C h a p t e r F o u r

DEVELOPMENT OF AN IN-HOUSE PREPARATION, PROCESSING, AND ANALYSIS WORKFLOW FOR NEXT GENERATION SEQUENCING DATA

David H. Case[1]

in collaboration with,

Alexis Pasulka[1], Elizabeth Trembath-Reichert[1], Stephanie Connon[1], Katherine Dawson[1], and Victoria Orphan[1]

[1]Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, CA, USA

This chapter will be made publically available through the Caltech thesis repository, as well as by request to the authors.

4.0 ABSTRACT

Next-generation sequencing ("iTag", "NGS") has revolutionized microbial surveys in the last decade, enabling orders of magnitude advances in sample throughput and per-sequence cost. As expected, many aspects of the upstream (i.e., sample preparation) and downstream (i.e., data processing) employment of iTag have required development in order to fully and accurately utilize new technologies. In this chapter I discuss the Orphan lab's development of a workflow for upstream and downstream iTag methods (Fig. 1). As a result of our methods testing, we are able to make estimates of iTag precision (i.e., reproducibility, estimated at 0.77-1.85% relative abundance) and accuracy (variable depending on taxa and polymerase enzyme; range: 20x underrepresentation to 7x overrepresentation). The field of next-generation sequencing remains in rapid development (and may be moving into so-called "next-next-generation sequencing" or "3rd generation sequencing"), such that workflows and analysis techniques are in constant need of development, improvement, and updating. What follows is in many ways a "state of the lab" snapshot that will change quickly as time passes. For example, as of this writing the Orphan lab is considering switching to a newer, more flexible approach of sample preparation that would allow simultaneous sequencing of multiple gene targets; this approach was not feasible even a year or two ago but will likely herald a new period of sequencing moving forward. At the end of this chapter, a primer is included to familiarize the reader with some of the quantitative ecological techniques most frequently employed in this thesis: Nonmetric Muldi-Dimensional Scaling (NMDS), Analysis of Similarity (ANOSIM), and Similarity Percentage (SIMPER).

4.1 INTRODUCTION AND PREPARING SAMPLES FOR SEQUENCING

The process of DNA extraction by definition isolates genomic material ("gDNA") from microbial cells (although not without biases depending on extraction method – see Morono et al. 2014). While the emergent fields of genomics and meta-genomics, enabled by next-generation sequencing technologies, interpret entire genomic content, in many cases a particular experiment only calls for analysis of specific genes. The most common assay is analysis of the 16S rRNA gene, a core ribosomal gene (Woese and Fox 1977). The 16S rRNA gene is conserved across all microbial life, is vertically inherited, and mutates at a slow but steady rate across evolutionary time. These characteristics make the 16S rRNA gene a high-quality marker of inter-organism relatedness, and therefore a metric of phylogenetic identity and whole-community diversity. In order to isolate the 16S rRNA gene from genomic content, a polymerase chain reaction (PCR) is employed (Kleppe et al. 1971; Mullis and Faloona 1987; Saiki et al. 1988). PCR is a flexible approach in which any region of genomic interest can be amplified in a series of repeating reactions ("cycles"). In principle, each cycle of a PCR doubles the concentration of the gene of interest; PCR is regularly performed for 20-35 cycles, thereby amplifying the gene of interest by many orders of magnitude over the initial gDNA template.

The entire 16S rRNA gene is ~1,500 base pairs (bp) in length and includes nine hypervariable regions interspersed within conserved regions. Sequencing of the hypervariable regions enables robust inter-species resolution, and ideally an experiment will include sequencing of all nine regions. Sequencing technology has evolved over the last four decades since the "molecular revolution" began with the advent of "Sanger sequencing" (Sanger and Coulson 1975; Sanger et al. 1977a; b). Sanger sequencing offers robust, high-quality, and most notably, nearly full-length sequencing of genes of interest, including hypervariable regions of the 16S rRNA gene. However, on a per-sequence basis Sanger sequencing is expensive and due to cost and time constraints is often limited to ~100 sequences per sample. Nonetheless, Sanger

sequencing remained the gold standard of DNA sequencing technology for three decades, until the advent of "next-generation sequencing", a new wave of DNA sequencing technologies which offered orders of magnitude more sequences per analysis as well as the ability to "multiplex" – to sequence hundreds of samples simultaneously. The development of next-generation sequencing is marked by numerous corporate acquisitions and rapid invention and obsolescence of technologies on a year-to-year timescale. Roughly, next-generation sequencers began to be commercially viable with the advent of 454/Roche Sequencing (Margulies et al. 2005), which offered thousands of sequences per sample, albeit at truncated length (a maximum length of several hundred bp – not enough for the entire 16S rRNA gene). Several years later, Illumina, Inc. released a commercially available sequencing platform that increased the number of sequences per sample by an order of magnitude and increased base calling accuracy, but again at the cost of shorter sequences – often limited to one hypervariable region of the 16S rRNA gene (Bentley et al. 2008). Illumina sequencing has remained the dominant method for massively parallel single-gene sequencing, enabling microbial ecology studies that span many environments and time points. The technology has additionally improved to allow longer sequences without sacrificing quality. However, a new "next-next-generation" of DNA sequencing is in development, incorporating technologies recently released by Ion Torrent Systems Inc. and Pacific Biosciences (Schadt et al. 2010).[1]

In order to prepare gDNA for sequencing of the 16S rRNA gene (the Orphan lab uses an Illumina MiSeq platform operated at Laragen, Inc.), two PCR steps are employed (see sections below for more details on the development of sample preparation protocols). First, duplicate PCRs are performed to amplify the V4 region of the 16S rRNA gene for each sample. These products, after being checked for quality by gel electrophoresis, are pooled and transferred to a second PCR reaction in which unique barcodes are appended to amplicons, as well as

[1] These paragraphs describing a brief overview of DNA sequencing are not comprehensive, do not describe all corporate and academic contributions to the development of the field, and are not an endorsement of any one particular DNA sequencing technology.

oligonucleotide adapters which bind the amplicon to the Illumina MiSeq flow cell upon sequencing. At this point, the barcoded amplicons from each sample are uniquely tagged and therefore may be combined into a single mixture. Prior to combination, each sample's amplicon pool is quantified by fluorescence assay. Samples are then combined by adding an equi-molar amount of each sample to one batch tube, so that no single sample swamps the signal from all others. This single aliquot, containing uniquely tagged 16S rRNA gene amplicons from hundreds of different samples, is passed through a PCR cleanup kit (Qiagen, Inc) and shipped to Laragen, Inc. for sequencing.

4.2 PROCESSING RAW DATA

The following section will describe the Orphan lab's workflow for processing raw Illumina MiSeq data, followed by a section detailing tests which were run in order to determine the best PCR practices for preparing environmental iTag samples. Briefly, this section will include discussion of three sample types: negative controls, plasmid mock communities, and genomic mock communities. Negative controls are PCR reactions which were run with zero gDNA template added (1 µL of PCR-grade water was used as a volume substitute) and amplified for enough cycles to produce a product which could then be sequenced and processed like any other sample. The plasmid mock communities (n=4) were generated by mixing known ratios of plasmids from uncultured methane seep organisms (Table 1). The genomic mock communitites (n=4) were generated by mixing known ratios of gDNA extract from cultured organismsm grown in the laboratory (Table 2). Both types of mock communities are employed to test iTag precision; the plasmid mock communities are additionally used to test iTag accuracy.

4.2.1 JOINING & QUALITY CHECKING, CHIMERA DETECTION, & SINGLETON REMOVAL

When sequences are first generated on an Illumina platform, they are produced as "paired-end reads". Paired-end reads consist of two separate sequences ("R1" and "R2") which represent sequencing from opposite ends of a single amplicon, respectively[2]. If the amplicon is short enough, the forward and reverse sequencing overlap one another. This allows the two reads to be "joined" into a single contig, in which they are overlapped and base calls are checked against one another. There are many algorithms for joining, and in practice our lab has settled on a commonly used software package, fastq-join (Aronesty 2011). Our implementation of this software requires a minimum overlap of 50 bp between the forward and reverse reads, with no more than an 8% difference in base calls within the overlapping region (therefore, no more than 4 mismatched base calls in a 50 bp overlap region). If a paired R1 and R2 read fail to meet these two criteria, they are removed from the dataset. In instances where 4 or fewer mismatches are identified, the contig is assigned at that position the base call which corresponds to the higher quality value from the R1 or R2 read[3]. In our datasets, there is wide range in the proportion of sequences removed during joining (17%±9%; Fig. 2).

Joining the R1 and R2 reads inherently provides initial quality filtering of the raw sequence data, by virtue of checking the forward read against the reverse read. However, further quality filtering is needed in order to assess non-overlapping regions of the contig. Therefore, a quality-filtering step is performed in which contigs must meet two criteria: first, a contig is not

---

[2] For an excellent description of Illumina MiSeq sequencing, see the following video introduction produced by Illumina, Inc.: https://www.youtube.com/watch?v=womKfikWlxM

[3] Quality values, termed "Q-scores" or "Phred scores", are a value indicating the confidence that an assigned base call is correct. Q-scores and probability are related by the following equation:

$$P = 10^{(-Q/10)}$$

For example, a Q-score of 30 corresponds to 10^{-3}, or a 1-in-1,000 probably the base call is incorrect (99.9% certainty). In practice, the most frequently used cutoff values for "acceptable" Q-scores are 20, 25, or 30.

allowed to have any "N" base calls in which a base is unidentified. Given even just one "N" base call, a contig is removed from the dataset. Second, every base call of a contig must have at least a Q-score of 30, meaning that after quality filtering every base of every contig is known to greater than or equal to 99.9% certainty. In practice only a small proportion of sequences are removed at this step (2%±1%, Fig. 2), but these quality filtering steps, applied with the QIIME software (Caporaso et al. 2010), improve our confidence in downstream interpretation of the sequencing profiles of our samples.

Quality filtering is an effective method of improving the dataset's veracity, but further steps are necessary before proceeding to data interpretation. After quality filtering, we apply a chimera checking step. Chimeras are oligonucleotide fragments which arise during a PCR reaction, in which two 16S rRNA gene fragments are erroneously combined into a single amplicon. These amplicons may be relatively abundant in the dataset and during sequencing may be assigned high quality scores; they are not, however, meaningful in the context of a given study and therefore ought to be removed. We achieve this by employing the UCHIME algorithm within the USEARCH software package (Edgar et al. 2011), which checks whether one half of a contig aligns strongly to a database sequence (Seq_A) while the other half of a contig aligns more strongly to a different database sequence (Seq_B). If so, the contig is identified as chimeric and removed from the dataset. If not (e.g., both halves of the contig align best to Seq_A), then the contig is deemed non-chimeric and retained in the dataset. For samples amplified with the 5-PRIME polymerase enzyme, this step usually removes a moderate proportion of sequences (5%±3%; Fig. 2); for samples amplified with the NEB Q5 polymerase, the removal rate is higher (~10-20%, data not shown). The specific parameters of chimera checking in UCHIME are highly modifiable; we apply the default stringencies for chimera detection, but future users could tune the parameters as desired. In particular, chimera detection on short Illumina sequences (in our case, <300 bp in length and covering only one hypervariable region of the 16S rRNA gene) presents novel challenges not faced when detecting chimeras on full-length 16S rRNA gene sequences. With

only one hypervariable region to analyze, is may not be feasible to identify two spliced 16S rRNA genes in a single contig (Nelson et al. 2014; Ruiz-Calderon et al. 2016), and some recent studies have not included chimera detection (Metcalf et al. 2016; Ruiz-Calderon et al. 2016). As iTag processing methodologies continue to develop, this question ought to be further addressed. Nonetheless, it remains common (arguably standard) in the community to apply chimera detection to short iTag sequences (Kozich et al. 2013; Nelson et al. 2014; Dominguez-Bello et al. 2016), as it has been for the published studies in this thesis (Mason et al. 2015; Case et al. 2015). Furthermore, as iTag technology improves and longer 16S rRNA sequences become feasible, covering multiple hypervariable regions, chimera detection will once more become unquestionably relevant and applicable, and therefore will remain a step in iTag processing workflows in the future.

It is possible at this point that hundreds or thousands of individual sequences are highly similar (or even 100% identical) to one another, and therefore these sequences may be grouped into clusters in order to save computational time. This is done by creating operational taxonomic units (OTUs): bins of highly similar sequences (Edgar, 2010). Most often, these OTUs are created by requiring 97% or 99% similarity of all sequences within the group. Because a fundamental interest of microbial ecologists is to identify the many species present in their dataset, a taxonomy is matched to each OTU (Wang et al., 2007), wherein as representative of each OTU the most frequently occurring oligonucleotide sequence is chosen. These steps are performed in the QIIME software package.

Even after these steps, "spurious" OTUs will remain which pass all quantitative quality thresholds but nonetheless are unlikely to represent genuine microbial community information. At this point a singleton-removal cutoff is applied, in which OTUs that occur one time in one sample in the entire dataset (i.e., a lone sequence dissimilar to any other sequence among the entire dataset). This step does not eliminate many sequences from the dataset (~1%; Fig. 2), but is nonetheless important for filtering out spurious sequencing data.

4.2.2 NEGATIVE CONTROLS

As with any method, it is critical to address contamination through analysis of negative controls. In iTag sequencing, a negative control is particularly important because the depth of sequencing allows unprecedented views into the "rare biosphere", those OTUs which are present at roughly less than 1-in-100 abundance in the microbial community, and thus would have been missed in the majority of clone library studies employing Sanger sequencing. However, iTag sequencing is fundamentally dependent on a PCR reaction at the very first step of sample preparation. By definition PCR reactions amplify low amounts of DNA into higher concentrations; while this is generally a benefit that has enabled the "molecular revolution" over the last ~40 years, it also presents challenges for iTag sequencing. Not even the most pure PCR-grade water is truly clean of genomic content, nor are the other components of a PCR reaction, i.e., dNTP mixes, polymerase enzymes, and oligonucleotide primers. This is not generally a problem, as the contaminant genomic content present in PCR reagents is usually swamped by the (relatively) high concentration of one's sample gDNA introduced as template. To the extent that contaminant genomic material is amplified, it is often in low enough relative bundance as to not be captured in clone libraries with low sequencing depth. However, if template gDNA concentration is low, approaching parity with contaminant DNA, then the resulting amplified product can be a mixture of genuine, sample-derived gene amplicons and a significant fraction of contaminant gene amplicons. Even in PCR reactions where the overall template gDNA concentration is orders of magnitude higher than contaminant genomic content, the 16S rRNA gene copies of a particular organism (e.g., a species present at 1-in-10,000 abundance in the bulk microbial community) may still approach parity with contaminant genomic content. In that case, it becomes critical to define which sequences are trusted as genuine and which are suspected to be

contaminants, as well as the threshold below which a user no longer trusts any iTag sequences to reflect genuine signal from low-abundance community members.

In order to address negative control contamination in iTag data, we have employed two approaches: one "pre-PCR" laboratory practice and one bioinformatics technique. The "pre-PCR" approach, while obvious, is unfortunately not always practiced in microbiology laboratories. Simply put, during iTag PCR preparation in our laboratory, we take extra care to maintain a working environment clean of possible exogenous sources of genomic contamination. When setting up PCR reactions with high-concentration template and for the purpose of generating clone libraries (where only community members greater than ~1% of the population will likely be observed), it is not uncommon for users to set up the PCR reaction on the bench-top after a wipe-down with ethanol and/or bleach and perhaps an open flame to sterilize surrounding air. However, these approaches are insufficient for iTag PCR setup. Instead, we perform all PCR setup in a Purifier Class II Biosafety Cabinet (Labconco) whose airflow is HEPA filtered to greater than >99.99% particulate purity. The cabinet itself is wiped before and after each use with RNAse AWAY (Molecular Bio-Products, Inc.), a surfactant specifically designed to clean PCR equipment. A dedicated set of pipets and pipet tips is left in the cabinet at all times, and these equipment are also wiped with RNAse AWAY before and after use. Furthermore, the pipets, pipet tips, plastic PCR strips, and the interior of the cabinet itself are all subjected to ultraviolet light sterilization before and after every use in order to break down contaminant gDNA. While none of these techniques are able to address genomic contamination in PCR reagents themselves (e.g., dNTPs, polymerase enzymes), they do at least minimize the chances of exogenous genomic contamination during PCR setup. As a part of this "pre-PCR" practice, with every iTag sequencing run we include at least one negative control PCR reaction. This is a reaction in which PCR-grade water is used as the "template", and amplification is run for enough cycles that an amplicon band is observed on electrophoresis gel (environmental samples are generally amplified for 30 cycles in the first PCR, while the negative control usually requires 35-37 cycles in order to

generate an observable product). From there on, the sample is treated like any regular environmental sample in the preparation and sequencing pipeline.

