Development of New Tools and Applications for High-Throughput Sequencing of Microbiomes in Environmental or Clinical Samples

By

Matthew Christopher Blackburn

B.S., Chemical Engineering (2008)

University of Florida

Submitted to the Department of Chemical Engineering in Partial Fulfillment of the Requirements for the Degree of

Master of Science in Chemical Engineering

at the

Massachusetts Institute of Technology

September 2010

Signature of Author.....
Department of Chemical Engineering
August 6, 2010

Certified by.....
Eric J. Alm
Assistant Professor of Biological Engineering & Civil and Environmental Engineering
Thesis Supervisor

Certified by..
Kristala J. Prather
Assistant Professor of Chemical Engineering
Co-advisor

Accepted by.....
William M. Deen
Professor of Chemical Engineering
Chairman, Committee for Graduate Students

Development of New Tools and Applications for High-Throughput Sequencing of Microbiomes in Environmental or Clinical Samples

By

Matthew Christopher Blackburn

Submitted to the Department of Chemical Engineering on August 6, 2010 in Partial Fulfillment of the Requirements for the Degree of Master of Science in Chemical Engineering

ABSTRACT

Novel sequencing technologies are rapidly advancing studies of microbial community structure and diversity. Sequencing platforms like the Illumina Genome Analyzer II (GA_{II}) and the Applied Biosystems SOLiD enable experiments that were previously too expensive or time-consuming by providing a very large number of short reads at a significantly lower cost per base pair (bp) than conventional longer-read systems like the Roche-454 GS FLX pyrosequencing instrument. Short-read platforms, however, are not readily amenable to some applications like metagenomics and meta-transcriptomics, and therefore pyrosequencing remains the dominant sequencing technique in these fields.

The primary reason short-read technologies have not been used for metagenomic analyses is due to the difficulty of confidently assigning phylogeny or putative gene function to short sequences. In an effort to overcome this limitation, a strategy was developed for preparing libraries from sheared genomic DNA with tunable size distributions using solid phase reversible immobilization (SPRI). This size selection captures DNA fragments of the necessary length to enable the generation of overlapping reads when sequenced from both ends. The lower-quality ends of mated reads were then used to produce a high-quality consensus sequence in the region of overlap. The fraction of composite reads that could be assigned to a taxon was similar to those from 454-FLX, despite the slightly shorter average read length of the composite Illumina reads. This technique successfully demonstrates a practical and economical alternative to 454-FLX for metagenomics.

In addition, a scalable, fully automated process for creating sequence-ready, barcoded libraries of 16S rDNA for microbial diversity studies was developed for the Illumina platform. This process will enable sequencing of hundreds of environmental samples on a single Illumina flowcell, greatly decreasing the cost per sample while providing thousands of short-reads for microbial ecology studies. The incorporation of error-correcting, short DNA "barcodes" (also called tags or indexes) during polymerase chain reaction (PCR) amplification of the 16S sequence facilitates sample multiplexing.

This process also utilizes the SPRI method to replace column-based reaction clean-ups, enabling the library preparation procedure to be performed almost entirely by a robotic liquid handling workstation. Finally, two unique PCR primer systems (*primer-clipping* and *primer-skipping*) were engineered to increase the informative read length of 16S sequence by either cutting the known universal tract out of the final-product to be sequenced, or by omitting sequencing of the universal regions using specially-crafted primers designed to be compatible with Illumina platform conditions.

By applying both the overlapping-read technique and multiplexed 16S library preparation workflow, a streamlined approach for efficient gene and species discovery has been assembled to accommodate new metagenomic applications for the Illumina sequencing platform.

