experiments underway during expedition 38/39, such as *Penicillium Growth Rate in Microgravity* (Pennsauken Phifer Middle School), but are common (ubiquitous) members of any environmental sample and have previously been identified on the ISS (Castro *et al*., 2004; Yamaguchi *et al*., 2014).

The prevalence of human GIT bacteria within isolated or repeatedly sterilized environments is very well documented within Lang *et al*. (Lang *et al*., 2017) as well as in the fascinating research performed by Mora *et al*. (Mora *et al*., 2016), which compared the ISS, intensive care units, operating rooms and cleanrooms. However, the extent to which the ISS environment here reflected human gastrointestinal microbiome samples surprised the authors. The long-term environmental and health impact of a persistent, solely human driven, habitat microbiome is hard to predict given that the technology only allows observation of some of the species within the community, does not distinguish between viable and nonviable bacteria, and, more generally, because the field of microbiome science is still in its infancy.

- *Differential abundance: U.S. laboratory vs sleeping stations.* Although no specific biological question was included in the original sampling design, we posed the hypothesis that location (Destiny Module US laboratory versus Harmony Module sleeping stations; Fig. 3) would comprise differentially abundant (DA) microbes and so grouped samples by this criteria for comparison using DESeq2 [a method for differential analysis of count data (Love *et al*., 2014)] (Supplementary file 1). Principal coordinates analysis (PCoA) on Bray Curtis distances (Fig. 3D) suggests samples separate by location with permutational multivariate analysis of variance (PERMANOVA) using an analysis of dissimilarity [Adonis function, R package vegan (Oksanen *et al*., 2007)] showing between group variance to be significantly greater than within group variance (Pr < 0.05). Shannon and Inverse Simpson alpha-diversity indices were both found to be significantly different (*t*-test, $p < 0.05$) [Fig. 3C; phyloseq (McMurdie and Holmes, 2013)]. Thirty-two OTUs

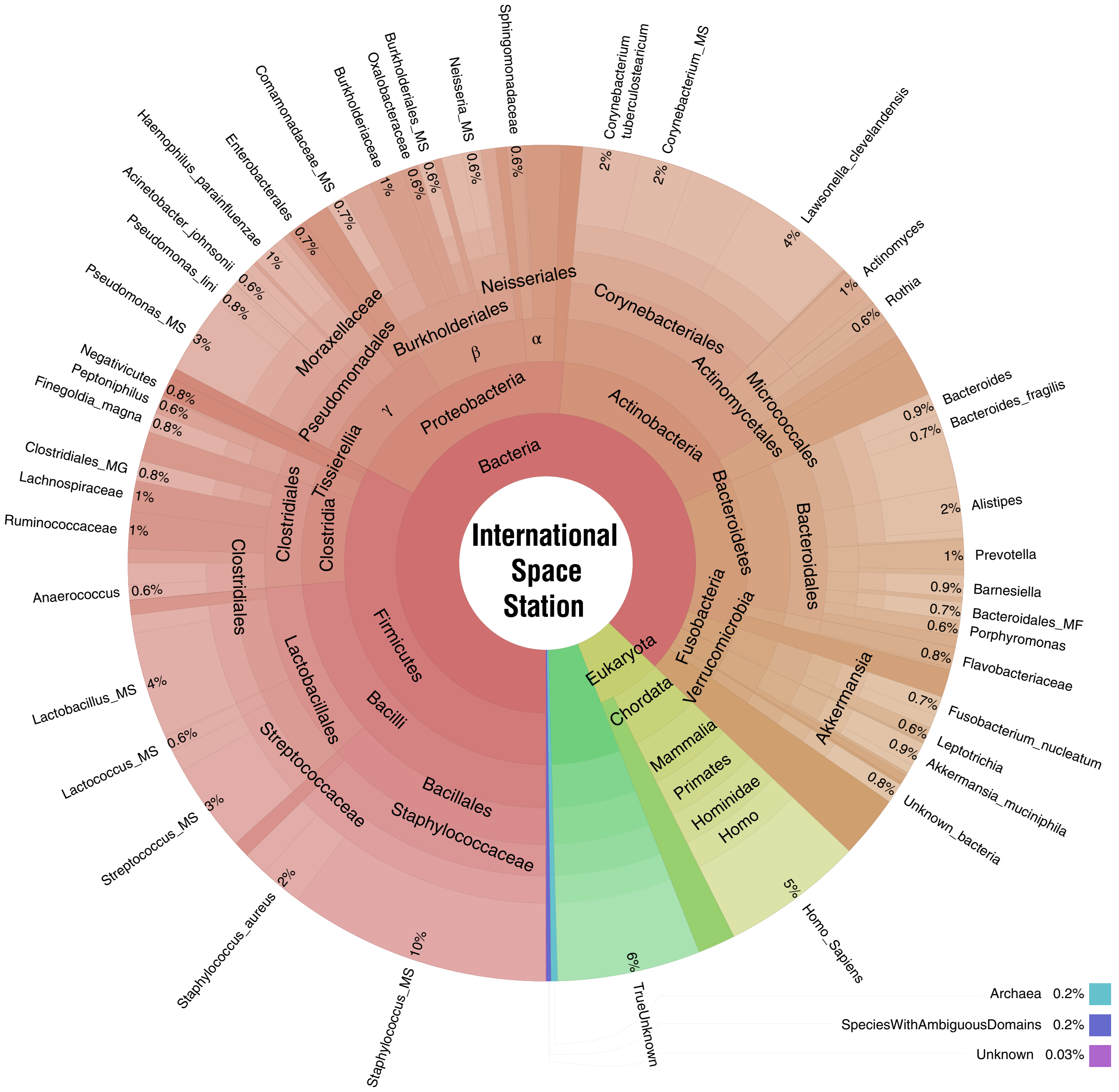

Fig. 3. Total community makeup from International Space Station Destiny and Harmony module surface swabs.
Krona graph [139] presenting the overview of OTUs and their abundance across all samples. The complete OTU table and including relative abundance, annotation, count distribution, blast statistics, alternative database hits, and sequences are provided in Supplementary file 5. MS, MG and MF refer to annotation as potentially multiple species, genera or families do to sequence conservation at the amplified region. Interactive figure available at https://github.com/gonzalezem/ANCHOR/tree/master/article. [Correction added on 18 June 2019, after first online publication: Figure 3 caption has been corrected in this version]. [Color figure can be viewed at wileyonlinelibrary.com]

were identified as DA between the samples taken from Destiny module (U.S. laboratory) and Harmony module (sleeping stations) (Fig. 4). Only 14 DA OTUs were annotated at the species level, eight of which were unique to a single species. Nineteen OTUs were in greater relative abundance within the sleeping stations, predominantly from the phylum Firmicutes (mostly Clostridiales, one Tissierellales) but also from Proteobacteria (Burkholderiales and Pseudomonadales), Bacteroidetes (Bacteroidales) and Actinobacteria (Bifidobacteriales). The remaining 13 DA OTUs, in higher relative abundance in the laboratory, were similarly from Firmicutes (Lactobacillales), Proteobacteria (Caulobacterales, Burkholderiales and Campylobacterales) and Bacteroidetes (Bacteroidales), but also from Cryptophyta (Cryptomonadales). Depending on the experimental design, it may be possible to establish or speculate as to