Including a negative control in the sequencing enables our bioinformatics approache to addressing contamination. This approach is simply to identify which OTUs are present in the sequences from the negative control and remove them from the samples, not unlike a blank subtraction (Fig. 3). Since our lab began iTag sequencing November 2013, we have performed six runs (2013-11, 2014-05, 2014-11, 2015-03, 2015-09, 2015-12) and have used two polymerase enzymes (5-PRIME Hot Master Mix and NEB Q5, see discussion below). Over those runs and with both enzymes, we have observed contaminant OTUs associated with a range of taxonomies. Negative controls amplified with the 5-PRIME enzyme have more consistent composition than the negative controls amplified with NEB Q5. 5-PRIME negative controls are rich in a Gammaproteobacterial OTU (#57 in Fig. 3) as well as OTUs associated with Betaproteobacteria, Firmicutes, and Bacteroidetes. Negative controls amplified with the NEB Q5 enzyme, while also occasionally containing OTUs associated with those taxa, also have exhibited OTUs associated with Acidobacteria, Actinobacteria, Planctomycetes, and Alphaproteobacteria. This wide range of OTUs associated with negative controls demonstrates the importance of including a negative control in every iTag run. This enables our bioinformatics approach to addressing contamination, which is not fundamentally different than a "blank correction" applied across a variety of disciplines: for each run, we identify the OTUs present in the corresponding negative control and subtract those OTUs from all environmental samples in that run's dataset. This is bioinformatically simple to do, and can be achieved by two methods. The first method, which is more conservative, is to identify the taxonomies associated with negative control OTUs for a given iTag run (e.g., in November 2014 the 5-PRIME negative control contained Clostridia-associated OTUs; Fig. 3). Then, all OTUs associated with that taxonomy can be removed from the environmental dataset. While this method conservatively removes data which might be contaminant-related, it has the potential to remove OTUs from the environmental dataset which

are genuine (e.g., Clostridia OTUs which are genuinely amplified from the environmental template, not from contamination). A more refined method, then, is to identify only the specific OTUs associated with the negative control for a given run and remove solely those OTUs from the environmental dataset. In a sample set of our environmental data, this second method removed 4%±2% of the dataset's sequences (Fig. 2). In earlier publications (Mason et al. 2015; Case et al. 2015), the conservative "taxa-removal" method was applied, but in future studies I recommend the more precise "OTU-removal" method as it still accurately addresses contamination while retaining as much genuine environmental data as possible. From a practical standpoint, the two methods do not make a significant difference in studies from the methane seep environment. This is because the methane seep microbial ecosystem is dominantly populated by taxa (e.g., ANME archaea, Deltaproteobacteria, Epsilonproteobacteria) which are rarely observed in the negative controls. Therefore removing, for example, an entire clade observed in the negative controls has a limited impact on the methane seep environmental samples. However, in other environments negative control contaminant taxa may be closely related to organisms of experimental interest. For example, human microbiome samples are often rich in Bacteroidetes- and Firmicutes-related organisms (Faith et al. 2013; Rosenbaum et al. 2015). Therefore it is not reasonable to remove all OTUs associated with these taxa, even if some Bacteroidetes- and Firmicutes-associated OTUs are observed in the negative control. In such a case, an "OTU-removal" method would be best suited to address the experimental goals.

At this stage, it is also logical to remove OTUs from the dataset which are undesirable for other reasons besides their presence in the negative controls. For example, OTUs which are "unassigned" a taxonomy (often these are sequences which have high enough quality to pass joining and quality criteria but have zero matches in the reference database – and therefore cannot by definition be identified as chimeric) are not useful and are often removed. Also, depending on a study's objectives, certain clades may simply not be desired for the scientific question at hand. For example, if only Bacteria are of interest for the study, all OTUs associated

with Archaea could be intentionally removed at this step. It is also not uncommon for a few OTUs to be identified as Eukaryotic; these are generally removed at this point.

4.2.3 TRESHOLD FILTERING

In addition to removing OTUs linked to contamination, it is also necessary to consider spurious OTUs which may have been generated either during the preparation PCR reactions or during sequencing itself (Fig. 4). This was already partially addressed by the removal of singletons (OTUs which occur once, in one sample, in the entire dataset; Fig. 2). However, further applying a threshold cutoff, below which OTUs are summarily removed, can minimize the impact of spurious OTUs and increase confidence in the remaining very-low-abundance sequences (Bokulich et al. 2013). To address this, we take advantage of our plasmid mock communities which have known composition. We know that 12 plasmids were introduced; therefore, in a perfect dataset we would expect to only see 12 OTUs recovered in the sequencing data. However, even after removing singleton OTUs and subtracting contaminant taxa as determined by the relevant negative control, we still observe hundreds of OTUs present in the plasmid mock communities (Fig. 4). This observation is not improved by using either the 5-PRIME or NEB Q5 polymerase. Ultimately, this suggests that interpretation of very low abundance OTUs in iTag data may be fraught, since many OTUs appear to be spurious. Thus, it is desirable to choose a threshold relative abundance below which OTUs are removed from the dataset. Clearly a higher threshold will remove more OTUs, but at the cost of eliminating data which might represent genuine, low abundance 16S rRNA gene copies. In our mock communities, a threshold cutoff of 0.001 (0.1% relative abundance) appears to be appropriate; such a cutoff would come very close to limiting our sequencing data to the 12 OTUs expected to be observed (Fig. 4). However, in practice environmental samples uniformly host greater 16S rRNA gene diversity than mock communities; therefore, removing data less abundant than 0.1% may eliminate OTUs which

genuinely represent microbial diversity. Furthermore, in practice the threshold cutoff is often applied to the entire dataset; that is, if the threshold value is 0.001, then of all the sequences across all samples, the sequences in OTU_x must be present at least above 0.1% of the entire dataset. It is easy to conceive of a scenario in which OTU_x is genuinely highly abundant in $Sample_x$, but not in other samples in the dataset. Therefore, OTU_x might fail the entire dataset threshold and be removed from the sample set. In such a scenario, genuine and significant information will have been lost from the sequencing run, and the resulting sequencing composition of $Sample_x$ will not be reflective of its actual 16S rRNA gene profile. An obvious solution to this issue would be to apply cutoff thresholds on a per-sample basis, rather than a per-dataset basis. Because the cutoff threshold is often applied on a whole-dataset basis, a lower cutoff threshold minimizes the likelihood of accidentally removing important, but not well distributed, OTUs. In order to strike a balance between addressing spurious OTUs and retaining genuine biologic information, we have settled on 0.0001 (0.01% relative abundance; horizontal dashed lines in Fig. 4) as an appropriate threshold cutoff in our iTag data (16%±8% of sequences are removed with this cutoff; Fig. 2). This somewhat liberal cutoff value is itself more conservative than a previously recommended value of 0.0005 (0.005% relative abundance; Bokulich et al. 2013). Of course, this value may be varied by each user during bioinformatic processing, and depending on experimental details a value between 0.0001 and 0.001 may be appropriate.

4.3 iTag SEQUENCING PRECISION

By applying the above set of processing steps to our mock community data over multiple sequencing runs, we have been able over time to generate sequencing data of the same samples repeatedly. This enables comparison across runs to estimate precision of iTag sequencing as applied in the Orphan lab. These results do not represent the precision of any individual step (e.g.

PCR, or MiSeq sequencing), but are rather a cumulative result of the entire workflow from template amplification to sequencing to data processing.

Precision is overall quite good, with little variation between taxa, mock communities, or amplification enzyme across runs (Fig. 5). For plasmid mock communities amplified with the 5-PRIME or NEB Q5 enzyme, precision is 0.77% or 0.84%, respectively. For gDNA mock communities amplified with the 5-PRIME or NEB Q5 enzyme, precision is 1.85% or 0.88%, respectively. Since environmental samples are amplified from natural mixes of gDNA, rather than mixed plasmids, the gDNA mock communities are likely more relevant for determining precision of iTag sequencing data. Depending on the enzyme used to amplify environmental samples, then, a precision of between 1% and 2% is conservative and reasonable.

Within the gDNA mock communities, precision is worst for OTUs of Streptococcus spp. when amplified with the 5-PRIME enzyme. This is due to an oddity, that in the September 2015 iTag run, zero sequences were recovered in any of the four gDNA mock communities for Streptococcus spp. amplified with the 5-PRIME enzyme. Strangely, sequences of Streptococcus spp. were recovered at high and consistent relative abundance in all gDNA mock communities from all other iTag runs, and even in the September 2015 run of gDNA mock communities amplified with the NEB Q5 enzyme. The lack of Streptococcus spp. sequences remains perplexing, but appears to have been an isolated incident.

4.4 iTag SEQUENCING ACCURACY

In addition to the precision of iTag sequencing, it is important to consider the accuracy of sequencing results. All methods have inherent bias, including PCR which can yield variable results depending on the primers employed, the polymerase enzyme used, the annealing temperature, the gDNA template concentration, and other factors. It is obviously critical to employ, as much as possible, the exact some conditions within a given study. However, the effects

of PCR bias become especially concerning when attempting to compare results across multiple microbial ecology studies. Some community efforts have attempted to address this. For example, the Earth Microbiome Project (EMP) advocates for one set of universal 16S rRNA gene primers to be employed in all surveys of environmental microbiology (Gilbert et al. 2011). However, these results are hampered by several factors. First, the EMP primers are known to have certain taxonomic biases (Parada et al. 2015; Trembath-Reichert et al. 2016). Second, as next-generation sequencing technology rapidly develops, it becomes necessary (and beneficial) to develop new primers targeting larger segments of the 16S rRNA gene, thereby compromising efforts to apply consistent primers across datasets and across time (the EMP has partially addressed this by recommending changes to their primer sets over time). Third, the EMP primers are theoretically universal in their coverage but in some studies only specific microbial community members are of experimental interest (e.g., Archaea but not Bacteria). In that case, the user must decide whether to employ domain-specific primers (at the cost of comparability to other studies) or employ universal EMP primers (at the cost of throwing away a large proportion of the expensive dataset). Ultimately the Orphan lab, given the diverse range of environments studied (e.g. marine methane seeps and deep terrestrial boreholes) has moved forward with applying the EMP primers with the hope of generating comparable data across studies both within the lab and outside of the lab.

With our plasmid mock communities, whose composition is very well characterized, we are able to evaluate the accuracy of iTag sequencing on taxa which are particularly relevant in marine methane seep ecosystems (Fig. 6). We find some taxa are faithfully recovered in sequencing data regardless of polymerase enzyme employed (e.g., Desulfococcus; Fig. 6a-b), although the NEB Q5 enzyme better represents more taxa than 5-PRIME (e.g., ANME-1b, Desulfobulbus, Sulfurovum; Fig. 6a-b). Of particular note are the biases for and against certain taxa. Sequencing results from both enzymes indicate underrepresentation of some SEEP-SRB1 bacteria as well as ANME-2a, -2b, and -2c (Fig. 6a-b). The 5-PRIME samples are especially biased against ANME-2c, an important taxon on marine methane seep settings, yielding ANME-

2c sequences at ~25% of the expected relative abundance (or, a 4x bias). The NEB Q5 enzyme does moderately better at representing ANME-2c, although still underrepresents ANME-2c by half (a 2x bias). The biases against ANME-2 archaea have been explored in a recent publication, and do not appear to be related to primer mismatch (Trembath-Reichert et al. 2016). Correspondingly, plasmid mock communities amplified with the 5-PRIME and NEB Q5 enzymes also overrepresent some taxa (this is not surprising, given that relative abundance is a zero-sum metric – if some taxa are underrepresented, some must by definition be overrepresented). Within our plasmid mock communities, the ANME-1 archaea are consistently overrepresented by a factor of ~2 (Fig. 6a-b).

One robust method for countering bias, common in macrofaunal ecological literature (e.g., Levin et al. 2015) but only recently applied to microbial datasets, is to apply down-weighting transformations to the relative abundance data. A commonly applied function is to take the square root of the relative abundance data. Because the square root function is non-linear with increasing values, taxa at low relative abundance will be relatively unimpacted by the square root function while taxa at high relative abundance will be more greatly impacted. This acts to "smear" the relative abundance data, reducing bias effects. Indeed, when applying the square root transformation to our plasmid mock communities, the representation is overall improved as compared to non-transformed data (Fig. 6c-d). In fact, a transformation can be applied as strongly as desired. A 4^{th} root transformation of our plasmid mock community data results in even better representation of the data (Fig. 6e-f).

The most severe transformation would be to simply count the presence/absence of OTUs in the dataset, forgoing relative abundance completely. In such a transformation, all the data in our plasmid mock communities would be "perfectly" represented. However, such a severe transformation sacrifices genuine differences in relative abundance between microbial community members, potentially curtailing ecological interpretations which might otherwise contribute to a study. Ultimately the square root transformation is a good compromise between addressing bias

while retaining relative abundance information. Of course, it is not necessary to limit a dataset to one single analysis. If broad inter-sample trends are consistent across multiple transformations, it strengthens the conclusions to be able to note that the data is consistent across data treatments (Pasulka et al. 2015). Moreover, if the inter-sample trends do differ upon variable transformation of the data, it does not imply that any particular transformation is "right" or "wrong", *per se*. Although care must be taken with regard to biases, if the user is confident that the dataset is inherently unbiased, then transformations can be a powerful tool to examine the contribution of high-relative-abundance or low-relative-abundance community members to inter-sample trends. Untransformed data will naturally be dominated by the high-relative-abundance community members, and thus conclusions drawn from untransformed data will represent those members. If the scientific question at hand is with regard to the rare or low-relative-abundance members of a biological community, then a moderate or severe transformation will draw out the effect of those species upon data interpretation.

4.5 TESTING PREPARATION METHODS ON ENVIRONMETANL SAMPLES

The above sections of this chapter have focused on developing a data processing pipeline for iTag data, with an emphasis on what could be learned from negative controls and mock communities in terms of data quality, precision, and accuracy. Having used those samples to inform a robust, high-quality processing workflow, it is possible to now examine data from environmental samples (i.e., complex microbial communities) as a function of various methods tests applied during the initial phase of iTag preparation: PCR. Below, I summarize results from five tests:

1) Performing PCR in one step vs two steps.
2) Pooling singlet, duplicate, or triplicate PCR products.
3) Template concentration across three orders of magnitude.
4) 5-PRIME vs NEB Q5 enzyme amplification.
5) Annealing temperature for the NEB Q5 enzyme.

4.5.1 1-STEP vs 2-STEP PCR

The Orphan lab employs three deviations from the standard EMP preparation protocol. Firstly, the original EMP protocol calls for only one PCR reaction to be employed per sample. During this reaction, which is recommended for 35 PCR cycles, primers are employed which contain, in addition to the primer itself (19 bp-long for the 515f primer, 20 bp-long for the 806r primer), a 24- to 29 bp-long Illumina adapter, a 10 bp-long primer pad, a 2 bp-long primer linker, and, for the 806r primer, a 12 bp-long barcode sequence:

```
515f_EMP: AATGATACGGCGACCACCGAGATCTACAC              TATGGTAATT GT GTGCCAGCMGCCGCGGTAA
806r_EMP: CAAGCAGAAGACGGCATACGAGAT      XXXXXXXXXXXX AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT
          |----------adapter----------| |-barcode--| |pad--linker| |------primer------|
```

Thus, the 515f primer is 60 bp in length and the 806r primer is 68 bp in length. Due to evidence that employing long primers over many cycles may lead to enhanced PCR bias (Berry et al. 2011), we decided to test a modification to the EMP protocol: rather than one amplification for 35 cycles with long primers, to instead perform first a 30 cycle amplification with only the raw primers ("PCR#1"):

```
515f: GTGCCAGCMGCCGCGGTAA
806r: GGACTACHVGGGTWTCTAAT
```

Followed by a second PCR for 5 cycles in which the full EMP primers are used in order to attach the pads, linkers, barcodes, and adapters ("PCR#2"). This test was performed on sediment #2687, from Eel River Basin, a marine methane seep offshore California. Besides this one protocol modification, the sample was treated identically in all other respects of the sample preparation and data processing workflow.

Results from this test indicated some differences in the recovered 16S rRNA gene profile (Fig. 7). Although the majority of OTUs were recovered in both the 1-step and 2-step treatments (the small number of OTUs not shared between the two treatments were very minor (<0.2% relative abundance) constituents of the gene profiles), the 2-Step PCR had decreased relative abundance of major OTUs and increased relative abundance of minor OTUs (Fig. 7b,d). Thus, a cross plot of OTU relative abundance between the two treatments resulted in a slope of just 0.69. It is hypothesized that 1-step PCR using long barcoded primers causes PCR bias because the overhanging barcode and/or adapter interact with template gDNA in a manner which varies based on template oligonucleotide sequences (Berry et al. 2011). Our data suggests that performing a 1-step PCR procedure decreases evenness across resulting sequence data as compared to a 2-step procedure. Our lab has moved forward with the 2-step approach, minimizing the chances for biased amplification of 16S rRNA genes from our environmental samples.

4.5.2 POOLING SINGLET, DUPLICATE, OR TRIPLICATE PCR PRODUCTS

The Orphan lab's second modification from the EMP protocol involves the pooling of replicated PCR products. The original EMP protocol calls for performing triplicate PCR reactions, which are all pooled before sequencing. Ostensibly, this is to buffer out the possible effect of PCR bias occurring in any single reaction series. However, performing triplicate PCR reactions is time consuming and costly, and so we tested the effect on sequencing data of performing just a single a PCR reaction, pooling duplicates, or pooling triplicate products. This test was performed on four marine methane seep samples representing a variety of environmental substrates: #2687 (sediment; the same sample on which we tested 1-step vs 2-step PCR), #5036 (a wood block colonized on the seafloor), #5193 (a carbonate), and #5472 (bottom water).

Results from all four samples were similar, indicating very little effect of performing and pooling one, two, or three PCR reactions (Fig. 8-11). In all cases the majority of OTUs, including all OTUs in all samples greater than 0.3% relative abundance, were recovered in all three treatments (Fig. 8-11, sub-plots a). Not only were the OTUs consistently recovered, they were represented in the sequencing data at very similar relative abundances regardless of pooling treatment. Cross plots of relative abundances therefore exhibited slopes very close to 1.0 with R^2 values also near 1.0 (Fig. 8-11, sub-plots c). Furthermore, there was no measurable effect on evenness; the major and minor OTUs were represented nearly equally across all pooling treatments (Fig. 8-11, sub-plot d). Not surprisingly, then, a non-metric multidimensional scaling analysis of the 12 sequencing profiles (4 samples x 3 pooling treatments each) revealed the samples to be well differentiated by substrate and not well differentiated by pooling treatment (Fig. 12). While these results suggest single PCR reactions would be sufficient to represent the data, our lab has taken the conservative approach of pooling duplicate PCR preparations for each sample; this saves time and cost from the originally recommended triplicate PCR reactions, while still minimizing the chances of spurious bias introduced in any single PCR reaction.

4.5.3 TEMPLATE CONCENTRATION ACROSS THREE ORDERS OF MAGNITUDE

As described above, PCR reactions can be compromised if gDNA template is low enough to be on par with the concentration of contaminant genetic material present in the PCR reactants or enzymes. In addition, it is possible for template concentration itself to affect the amplicon profile of a PCR reaction, likely due to an effect on the efficiency of each PCR cycle (Chandler et al. 1997). In order to test this affect, we applied our iTag pipeline to one sample (#5122, a carbonate) in which the gDNA template was undiluted ("1x"), diluted 10-fold ("10x"), and diluted 100-fold ("100x"). Besides these variations in template concentration, all other aspects of PCR preparation, sequencing, and data analysis were identical for the three samples.