Thesis Supervisor: Eric J. Alm
Title: Assistant Professor of Biological Engineering & Civil and Environmental Engineering

Acknowledgements

For all of his patience, generosity, mentoring and friendship, I'd like to thank Arne Materna for teaching me everything I knew about molecular biology prior to taking any formal molecular biology class. I am also indebted to Eric Alm for accepting me into his lab, helping me along my scientific career path and in preparing this thesis. I would also like to thank the other members of the Alm Lab (Sarah Preheim, Inês Baptista, Jesse Shapiro, Sean Clarke, Lawrence David, Jonathan Friedman, Sonia Timberlake, Mark Smith, and Chris Smilie) for the conversations, words of wisdom, and trips to the brewery. I am also thankful for the friendship of Ali Perrotta, who first taught me how to use the Bioanalyzer, then how to program a robot over beers at the Muddy Charles pub.

I've made many friends while at MIT, but I have particularly appreciated being friends with my old roommate, Micah Sheppard, and Shawn Finney-Manchester. Both of them were amazing p-set and study partners, and I don't think I would have survived the first semester of MIT without them. Also, for all of the fresh air, scenery, exercise, and especially friends, I'd like to thank the MIT Cycling Team for making me unafraid of the streets of Boston while teaching me how to handle my bike like a pro.

I also need to thank all of my research mentors from the past six years, who have significantly influenced my path through life, and shared significant amounts of their time and knowledge with me: Iris Schumacher, Katie Dziak, David Myers, Chelsea Magin, Thomas Estes, Prof. Tony Brennan, Shanna Ratnesar-Shumate, and Prof. Chang-Yu Wu.

Finally, I would like to thank my parents and the rest of my family for supporting me through the years with their words of encouragement.

Table of Contents

List of Figures

List of Tables

1. *Introduction*

1.1 *Metagenomics*

Metagenomics is broadly defined as the use of DNA sequencing to identify and catalog all of the microorganisms present in an environmental sample (1, 2, 3, 4). Since ~99% of all microbes in the biosphere are not readily obtainable in pure culture, sequence-based techniques circumvent the constraints of culture-based methods by enabling studies of microbial communities sampled directly from their natural habitats. Driven primarily by gene discovery and phylogenetic classification of species from uncharacterized environments, sequencing projects have explored microbial communities from the Sargasso Sea (5), acid mine drainage biofilms (6), and the human body (7, 8, 9, 10, 11), among countless other locations and habitats.

To date, a majority of metagenomic studies have relied on whole-genome shotgun sequencing approaches, which involved preparing small-insert DNA clone libraries with DNA from environmental samples, followed by Sanger sequencing. While this method yields highly accurate sequence reads with lengths up to 1000 bp, biases introduced during the creation of clone libraries combined with the high cost per base pair have motivated researchers to adopt new, high-throughput sequencing-by-synthesis technologies. Concurrent with the dramatic improvements in the speed and ease of DNA-sequence data collection, the field of metagenomics is rapidly growing as scientists begin to develop a systems-level understanding of microbial community interactions, how they have evolved, and their gene ensembles.

A typical high-throughput sequencing run can produce several gigabytes of short-read data (~0.1-10 Gbp of sequence). This presents a serious challenge to bioinformaticians who attempt to identify sequence fragments by their homology to known genes, or by using *ab initio* gene-prediction algorithms to aid the discovery of novel genes. One drawback to applying short-read technologies to metagenomics is the difficulty of detecting homologs or inferring gene function (12, 13, 14, 15, 16).