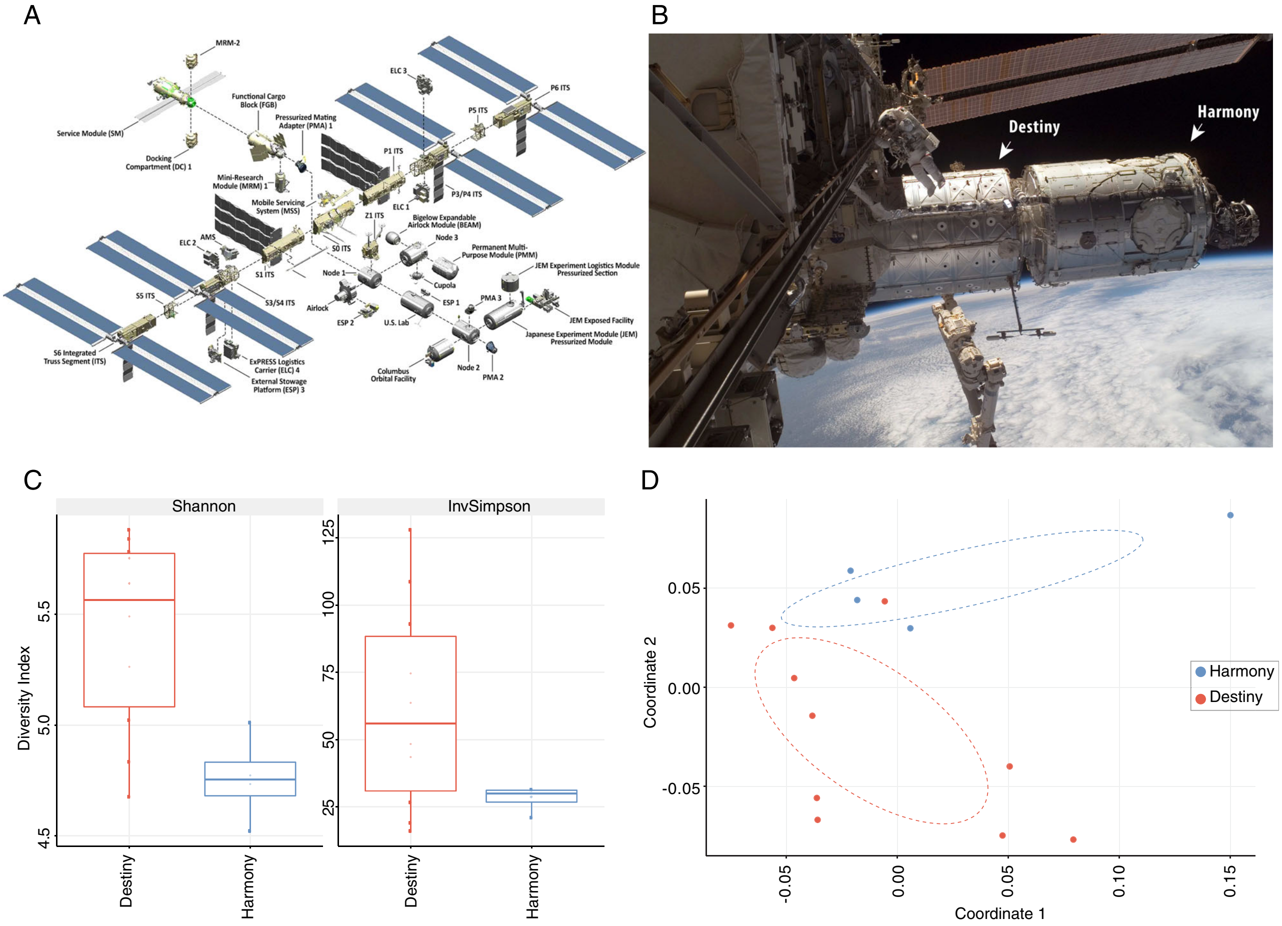

Fig. 4. Destiny and Harmony module community comparison.

A. Diagram of the ISS (https://www.nasa.gov/feature/facts-and-figures) with the Destiny Module is labelled as U.S. Lab while Harmony Module is labelled as Node 2 (includes sleeping stations).

B. Photograph (ISS016-E-012617, 24 Nov. 2007) of the Destiny Module and Harmony Module; Astronaut Peggy Whitson (expedition 16 commander, in frame) works over a 7-h, 4-min spacewalk with astronaut Daniel Tani (out of shot) outfitting Harmony module in position in front of the Destiny module.

C. Destiny and Harmony Module microbial community richness as measured by Shannon and Inverse Simpson were found to be significantly different (t-test, $p < 0.05$).

D. Composition of ISS communities in Harmony and Destiny modules represented by PCoA on Bray Curtis distances (PERMANOVA, Pr < 0.05). The first coordinate explains 22.3% of the total variation and the second 17.0%. Destiny n = 4 and Harmony n = 10 samples. Further richness and ordination is available at https://github.com/gonzalezem/ANCHOR/tree/master/article. [Correction added on 18 June 2019, after first online publication: Figure 4 caption has been corrected in this version]. [Color figure can be viewed at wileyonlinelibrary.com]

whether an increase or decrease in relative abundance of an OTU is driven by a specific factor, as well as the abolishment or creation of a novel niche for a species (Kou *et al*., 2018). While this is challenging within the design of the ISS sampling, speculation is made here as to the cause of change (presented only from the perspective of potential causal *increase* in abundance).

- *Differential abundance: higher in* Sleeping Station. Five of the 12 Firmicutes OTUs in higher relative abundance in sleeping quarters could be annotated at species level with four of those being unique species: *F. magna* (formerly *Peptostreptococcus magnus*), *Gemmiger formicilis*, *Ruminiclostridium leptum* (formerly *Clostridium leptum*) and *Levyella massiliensis* (Fig. 4; Supplementary file 5). All are relatively well characterized, often found (usually highly abundant) in the GIT and present in faecal matter of healthy humans (Gossling and Moore, 1975; Louis and Flint, 2009; La Scola *et al*., 2011; Rosenthal *et al*., 2012; Kabeerdoss *et al*., 2013); although *Levyella massiliensis* has also been found in koalas [with 'wet bottom' (Legione *et al*., 2018)]. Clostridiales MS_1 was ambiguously annotated as either *Butyricicoccus faecihominis* or *Agathobaculum butyriciproducens.* Both are butyrate-producing bacteria isolated from human faeces, maintaining the common pattern of GIT bacteria, but may be the same species as they share highly similar 16S rRNA gene sequences (100% at amplified region) and were both reported with International Journal of

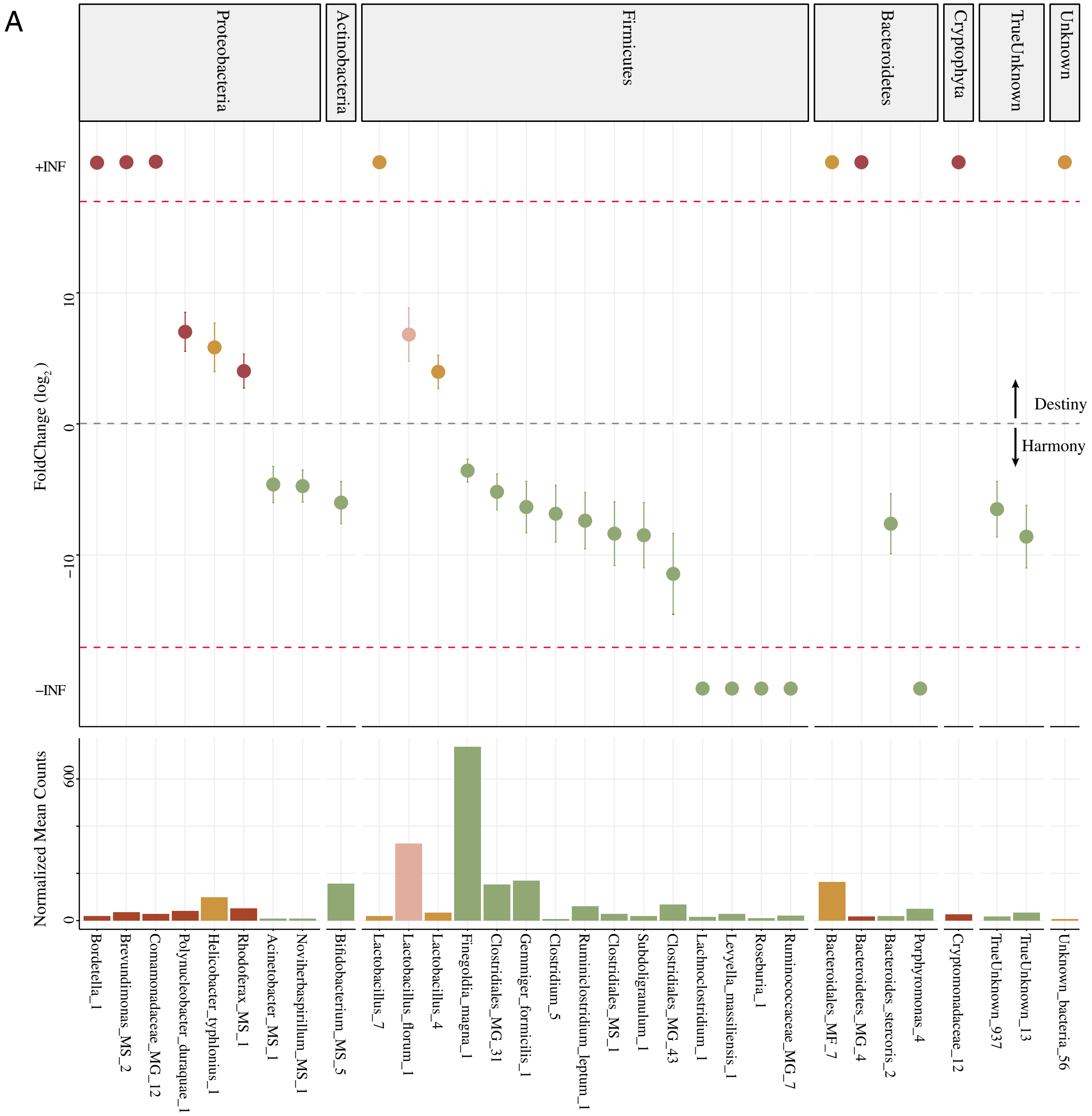

Fig. 5. Destiny and Harmony Module differential abundance.
A. Fold change and normalized mean counts. Fold change (FC Log2) is relative differences in abundance between locations. +/– INF (demarcated by the dashed red line) indicates 'infinite' fold change, where an OTU had detectable counts in samples from only a single location. Normalized mean counts originate from DESeq2 basemean output. Species are grouped by phylum.
B. Chord diagram illustrates the putative association of each DA OTU alongside the location where they were detected in the greatest abundance. The complete differential abundance table including relative abundance, fold change, annotation, count distribution, blast statistics, alternative database hits and sequences are provided in Supplementary file 5. Interactive figures are available at https://github.com/gonzalezem/ANCHOR/tree/master/article. [Correction added on 18 June 2019, after first online publication: Figure 5 caption has been corrected in this version]. [Color figure can be viewed at wileyonlinelibrary.com]