It appeared from our data that template dilution had a measureable effect on the resulting sequence profiles (Fig. 13). Once again the majority of OTUs were shared across all treatments (Fig. 13a), but the relative abundances of the constituent OTUs was different depending on the dilution factor. The most dilute (100x) sample displayed the most even profile, with major OTUs exhibiting lower relative abundance and minor OTUs exhibiting higher relative abundance as compared to the 10x and 1x treatments (Fig. 13b,d). Despite these differences, a cross plot of relative abundance of all OTUs suggests only a moderately magnitude of effect on the sequence profile, with slope of 1.2 and R^2=0.92. Thus, while the effect of gDNA template concentration is measurable, it does not seem to impact the recovered sequence profile too adversely. While it is certainly preferable to perform PCR reactions with an equimolar amount of gDNA template in each reaction, it also appears that in cases where this is impracticable the resulting sequencing data may still be cross-compared between samples. An example of such a scenario includes when template concentrations are so low as to be immeasurable by even high-sensitivity fluorescence assays (Trembath-Reichert et al. 2016).

4.5.4 5-PRIME vs NEB Q5 enzyme amplification

The specific polymerase enzyme employed during a PCR reaction can have a dramatic effect on the resulting amplicon profile (Brandariz-Fontes et al. 2015). High-fidelity polymerase enzymes are less prone to accidentally synthesize an incorrect base during a PCR cycle, an error which can be propogated to high concentration over the course of many cycles in a PCR reaction. Errors introduced during PCR cycles affect the sequencing results and, depending on the extent of the error, may influence taxonomy assignments downstream.

The third and final deviation the Orphan lab has explored from the EMP protocol has been the choice of polymerase enzyme. The original protocol suggests using 5-PRIME HotMasterMix (catalog #2200410), but we have also explored use of the high-fidelity New

England Biolabs Q5 enzyme (catalog #M0491L). This enzyme is similar to the Finnzymes Phusion High Fidelity DNA Polymerase (S. Connon, personal communication from New England Biolabs), which has been independently shown to produce highly accurate PCR products in next-generation sequencing data (Brandariz-Fontes et al. 2015). In our tests, we employed the 5-PRIME HotMasterMix and NEB Q5 enzymes on five separate environmental samples in order to compare sequencing results (note, we also applied both enzymes to our plasmid mock communities and found the NEB Q5 enzyme to produce more accurate results – see above sections). These samples included four carbonates (#3622, #3624, #5104d, #5122) and one sediment (#5133).

As was the case with our plasmid mock communities, we observed a measurable difference in the resulting sequence profiles when amplified with 5-PRIME or NEB Q5 enzymes (Fig. 5,6,14). Although a large proportion of OTUs were reproducibly recovered in both treatments, the slopes and R^2 values of cross plots varied substantially from values of 1.0 (slope range: 0.78-1.31; R^2 range: 0.52-0.85). This suggests the choice of polymerase enzyme is important when preparing samples for next-generation sequencing; given that our plasmid mock communities were more accurately represented by the NEB Q5 enzyme, we made the decision to employ that enzyme moving forward from November 2014 (Orphan lab runs in November 2013 and May 2014 exclusively used the 5-PRIME enzyme and some continuing projects continue to employ the enzyme for continuity across datasets).

4.5.5 ANNEALING TEMPERATURE FOR THE NEB Q5 ENZYME

One of our observations from sequencing of plasmid mock communities was the bias against ANME-2 archaea which appeared to be improved by switching to the NEB Q5 polymerase enzyme. In order to further test if we could improve representation of ANME-2 archaea in the dataset, we applied the NEB Q5 enzyme at two different annealing temperatures

to an environmental sample (#5133, sediment) which was known to be rich in ANME-2 from previous studies (Trembath-Reichert et al. 2013). In a PCR reaction, annealing is the step at which the primer adheres to the template, after which extension (i.e., amplification) will occur. Only those genomic fragments will be amplified which successfully adhere with a primer. In general, specificity is a good quality at this step, as a user does not want PCR primers to incorrectly adhere to genomic fragments for which they were not intended. Specificity is improved by raising the temperature of the annealing reaction, requiring a better match between the primer and the template in order for successful adhesion (otherwise the primer will denature off the template). However, in the case that a user wishes to decrease specificity (e.g., to increase amplification of ANME-2, if primer adhesion were an issue), the annealing temperature can be lowered. This lowers the energetic requirements for adhesion, helping the primers adhere to genomic template that may not be a perfect nucleotide match. We tested the NEB Q5 enzyme on sample #5133 at annealing temperatures of 50˚C and 54˚C, in addition to 5-PRIME enzyme amplification at 50˚C. We note that the NEB Q5 enzyme is optimized to operate at 57-64˚C, but we employed lower temperatures in order to decrease specificity.

As hypothesized, the NEB Q5 enzyme at 50˚C did provide better representation of ANME-2 archaea than the NEB Q5 enzyme at 54˚C (and much better than the 5-PRIME enzyme; Fig. 15). However, overall the differences in NEB Q5 annealing at 50˚C or 54˚C were relatively small (Fig. 15a,d), and we chose to employ the NEB Q5 enzyme at an annealing temperature of 54˚C as it strikes a balance between the optimized temperature for the enzyme (57-64˚C) and a cooler temperature in order to increase adhesion with a wide range of genomic templates.

4.5.6 SUMMARY OF PCR TEST RESULTS

We observed that variations in the sample preparation procedure induced a range of differences in the resulting sequence profiles for environmental samples. Some variations

produced relatively large differences (e.g., 5-PRIME vs NEB Q5 polymerase enzyme) while others made little difference (e.g., pooling one, two, or three replicate PCR reactions). In order to visualize these differences at a broad level, we included all preparation tests in a non-metric multidimensional scaling ordination with a variety of other environmental samples from marine methane seep worldwide (Fig. 16). These results showed that although intra-sample differences were observed as a function of preparation method, the differences were uniformly small as compared to inter-sample difference in 16S rRNA gene profiles. Therefore, we are confident that the ecological interpretations made from iTag datasets are not compromised by small differences in sample preparation protocols. Of course, we recommend run-to-run consistency and following the best possible protocol at all times (e.g., more replicates is better than fewer, and high-fidelity polymerase enzymes are better than lower-fidelity ones), but we find these effects to be secondary to genuine microbial community differences as a function of environmental factors such as habitat substrate.

4.6 PRIMER ON QUANTITATIVE ECOLOGICAL TOOLS EMPLOYED IN DOWNSTREAM DATA INTERPRETATION

The field of ecology employs a number of mathematical and statistical tools in order to probe the relationships between biological communities and the environment. Here I will briefly describe the principle techniques used in this thesis, acknowledging that the full array of methods available to ecologists would (and does) encompass entire textbooks (e.g., Clarke and Warwick 2001; Legendre and Legendre 2012). When parsing ecological datasets, the investigator's goals are often divided into two aims: 1) to visualize inter-sample relationships by translating biological similarity into spatial distance, producing a graphic, or ordination, which enables hypothesis development, and 2) to apply statistical metrics to assess both whether apparent inter-sample or inter-group differences are robust and, if so, which biological community members contribute to

inter-sample or inter-group differences. In this thesis, the primary ordination method employed is Nonmetric Multi-Dimensional Scaling (NMDS). Inter-group differences are probed with the Analysis of Similarity (ANOSIM) statistical test, and biological community members contributing to inter-group differences are identified by the Similarity Percentage (SIMPER) routine.

NMDS is a more appropriate ordination technique for biological datasets, and in particular large next-generation sequencing datasets, than other familiar techniques such as Principal Components Analysis (PCA) or Principal Coordinates Analysis (PCoA). PCA, perhaps the most widely familiar ordination technique, plots sample data from *m* species onto *m* axes (aka, components), and then projects the resulting multidimensional field onto a lower dimensional graph. As a simple analogy, imagine holding up a ball-and-stick crystal lattice structure, then shining a flashlight on it and viewing the resulting projection against the wall. In this example, a 3-dimensional structure is collapsed to a 2-dimensional structure. In next-generation sequencing datasets, however, it is common to have hundreds or thousands of taxa represented. Genuine inter-sample relationships therefore exist in a multidimensional space with hundreds or thousands of axes, and collapsing the ordination down to two or three dimensions has the potential to drastically distort inter-sample relationships. Furthermore, PCA is highly sensitive to "joint absences" – zero values in a species abundance matrix. Imagine, for example, collecting species data from a forest, a desert, and a beach. Several dolphins are counted at the beach site, but zero dolphins are counted in the forest or in the desert. Intuitively, an ecologist would not claim that because the forest and the desert lack dolphins, they are more similar to one another. (Similarity is better defined as *joint presence* than *joint absence*. For example the forest and the desert are more similar because they both contain rabbits, not because they both lack dolphins.) PCoA is better at addressing joint absences than PCA, but suffers from the same issue in collapsing multidimensional ordination into lower dimensions by projection.

The algorithm supporting NDMS is fundamentally different than PCA or PCoA. First introduced in the 1960s in the field of Psychology (Shepard 1962; Kruskal 1964a; b), NMDS

avoids the constraint of attempting to project Euclidian distance onto lower dimensions (the crystal lattice and flashlight example above). Instead, NMDS is an iterative algorithm in which the number of dimensions is set *a priori* (usually 2 or 3) and the locations of samples in the ordination space are repeatedly adjusted in order to maximize the accuracy of the depicted inter-sample relationships (functionally, this involves minimizing a stress function that measures how accurately inter-sample relationships are represented). Importantly, because NMDS attempts to preserve the rank ordering of sample similarities, rather than Euclidian distance, the graphical depiction of inter-sample differences will never be perfect, but in practice the aggregate representation of the samples' data is often more accurate than in a PCA or PCoA plot. Furthermore, NMDS is not constrained by the problem of joint absences as was PCA. This is because the NDMS algorithm does not work on a raw sample-by-species table of abundance (or relative abundance), but rather on a triangular matrix of inter-sample similarity (this is the same workaround that PCoA uses to avoid the joint absence problem, although PCoA still attempts to project high-dimensional Euclidian distance onto lower dimensional space).

A common inter-sample similarity metric applied, and the one employed in this thesis, is Bray-Curtis similarity (S_{jk}), which is immune to joint absences and which results in intuitive values of 0 (if two samples share no species) and 100 (if two samples share all the same species at the same abundances or relative abundances; Bray and Curtis 1957; Oksanen et al. 2013). Bray-Curtis dissimilarity is also frequently employed in ecological approaches, and is simply equivalent to 100-S_{jk}, denoted as δ_{jk}:

$$S_{jk} = 100 \bullet (1 - (\Sigma^{p}_{i=1} \, |y_{ij} - y_{ik}|) \, / \, (\Sigma^{p}_{i=1} \, (y_{ij} + y_{ik}))) \qquad \text{Eq. 1}$$

$$\delta_{jk} = 100 \bullet ((\Sigma^{p}_{i=1} \, |y_{ij} - y_{ik}|) \, / \, (\Sigma^{p}_{i=1} \, (y_{ij} + y_{ik}))) \qquad \text{Eq. 2}$$

where p is the number of the species observed in the dataset and y is the relative abundance of species i in sample k or j. For a sample set with n samples, the number of sample-

sample combinations (i.e., the number of calculated S_{jk} values) will be equal to $n(n\text{-}1)/2$. As an example, take the synthetic relative abundance data described in Table 3. The Bray-Curtis similarities for the five samples are calculated and presented in Table 4. As expected from the raw relative abundance data, Samples 1 and 2 are highly similar. The other samples demonstrate a range of similarity values from 44 to 75. A 2-dimensional NMDS ordination would begin by placing five points on a plane (representing the five samples), and checking how close on the plane each sample is to each other sample. Sample 1 and 2 ought to be closer to each other than to any other samples, for example, and Sample 4 ought to be closest to samples 5, 2, 1, and 3 in that order. In other words, in a perfect ordination the distance between any two samples in low-dimensional space would increase monotonically with dissimilarity (defined as $100\text{-}S_{jk}$, or δ_{jk}). A plot of distance between two points in ordination space vs distance predicted from a monotonic regression line is a *Shepard Diagram* (Fig. 17B,D). The *stress* of an NMDS plot is defined as the extent to which ordination distance does not increase monotonically with $100\text{-}S_{jk}$:

$$\text{Stress} = \Sigma(d_{jk} - \hat{d}_{jk})^2 \,/\, \Sigma(d^2_{jk})` \qquad \text{Eq. 3}$$

where d is the actual distance between samples j and k on the ordination plot, and $\hat{d}$ is the predicted distance between samples j and k from the monotonic regression line (Kruskal 1964b; Clarke and Warwick 2001; Oksanen et al. 2013; Fig. 17). In the second iteration of the NMDS algorithm, the five data points will be relocated in order to minimize the stress value. Once the stress value is minimized to a pre-set threshold, or once the stress value no longer decreases, the ordination algorithm terminates and the resulting ordination is saved. NMDS, by virtue of handling Bray-Curtis similarity matrices rather than raw data (and therefore being immune to joint absences), and by applying a stress minimization function to achieve low-dimensional ordination rather than attempting to project high dimensional space onto lower dimensions, is considered a robust ordination technique for representing inter-sample relationships in biological

(including microbiological) datasets (Clarke and Warwick 2001; Ramette 2007). However, NMDS is not more than a visualization technique in order to generate hypotheses.

It is conceivable that once a study's samples are ordinated, some apparent trends will emerge. For example, in Fig. 17C it is apparent that the 16S rRNA gene profiles of carbonate and non-carbonate habitat substrates may differ. In order to test this hypothesis, the ANOSIM test is applied. This tests whether the inter-sample similarities are higher among samples of a defined group vs among samples of different groups. In this example, the test will be whether the 16S rRNA gene profiles between two samples of the same habitat substrate (sediment, nodule, bottom water, or carbonate) display, on average, higher Bray-Curtis similarity than the 16S rRNA gene profiles between two samples of differing habitat substrate (e.g., a sediment vs a carbonate). In order to discern this, the triangular Bray-Curtis similarity matrix is converted into a triangular rank similarity matrix; that is, every similarity is ranked with #1 being the two samples which have the highest Bray-Curtis similarity and #$n(n\text{-}1)/2$ being the two samples which have the lowest Bray-Curtis similarity. An example of this conversion is provided in Table 5. Then, the calculation of the ANOSIM test is straightforward. A metric, R, is calculated as described in Clarke and Warwick 2001 and Oksanen et al. 2013:

$$R = (r_{between} - r_{within}) \,/\, (n(n\text{-}1)/4) \qquad \text{Eq. 4}$$

Where r_{within} is the average rank similarity of all like-type sample-sample combinations (e.g., sediment-sediment or carbonate-carbonate) and $r_{between}$ is the average rank similarity of all sample-sample combinations which are of differing type (e.g., sediment-carbonate). As before, n is the number of samples in the study. If within-group similarity is higher than between-group similarity, an R value greater than 0 is calculated (R=1 would indicate every possible sample-sample pair within groups is more similar to every possible sample-sample pair between groups). By recalculating R many hundreds of times using randomly rearranged sample assignments to be

within or between groups, a significance value, *p*, can be calculated. This tests whether observed inter-group differences are likely to be encountered by chance. In the example from Fig. 17C, R and p values of 0.49 and <0.001, respectively, confirm that the 16S rRNA gene profiles are significantly distinguished by habitat substrate; this was explored further in Chapter Two of this thesis.

Once it has been determined that two (or more) sample groups differ in their biological communities, it is logical to probe which species contribute to inter-group differences. This problem can be computationally approached using the SIMPER routine (Clarke and Warwick 2001; Oksanen et al. 2013). The SIMPER algorithm is conceptually simple, but also susceptible to yielding misleading results in cases when standard deviations of species distributions are high. For this reason, SIMPER results must be interpreted with a critical eye. The safest use of SIMPER is to use the routine for identifying *possible* species contributing to inter-group differences, and then to always return to the raw data to confirm the trends. Depending on the hypotheses being tested, in some cases it is possible to discern the most important species in a dataset simply by examining the raw species-sample abundance table, and foregoing the SIMPER routine altogether.

SIMPER works by deconstructing Bray-Curtis dissimilarity (recall dissimilarity is defined as δ_{jk}, equal to $100\text{-}S_{jk}$) into contributions from each individual species. Recall from Eq. 2 that for each pair of samples *j* and *k*, each species *i* represents one term in the numerator summation. Therefore, in order to calculate each species' individual contribution to the Bray-Curtis dissimilarity between a sample-sample pair, the dissimilarity equation is applied without the numerator summation (Clarke and Warwick 2001; Oksanen et al. 2013):

$$\delta_{jk}(i) = 100 \bullet ((|y_{ij} - y_{ik}|) / (\Sigma^{p}_{i=1} (y_{ij} + y_{ik}))) \qquad \text{Eq. 5}$$

This calculation is applied to every species *i* for every pair of *j* and *k*, where *j* and *k* are samples between the two groups being tested (e.g., sediments *vs* carbonates). For each species, the resulting values across all sample pairs of $\delta_{jk}(i)$ are averaged, resulting in an average contribution of species *i* to inter-group differences in the dataset. The susceptibility of SIMPER to misleading results comes from the fact that over all possible combinations of *j* and *k*, there is sometimes very wide standard deviation among the calculated values of $\delta_{jk}(i)$. For this reason, in addition to an average contribution of each species *i* to inter-group dissimilarity, the standard deviation of the contribution is also always reported. This standard deviation is critically important for assessing the significance, or consistency, of species *i* contributing to inter-group differences. It is for this reason that I also strongly recommend using SIMPER only as a tool to identify *possible* species contributing to inter-group differences, and then always following up by double checking the raw species abundance (or relative abundance) data to confirm the inter-group differences are robust. Similar caution is recommended in Clarke and Warwick 2001.