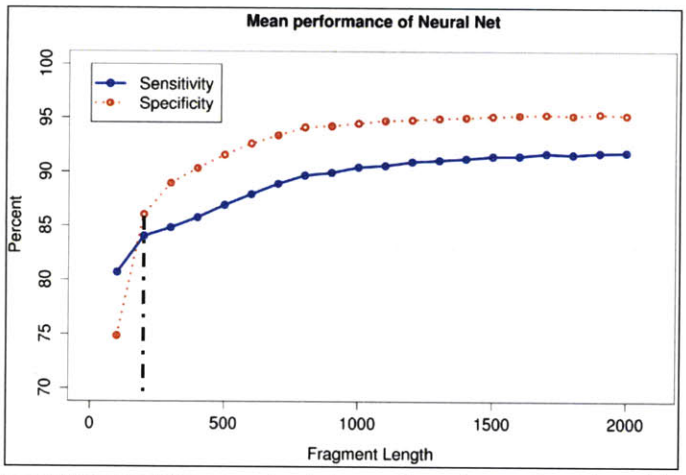

Figure 1. 200 bp fragments are important for accurate phylogenetic assignments. Adapted from Hoff et al. (14); depicts average gene prediction of a neural network as a function of sequence fragment lengths (from 100-2000 bp). The largest increase in sensitivity and specificity associated with the change from 100 to 200 bp fragment length provides the motivation for development of the over-lapping reads.

Wommack et al. (12) performed simulations concluding that reads shorter than 400 bp miss a significant amount of the homologs found with longer reads, and also restricts detection of gene function classes. It is evident from their work that overall sequencing depth does not necessarily compensate for the short-read lengths. Additionally, Mitra et al. (16) performed simulations and found longer reads unequivocally allow for more specific taxonomic assignments than short-reads. An approach to amending this limitation of short-reads is to extend the total read length of high-throughput sequencing systems. For the Illumina platform, the target sequence length for libraries is 200 bp (Figure 1). These fragments have the length necessary to enable the generation of overlapping reads when sequenced from both ends and correspond to a significant increase in the specificity of phylogenetic and gene function assignments (13, 14, 17, 18, 19).

A widely-used metric for determining the base-calling accuracy or quality of sequencing reads is the Phred quality score, which was originally developed for Sanger sequencing (20). Phred is a program that is used to calculate quality scores logarithmically linked with error probabilities. The quality of a sequencing read, Q, is a function of noise, signal intensity, and peak spacing. Quality scoring is also highly context-dependent since error distributions vary across the different sequencing technologies.

$$\mathrm{Q} = -10\ log_{10} P_{\mathrm{incorrect}}$$

alternatively,

$$\mathrm{P_{incorrect}} = 1 - \mathrm{P_{correct}} = 10^{\left(-\frac{\mathrm{Q}}{10}\right)}$$

For example, if Phred assigns a quality score of 20 to a base, the probability that the base was called incorrectly is $10^{-2} = 0.01$ or 1 in 100 (the base call was 99% accurate).

The key engineering challenge in enabling metagenomics on the Illumina platform is increasing read-length while maintaining satisfactory quality scores.

1.2 *Phylogenetic Markers*

Species-specific, phylogenetic markers are a useful tool for distinguishing between different species in a heterogeneous population. In most phylogenetic surveys of bacteria, the 16S small ribosomal subunit rDNA (rRNA gene) is used to estimate species diversity (21, 22, 23, 24). The 16S rRNA gene has been studied in great detail, and the database of 16S rDNA sequences (the Ribosome Database Project) has grown immensely as more species are discovered. Polymerase chain reaction (PCR) of 16S rDNA from unculturable microorganisms requires the judicious selection and use of oligonucleotide primers complementary to a universally-conserved region of 16S rDNA. There exist two significant drawbacks to amplification of the 16S rDNA with universal, degenerate primers. Firstly, simultaneous amplification of a specific gene from a heterogeneous mixture of templates can introduce PCR bias (to be addressed later). Secondly, "universal" primers are not necessarily complementary to all of the conserved regions of all taxa, and therefore do not ensure an accurate representation of all species present. Limitations aside, sequence analysis of the 16S marker gene has become the gold standard for providing insight into the diversity of genomic content and composition of microbial communities. For our phylogenetic studies, we developed primers complementary to the universally conserved C3 and C5 regions at positions 515 and 926, for the forward and reverse primers, respectively (25). These positions were chosen for their ability to capture the genomic content of the V4 and V5 hypervariable regions that are used in bacterial phylogenetic classification (Figure 2).

The key engineering challenge in using PCR of 16S rDNA for phylogenetic analysis is the rational design of universal primers.