Systematic and Evolutionary Microbiology publications (within a month of each other) claiming to reclassify *Eubacterium desmolans* as either *Butyricicoccus desmolans* or *Agathobaculum desmolans* (Ahn *et al*., 2016; Takada *et al*., 2016). While the biology relating to microbial species represented by genus level OTUs is less precise, species within the genra *Roseburia* (Roseburia_1), *Subdoligranulum* (Subdoligranulum_1), *Clostridium* (Clostridium_5) and *Lachnoclostridium* (Lachnoclostridium_1) are also consistent with the pattern of GIT inhabiting (many butyrate

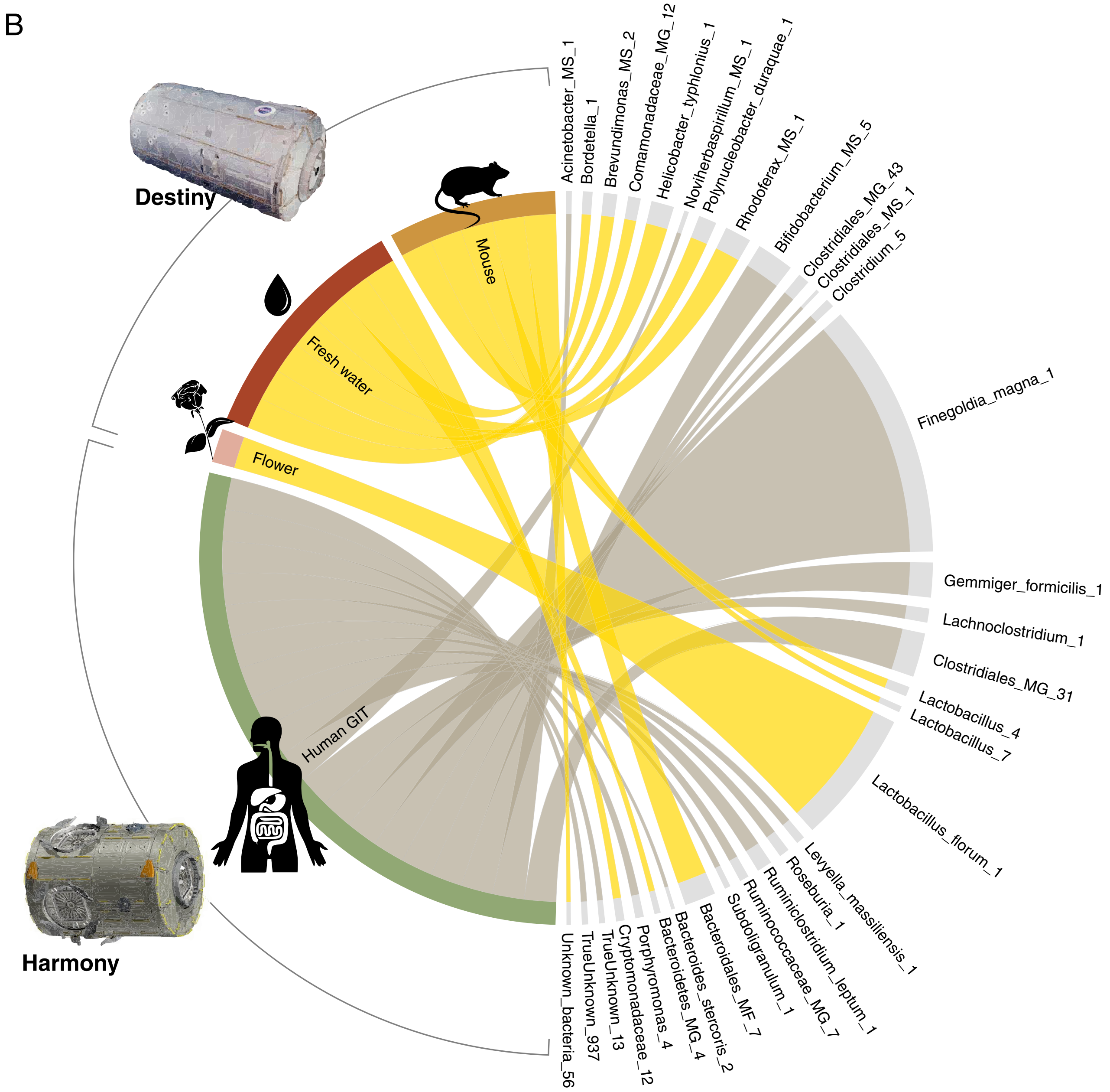

Fig. 5. (Continued). [Color figure can be viewed at wileyonlinelibrary.com]

producing) bacteria (Holmstrøm *et al*., 2004; Louis and Flint, 2009; Yutin and Galperin, 2013; Tamanai-Shacoori *et al*., 2017).

Two proteobacteria OTUs had increased relative abundance in samples from sleeping stations in the harmony module, Noviherbaspirillum MS_1 and Acinetobacter MS_1 (Fig. 4). Noviherbaspirillum MS_1 could be annotated (100% identity) as the very closely related (Ishii *et al*., 2017) *Noviherbaspirillum autotrophicum*, *Noviherbaspirillum denitrificans* and *Noviherbaspirillum massiliense* (β-proteobacteria). Interestingly, in the context highly abundant GIT clostriales bacteria, all three favour organic acids as carbon sources (including acetate, butyrate and succintate), and while *N. autotrophicum* and *N. denitrificans* were first isolated from soil and have an optimal temperature of 30°C, *N. massiliense* was first isolated from faecal samples (Lagier *et al*., 2012) and has an optimal temperature of 37°C (suggesting *N. massiliense* may be present when considered against the background of GIT flora). Acinetobacter MS_1 could be annotated as one of six Acinetobacter species (100% identity); while Acinetobacter species are highly diverse in the environment, these specific species have been isolated from clinical samples and have haemolytic capability (Bouvet and Grimont, 1986;

Nemec *et al*., 2009) or from human sewage plants (Carr *et al*., 2001). Two Bacteroidete OTUs were identified as in higher relative abundance in sleep station samples Bacteroides_stercoris_2 and Porphyromonas_4. *Bacteroides stercoris* is a normal GIT bacteria commonly isolated from human faeces (Johnson *et al*., 1986; Hong *et al*., 2008). Similarly, most species within the genus *Porphyromonas* are common GIT bacteria (in particular found in the oral cavity) (Wexler, 2007; Wang *et al*., 2016).

Only a single OTU from the phylum Actinobacteria, Bifidobacterium MS_5, was identified as DA with higher abundance in samples from sleep stations. The OTU sequence is conserved (100% identity) across two well-characterized *Bifidobacterium* species: *Bifidobacterium breve* and *Bifidobacterium longum.* In keeping with the domination of this environment with human GIT bacteria, the genus *Bifidobacterium* is found ubiquitously within the GIT and is readily cultured from faecal samples (Langendijk *et al*., 1995). Ambiguous annotation can allow for interesting interpretation in RNASeq (Gonzalez *et al*., 2018) and 16S rRNA barcoding studies (Kou *et al*., 2018); although other species across the genus would have to be assessed to confirm any biologically relevant hypotheses, the important bioinformatics step here is to not obscure any clues to a pattern of biological interest, which might drive further research. Bifidobacterium MS_5 is a useful example to illustrate the potential biological value of simply reporting the annotation for an observed sequence as opposed to the practices of either reporting a single species (often the first, alphabetically, within a blast return) or stepping up the taxonomy to report the genus. Many of the roughly 67 known species genus *Bifidobacterium* (NCBI taxonomy 04/2018) could very confidently *not* be potential annotation here (*Bifidobacterium psychraerophilum* strain T16, 95% identity, NR_029065.1 or Bifidobacterium magnum strain JCM 1218, 94% identity, NR_115644.1).

As well as relatively well-characterized bacteria, two DA OTUs observed in higher relative abundance in sleeping stations could not be annotated at >99% identity across any of the four databases queried: TrueUnknown_13 and TrueUnknown_937. A more detailed sequence investigation revealed TrueUnkown13 shared 97% similarity to *Prevotella buccalis* (formally Bacteroides), one of a number of *Prevotella* species found in the oral microbiome (Shah and Collins, 1990), while TrueUnknown937 shared 98% similarity to the newly described *Fenollaria massiliensis* and *F. timonensis* (Pagnier *et al*., 2014; Durand *et al*., 2017), observed in a variety of human microbiome samples including oral, bone, intestine and stool (from a variety of blast submissions).