A BRIEF SUMMARY OF HIGHLIGHTED QUANTITATIVE ECOLOGICAL TOOLS

An NMDS plot is a low dimension graph (in this thesis, always 2 dimensions) in which each point represents the entire microbial community profile (all species observed, including their relative abundances). Distances between points indicate sample-sample similarity, with closer points being having more similar biomarker profiles to one another. Since by nature an NMDS plot only attempts to preserve relative inter-sample differences (i.e., rank ordered differences), the units of the x and y axes (as well as orientation, rotation, and scaling) are arbitrary and therefore not generally reported. Lower stress values indicate better representation of the cumulate data, and stress values of <0.20 are generally considered sufficient for interpretation (<0.1 is ideal but often not achieved in large environmental datasets; Clarke and Warwick 2001).

The ANOSIM test evaluates whether groups of samples are statistically distinct from other groups of samples. Furthermore, the R value produced from the ANOSIM test is a measure of how strongly the defined groups differentiate the samples being interrogated. In practice R values >0.60 are rarely computed in diverse environmental 16S rRNA gene datasets presented in this thesis, but statistically significant R values in the range of 0.30 to 0.60 are not uncommon.

The SIMPER test identified species that contribute to differences in biological communities between sample groups. The SIMPER algorithm parses the Bray-Curtis similarity calculation to assess the contribution of each species, but is susceptible to wide standard deviations due to the large number of possible sample combinations. SIMPER should be carefully evaluated and always double checked against the raw abundance (or relative abundance) data.

Many more details of these tests, including variable implementations, can be found in numerous literature resources (Clarke and Warwick 2001; Legendre and Legendre 2012). This section is intended to briefly familiarize the reader with some of the types of ecological techniques which were frequently employed in this thesis, especially in Chapters One through Three. With the exception of Chapter One, all ecological calculations in this thesis were performed with the 'vegan' (v2.0.10) package of the R environment (Oksanen et al. 2013; R Core Team 2014), with frequent conceptual clarity gleaned from the methods manual produced in conjunction with the Primer-E software package (Clarke and Warwick 2001). Analyses in Chapter One were performed with the Primer-E software package (Clarke and Warwick 2001).

Table 1. Relative abundances of taxa included in the plasmid mock communities. Mock communities were generated by mixing quantified amounts of 16S rRNA plasmids clone libraries in lab. These mock communities were used to assess both precision and accuracy of iTag sequencing.

Phylogeny	Comm. 1	Comm. 2	Comm. 3	Comm. 4
Archaea/Euryarchaeota/Methanomicrobia/ANME-1/ANME-1a	0.01	0.04	0.05	0.20
Archaea/Euryarchaeota/Methanomicrobia/ANME-1/ANME-1b	0.01	0.04	0.05	0.20
Archaea/Euryarchaeota/Methanomicrobia/Methanosarcinales/ANME-2a-2b/ANME-2b	0.04	0.01	0.15	0.05
Archaea/Euryarchaeota/Methanomicrobia/Methanosarcinales/ANME-2c	0.04	0.01	0.15	0.05
Archaea/Euryarchaeota/Methanomicrobia/Methanosarcinales/ANME-2a-2b/ANME-2a	0.03	0.01	0.15	0.05
Bacteria/Proteobacteria/Deltaproteobacteria/Desulfobacterales/Desulfobulbaceae/Desulfobulbus	0.03	0.03	0.11	0.11
Bacteria/Proteobacteria/Epsilonproteobacteria/Campylobacterales/Helicobacteraceae/Sulfurovum	0.25	0.25	0.01	0.01
Bacteria/Proteobacteria/Deltaproteobacteria/Desulfobacterales/Desulfobacteraceae/Desulfococcus	0.03	0.03	0.11	0.11
Bacteria/Proteobacteria/Deltaproteobacteria/Desulfobacterales/Desulfobacteraceae/SEEP-SRB1	0.03	0.03	0.11	0.10
Archaea/Thaumarchaeota/Miscellaneous Crenarchaeotic Group	0.10	0.10	0.01	0.01
Archaea/Euryarchaeota/Thermoplasmata/Thermoplasmatales/Marine Benthic Group D and DHVEG-1	0.42	0.42	0.01	0.01
Bacteria/Proteobacteria/Deltaproteobacteria/Desulfobacterales/Desulfobacteraceae/SEEP-SRB1	0.03	0.03	0.09	0.10

Table 2. Relative abundances of taxa included in the genomic mock communities. Mock communities were generated by mixing known amounts of genomic extract (gDNA) from cultures grown in lab. Compositions are corrected to include copy number of the 16S rRNA gene for each organism.

Phylogeny	Comm. 1	Comm. 2	Comm. 3	Comm. 4
Bacteria/Firmicutes/Clostridia/Clostridiales/Eubacteriaceae/Acetobacterium/Woodii	0.18	0.13	0.08	0.02
Bacteria/Bacteroidetes/Bacteroidia/Bacteroidales/Bacteroidaceae/Bacteroides/Coprosuis	0.11	0.08	0.05	0.03
Bacteria/Firmicutes/Bacilli/Lactobacillales/Streptococcaceae/Streptococcus/spp.	0.39	0.29	0.17	0.05
Bacteria/Proteobacteria/Deltaproteobacteria/Desulfobacterales/Desulfobacteraceae/Desulfococcus/Multivorans	0.04	0.05	0.04	0.02
Bacteria/Proteobacteria/Deltaproteobacteria/Desulfobacterales/Desulfobulbaceae/Desulfobulbus/Propionicus	0.07	0.14	0.22	0.13
Archaea/Euryarchaeota/Methanomicrobia/Methanosarcinales/Methanosarcinaceae/Methanosarcina/Acetivorans	0.11	0.21	0.39	0.61
Archaea/Euryarchaeota/Methanomicrobia/Methanosarcinales/Methanosarcinaceae/Methanolobus/Zinderi	0.04	0.03	0.02	0.11
Archaea/Euryarchaeota/Thermoplasmata/Thermoplasmatales/Thermoplasmataceae/Thermoplasma/Acidophilum	0.04	0.03	0.02	0.01
Bacteria/Proteobacteria/Gammaproteobacteria/Methylococcales/Methylococcaceae/Methyloprofundus/Sedimenti	0.04	0.03	0.02	0.02

Table 3. Synthetic relative abundance data of five species recovered from five samples.

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
Species A	0.23	0.20	0.62	0.06	0.31
Species B	0.14	0.11	0.22	0.43	0.29
Species C	0.34	0.30	0.01	0.19	0.16
Species D	0.20	0.24	0.02	0.13	0.12
Species E	0.09	0.15	0.13	0.19	0.12

Table 4. Bray-Curtis similarity values (S_{jk}) as calculated from Eq. 1 for samples 1 through 5 in Table 3.

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
Sample 1					
Sample 2	90				
Sample 3	49	47			
Sample 4	61	64	44		
Sample 5	74	71	68	75	

Table 5. Ranked similarities as determined from Table 4.

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
Sample 1					
Sample 2	#1				
Sample 3	#8	#9			
Sample 4	#7	#6	#10		
Sample 5	#3	#4	#5	#2	

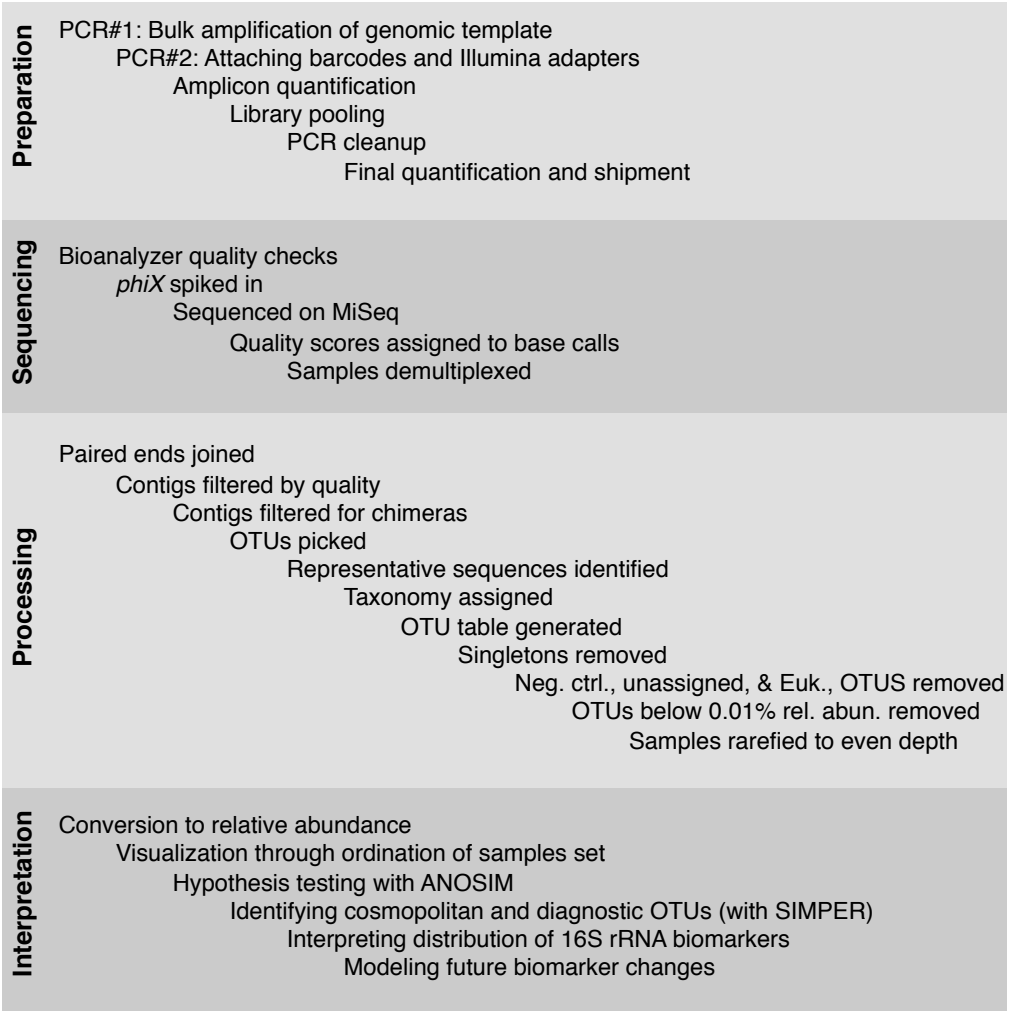

Figure 1. Flow chart of procedures from amplification of bulk genomic template to interpretation of Illumina sequencing data. The workflow is divided into four categories. Preparation, Processing, and Interpretation and performed at Caltech; Sequencing is performed at Laragen, Inc. At every step of the workflow, options are variations are available and sometimes appropriate; each step requires advanced knowledge by the user in order to make informed decisions about how to prepare, process, and interpret the data.

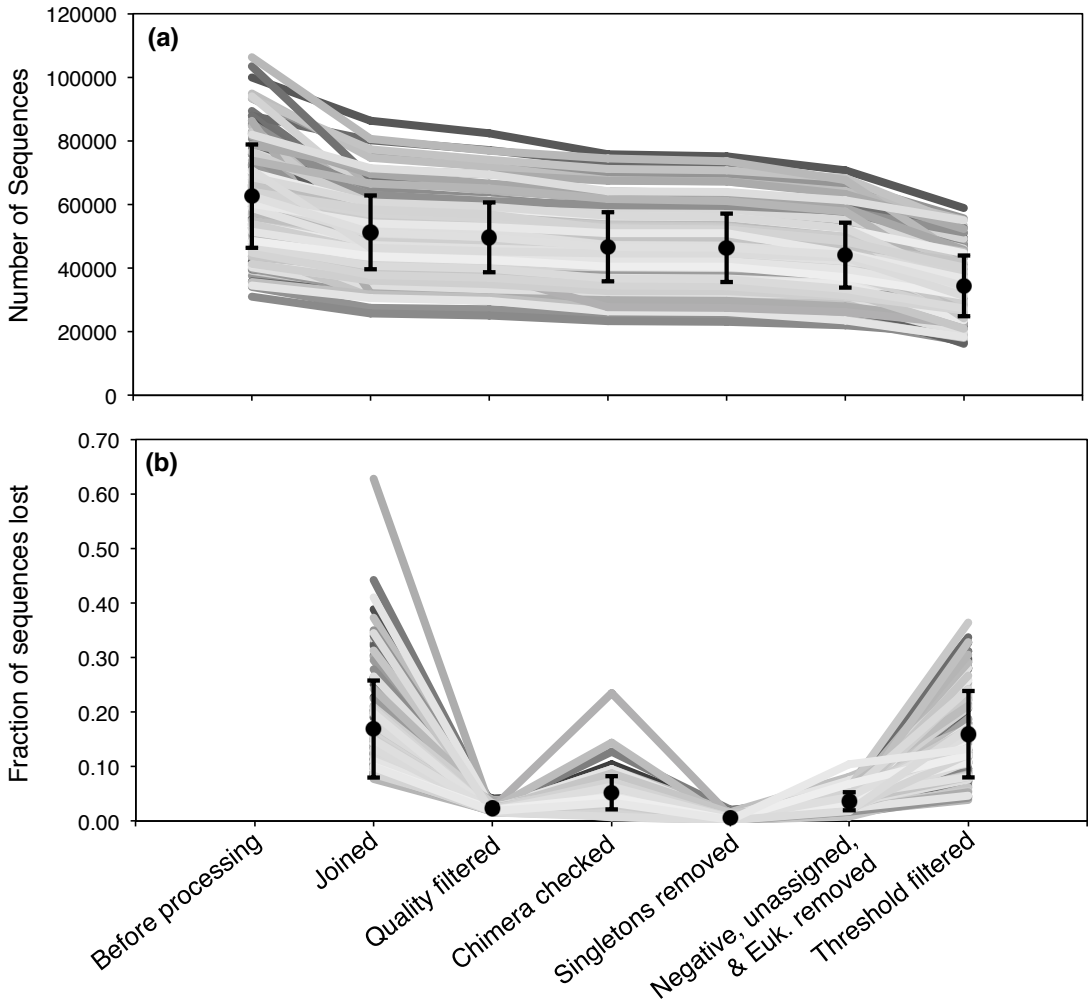

Figure 2. Detailed examination of the loss of sequences at each step of data processing. Panel (a) gives the raw number of sequences at each step, i.e. the value given for "Joined" is the number of remaining sequences after joining. Panel (b) gives the fraction of initial sequences which are lost at each processing step. In both panels, average values for each step are denoted with a black circle (plus/minus one standard deviation). Most sequences are lost at the joining step or at the threshold filtering step. Trimming, singleton removal, and filtering of negative control, unassigned, or eukaryotic sequences have a relatively minor effect on the dataset.

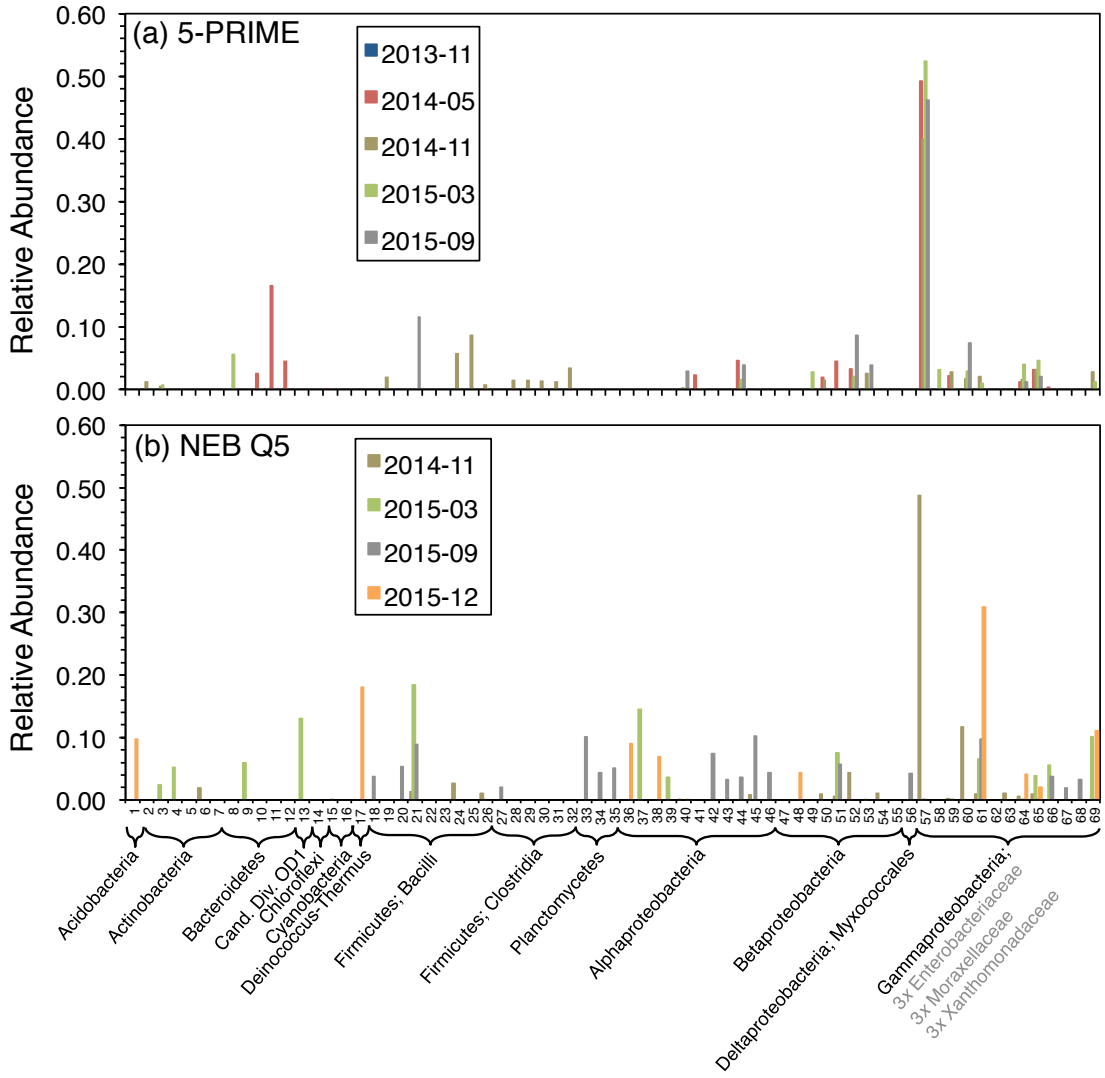

Figure 3. Relative abundance of taxa in negative controls. (a) Negative controls amplified with the 5-PRIME Hot Master Mix *Taq* (Item# 2200410). (b) Negative controls amplified with the Q5 Hot Start High-Fidelity 2X Master Mix (Item# M0494s). Negative controls are included in each sequencing batch (here sequencing batches are identified in the legends). In samples amplified with the 5-PRIME enzyme, an OTU identified generally as *Gammaprotoebacteria* is highly abundant in all negative controls. Other OTUs are also present at up to ~20% relative abundance, representing a range of taxonomies. These OTUs and their associated taxonomies vary within the negative control from each sequencing batch, emphasizing the importance of including negative controls in iTag sequencing batches. Variability is even greater in the Q5-amplified negative controls, with no particular OTU or taxa being consistently observed across all negative controls.