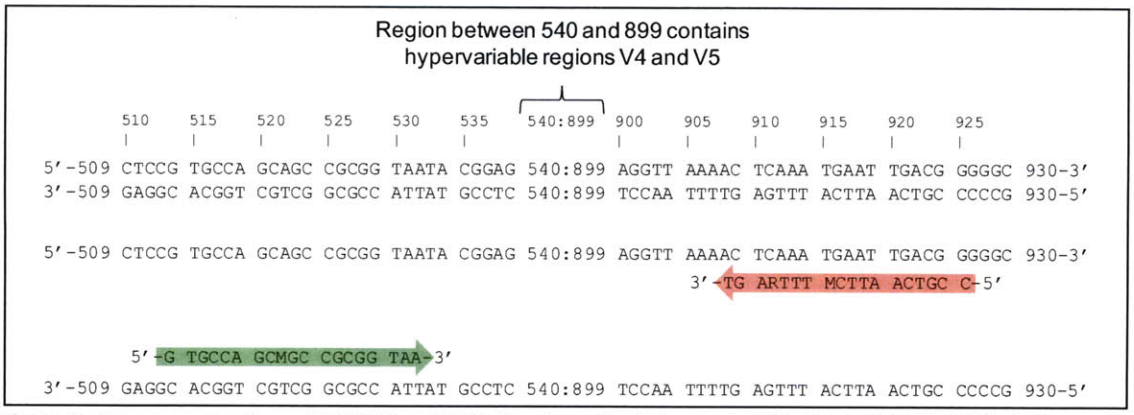

Figure 2. Sequence structure of U515-F and U926-R universal priming regions for bacterial 16S rDNA phylogenetic classification.

1.3 *Next Generation Sequencing (NGS)*

Up until the past 5 years, the automated chain-terminator method developed by Frederick Sanger was the dominant tool employed for genome analysis. Despite many improvements in this technique since its invention in 1977, the limitations of the cost- and labor-intensive automated Sanger ('first-generation') sequencing has spurred the development of advanced 'next-generation' technologies that provide fast, affordable genome sequencing. This movement towards more economical, high-throughput (massively parallelized) sequencing is democratizing the ability of individual labs to generate DNA sequence data at a scale that was previously only available to large genome centers (Table 1). Important metrics to consider for next-generation sequencing are:

1. (Bases/read)*(Reads/run) = Bases/run
2. (Bases/run)*(Cost/base) = Cost/run
3. Accuracy
4. Coverage
5. Hidden costs: Instrument cost, technician

The most prominent 'second-generation' strategy for DNA sequencing involves sequencing-by-synthesis (or cyclic reversible termination, (26)), in which the incorporation of a base by polymerase is indicated by a fluorescent signal that is detected by a CCD (camera). This technique consists of multiple rounds of fluorescent nucleotide incorporation, fluorescence imaging, and cleavage of the inhibiting moiety and fluorescent dye. There are several second generation technologies available (26, 27, 28, 29, 30, 31), but only the Illumina/Solexa Genome Analyzer platform will be described in detail.

Table 1. Comparison of Sequencing Platforms

Platform	Read length (bp)	Total number of reads	Cost per run ($)	Cost per bp
Sanger sequencing	550-900[c]	1	$4-11[e]	~ $0.01
Roche-454 GS FLX	330[a], 400[c], 400-500[d]	0.4×10^{6} [c], 1×10^{6} [d]	$1000 (1/16 plate) to $10000 (full plate, Titanium chemistry)[h]	~ $0.00001
Applied Biosystems SOLiD	35[c], 30[d]	100×10^{6} [d]	$5000	~ $0.000001
Illumina GAIIx	75-100[a, g], 36[b, c], 35 per direction[d]	50×10^{6} [d] (per lane)	$1500 per lane (36 paired-end read) to $5000 per lane (144 paired-end read)[f], 7 lanes per flowcell	~ $0.000001

a. Metzker ML (2010) Sequencing technologies – the next generation. Nature Reviews Genetics 11(1):31-46.