- *Differential abundance: higher in U.S. Laboratory.* Three Firmicutes OTUs were identified as present in higher abundance in the US laboratory when than the samples from the sleeping stations: Lactobacillus florum_1, Lactobacillus_7 and Lactobacillus_4 (Fig. 4). *Lactobacillus florum* (F9-1) is commonly found in flowering plants and was first isolated from flowers of peony (Endo *et al*., 2009), which could be unexpected in the space station; however, astronauts were conducting a number of plant growth experiments during this period including: Resist Tubule, NanoRacks-WA-Resurrection Plant Growth, NanoRacks-VCHS-Improved Multiple Plant Growth and NanoRacks-GSH-Arugula Plant Growth, CARA and BRIC 18–2 (substantial, highly ambiguous plant chloroplast and mitochondrial 16S rRNA genes were also identified, although not DA, Supplementary file 5). The OTUs Lactobacillus_4 and Lactobacillus_7 (only detected in laboratory samples) shared 98% sequence identity but both were 98.8% similar to five lactobacillus strains (NCBI 16S refseq): *Lactobacillus apodemi* ASB1/DSM 16634 isolated from Japanese wood mouse faeces (Osawa *et al*., 2006), *Lactobacillus faecis* strain AFL13-2 isolated from animal faeces [a jackal (Endo *et al*., 2013)], Lactobacillus animalis strain KCTC 3501 isolated from animal teeth [a baboon (Dent and Williams, 1982)] and *Lactobacillus murinus* strains NBRC 14221/DSM 20452 and LMG 14189 both isolated from rat GIT (Hemme *et al*., 1980).

Six OTUs were annotated as Proteobacteria (2 α, 3 β and 1 ε): Rhodoferax MS_1, Brevundimonas MS_2, Polynucleobacter duraquae_1, Comamonadaceae MG_12, Bordetella_1 and Helicobacter_typhlonius_1. Rhidoferax MS_1 could be annotated as either *Rhodoferax ferrireducens* or *Rhodoferax saidenbachensis* [strains T118 and ED16 respectively (Kaden *et al*., 2014)], which share common 16S rRNA gene sequence at this amplified region. Both are psychrotolerant (can grow at 4 °C, although not pyschorophilic) bacteria commonly isolated from water; however, it is interesting that *R. ferrireducens* was transported to the space station as part of the first bacterial fuel cell experiments on the ISS [exhibition 8 (De Vet and Rutgers, 2007)]. Brevundimonas MS_2 was unique to the laboratory environment (Fig. 4) and could be annotated as either *Brevundimonas diminuta* (strains NBRC12697, ATCC11568, JCM2788 and LMG2089) or *Brevundimonas naejangsanensis* (strain Bio-TAS2-2) at 100% identity. Due to very a small size, *B. diminuta* has been extensively used to test point-of-use filters (0.2 μm) (Lee *et al*., 2002), including by NASA investigating drinking water storage (Tuan and Vega, 2010). The presence of *B. diminuta* has been previously observed extensively in highly isolated/filtered environments, including the ISS(Castro *et al*., 2004) and, more recently, MARS500 project [Microbial ecology of confined habitats an human health, MICHA (Schwendner *et al*., 2017)]. *Polynucleobacter duraquae* is also most often found in fresh water samples (and is free-living unlike many host-associated *Polynucleobacter* species) (Hahn *et al*., 2016). Similarly, the OTUs annotated as Comamonadaceae MG_12 (placed as within the genera *Acidovorax* or *Limnohabitans*) and Bordetella_1 did not correspond to any well characterized bacterial species, but have previously been identified or isolated as unknown bacteria (100% similar) from numerous samples deriving from fresh water (Shaw *et al*., 2008; Mueller-Spitz *et al*., 2009; Wu *et al*., 2012; Elser *et al*., 2014; Balmonte *et al*., 2016; Huang *et al*., 2016) and were also only detected in laboratory samples.

Helicobacter typhlonius, an ε-proteobacteria, was originally isolated independently from two laboratory mice (GIT and faeces) (Franklin *et al*., 2001) and has since been shown to be an endemic infection to terrestrial rodent research facilities (Chichlowski *et al*., 2008); although not as common as *Helicobacter ganmani* or *Helicobacter hepaticus* (Johansson *et al*., 2006), it could potentially be better adapted to microgravity environments. While microbial adaptation to microgravity has been studied (Nickerson *et al*., 2004; Chopra *et al*., 2006; Tirumalai *et al*., 2017), it is important to remember how little is known regarding the impact of the extraterrestrial environment on biology and therefore ecology. Extensive research has been conducted using mice as a model species in the ISS (investigating bone loss due to microgravity conditions, amongst other queries). As the swabs were taken during expedition 39, Nov 2013–May 2014, experiments would have been underway in the predecessor of the Rodent Research Facility, the Mice Draw System (Apr 2009–Sept 2014: https://www.nasa.gov/mission_pages/station/research/experiments/665.html), so it is perhaps not surprising then that one of the most prominent species identified as DA within the laboratory was *H. typhlonius.* On further investigation of OTUs not identified as DA, uniquely annotated OTUs representing *H. ganmani*, *H. hepaticus* and *H. rodentium* were also identified as present in relatively high abundance in laboratory samples (absent from sleeping station samples) but present in too few samples to overcome ANCHOR-applied DA sparsity filters (sequences putatively representing bacteria must be present in three or more samples for presumed relevancy to the biological question).

Two Bacteroidete OTUs were identified in higher abundance in the U.S. laboratory samples and absent (below detection limit or not present) from sleeping station samples: Bacteroidales MF_7 and Bacteroidetes MG_4. While these sequences are not currently associated to known species, the Bacteroidales MF_7 sequence was independently identified as present in mouse faecal samples [AJ400254 and AB606319 (Salzman *et al*., 2002; Matsumoto *et al*., 2005)] and placed in either Porphyromonadaceae or Muribaculaceae [mouse GIT bacteria family (Lagkouvardos *et al*., 2016)]. Bacteroidetes MG_4 has previously been independently identified (100% similar) in freshwater samples [JN634145.1, HQ663099.1 (Martinez-Garcia *et al*., 2012)] and placed in either *Dinghuibacter* or *Sphingobacterium.* The entirely consistent patterns of water- or mouse-associated bacteria in higher abundance in the U.S. laboratory samples can be extended to both the DA eukaryote OTU, Cryptomonadaceae_12 and the OTU identified as Unknown_bacteria_56 (also absent from sleep station samples). Cryptomonadaceae is an algal family containing genera such as *Cryptomonas*, which inhabit bodies of freshwater(Tranvik *et al*., 1989) and this specific Cryptomonadaceae_12 sequence has indeed been previously identified (100% similarity) in fresh water samples as an uncultured bacterial clone [interestingly from the same experiment identifying Comamonadaceae MG_12 at 100% similarity as an unknown bacteria (Elser *et al*., 2014)]. The Unknown_bacteria_56 sequence was most similar to the known species *Lactobacillus murinus* (98% similarity) and to a large number of unknown NCBI nt/nr hits (uncultured bacteria at 99% identity), which all derived from mouse microbiome samples taken during two independent experiments [studying the vagina of promiscuous mice and the gut of exercising mice on a high fat diet (MacManes, 2011; Evans *et al*., 2014)].

These results generally agree with the work of Lang *et al*. (Lang *et al*., 2017) identifying that bacteria within the Destiny and Harmony modules are dominated by those deriving from the human microbiome, and more specifically, the human GIT. The comparison between the two modules and across all the ISS samples yielded some starkly similar bacteria to those revealed during the MICHA experiment, namely relating to abundance of eOTU representing clostridium sp., Prevotella sp., Bifidobacterium sp., *Polynucleobacter* sp. and *Finegoldia* sp. in crew quarters, illustrating the value and strong design of the Mars500 project (Schwendner *et al*., 2017). Beyond this, while previous research highlights that domination of the ISS by human GIT bacteria is unsurprising, given humans are the only source of bacteria entering the environment, ANCHOR reveals laboratory surfaces also harbour bacteria deriving from those other microbiome carrying animals travelling upon the ISS, research facility rodents.

Conclusion

The purpose of ANCHOR development was to produce a microbial barcoding bioinformatics approach for multiple complex samples using 16S rRNA genes and with utility for users answering biological questions in-mind. As such, ANCHOR output aims to provide the best possible taxonomic resolution of microbial communities as well as maximize the information associated with each OTU.

ANCHOR performed well with very simple single sample data, identifying species when the marker was unique and equally well or better than contemporary pipelines when replicated samples are used. Surprisingly, the majority of gene copies that varied at the amplified region were distinguished as separate OTUs in mock data sets without OTU inflation, even when sequences differed by only one nucleotide.