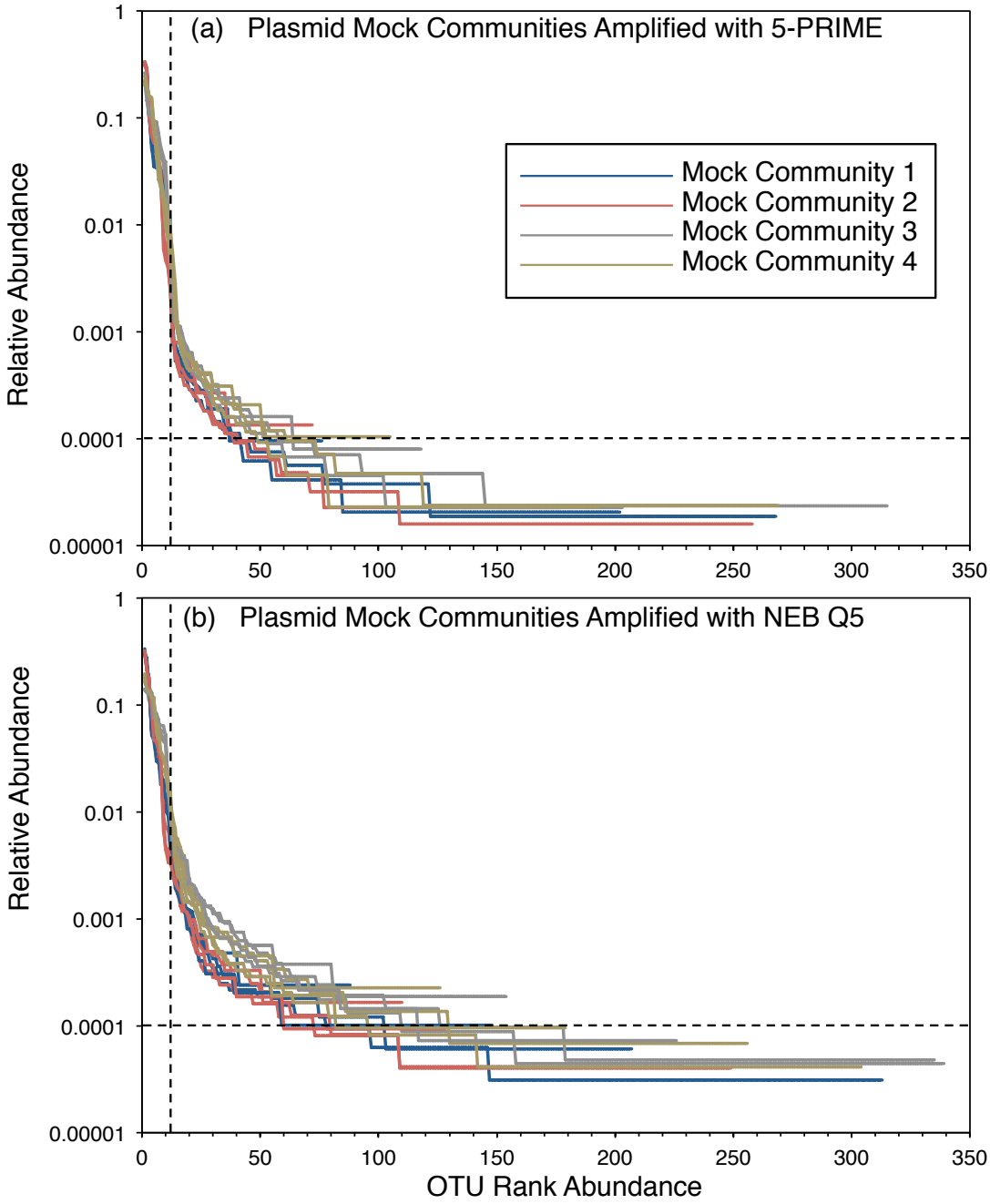

Figure 4. Relative abundance of OTUs detected in the plasmid mock communities, ordered by rank abundance of OTUs. Legend (a) applies also to (b). In both panels, the vertical dashed line indicates the expected number of OTUs, 12, which was the number of plasmids mixed into the mock communities. In both panels, the horizontal dashed line denotes a relative abundance of 0.0001 (0.01%), the threshold cutoff value applied to environmental samples. Although a threshold cutoff of 0.001 (0.1%) would have been more conservative, in practice this removed an undesirable fraction of the total dataset: environmental samples host richer OTU diversity than mock communities by several orders of magnitude, making us hesitant to remove too many OTUs which might be genuine signals from environmental samples. Furthermore, our environmental analyses exclusively focus on OTUs which are relatively highly abundant in the datasets (generally >0.1%), so our conclusions regarding microbial distribution and ecology are rarely dependent on the veracity of OTUs present at only ~0.01% (i.e., our conclusions are not impacted by the retention of OTUs present at 0.1%-0.01% relative abundance). Finally, when the threshold cutoff is applied to environmental, it is applied simultaneously to the *whole dataset*. That is, among *all sequences* from all samples in the dataset, an OTU must be present at 0.01% or greater relative abundance among *all sequences* in order to be retained. This could be achieved by a single sample genuinely hosting very many sequences of an OTU, while all other samples lack the OTU. In that case, the OTU ought to be retained since it represents real microbial presence. Setting a threshold cutoff at 0.01% relative abundance therefore minimizes our chances of accidentally removing real and relevant OTUs.

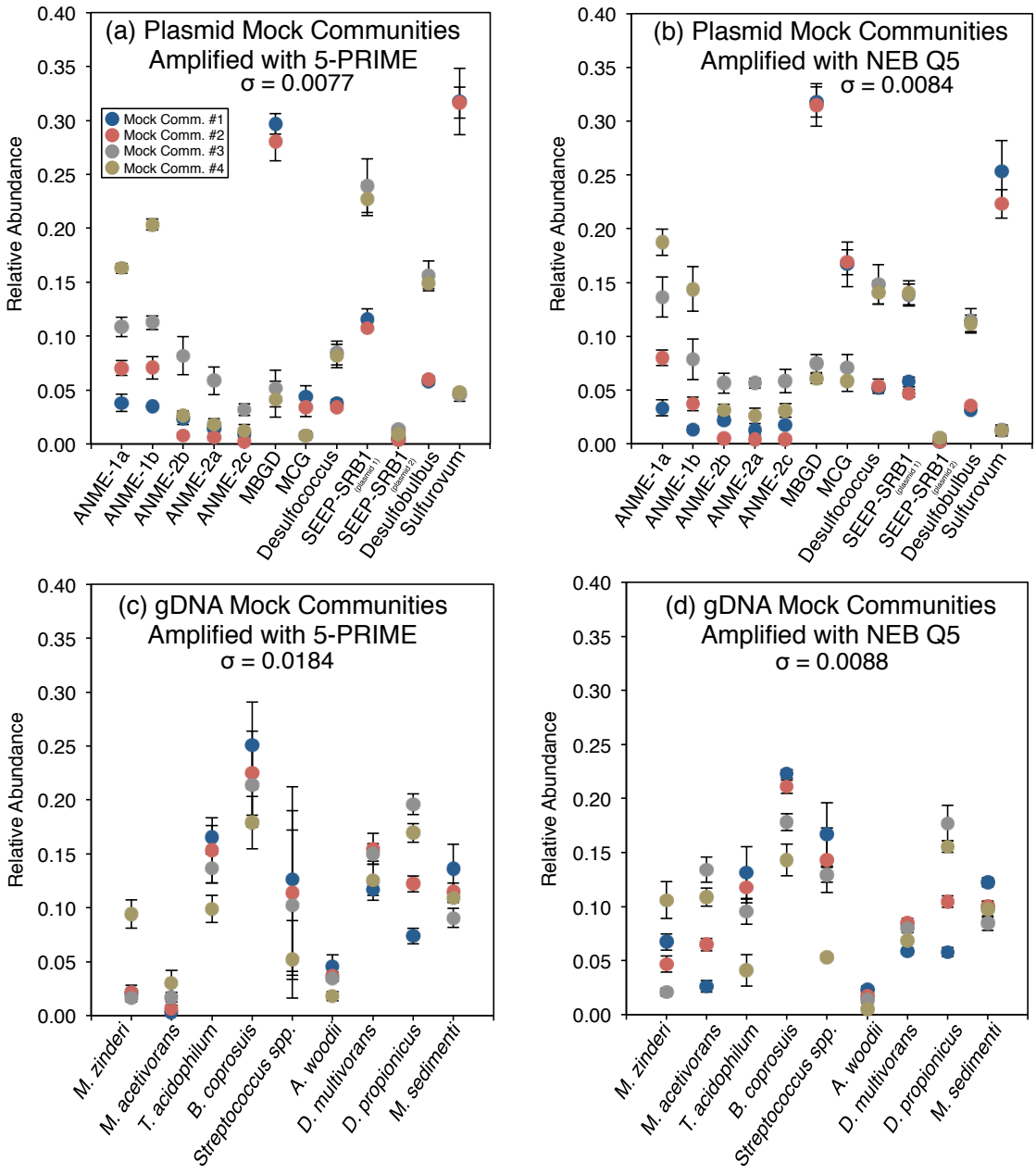

Figure 5. Reproducibility (precision) of iTag sequencing of mock communities. Data of plasmid mock communities is given in (a) and (b). Data of gDNA mock communities is given in (c) and (d). Reproducibility is better for the plasmid mock communities than the gDNA mock communities, possibly due to better-prepared template. For plasmid mock communities, 1-sigma precision is 0.77% (5-PRIME) and 0.84% (NEB Q5). For gDNA mock communities, 1-sigma precision is 1.85% (5-PRIME) and 0.88% (NEB Q5). Thus, it appears that iTag sequencing is precise to ~1-2%, depending on specific of sample preparation such as amplification enzyme. Legend in (a) applies to all panels.

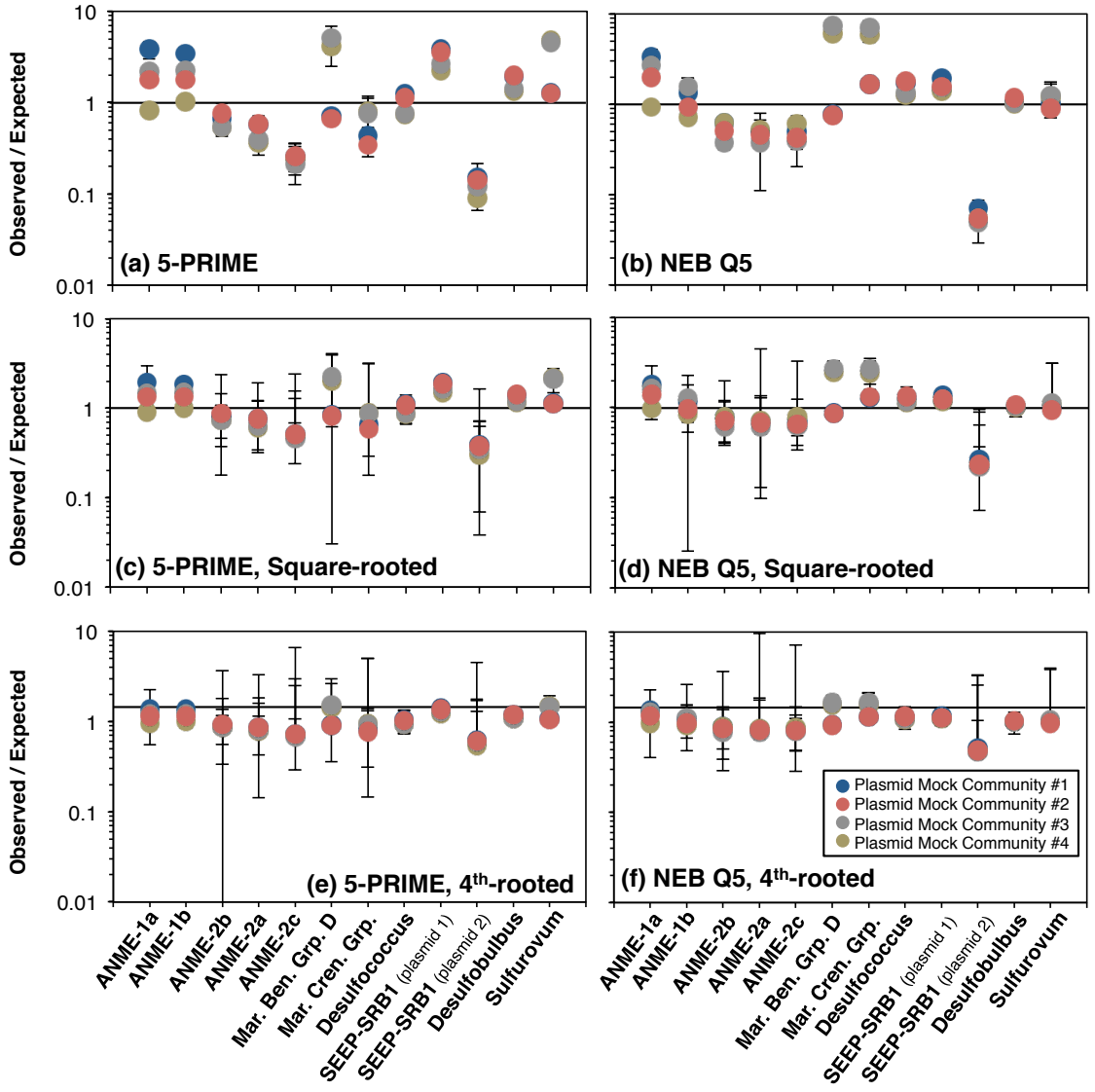

Figure 6. Accuracy of iTag sequencing of plasmid mock communities. In the raw relative abundance data (a, b), the accuracy of representation of OTUs varies widely. Some OTUs are well represented, especially after amplification with NEB Q5 (c.f. Desulfobulbus and Sulfurovum in (b)), while others are overrepresented by as much as ~8*x* (c.f. MBGD and MCG in (b)) or underrepresented by as much as ~10*x* (c.f. one SEEP-SRB1 plamid in (a) and (b)). Notably, the ANME-1 archaea are slightly overrepresented by a factor of ~2-3*x* while the ANME-2 archaea are generally underrepresented by a factor of ~2-3*x*. This is faiirly consistent between 5-PRIME and NEB Q5, with the exception that ANME-1b and ANME-2c are markedly better represented in NEB Q5 data than 5-PRIME data. In general, all OTUs are better represented when a square-root normalization is applied to the relative abundance data. The reason for this is that the square-root function mitigates PCR bias by preferentially down-weighting the OTUs which appear at high relative abundance compared to those at relative abundance. A more severe correction, such as a 4^{th} root normalization (e,f), further mitigates PCR bias, but at the cost of lost information regarding which OTUs are genuinely more or less abundant in the dataset (the most severe transformation possible is toexamine only presence/absence). Ultimately, the square-root transformation is a good compromise between addressing PCR bias while still retaining valuable information about the relative abundance of OTUs.