b. MacLean D, Jones JDG, Studholme DJ (2009) Application of 'next-generation'sequencing technologies to microbial genetics. Nature Reviews Microbiology 7(4):287-96.

c. Cardenas E, Tiedje JM (2008) New tools for discovering and characterizing microbial diversity. Current Opinion in Biotechnology 19:544-49.

d. Aranout, RA. Presentation - Next-generation Sequencing Training (2009).

e. http://web.mit.edu/ki/facilities/biopolymers/fees_biopolymers.html

f. http://openwetware.org/wiki/BioMicroCenter:Pricing

g. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, Fierer N, Knight R (2010) Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. PNAS Early-Edition.

h. http://www.med.upenn.edu/dnaseq/454sequencer.shtml

The Illumina Genome Analyzer sequencing platform has its technology based in research by Turcatti and colleagues (32, 33) as well as work patented by Mayer, Farinelli and Kawashima (34). Briefly, DNA insert libraries are prepared such that the fragments are flanked on each end with one of 2 adapters which allow the sequences to be immobilized on a solid substrate by annealing to either forward or reverse PCR primers covalently fixed to the floor of a sequencing flowcell. The DNA inserts can contain several hundred base pairs and are randomly distributed on the flowcell. In order to visualize the fluorescent signal of each DNA insert, hundreds of clonal amplicons are produced via bridge PCR in a process referred to as 'clustering'. Bridge amplification uses the substrate-attached forward and reverse PCR primers to create 'clusters' containing amplicons arising from a single template molecule. Extreme care must be taken to accurately quantify the concentration of the completed template library prior to its application to the flowcell to maximize the cluster density but prevent over-crowding. Ideally, the completion of solid-phase amplification produces several million clusters, each in spatially discrete locations to prevent interference with fluorescent signal resolution during imaging.

Following cluster generation, a solution containing sequencing primer is flowed over the lawn of single-stranded amplicons which hybridizes to a universal sequence contained in the adaptors flanking each DNA insert. Successive cycles of base-interrogation are conducted by single-base addition of nucleotides containing reversible terminators at the 3'-hydroxyl position and one of four fluorescent labels, both of which are chemically cleavable. After single-base extension with a modified DNA polymerase, images are taken in 4 different color channels to determine which nucleotide was incorporated, the terminator and fluorophore are cleaved, and the next cycle begins. Color detection utilizes total internal reflection fluorescence (TIRF) imaging with 2 lasers.

The key engineering challenge in using the Illumina platform for 16S sequencing is developing a primer set that is compatible with current Illumina technology.

1.4 *Bias and Artifacts in Multitemplate PCR & their Prevention*

Following environmental sampling, extraction methods for obtaining genomic DNA must be thoughtfully constructed to limit extraction bias (35, 36, 37). Filtration methods are used to harvest the microorganisms of interest from aquatic/mucosal/saliva samples, whereas microbes in soil/tissues/feces must be isolated from enzyme inhibitors and nucleases. Cell-lysis techniques need to be gentle enough to prevent DNA degradation while being aggressive enough to penetrate the peptidoglycan barrier of gram-positive bacteria and other difficult-to-lyse organisms (37). By combining physical, chemical, thermal, and enzymatic lysis techniques, DNA can be obtained with minimal discrimination between species (36).

PCR amplification of rRNA genes combined with high-throughput sequencing has become an invaluable tool for identifying organisms that are uncultivable *in vitro*. Ideally, these studies would provide a quantitative output that accurately reports the abundance of each species proportional to the genes present in the natural environment, maintaining the template-to-product ratio. The use of "universal" primers to probe genetic diversity cannot, however, provide an unbiased representation of an environmental sample because of limitations intrinsic to PCR. The first source of bias occurs with the selection and design of primers, which automatically comes with the caveat that no single set of primers can ensure amplification of all species present. The design of universal primers necessitates a compromise between universal complementarity and other characteristics like melting temperature, annealing temperature, G/C content, self-annealing score (measured in number of hydrogen-bonds between 2 copies of primer molecules, where A-T and G-C pairs contribute 2 and 3 bonds, respectively), self-end-annealing score (also calculated in units of H-bonds), and secondary structure score

(http://www.clcbio.com/index.php?id=569). Poor complementarity and amplification biases can be remedied by using a mix of primers that incorporate different nucleotides at degenerate positions. Both Watanabe et al. (38) and Baker et al. (25) explored the use of inosine residues to improve complementarity in universal primers where a single base was triply degenerate, and recommended them for analysis of diverse environmental populations.