By benchmarking technology intended to query biological hypotheses in complex systems against real data, the common challenges and compromises that are often not present or necessary within artificial or simple biological systems can be addressed. While such benchmarking can be challenging, there is no shortage of uncharacterised biology to test technology designed to explore the unknown and, importantly, such benchmarking ensures the obstacles sometimes separating biology and informatics are confronted. Using complex real-world data derived from swabs taken from the ISS, ANCHOR output agreed with previous findings as well as built upon them through novel biological discovery. These discoveries included confident identification of bacterial species associated with the human GIT, which were DA within the crew's sleeping quarters as well as the prevalence of DA mouse associated bacteria in samples from surfaces of the US laboratory.

The design of ANCHOR around human-based decisions should provide accessibility and flexibility to respond to diverse biological scenarios as well as maximize the meaningfulness of data deriving from poorly understood environments.

Acknowledgements

Thanks are extended to Evguenia Kopylova, Jose Navas-Molina, Céline Mercier, Zhenjiang Zech Xu, Frédéric Mahé, Yan He, Hong-Wei Zhou, Torbjørn Rognes, Gregory Caporaso, Rob Knight, James Kozich, Sarah Westcott, Nielson Baxter, Sarah Highlander, Patrick Schloss, Manuel Kleiner, Erin Thorson, Christine Sharp, Xiaoli Dong, Dan Liu, Carmen Li, Marc Strous, Jenna Lang, David Coil, Russell Neches, Wendy Brown, Darlene Cavalier, Mark Severance, Jarrad Hampton-Marcell, Jack Gilbert and Jonathan Eisen for making the data from their research freely available and easily navigable for other researchers to explore. The authors hope to have built upon the strong research of these authors in a constructive and respectful manner. We would like to thank David Anderson from the Canadian Centre for Computational Genomics for advising on figure design. Special thanks are also given to the astronauts of the ISS's 39th expedition, Alexander Skvortsov, Steve Swanson, Oleg Artemyev, Richard Mastracchio, Mikhail Tyurin and, in particular, Commander Koichi Wakata for so carefully taking surface swabs.

Data availability

Data generated from novel analyses during this study is included in supplementary information files. Additional material as well as all custom scripts are available at https://github.com/gonzalezem/ANCHOR.

Funding

The research was funded by Prof Pitre's NSERC Discovery Grant (RGPIN-2017-05452).

References

Acinas, S.G., Marcelino, L.A., Klepac-Ceraj, V., and Polz, M. F. (2004) Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons. *J Bacteriol* **186**: 2629–2635.

Ahn, S., Jin, T.-E., Chang, D.-H., Rhee, M.-S., Kim, H.J., Lee, S.J., *et al*. (2016) Agathobaculum butyriciproducens gen. nov. sp. nov., a strict anaerobic, butyrate-producing gut bacterium isolated from human faeces and reclassification of Eubacterium desmolans as Agathobaculum desmolans comb. nov. *Int J Syst Evol Microbiol* **66**: 3656–3661.

Anders, S., McCarthy, D.J., Chen, Y., Okoniewski, M., Smyth, G.K., Huber, W., and Robinson, M.D. (2013) Count-based differential expression analysis of RNA sequencing data using R and bioconductor. *Nat Protoc* **8**: 1765–1786.

Bacteriology, J.C.o.t.I.C.o.S. (1982) Opinion 54: rejection of the species name Pseudomonas denitrificans (Christensen) Bergey et al. 1923. *Int J Syst Evol Microbiol* **32**: 466.

Balmonte, J.P., Arnosti, C., Underwood, S., McKee, B.A., and Teske, A. (2016) Riverine bacterial communities reveal environmental disturbance signatures within the Betaproteobacteria and Verrucomicrobia. *Front Microbiol* **7**: 1441.

Bartram, A.K., Lynch, M.D., Stearns, J.C., Moreno-Hagelsieb, G., and Neufeld, J.D. (2011) Generation of multi-million 16S rRNA gene libraries from complex microbial communities by assembling paired-end Illumina reads. *Appl Environ Microbiol* **77**: 3846–3852.

Bécavin, C., Bouchier, C., Lechat, P., Archambaud, C., Creno, S., Gouin, E., *et al*. (2014) Comparison of widely used listeria monocytogenes strains EGD, 10403S, and EGD-e highlights genomic differences underlying variations in pathogenicity. *MBio* **5**: e00969–e00914.

Bell, M.E., Bernard, K.A., Harrington, S.M., Patel, N.B., Tucker, T.-A., Metcalfe, M.G., and McQuiston, J.R. (2016) Lawsonella clevelandensis gen. nov., sp. nov., a new member of the suborder Corynebacterineae isolated from human abscesses. *Int J Syst Evol Microbiol* **66**: 2929–2935.

Bisiaux-Salauze, B., Perez, C., Sebald, M., and Petit, J. (1990) Bacteremias caused by Selenomonas artemidis and Selenomonas infelix. *J Clin Microbiol* **28**: 140–142.

Bouvet, P.J., and Grimont, P.A. (1986) Taxonomy of the genus Acinetobacter with the recognition of Acinetobacter baumannii sp. nov., Acinetobacter haemolyticus sp. nov., Acinetobacter johnsonii sp. nov., and Acinetobacter junii sp. nov. and emended descriptions of Acinetobacter calcoaceticus and Acinetobacter lwoffii. *Int J Syst Evol Microbiol* **36**: 228–240.

Brereton, N.J., Gonzalez, E., Marleau, J., Nissim, W.G., Labrecque, M., Joly, S., and Pitre, F.E. (2016)

Comparative transcriptomic approaches exploring contamination stress tolerance in Salix sp. reveal the importance for a Metaorganismal de Novo Assembly Approach for nonmodel plants. *Plant Physiol* **171**: 3–24.

Brown, S., Lanéelle, M.-A., Asselineau, J., and Barksdale, L. (1984) Description of Corynebacterium tuberculostearicum sp. nov., a leprosy-derived Corynebacterium. In *Annales de l'Institut Pasteur/Microbiologie*. Paris: Masson, pp. 251–267.

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., and Holmes, S.P. (2016) DADA2: high-resolution sample inference from Illumina amplicon data. *Nat Methods* **13**: 581–583.

Callahan, B.J., McMurdie, P.J., and Holmes, S.P. (2017) Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J* **11**: 2639–2643.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., *et al*. (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7**: 335–336.

Carr, E., Eason, H., Feng, S., Hoogenraad, A., Croome, R., Soddell, J., *et al*. (2001) RAPD-PCR typing of Acinetobacter isolates from activated sludge systems designed to remove phosphorus microbiologically. *J Appl Microbiol* **90**: 309–319.

Castro, V.A., Thrasher, A.N., Healy, M., Ott, C.M., and Pierson, D.L. (2004) Microbial characterization during the early habitation of the International Space Station. *Microb Ecol* **47**: 119–126.

Chichlowski, M., Sharp, J.M., Vanderford, D.A., Myles, M.H., and Hale, L.P. (2008) Helicobacter typhlonius and helicobacter rodentium differentially affect the severity of colon inflammation and inflammation-associated neoplasia in IL10-deficient mice. *Comp Med* **58**: 534–541.

Chopra, V., Fadl, A., Sha, J., Chopra, S., Galindo, C., and Chopra, A. (2006) Alterations in the virulence potential of enteric pathogens and bacterial–host cell interactions under simulated microgravity conditions. *J Toxicol Environ Health A* **69**: 1345–1370.

Collins, M., Wallbanks, S., Lane, D., Shah, J., Nietupski, R., Smida, J., *et al*. (1991) Phylogenetic analysis of the genus listeria based on reverse transcriptase sequencing of 16S rRNA. *Int J Syst Evol Microbiol* **41**: 240–246.

De Vet, S., and Rutgers, R. (2007) From waste to energy: first experimental bacterial fuel cells onboard the international space station. *Microgravity Sci Technol* **19**: 225–229.

Dent, V., and Williams, R. (1982) Lactobacillus animalis sp. nov., a new species of lactobacillus from the alimentary canal of animals. *Zentralblatt für Bakteriologie Mikrobiologie und Hygiene: I Abt Originale C: Allgemeine, angewandte und ökologische Mikrobiologie* **3**: 377–386.

Derrien, M., Vaughan, E.E., Plugge, C.M., and de Vos, W.M. (2004) Akkermansia muciniphila gen. nov., sp. nov., a human intestinal mucin-degrading bacterium. *Int J Syst Evol Microbiol* **54**: 1469–1476.

Dong, X., Kleiner, M., Sharp, C.E., Thorson, E., Li, C., Liu, D., and Strous, M. (2017) Fast and simple analysis of MiSeq amplicon sequencing data with MetaAmp. *Front Microbiol* **8**: 1461.

Durand, G., Cadoret, F., Lagier, J., Fournier, P., and Raoult, D. (2017) Description of 'Gorbachella massiliensis' gen. nov., sp. nov.,'Fenollaria timonensis' sp. nov.,'-Intestinimonas timonensis' sp. nov. and 'Collinsella ihuae'sp. nov. isolated from healthy fresh stools with culturomics. *New Microbes New Infect* **16**: 60–62.

Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**: 2460–2461.

Edgar, R.C. (2016) UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *BioRxiv* : 081257.

Edgar, R.C. (2017) Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *bioRxiv* : 192211.

Edgar, R. (2018) Taxonomy annotation and guide tree errors in 16S rRNA databases. *PeerJ* **6**: e5030.

Elser, J.J., Bastidas, M., Corman, J.R., Emick, H., Kellom, M., Laspoumaderes, C., *et al*. (2014) Community structure and biogeochemical impacts of microbial life on floating pumice. *Appl Environ Microb* **81**: 03160–03114.

Endo, A., Futagawa-Endo, Y., and Dicks, L.M. (2009) Isolation and characterization of fructophilic lactic acid bacteria from fructose-rich niches. *Syst Appl Microbiol* **32**: 593–600.

Endo, A., Irisawa, T., Futagawa-Endo, Y., Salminen, S., Ohkuma, M., and Dicks, L. (2013) Lactobacillus faecis sp. nov., isolated from animal faeces. *Int J Syst Evol Microbiol* **63**: 4502–4507.

Escapa, I.F., Chen, T., Huang, Y., Gajare, P., Dewhirst, F.E., and Lemon, K.P. (2018) New insights into human nostril microbiome from the expanded Human Oral Microbiome Database (eHOMD): a resource for species-level identification of microbiome data from the aerodigestive tract. *bioRxiv* : 347013.

Evans, C.C., LePard, K.J., Kwak, J.W., Stancukas, M.C., Laskowski, S., Dougherty, J., *et al*. (2014) Exercise prevents weight gain and alters the gut microbiota in a mouse model of high fat diet-induced obesity. *PLoS One* **9**: e92193.

Feurer, C., Clermont, D., Bimet, F., Candrea, A., Jackson, M., Glaser, P., *et al*. (2004) Taxonomic characterization of nine strains isolated from clinical and environmental specimens, and proposal of Corynebacterium tuberculostearicum sp. nov. *Int J Syst Evol Microbiol* **54**: 1055–1061.

Fox, G.E., Pechman, K.R., and Woese, C.R. (1977) Comparative cataloging of 16S ribosomal ribonucleic acid: molecular approach to procaryotic systematics. *Int J Syst Evol Microbiol* **27**: 44–57.

Frandsen, E.V., Poulsen, K., Könönen, E., and Kilian, M. (2008) Diversity of Capnocytophaga species in children and description of Capnocytophaga leadbetteri sp. nov. and Capnocytophaga genospecies AHN8471. *Int J Syst Evol Microbiol* **58**: 324–336.

Franklin, C.L., Gorelick, P.L., Riley, L.K., Dewhirst, F.E., Livingston, R.S., Ward, J.M., *et al*. (2001) Helicobacter typhlonius sp. Nov., a novel murine urease-negative Helicobacterspecies. *J Clin Microbiol* **39**: 3920–3926.

Gonzalez, E., Brereton, N.J., Marleau, J., Nissim, W.G., Labrecque, M., Pitre, F.E., and Joly, S. (2015) Meta-transcriptomics indicates biotic cross-tolerance in willow trees cultivated on petroleum hydrocarbon contaminated soil. *BMC Plant Biol* **15**: 1.

Gonzalez, E., Pitre, F., Pagé, A., Marleau, J., Nissim, W.G., St-Arnaud, M., *et al*. (2018) Trees, fungi and bacteria: tripartite metatranscriptomics of a root microbiome responding to soil contamination. *Microbiome* **6**: 53.

Goodhew, C.F., Pettigrew, G.W., Devreese, B., Van Beeumen, J., Van Spanning, R.J., Baker, S.C., *et al*. (1996) The cytochromes c-550 of Paracoccus denitrificans and Thiosphaera pantotropha: a need for re-evaluation of the history of Paracoccus cultures. *FEMS Microbiol Lett* **137**: 95–101.

Gossling, J., and Moore, W. (1975) Gemmiger formicilis, n. gen., n. sp., an anaerobic budding bacterium from intestines. *Int J Syst Evol Microbiol* **25**: 202–207.

Gyllenberg, H.G. (1963) A genreal method for deriving determination schemes for random collections of microbial isolates. Annales Academiae Scientarum Fennicase, Series A (IV Biologica).

Hahn, M.W., Schmidt, J., Pitt, A., Taipale, S.J., and Lang, E. (2016) Reclassification of four Polynucleobacter necessarius strains as representatives of Polynucleobacter asymbioticus comb. nov., Polynucleobacter duraquae sp. nov., Polynucleobacter yangtzensis sp. nov. and Polynucleobacter sinensis sp. nov., and emended description of Polynucleobacter necessarius. *Int J Syst Evol Microbiol* **66**: 2883–2892.

Han, Y.W. (2015) Fusobacterium nucleatum: a commensal-turned pathogen. *Curr Opin Microbiol* **23**: 141–147.

Hansen, E.E., Lozupone, C.A., Rey, F.E., Wu, M., Guruge, J.L., Narra, A., *et al*. (2011) Pan-genome of the dominant human gut-associated archaeon, Methanobrevibacter smithii, studied in twins. *Proc Natl Acad Sci U S A* **108**: 201000071.

Harrington, S.M., Bell, M., Bernard, K., Lagacé-Wiens, P., Schuetz, A., Hartman, B., *et al*. (2013) Novel fastidious, partially acid fast, anaerobic gram positive bacillus associated with abscess formation and recovered from multiple medical centers. *J Clin Microbiol* **51**: 01497–01413.

Hemme, D., Raibaud, P., Ducluzeau, R., Galpin, J., Sicard, P., and Van, J.H. (1980) “Lactobacillus murinus” n. sp., a new species of the autochtoneous dominant flora of the digestive tract of rat and mouse (author’s transl). *Ann Microbiol (Paris)* **131**: 297–308.

Henriksen, N., Meyhoff, C., Wetterslev, J., Wille-Jørgensen, P., Rasmussen, L., Jorgensen, L., and Group, P.T. (2010) Clinical relevance of surgical site infection as defined by the criteria of the Centers for Disease Control and Prevention. *J Hosp Infect* **75**: 173–177.

Hinić, V., Lang, C., Weisser, M., Straub, C., Frei, R., and Goldenberger, D. (2012) Corynebacterium tuberculostearicum: a potentially misidentified and multiresistant Corynebacterium species isolated from clinical specimens. *J Clin Microbiol* **50**: 2561–2567.

Holmstrøm, K., Collins, M.D., Møller, T., Falsen, E., and Lawson, P.A. (2004) Subdoligranulum variabile gen. nov., sp. nov. from human feces. *Anaerobe* **10**: 197–203.

Hong, P.-Y., Wu, J.-H., and Liu, W.-T. (2008) Relative abundance of Bacteroides spp. in stools and wastewaters as determined by hierarchical oligonucleotide primer extension. *Appl Environ Microbiol* **74**: 2882–2893.

Huang, Y., Zeng, Y., Lu, H., Feng, H., Zeng, Y., and Koblížek, M. (2016) Novel acsF gene primers revealed a diverse phototrophic bacterial population including Gemmatimonadetes in the Lake Taihu. *Appl Environ Microbiol* **82**: 01063–01016.

Hugenholtz, P., Tyson, G.W., Webb, R.I., Wagner, A.M., and Blackall, L.L. (2001) Investigation of candidate division TM7, a recently recognized major lineage of the domain bacteria with no known pure-culture representatives. *Appl Environ Microbiol* **67**: 411–419.

Ichijo, T., Yamaguchi, N., Tanigaki, F., Shirakawa, M., and Nasu, M. (2016) Four-year bacterial monitoring in the international Space Station—Japanese experiment module “Kibo” with culture-independent approach. *npj Microgravity* **2**: 16007.

Ishii, S., Ashida, N., Ohno, H., Segawa, T., Yabe, S., Otsuka, S., *et al*. (2017) Noviherbaspirillum denitrificans sp. nov., a denitrifying bacterium isolated from rice paddy soil and Noviherbaspirillum autotrophicum sp. nov., a denitrifying, facultatively autotrophic bacterium isolated from rice paddy soil and proposal to reclassify Herbaspirillum massiliense as Noviherbaspirillum massiliense comb. nov. *Int J Syst Evol Microbiol* **67**: 1841–1848.

Janssen, S., McDonald, D., Gonzalez, A., Navas-Molina, J.A., Jiang, L., Xu, Z.Z., *et al*. (2018) Phylogenetic placement of exact amplicon sequences improves associations with clinical information. *mSystems* **3**: e00021–e00018.

Johansson, S.K., Feinstein, R.E., Johansson, K.-E., and Lindberg, A.V. (2006) Occurrence of helicobacter species other than H. hepaticus in laboratory mice and rats in Sweden. *Comp Med* **56**: 110–113.

Johnson, J.L., Moore, W., and Moore, L.V. (1986) Bacteroides caccae sp. nov., Bacteroides merdae sp. nov., and Bacteroides stercoris sp. nov. isolated from human feces. *Int J Syst Evol Microbiol* **36**: 499–501.