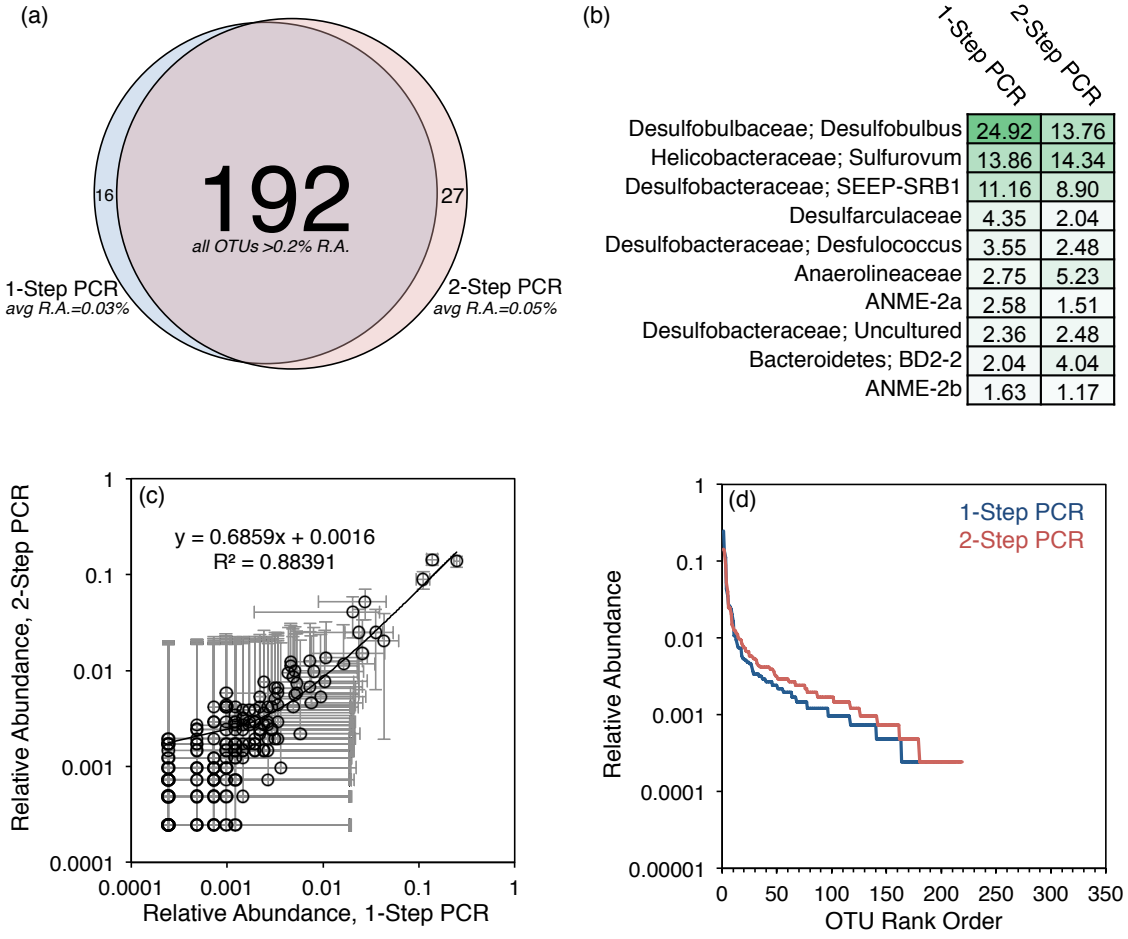

Figure 7. Comparison of amplification approach: a single PCR reaction employing long primers containing Illumina adapters and barcodes ("1-Step PCR") vs two PCR reactions, in which the first employs short primers and the second attaches the adapters and barcodes ("2-Step PCR"). This test was applied to a single methane seep sediment sample, #2687, in the November 2013 Illumina run. (a) shows the OTU overlap between the two samples. The majority of OTUs are shared, including all the major OTUs (any OTU >0.2% relative abundance in either sample is shared between the two samples). Only a small number of OTUs, with very low relative abundance, are unique to either the 1-Step or 2-Step PCR sample. (b) tabulates the relative abundances of the top 10 most abundant OTUs between the two samples, while (c) is a cross-plot of the relative abundance of the 192 OTUs shared between the two samples. Error bars in (c) are 1.85% relative abundance (c.f. precision for samples amplified with 5-PRIME *Taq*). (d) gives the OTU rank abundance curve for both samples. Overall the curves are quite similar. The fact that low-abundance OTUs (rank: ~25-250) are at slightly higher relative abundance in the 2-Step PCR suggests that rare community members are better represented by applying a 2-Step PCR preparation protocol than a 1-Step PCR preparation protocol.

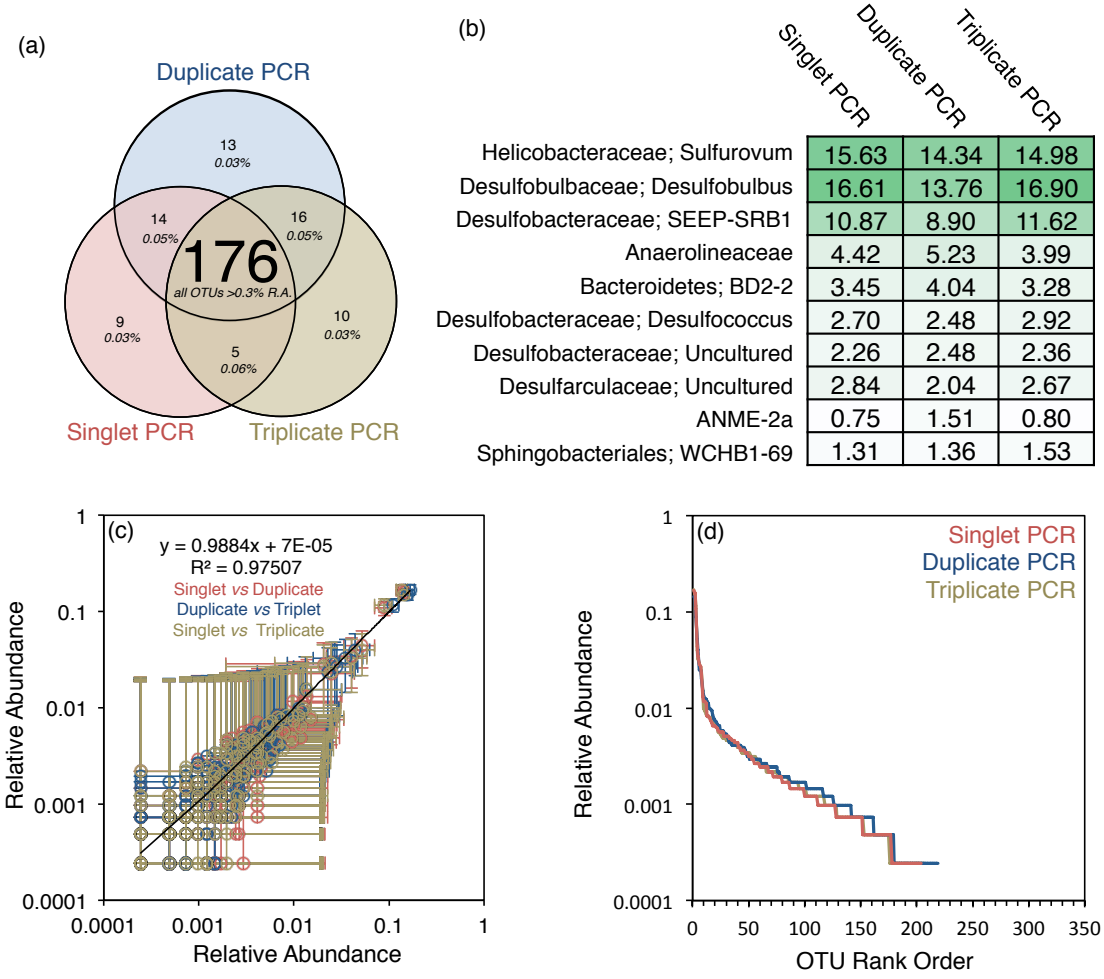

Figure 8. Comparison of amplification approach: pooling single, double, or triple PCR products during preparation of sample #2687 (sediment) for iTag sequencing. (a) shows the OTU overlap between results from the three preparations. The majority of OTUs are shared, including all the major OTUs (any OTU >0.3% relative abundance in either sample is shared between the two samples). Only a small number of OTUs, with very low relative abundance, are not shared between preparations. (b) tabulates the relative abundances of the top 10 most abundant OTUs between the preparations, while (c) is a cross-plot of the relative abundance of the OTUs shared between the three preparations. Error bars in (c) are 1.85% relative abundance (c.f. precision for samples amplified with 5-PRIME *Taq*). (d) gives the OTU rank abundance curve for all three preparations.

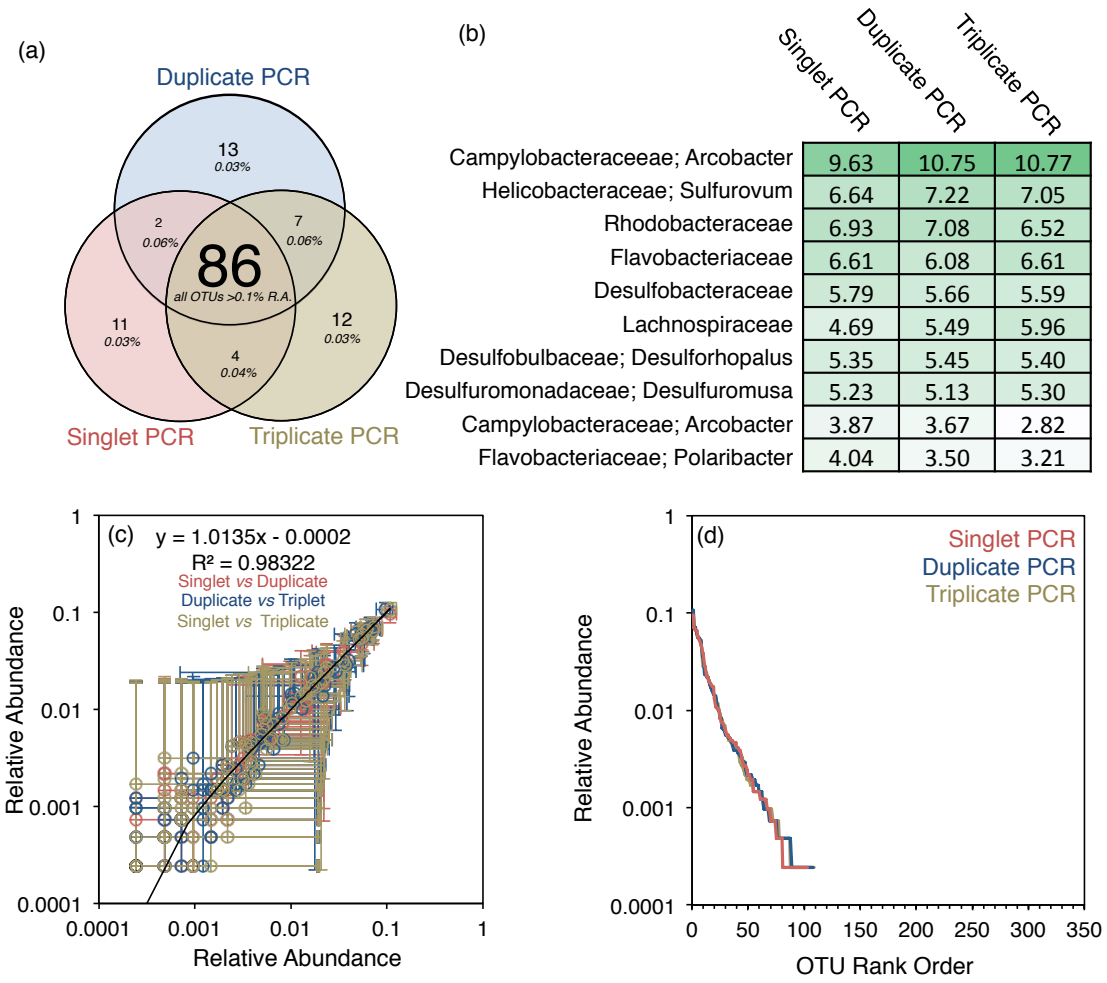

Figure 9. Comparison of amplification approach: pooling single, double, or triple PCR products during preparation of sample #5036 (colonized wood) for iTag sequencing. (a) shows the OTU overlap between results from the three preparations. The majority of OTUs are shared, including all the major OTUs (any OTU >0.1% relative abundance in either sample is shared between the two samples). Only a small number of OTUs, with very low relative abundance, are not shared between preparations. (b) tabulates the relative abundances of the top 10 most abundant OTUs between the preparations, while (c) is a cross-plot of the relative abundance of the OTUs shared between the three preparations. Error bars in (c) are 1.85% relative abundance (c.f. precision for samples amplified with 5-PRIME *Taq*). (d) gives the OTU rank abundance curve for all three preparations.

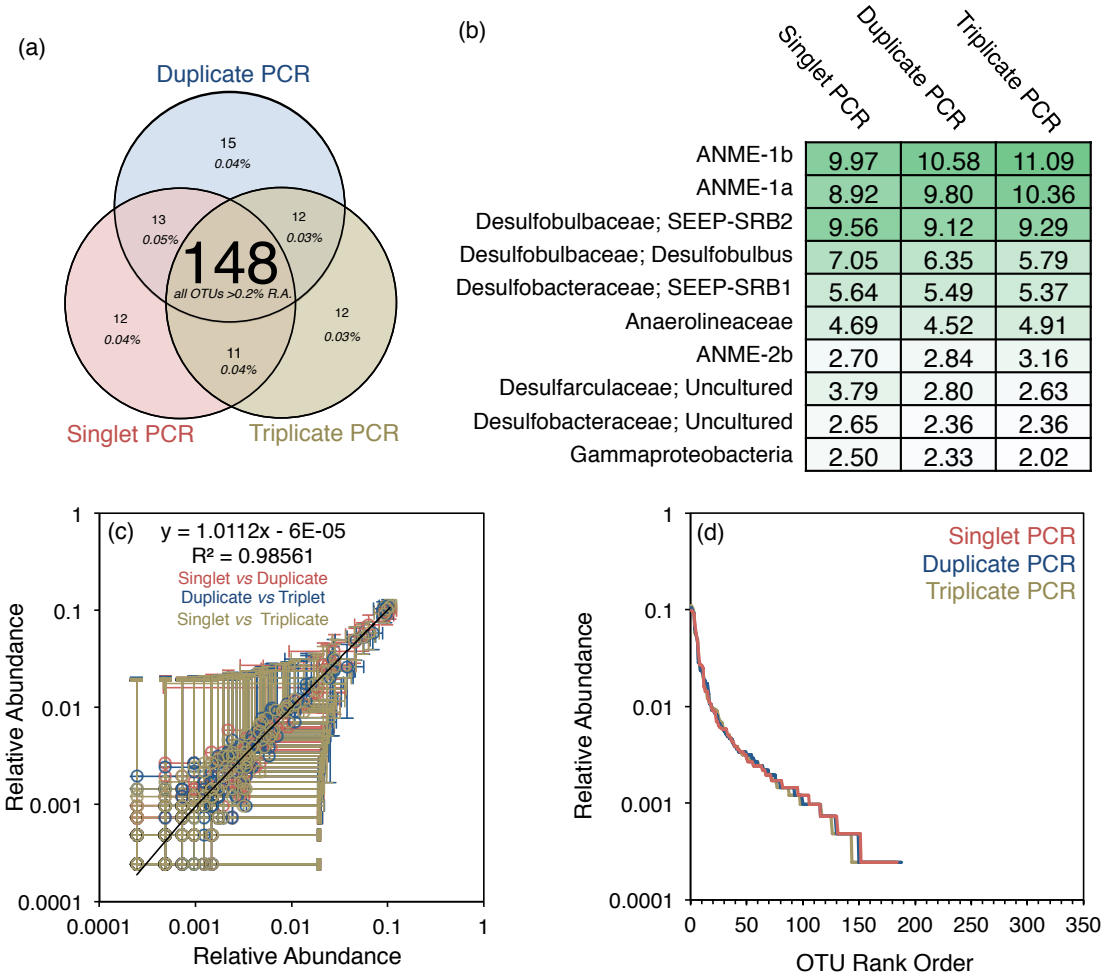

Figure 10. Comparison of amplification approach: pooling single, double, or triple PCR products during preparation of sample #5193 (transplanted carbonate) for iTag sequencing. (a) shows the OTU overlap between results from the three preparations. The majority of OTUs are shared, including all the major OTUs (any OTU >0.2% relative abundance in either sample is shared between the two samples). Only a small number of OTUs, with very low relative abundance, are not shared between preparations. (b) tabulates the relative abundances of the top 10 most abundant OTUs between the preparations, while (c) is a cross-plot of the relative abundance of the OTUs shared between the three preparations. Error bars in (c) are 1.85% relative abundance (c.f. precision for samples amplified with 5-PRIME *Taq*). (d) gives the OTU rank abundance curve for all three preparations.

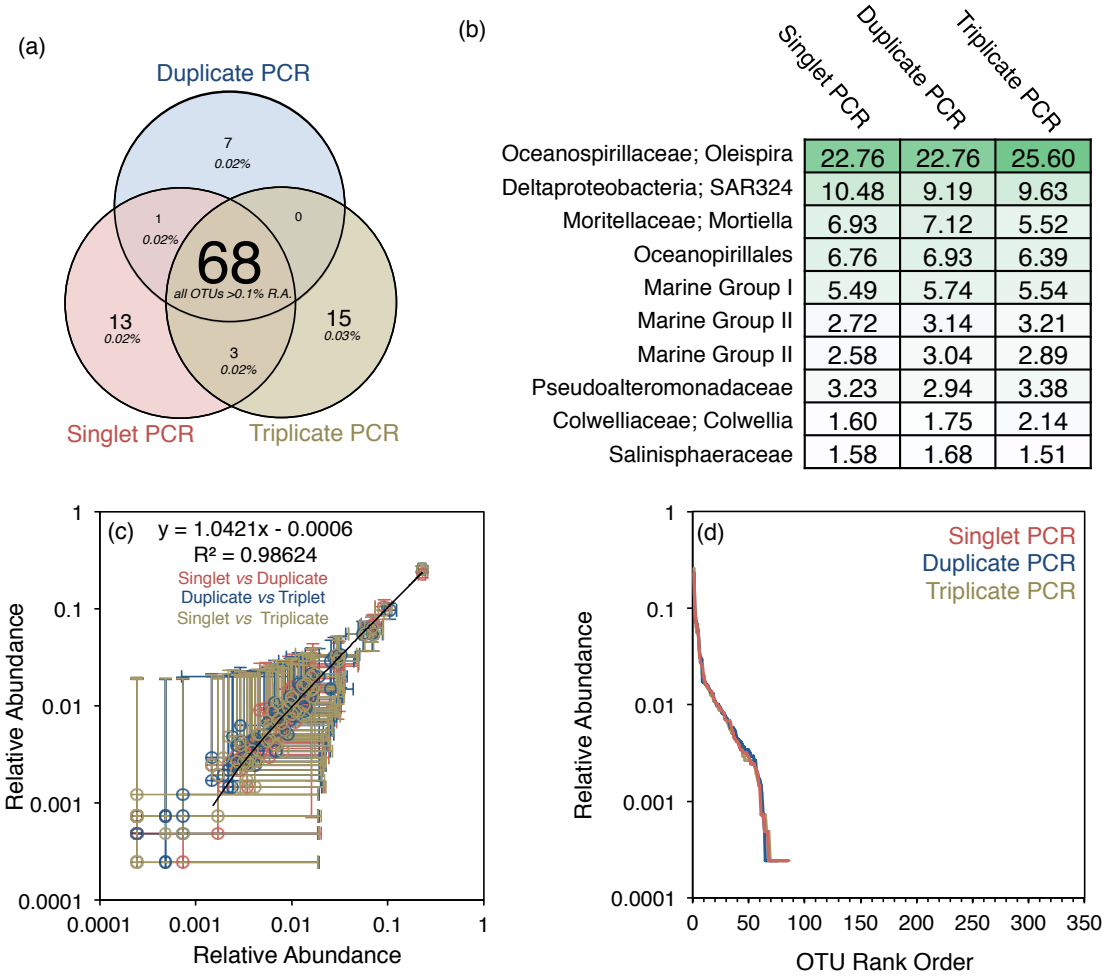

Figure 11. Comparison of amplification approach: pooling single, double, or triple PCR products during preparation of sample #5472 (bottom water) for iTag sequencing. (a) shows the OTU overlap between results from the three preparations. The majority of OTUs are shared, including all the major OTUs (any OTU >0.1% relative abundance in either sample is shared between the two samples). Only a small number of OTUs, with very low relative abundance, are not shared between preparations. (b) tabulates the relative abundances of the top 10 most abundant OTUs between the preparations, while (c) is a cross-plot of the relative abundance of the OTUs shared between the three preparations. Error bars in (c) are 1.85% relative abundance (c.f. precision for samples amplified with 5-PRIME *Taq*). (d) gives the OTU rank abundance curve for all three preparations.