Bias can also result from differences in PCR kinetics, which is a function of the number of homologous templates (species) present; the concentration of each template (genome dosage); the number of rRNA gene copies in each species (the mean number of ribosomal operons in bacteria is 4.1, although it has been shown 16S rDNA gene copy number can vary between 1 and 15 (3)); the processivity and fidelity of the DNA polymerase used; the selection of denaturation, annealing and elongation temperatures and times; the number of thermal cycles; primer concentration; and buffer chemistry (e.g., cosolvents such as acetamide (39, 40), betaine (41), formamide, and DMSO can be added to improve product specificity and yield). Farrelly et al. (42) stressed the importance of knowing 16S rRNA gene copy number data in order to accurately quantify the number of species in an environmental sample. Collectively, all of the mechanisms that preferentially favor the amplification of particular templates because of their sequence structure (e.g. overall low GC content to favor denaturation, high GC content in the priming region, restricted accessibility of rRNA genes due to template folding), are referred to as PCR selection (39).

In high efficiency, mixed-template amplifications, as the concentrations of product molecules increase, the frequency with which homologous, single-stranded template molecules hybridize with each other will also increase. This 'plateau effect' was documented by Suzuki and Giovannoni (40) by comparing a reaction containing two different templates at different concentrations with the same primer set. They found that the template with higher initial concentration reached self-inhibiting concentrations first, removing itself from the competition for dNTPs and primers, enabling the other less abundant template to amplify efficiently enough such that both templates obtained equivalent final concentrations by the end of the reaction. This 'competitive PCR' (Figure 3) primarily occurs once the reaction proceeds into the plateau phase (43).

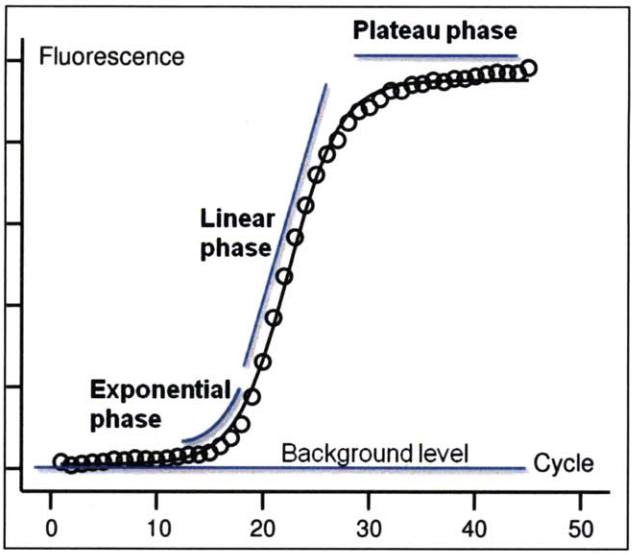

Figure 3. Plot of different reaction phases present in PCR

PCR amplification of highly conserved genes with mixed-genomic templates can also lead to the production of PCR artifacts such as chimeric and heteroduplex molecules (44, 45, 46, 47). Both of these PCR side-products are unavoidable in mixed-template PCR, and their presence in phylogenetic studies can lead to an overestimation of microbial diversity (Figure 4).