Kabeerdoss, J., Sankaran, V., Pugazhendhi, S., and Ramakrishna, B.S. (2013) Clostridium leptum group bacteria abundance and diversity in the fecal microbiota of patients with inflammatory bowel disease: a case–control study in India. *BMC Gastroenterol* **13**: 20.

Kaden, R., Spröer, C., Beyer, D., and Krolla-Sidenstein, P. (2014) Rhodoferax saidenbachensis sp. nov., a psychrotolerant, very slowly growing bacterium within the family Comamonadaceae, proposal of appropriate taxonomic position of Albidiferax ferrireducens strain T118T in the genus Rhodoferax and emended description of the genus Rhodoferax. *Int J Syst Evol Microbiol* **64**: 1186–1193.

Kelly, D.P., Euzeby, J.P., Goodhew, C.F., and Wood, A.P. (2006) Redefining Paracoccus denitrificans and Paracoccus pantotrophus and the case for a reassessment of the strains held by international culture collections. *Int J Syst Evol Microbiol* **56**: 2495–2500.

Klappenbach, J.A., Saxman, P.R., Cole, J.R., and Schmidt, T.M. (2001) Rrndb: the ribosomal RNA operon copy number database. *Nucleic Acids Res* **29**: 181–184.

Kleiner, M., Thorson, E., Sharp, C.E., Dong, X., Liu, D., Li, C., and Strous, M. (2017) Assessing species biomass contributions in microbial communities via metaproteomics. *Nat Commun* **8**: 1558.

Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., and Glöckner, F.O. (2013) Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res* **41**: e1–e1.

Knight, R., Vrbanac, A., Taylor, B.C., Aksenov, A., Callewaert, C., Debelius, J., *et al*. (2018) Best practices for analysing microbiomes. *Nat Rev Microbiol* **1**: 410–422.

Kopylova, E., Navas-Molina, J.A., Mercier, C., Xu, Z.Z., Mahé, F., He, Y., *et al*. (2016) Open-source sequence clustering methods improve the state of the art. *mSystems* **1**: e00003–e00015.

Kosikowska, U., Biernasiuk, A., Rybojad, P., Łoś, R., and Malm, A. (2016) Haemophilus parainfluenzae as a marker of the upper respiratory tract microbiota changes under the influence of preoperative prophylaxis with or without postoperative treatment in patients with lung cancer. *BMC Microbiol* **16**: 62.

Kou, S., Vincent, G., Gonzalez, E., Pitre, F.E., Labrecque, M., and Brereton, N.J. (2018) The response of a 16S ribosomal RNA gene fragment amplified community to Lead, zinc, and copper pollution in a Shanghai field trial. *Front Microbiol* **9**: 366.

Kozich, J.J., Westcott, S.L., Baxter, N.T., Highlander, S.K., and Schloss, P.D. (2013) Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl Environ Microbiol* **79**: 5112–5120.

La Scola, B., Fournier, P.-E., and Raoult, D. (2011) Burden of emerging anaerobes in the MALDI-TOF and 16S rRNA gene sequencing era. *Anaerobe* **17**: 106–112.

Lagier, J.-C., Gimenez, G., Robert, C., Raoult, D., and Fournier, P.-E. (2012) Non-contiguous finished genome sequence and description of Herbaspirillum massiliense sp. nov. *Stand Genomic Sci* **7**: 200.

Lagkouvardos, I., Pukall, R., Abt, B., Foesel, B.U., Meier-Kolthoff, J.P., Kumar, N., *et al*. (2016) The mouse intestinal bacterial collection (miBC) provides host-specific insight into cultured diversity and functional potential of the gut microbiota. *Nat Microbiol* **1**: 16131.

Lang, J.M., Coil, D.A., Neches, R.Y., Brown, W.E., Cavalier, D., Severance, M., *et al*. (2017) A microbial survey of the International Space Station (ISS). *PeerJ* **5**: e4029.

Langendijk, P.S., Schut, F., Jansen, G.J., Raangs, G.C., Kamphuis, G.R., Wilkinson, M., and Welling, G.W. (1995) Quantitative fluorescence in situ hybridization of Bifidobacterium spp. with genus-specific 16S rRNA-targeted probes and its application in fecal samples. *Appl Environ Microbiol* **61**: 3069–3075.

Lapage, S., Bascomb, S., Willcox, W., and Curtis, M. (1973) Identification of bacteria by computer: general aspects and perspectives. *Microbiology* **77**: 273–290.

Lawson, A.J., On, S., Logan, J., and Stanley, J. (2001) Campylobacter hominis sp. nov., from the human gastrointestinal tract. *Int J Syst Evol Microbiol* **51**: 651–660.

Lee, S.-H., Lee, S.-S., and Kim, C.-W. (2002) Changes in the cell size of Brevundimonas diminuta using different growth agitation rates. *PDA J Pharm Sci Technol* **56**: 99–108.

Legione, A.R., Amery-Gale, J., Lynch, M., Haynes, L., Gilkerson, J.R., Sansom, F.M., *et al.* (2018) Variation in themicrobiome of the urogenital tract of Chlamydia-free female koalas (Phascolarctos cinereus) with and without 'wet bottom'. *PloS one* **13**: e0194881.

Lenk, S., Moraru, C., Hahnke, S., Arnds, J., Richter, M., Kube, M., *et al*. (2012) Roseobacter clade bacteria are abundant in coastal sediments and encode a novel combination of sulfur oxidation genes. *ISME J* **6**: 2178–2187.

Li, W., and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**: 1658–1659.

Louis, P., and Flint, H.J. (2009) Diversity, metabolism and microbial ecology of butyrate-producing bacteria from the human large intestine. *FEMS Microbiol Lett* **294**: 1–8.

Love, M., Anders, S., and Huber, W. (2014) Differential analysis of count data–the DESeq2 package. *Genome Biol* **15**: 550.

Love, M.I., Anders, S., Kim, V., and Huber, W. (2015) RNA-Seq workflow: gene-level exploratory analysis and differential expression. *F1000Research* **4**: 1070.

MacManes, M.D. (2011) Promiscuity in mice is associated with increased vaginal bacterial diversity. *Naturwissenschaften* **98**: 951–960.

Magoč, T., and Salzberg, S.L. (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**: 2957–2963.

Martinez-Garcia, M., Swan, B.K., Poulton, N.J., Gomez, M. L., Masland, D., Sieracki, M.E., and Stepanauskas, R. (2012) High-throughput single-cell sequencing identifies photoheterotrophs and chemoautotrophs in freshwater bacterioplankton. *ISME J* **6**: 113–123.

Martinez-Porchas, M., Villalpando-Canchola, E., Suarez, L. E.O., and Vargas-Albores, F. (2017) How conserved are the conserved 16S-rRNA regions? *PeerJ* **5**: e3036.

Matsumoto, M., Sakamoto, M., Hayashi, H., and Benno, Y. (2005) Novel phylogenetic assignment database for terminal-restriction fragment length polymorphism analysis of human colonic microbiota. *J Microbiol Methods* **61**: 305–319.

McMurdie, P.J., and Holmes, S. (2013) phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* **8**: e61217.

Menezes, M.F., Sousa, M.J., Paixão, P., Atouguia, J., Negreiros, I., and Simões, M. (2018) Lawsonella clevelandensis as the causative agent of a breast abscess. *IDCases* **12**: 95–96.

Miller, T.L., and Wolin, M.J. (1985) Methanosphaera stadtmaniae gen. nov., sp. nov.: a species that forms methane by reducing methanol with hydrogen. *Arch Microbiol* **141**: 116–122.

Mora, M., Mahnert, A., Koskinen, K., Pausan, M.R., Oberauner-Wappis, L., Krause, R., *et al*. (2016) Microorganisms in confined habitats: microbial monitoring and control of intensive care units, operating rooms, cleanrooms and the International Space Station. *Front Microbiol* **7**: 1573.

Mueller-Spitz, S.R., Goetz, G.W., and McLellan, S.L. (2009) Temporal and spatial variability in nearshore bacterioplankton communities of Lake Michigan. *FEMS Microbiol Ecol* **67**: 511–522.

Muyzer, G., De Waal, E.C., and Uitterlinden, A.G. (1993) Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Appl Environ Microbiol* **59**: 695–700.

Nemec, A., Musilek, M., Maixnerova, M., De Baere, T., van der Reijden, T.J., Vaneechoutte, M., and Dijkshoorn, L. (2009) Acinetobacter beijerinckii sp. nov. and Acinetobacter gyllenbergii sp. nov., haemolytic organisms isolated from humans. *Int J Syst Evol Microbiol* **59**: 118–124.

Nguyen, N.-P., Warnow, T., Pop, M., and White, B. (2016) A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity. *NPJ Biofilms Microbiomes* **2**: 16004.