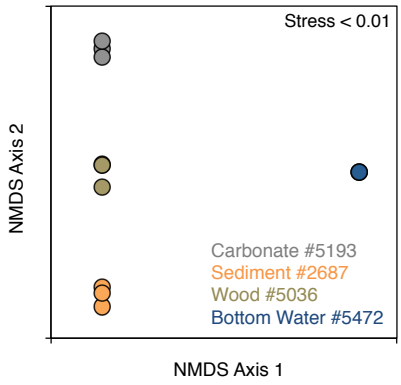

Figure 12. Non-metric multidimensional scaling analysis of PCR pooling treatments (single, double, triple) of four marine methane seep samples representing various substrates. The samples are well-differentiated by substrate but not by pooling treatment, indicating pooling treatment is not likely to alter ecological interpretations of datasets.

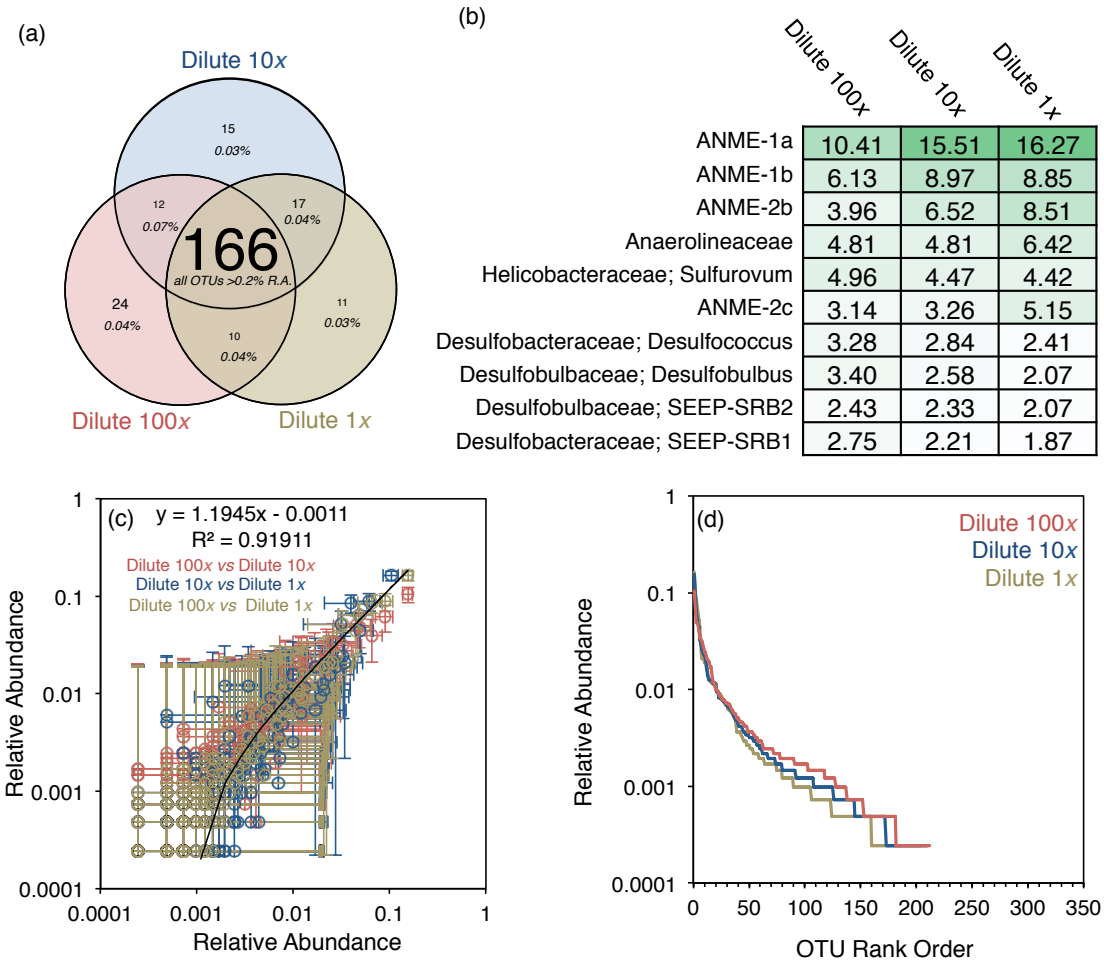

Figure 13. Comparison of amplification approach: gDNA template diluted 1X, 10X, or 100X prior to PCR amplification of sample #5122 (carbonate) for iTag sequencing. (a) shows the OTU overlap between results from the three preparations. The majority of OTUs are shared, including all the major OTUs (any OTU >0.2% relative abundance in either sample is shared between the two samples). Only a small number of OTUs, with very low relative abundance, are not shared between preparations. (b) tabulates the relative abundances of the top 10 most abundant OTUs between the preparations, while (c) is a cross-plot of the relative abundance of the OTUs shared between the three preparations. Error bars in (c) are 1.85% relative abundance (c.f. precision for samples amplified with 5-PRIME *Taq*). (d) gives the OTU rank abundance curve for all three preparations.

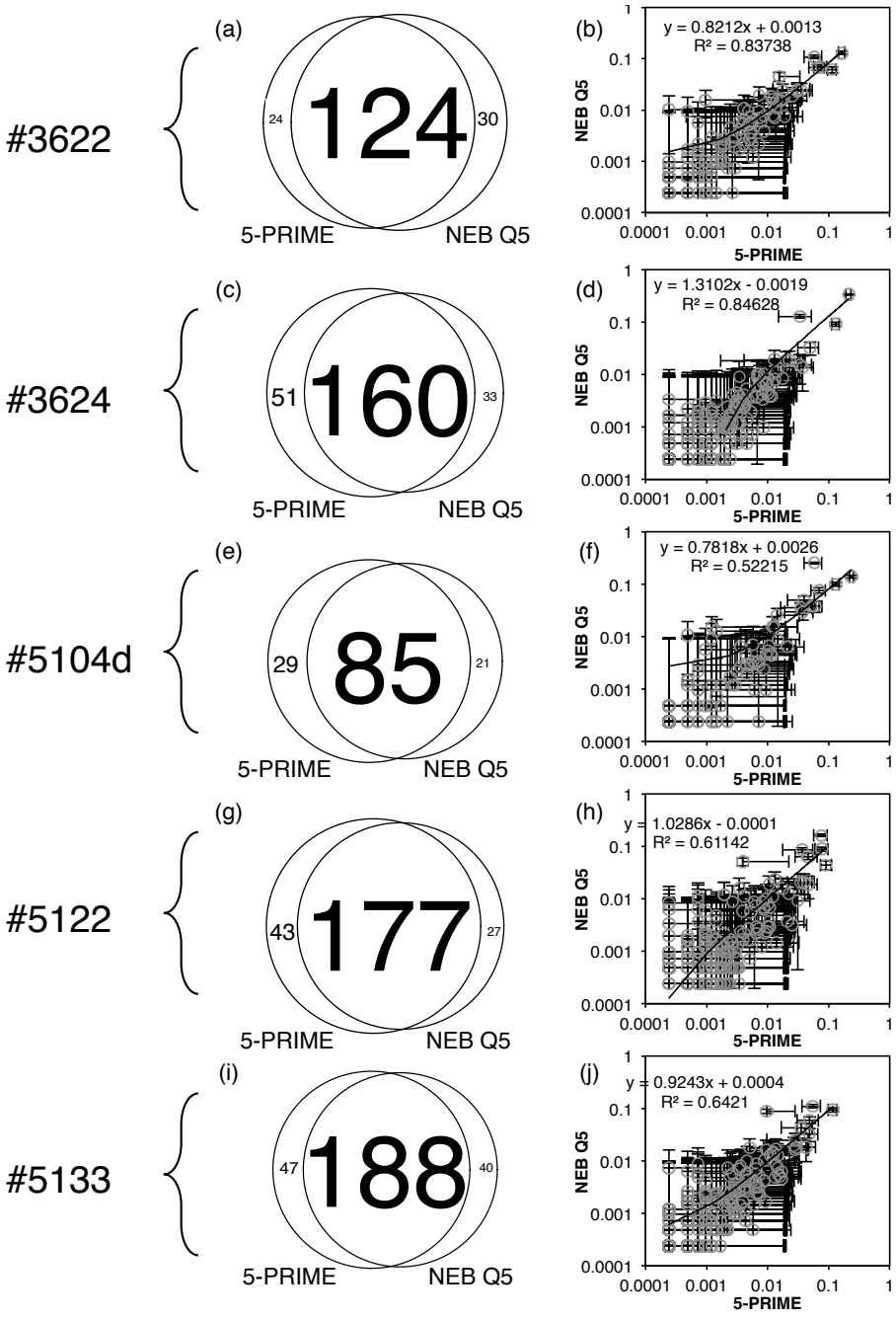

Figure 14. Comparison of amplification approach: 5-PRIME Hot Master Mix vs NEB Q5. (a,c,e,g,i) show the OTU overlap between the two amplifications. (b,d,f,g,j) show a cross plot of the shared OTUs. Although the majority of OTUs are shared, due to known biases between the enzymes (c.f. Mock Communities), the R^2 values in the cross plots relatively low with slopes of the regression line deviating from a value of 1.

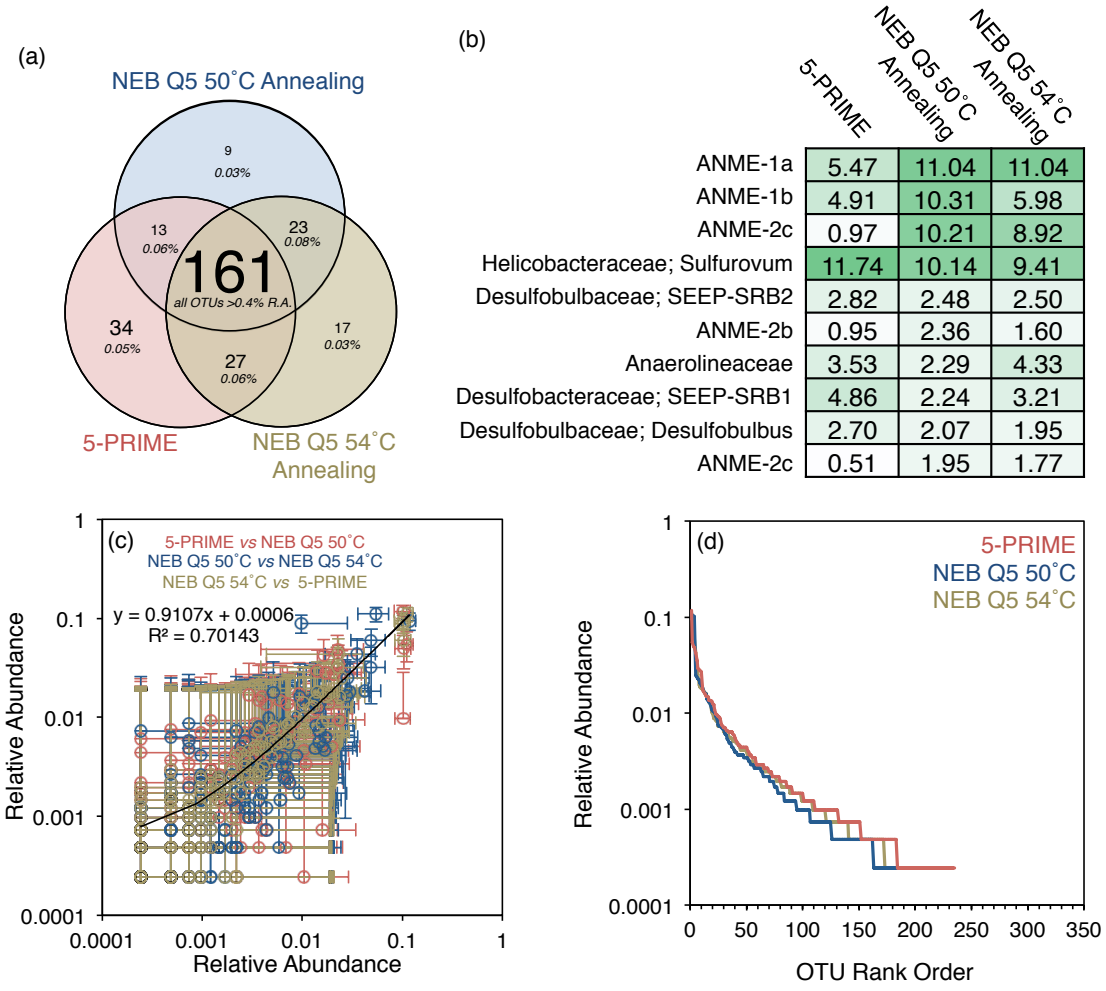

Figure 15. Comparison of amplification approach: amplification using 5-PRIME *Taq*, NEB Q5 annealing at 50˚C, or NEB Q5 annealing at 54˚C for sample #5133 (sediment) for iTag sequencing. (a) shows the OTU overlap between results from the three preparations. The majority of OTUs are shared, including all the major OTUs (any OTU >0.4% relative abundance in either sample is shared between the two samples). Only a small number of OTUs, with very low relative abundance, are not shared between preparations. (b) tabulates the relative abundances of the top 10 most abundant OTUs between the preparations, while (c) is a cross-plot of the relative abundance of the OTUs shared between the three preparations. Error bars in (c) are 1.85% relative abundance (c.f. precision for samples amplified with 5-PRIME *Taq*). (d) gives the OTU rank abundance curve for all three preparations.

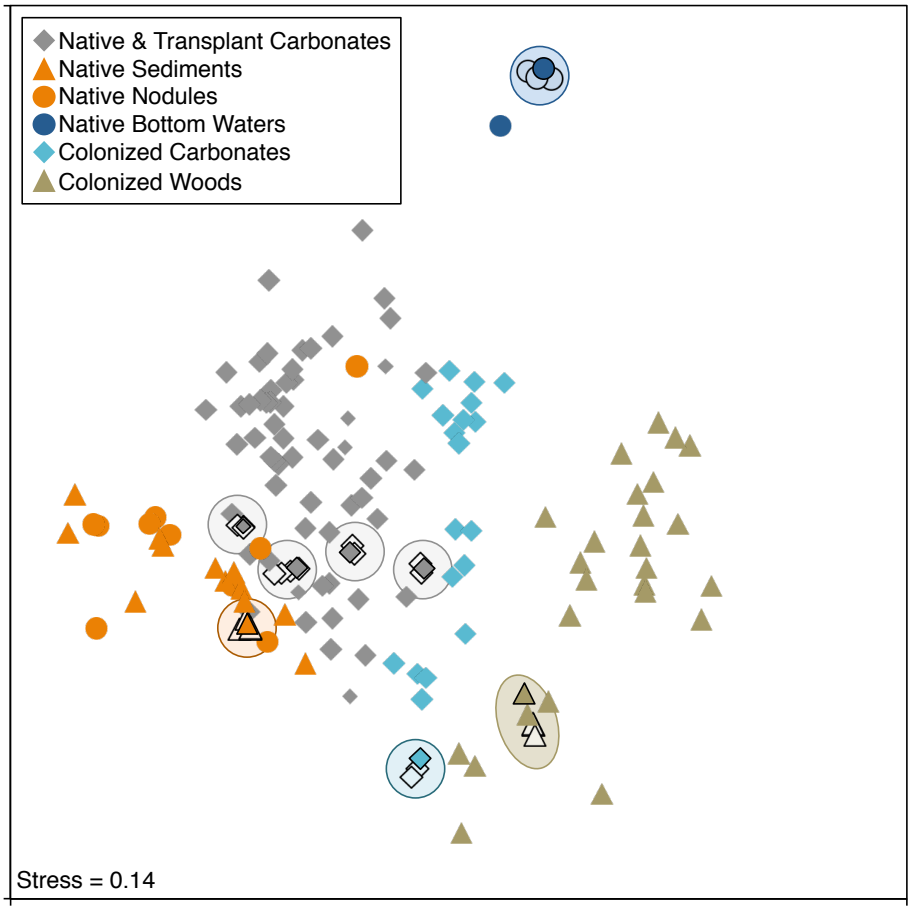

Figure 16. Nonmetric multidimensional scaling of 134 samples published in Case *et al.,* 2015, along with samples which were subjected to various methodology tests. Overall the methodology tests, despite differences in recovered 16S rRNA gene profiles, do not exhibit a large difference when compared to other samples in a large environmental dataset. Samples published in Case *et al.*, 2015 (using the "default" preparation and processing methodology) are given symbols with bold colors and a black border. Their corresponding samples which were subjected to methodology tests are identified by symbols with pale colors and a black border. Colored ovals are drawn by hand to guide the reader's eye to these groupings.