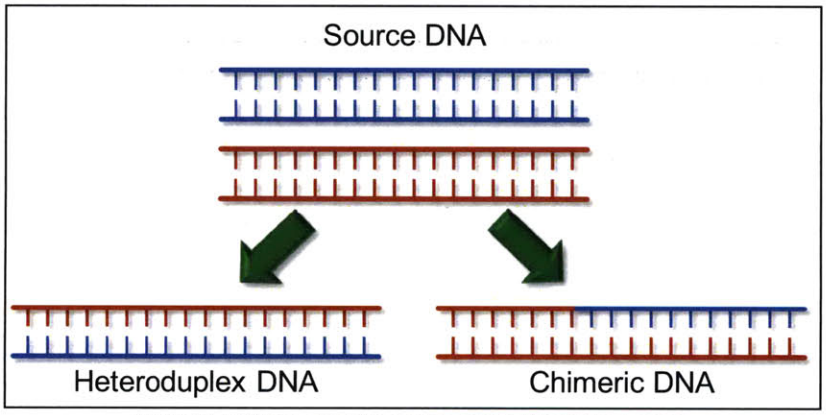

Figure 4. Cartoon depiction of non-specific PCR products

Qiu et al. (44) conducted studies using a four-species mock community to determine the effects of polymerase, cycle number, and elongation times on the frequency of PCR artifact formation. They concluded that using DNA polymerases with higher processivity, increasing the elongation time, and reducing the total number of cycles reduces the frequency of chimeras. Wintzingerode et al. (35) and Wang et al. (45) also achieved results supporting the effect of increasing elongation time and decreasing cycle number on restricting chimera production. The formation of chimeras is primarily

caused by incomplete primer extension during PCR, which produces fragments of DNA that can participate in subsequent PCR cycles by annealing to a heterologous target sequence and act as a primer, forming artificial hybrid products from 2 different template molecules (48). DNA damage has also been suggested as a source of chimeric molecules, since severe cell lysis conditions during DNA extraction may introduce breaks (35). Heteroduplexes are formed during the plateau phase of mixed-template PCR when the concentration of primers becomes limiting, and heterologous sequences favor cross-hybridization (46, 49). Thompson et al. (46) developed a 'reconditioning' PCR which involves diluting amplification products into a fresh PCR mixture with secondary amplification at a low number of cycles to successfully reduce the frequency of heteroduplexes.

The *in vitro* compartmentalization of PCR with water-in-oil emulsion microdroplets permits efficient amplification of complex template mixtures, nearly eliminating PCR bias and recombination artifacts (50, 51, 52, 53, 54). By creating minute reaction volumes, emulsion-based systems can be tailored to produce dilution conditions that accommodate a single template molecule per 'microreactor'. Water-in-oil emulsions use a ratio of DNA to droplets such that many of the droplets contain one template copy or are empty, effectively producing an unbiased, cell-free cloning system (Figure 5).