Nickerson, C.A., Ott, C.M., Wilson, J.W., Ramamurthy, R., and Pierson, D.L. (2004) Microbial responses to microgravity and other low-shear environments. *Microbiol Mol Biol Rev* **68**: 345–361.

Oksanen, J., Kindt, R., Legendre, P., O'Hara, B., Stevens, M.H.H., Oksanen, M.J., and Suggests, M. (2007) The vegan package. Community Ecolossgy Package **10**: 631–637.

Olsen, G.J., Lane, D.J., Giovannoni, S.J., Pace, N.R., and Stahl, D.A. (1986) Microbial ecology and evolution: a ribosomal RNA approach. *Annu Rev Microbiol* **40**: 337–365.

Osawa, R., Fujisawa, T., and Pukall, R. (2006) Lactobacillus apodemi sp. nov., a tannase-producing species isolated from wild mouse faeces. *Int J Syst Evol Microbiol* **56**: 1693–1696.

Pace, N.R., Stahl, D.A., Lane, D.J., and Olsen, G.J. (1986) The analysis of natural microbial populations by ribosomal RNA sequences. In *Advances in Microbial Ecology*. Boston, MA: Springer, pp. 1–55.

Pagnier, I., Croce, O., Robert, C., Raoult, D., and Scola, B. (2014) Non-contiguous finished genome sequence and description of Fenollaria massiliensis gen. nov., sp. nov., a new genus of anaerobic bacterium. *Stand Genomic Sci* **9**: 704.

Palleroni, N.J., and Bradbury, J.F. (1993) Stenotrophomonas, a new bacterial genus for Xanthomonas maltophilia (Hugh 1980) Swings et al. 1983. *Int J Syst Evol Microbiol* **43**: 606–609.

Pei, A.Y., Oberdorf, W.E., Nossa, C.W., Agarwal, A., Chokshi, P., Gerz, E.A., *et al*. (2010) Diversity of 16S rRNA genes within individual prokaryotic genomes. *Appl Environ Microbiol* **76**: 3886–3897.

Rainey, F.A., Kelly, D.P., Stackebrandt, E., Burghardt, J., Hiraishi, A., Katayama, Y., and Wood, A.P. (1999) A re-evaluation of the taxonomy of Paracoccus denitrificans and a proposal for the combination Paracoccus pantotrophus comb. nov. *Int J Syst Evol Microbiol* **49**: 645–651.

Rehakova, K., Johansen, J.R., Bowen, M.B., Martin, M.P., and Sheil, C.A. (2014) Variation in secondary structure of the 16S rRNA molecule in cyanobacteria with implications for phylogenetic analysis. *Fottea* **14**: 161–178.

Rosenthal, M.E., Rojtman, A.D., and Frank, E. (2012) Finegoldia magna (formerly Peptostreptococcus magnus): an overlooked etiology for toxic shock syndrome? *Med Hypotheses* **79**: 138–140.

Salzman, N.H., de Jong, H., Paterson, Y., Harmsen, H.J., Welling, G.W., and Bos, N.A. (2002) Analysis of 16S libraries of mouse gastrointestinal microflora reveals a large new group of mouse intestinal bacteriab. *Microbiology* **148**: 3651–3660.

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., *et al*. (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75**: 7537–7541.

Schwendner, P., Mahnert, A., Koskinen, K., Moissl-Eichinger, C., Barczyk, S., Wirth, R., *et al*. (2017) Preparing for the crewed Mars journey: microbiota dynamics in the confined Mars500 habitat during simulated Mars flight and landing. *Microbiome* **5**: 129.

Shah, H.N., and Collins, D.M. (1990) Prevotella, a new genus to include Bacteroides melaninogenicus and related species formerly classified in the genus Bacteroides. *Int J Syst Evol Microbiol* **40**: 205–208.

Shaw, A.K., Halpern, A.L., Beeson, K., Tran, B., Venter, J. C., and Martiny, J.B. (2008) It's all relative: ranking the diversity of aquatic bacterial communities. *Environ Microbiol* **10**: 2200–2210.

Sneath, P.H. (1957) Some thoughts on bacterial classification. *Microbiology* **17**: 184–200.

Sneath, P. (1964) New approaches to bacterial taxonomy: use of computers. *Annu Rev Microbiol* **18**: 335–346.

Sokal, R.R. (1965) Statistical methods in systematics. *Biol Rev* **40**: 337–389.

Takada, T., Watanabe, K., Makino, H., and Kushiro, A. (2016) Reclassification of Eubacterium desmolans as Butyricicoccus desmolans comb. nov., and description of Butyricicoccus faecihominis sp. nov., a butyrate-producing bacterium from human faeces. *Int J Syst Evol Microbiol* **66**: 4125–4131.

Tamanai-Shacoori, Z., Smida, I., Bousarghin, L., Loreal, O., Meuric, V., Fong, S.B., *et al*. (2017) Roseburia spp.: a marker of health? *Future Microbiol* **12**: 157–170.

Tirumalai, M.R., Karouia, F., Tran, Q., Stepanov, V.G., Bruce, R.J., Ott, C.M., *et al*. (2017) The adaptation of Escherichia coli cells grown in simulated microgravity for an extended period is both phenotypic and genomic. *npj Microgravity* **3**: 15.

Tranvik, L.J., Porter, K.G., and Sieburth, J.M. (1989) Occurrence of bacterivory in Cryptomonas, a common freshwater phytoplankter. *Oecologia* **78**: 473–476.

Tuan, G., and Vega, L. (2010) Crew exploration vehicle potable water system verification description. In *40th International Conference on Environmental Systems.* p. 6156.

Vetrovsky, T., and Baldrian, P. (2013) The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. *PLoS One* **8**: e57923.

Walker, C., De La Torre, J., Klotz, M., Urakawa, H., Pinel, N., Arp, D., *et al*. (2010) Nitrosopumilus maritimus genome reveals unique mechanisms for nitrification and autotrophy in globally distributed marine crenarchaea. *Proc Natl Acad Sci* **107**: 8818–8823.

Wang, K., Lu, W., Tu, Q., Ge, Y., He, J., Zhou, Y., *et al*. (2016) Preliminary analysis of salivary microbiome and their potential roles in oral lichen planus. *Sci Rep* **6**: 22943.

Weiss, S., Xu, Z.Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., *et al*. (2017) Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* **5**: 27.

Wexler, H.M. (2007) Bacteroides: the good, the bad, and the nitty-gritty. *Clin Microbiol Rev* **20**: 593–621.

Woese, C.R., Gutell, R., Gupta, R., and Noller, H. (1983) Detailed analysis of the higher-order structure of 16S-like ribosomal ribonucleic acids. *Microbiol Rev* **47**: 621.

Wu, L., Ge, G., Zhu, G., Gong, S., Li, S., and Wan, J. (2012) Diversity and composition of the bacterial community of Poyang Lake (China) as determined by 16S rRNA gene sequence analysis. *World J Microbiol Biotechnol* **28**: 233–244.

Yabuuchi, E., Yano, I., Oyaizu, H., Hashimoto, Y., Ezaki, T., and Yamamoto, H. (1990) Proposals of Sphingomonas paucimobilis gen. nov. and comb. nov., Sphingomonas parapaucimobilis sp. nov., Sphingomonas yanoikuyae sp. nov., Sphingomonas adhaesiva sp. nov., Sphingomonas capsulata comb, nov., and two Genospecies of the genus Sphingomonas. *Microbiol Immunol* **34**: 99–119.

Yamaguchi, N., Roberts, M., Castro, S., Oubre, C., Makimura, K., Leys, N., *et al*. (2014) Microbial monitoring of crewed habitats in space—current status and future perspectives. *Microbes Environ* **29**: 250–260.

Yutin, N., and Galperin, M.Y. (2013) A genomic update on clostridial phylogeny: G ram-negative spore formers and other misplaced clostridia. *Environ Microbiol* **15**: 2631–2641.

Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2013) PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**: 614–620.

Ziesemer, K.A., Mann, A.E., Sankaranarayanan, K., Schroeder, H., Ozga, A.T., Brandt, B.W., *et al*. (2015) Intrinsic challenges in ancient microbiome reconstruction using 16S rRNA gene amplification. *Sci Rep* **5**: 16498.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Supplementary file 1 – Dataset specifics (includes references: (DeSantis *et al*., 2006; Schloss *et al*., 2009; Caporaso *et al*., 2010; Haas *et al*., 2011; Anders *et al*., 2013; Kozich *et al*., 2013; Love *et al*., 2014; Gonzalez *et al*., 2015; Love *et al*., 2015; Zhbannikov and Foster, 2015; Brereton *et al*., 2016; Callahan *et al*., 2016; Thorsen *et al*., 2016; Kleiner *et al*., 2017; Lang *et al*., 2017; Gonzalez *et al*., 2018; Kou *et al*., 2018; Parks *et al*., 2018))
Supplementary file 2 – Even and Staggered data
Supplementary file 3 – Kozich's Mock data
Supplementary file 4 – Kleiner's Mock data
Supplementary file 5 – ISS data