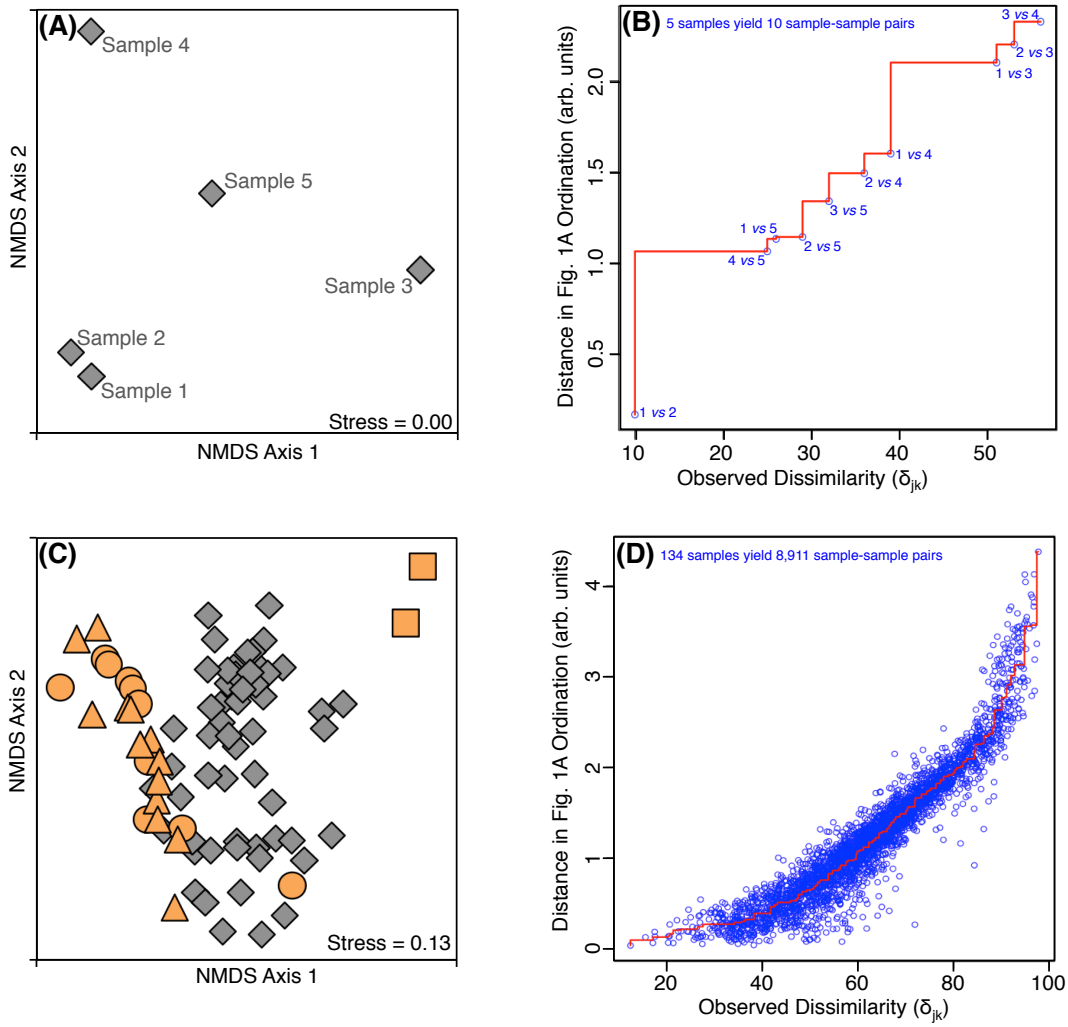

Figure 17. NMDS and Shepard plots. Panels (A) and (B) are calculated from synthetic data of five samples given in Tables 3-5. Panels (C) and (D) are calculated from 16S rRNA gene data from 134 methane seep samples (c.f., Chapter Two of this thesis; Case et al., 2015). In NMDS plots (panels (A) and (C)), each point represents the entire microbial assemblage from one sample. Data points closer to one another are more biologically similar. In (C), gray and orange indicate carbonate and non-carbonate habitats, respectively. Also in (C), circles, triangles, squares, and diamonds represent nodules, circles, bottom waters, and carbonates, respectively. In Shepard plots (panels (B) and (D)), the x-axis is calculated using Equation 2 of Chapter 4 and the y-axis is calculated by Euclidian distance on the accompanying NMDS plots. Every blue circle represents the dissimilarity and ordination distance between one pair of samples in the dataset. The red line is a monotonic regression to the blue circles. Stress, reported in the lower right corners of (A) and (C), is calculated as in Equation 3 of Chapter 4 by summing the differences between blue data points and the red regression line.

4.9 REFERENCES

Aronesty, E. 2011. ea-utils: Command-line tools for processing biological sequencing data.

Bentley, D. R., S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, J. M. Boutell, J. Bryant, R. J. Carter, R. K. Cheetham, A. J. Cox, D. J. Ellis, M. R. Flatbush, N. A. Gormley, S. J. Humphray, L. J. Irving, M. S. Karbelashvili, S. M. Kirk, H. Li, X. Liu, K. S. Maisinger, L. J. Murray, B. Obradovic, T. Ost, M. L. Parkinson, M. R. Pratt, I. M. J. Rasolonjatovo, M. T. Reed, R. Rigatti, C. Rodighiero, M. T. Ross, A. Sabot, S. V. Sankar, A. Scally, G. P. Schroth, M. E. Smith, V. P. Smith, A. Spiridou, P. E. Torrance, S. S. Tzonev, E. H. Vermaas, K. Walter, X. Wu, L. Zhang, M. D. Alam, C. Anastasi, I. C. Aniebo, D. M. D. Bailey, I. R. Bancarz, S. Banerjee, S. G. Barbour, P. A. Baybayan, V. A. Benoit, K. F. Benson, C. Bevis, P. J. Black, A. Boodhun, J. S. Brennan, J. A. Bridgham, R. C. Brown, A. A. Brown, D. H. Buermann, A. A. Bundu, J. C. Burrows, N. P. Carter, N. Castillo, M. C. E. Catenazzi, S. Chang, R. N. Cooley, N. R. Crake, O. O. Dada, K. D. Diakoumakos, B. Dominguez-Fernandez, D. J. Earnshaw, U. C. Egbujor, D. W. Elmore, S. S. Etchin, M. R. Ewan, M. Fedurco, L. J. Fraser, K. V. F. Fajardo, W. S. Furey, D. George, K. J. Gietzen, C. P. Goddard, G. S. Golda, P. A. Granieri, D. E. Green, D. L. Gustafson, N. F. Hansen, K. Harnish, C. D. Haudenschild, N. I. Heyer, M. M. Hims, J. T. Ho, A. M. Horgan, K. Hoschler, S. Hurwitz, D. V. Ivanov, M. Q. Johnson, T. James, T. A. H. Jones, G.-D. Kang, T. H. Kerelska, A. D. Kersey, I. Khrebtukova, A. P. Kindwall, Z. Kingsbury, P. I. Kokko-Gonzales, A. Kumar, M. A. Laurent, C. T. Lawley, S. E. Lee, X. Lee, A. K. Liao, J. A. Loch, M. Lok, S. Luo, R. M. Mammen, J. W. Martin, P. G. McCauley, P. McNitt, P. Mehta, K. W. Moon, J. W. Mullens, T. Newington, Z. Ning, B. L. Ng, S. M. Novo, M. J. O'Neill, M. A. Osborne, A. Osnowski, O. Ostadan, L. L. Paraschos, L. Pickering, A. C. Pike, A. C. Pike, D. C. Pinkard, D. P. Pliskin, J. Podhasky, V. J. Quijano, C. Raczy, V. H. Rae, S. R. Rawlings, A. C. Rodriguez, P. M. Roe, J. Rogers, M. C. R. Bacigalupo, N. Romanov, A. Romieu, R. K. Roth, N. J. Rourke, S. T. Ruediger, E. Rusman, R. M. Sanches-Kuiper, M. R. Schenker, J. M. Seoane, R. J. Shaw, M. K. Shiver, S. W. Short, N. L. Sizto, J. P. Sluis, M. A. Smith, J. E. S. Sohna, E. J. Spence, K. Stevens, N. Sutton, L. Szajkowski, C. L. Tregidgo, G. Turcatti, S. vandeVondele, Y. Verhovsky, S. M. Virk, S. Wakelin, G. C. Walcott, J. Wang, G. J. Worsley, J. Yan, L. Yau, M. Zuerlein, J. Rogers, J. C. Mullikin, M. E. Hurles, N. J. McCooke, J. S. West, F. L. Oaks, P. L. Lundberg, D. Klenerman, R. Durbin, and A. J. Smith. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature **456**: 53–59.

Berry, D., K. Ben Mahfoudh, M. Wagner, and A. Loy. 2011. Barcoded Primers Used in Multiplex Amplicon Pyrosequencing Bias Amplification. Applied and Environmental Microbiology **77**: 7846–7849.

Bokulich, N. A., S. Subramanian, J. J. Faith, D. Gevers, J. I. Gordon, R. Knight, D. A. Mills, and J. G. Caporaso. 2013. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. Nature Methods **10**: 57–59.

Brandariz-Fontes, C., M. Camacho-Sanchez, C. Vilà, J. L. Vega-Pla, C. Rico, and J. A. Leonard. 2015. Effect of the enzyme and PCR conditions on the quality of high-throughput DNA sequencing results. Scientific Reports **5**: 8056.

Bray, J. R., and J. T. Curtis. 1957. An Ordination of the Upland Forest Communities of Southern Wisconsin. Ecological Monographs **27**: 325–349.

Caporaso, J. G., J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Peña, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld, and R. Knight. 2010. QIIME allows analysis of high-throughput community sequencing data. Nature Methods **7**: 335–336.

Case, D. H., A. L. Pasulka, J. J. Marlow, B. M. Grupe, L. A. Levin, and V. J. Orphan. 2015. Methane Seep Carbonates Host Distinct, Diverse, and Dynamic Microbial Assemblages. mBio **6**: e01348–15.

Chandler, D. P., J. K. Fredrickson, and F. J. Brockman. 1997. Effect of PCR template concentration on the composition and distribution of total community 16S rDNA clone libraries. Mol Ecol **6**: 475–482.

Clarke, K. R., and R. M. Warwick. 2001. Change in Marine Communities, 2nd ed. PRIMER-E Ltd.

Dominguez-Bello, M. G., K. M. De Jesus-Laboy, N. Shen, L. M. Cox, A. Amir, A. González, N. A. Bokulich, S. J. Song, M. Hoashi, J. I. Rivera-Vinas, K. Mendez, R. Knight, and J. C. Clemente. 2016. Partial restoration of the microbiota of cesarean-born infants via vaginal microbial transfer. Nature Medicine **22**: 250–253.

Edgar, R. C., B. J. Haas, J. C. Clemente, C. Quince, and R. Knight. 2011. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics **27**: 2194–2200.

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

Faith, J. J., J. L. Guruge, M. Charbonneau, S. Subramanian, H. Seedorf, A. L. Goodman, J. C. Clemente, R. Knight, A. C. Heath, R. L. Leibel, M. Rosenbaum, and J. I. Gordon. 2013. The Long-Term Stability of the Human Gut Microbiota. Science **341**: 1237439.

Gilbert, J. A., F. Meyer, J. Jansson, J. Gordon, N. R. Pace, J. M. Tiedje, R. E. Ley, N. Fierer, D. Field, N. C. Kyrpides, F. O. Gloeckner, H. P. Klenk, K. E. Wommack, E. Glass, K. Docherty, R. Gallery, R. Stevens, and R. Knight. 2011. The Earth Microbiome Project: Meeting report of the "1st EMP meeting on sample selection and acquisition" at Argonne National Laboratory October 6th 2010. 1–5.

Kleppe, K., E. Ohtsuka, R. Kleppe, I. Molineux, and H. G. Khorana. 1971. Studies on polynucleotides: Repair Replication of Short Synthetic DNAs as catalyzed by DNA Polymerases. Journal of Molecular Biology **56**: 341–361.

Kozich, J. J., S. L. Westcott, N. T. Baxter, S. K. Highlander, and P. D. Schloss. 2013. Development of a Dual-Index Sequencing Strategy and Curation Pipeline for Analyzing Amplicon Sequence Data on the MiSeq Illumina Sequencing Platform. Applied and Environmental Microbiology **79**: 5112–5120.

Kruskal, J. B. 1964a. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika **29**: 1–27.

Kruskal, J. B. 1964b. Nonmetric multidimensional scaling: A numerical method. Psychometrika **29**: 115–129.

Legendre, P., and L. Legendre. 2012. Numerical ecology.

Levin, L. A., G. F. Mendoza, B. M. Grupe, J. P. Gonzalez, B. Jellison, G. W. Rouse, A. R. Thurber, and A. Waren. 2015. Biodiversity on the Rocks: Macrofauna Inhabiting Authigenic Carbonate at Costa Rica Methane Seeps. PLoS ONE 1–31.

Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y.-J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. I. Alenquer, T. P. Jarvie, K. B. Jirage, J.-B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature **437**: 376–380.

Mason, O. U., D. H. Case, T. H. Naehr, R. W. Lee, R. B. Thomas, J. V. Bailey, and V. J. Orphan. 2015. Comparison of Archaeal and Bacterial Diversity in Methane Seep Carbonate Nodules and Host Sediments, Eel River Basin and Hydrate Ridge, USA. Microbial Ecology **70**: 776–784.

Metcalf, J. L., Z. Z. Xu, S. Weiss, S. Lax, W. Van Treuren, E. R. Hyde, S. J. Song, A. Amir, P. Larsen, N. Sangwan, D. Haarmann, G. C. Humphrey, G. Ackermann, L. R. Thompson, C. Lauber, A. Bibat, C. Nicholas, M. J. Gebert, J. F. Petrosino, S. C. Reed, J. A. Gilbert, A. M. Lynne, S. R. Bucheli, D. O. Carter, and R. Knight. 2016. Microbial community assembly and metabolic function during mammalian corpse decomposition. Science **351**: 158–162.

Morono, Y., T. Terada, T. Hoshino, and F. Inagaki. 2014. Hot-Alkaline DNA Extraction Method for

Deep-Subseafloor Archaeal Communities. Applied and Environmental Microbiology **80**: 1985–1994.

Mullis, K. B., and F. A. Faloona. 1987. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction, p. 335–350. *In* Recombinant DNA Part F. Elsevier.

Nelson, M. C., H. G. Morrison, J. Benjamino, S. L. Grim, and J. Graf. 2014. Analysis, Optimization and Verification of Illumina-Generated 16S rRNA Gene Amplicon Surveys M.M. Heimesaat [ed.]. PLoS ONE **9**: e94249.

Oksanen, J., F. G. Blanchet, R. Kindt, P. Legendre, P. Minchin, R. B. OHara, G. Simpson, P. Solymos, M. H. H. Stevens, and H. Wagner. 2013. vegan: Community Ecology Package.

Parada, A. E., D. M. Needham, and J. A. Fuhrman. 2015. Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. Environmental microbiology, doi:10.1111/1462-2920.13023

Pasulka, A. L., L. A. Levin, J. A. Steele, D. H. Case, M. R. Landry, and V. J. Orphan. 2015. Microbial eukaryotic distributions and diversity patterns in a deep-sea methane seep ecosystem. Environmental microbiology doi:10.1111/1462-2920.13185.

R Core Team. 2014. R: A language and environment for statistical computing.

Ramette, A. 2007. Multivariate analyses in microbial ecology. FEMS Microbiology Ecology **62**: 142–160.

Rosenbaum, M., R. Knight, and R. L. Leibel. 2015. The gut microbiota in human energy homeostasis and obesity. Trends in Endocrinology & Metabolism **26**: 493–501.

Ruiz-Calderon, J. F., H. Cavallin, S. J. Song, A. Novoselac, L. R. Pericchi, J. N. Hernandez, R. Rios, O. H. Branch, H. Pereira, L. C. Paulino, M. J. Blaser, R. Knight, and M. G. Dominguez-Bello. 2016. Walls talk: Microbial biogeography of homes spanning urbanization. Science Advances **2**: e1501061–e1501061.

Saiki, R. K., D. H. Gelfand, S. Stoffel, S. J. Scharf, R. Higuchi, G. T. Horn, K. B. Mullis, and H. A. Erlich. 1988. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science **239**: 487–491.

Sanger, F., and A. R. Coulson. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. Journal of Molecular Biology **94**: 441–448.

Sanger, F., G. M. Air, B. G. Barrell, N. L. Brown, A. R. Couson, J. C. Fiddes, C. A. Hutchinson III, P. M. Slocombe, and M. Smith. 1977a. Nucleotide sequence of bacteriophage ΦX174 DNA. Nature **265**: 687–695.

Sanger, F., S. Nicklen, and A. R. Coulson. 1977b. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences **74**: 5463–5467.

Schadt, E. E., S. Turner, and A. Kasarskis. 2010. A window into third-generation sequencing. Human Molecular Genetics **19**: 227-240.

Shepard, R. N. 1962. The analysis of proximities: Multidimensional scaling with an unknown distance function. I. Psychometrika **27**: 125–140.

Trembath-Reichert, E., A. Green-Saxena, and V. J. Orphan. 2013. Whole Cell Immunomagnetic Enrichment of Environmental Microbial Consortia Using rRNA-Targeted Magneto-FISH, 1st ed. Elsevier Inc.

Trembath-Reichert, E., D. H. Case, and V. J. Orphan. 2016. Characterization of microbial associations with methanotrophic archaea and sulfate-reducing bacteria through statistical comparison of nested Magneto-FISH enrichments. PeerJ **4**: e1913–31.

Wang, Q., G. M. Garrity, J. M. Tiedje, and J. R. Cole. 2007. Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. Applied and Environmental Microbiology **73**: 5261-5267.

Woese, C. R., and G. E. Fox. 1977. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. Proceedings of the National Academy of Sciences **74**: 5088–5090.