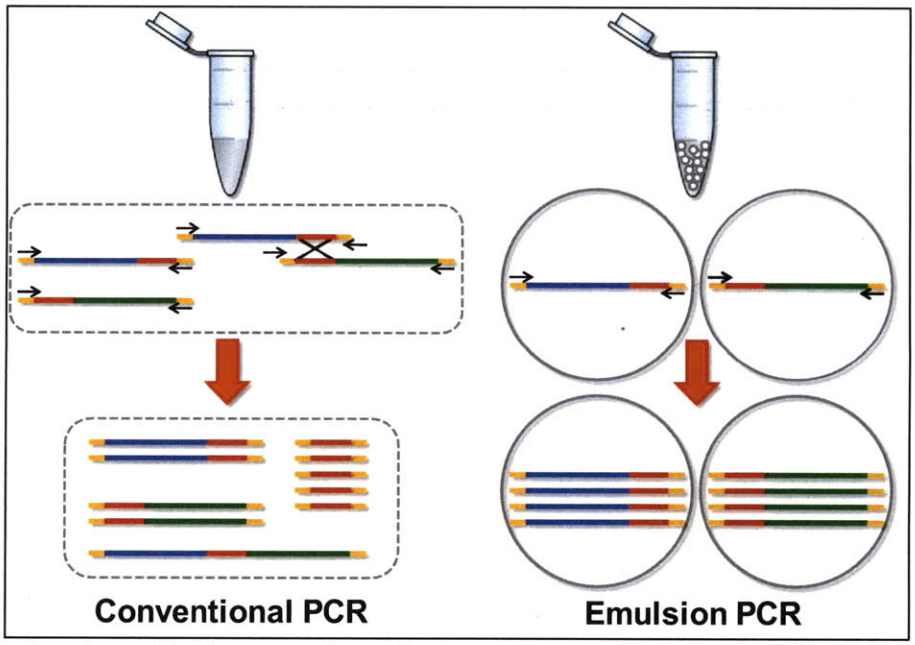

Figure 5. Reduction of recombination events through *in vitro* compartmentalization. Adapted from Williams et al. (53); displays the preventative effect of dilution-limiting emulsion PCR on recombination events when compared to traditional PCR carried out in a single bulk volume.

Another method for preventing nonspecific PCR products, is to incorporate touchdown-PCR (TD-PCR) into the library-preparation scheme (55). The fundamental principle behind TD-PCR is that it favors

the most specific primer-template interaction by beginning with an annealing temperature above the expected T_m (melting temperature) for the primers being used. Successive cycles transition to lower, more permissive T_m's that allow for imperfect primer annealing. This method takes advantage of the exponential PCR phase, in which the first rounds of amplification are the most important for producing the correct product. A typical thermal cycling program incorporating TD-PCR has 2 phases. Phase 1 lasts 10-15 cycles and begins at an initial annealing temperature of T_m+10°C, which decreases 0.5-1°C per cycle until the estimated T_m of the primers is reached. Phase 2 consists of another 20-25 cycles using the lowest annealing temperature reached during Phase 1.

Key engineering challenges to reducing PCR bias are developing a method for compartmentalization of PCR in emulsions and also a method for harvesting PCR products from emulsions (emulsion-breaking)

1.5 *Throughput and Automation*

The automation of DNA library construction for high-throughput sequencing platforms is a necessary development towards obtaining rigorous sampling of microbial ecosystems and capturing their evolution with time. By incorporating error-correcting nucleotide barcodes into PCR primers before amplification (56, 57, 58, 59, 60), hundreds of distinct DNA libraries can be pooled together and sequenced in a single run. By multiplexing samples, and segregating reads based on their barcodes for reference back to metadata, the cost of sequencing per sample drops rapidly, while still allowing for thousands of reads per unique sample.

There are several bottlenecks in standard Illumina library preparation protocol that resist automation. Briefly, the workflow for typical Illumina libraries consists of:

i) fragmenting genomic DNA using a nebulizer

ii) performing gel electrophoresis to isolate products of 800 bp or less

iii) using an enzyme mix of T4 DNA polymerase and T4 polynucleotide kinase to create blunt-ended fragments with 5'-phosphorylated ends

iv) using Klenow (exo-) with dATP to add a 3'-dA overhang

v) adapter ligation (for hybridization to sequencing flowcell)

vi) gel-purification and selection of ligation products to remove unligated adapters

vii) PCR amplification of the library, followed by PCR clean-up and quantification of library concentration

Any of the steps involving gel purification and size selection are not readily scalable or amenable to automation. To overcome this limitation, a technique called solid phase reversible immobilization (SPRI) was implemented. SPRI involves the use of carboxyl-coated paramagnetic particles to selectively bind nucleic acids in the presence of a buffer containing high concentrations of polyethylene glycol (PEG) and salt (61, 62). DNA associates with these particles in a strictly size-dependent manner. By manipulating the volume ratio of SPRI bead suspension to DNA solution, different length DNA fragments can be selectively isolated. This technique avoids the use of column-based reaction clean-ups and the use of gel electrophoresis for fragment size purification, and can be readily automated using a liquid-handling robotic workstation.

Key engineering challenges to developing a high-throughput library preparation system are designin a PCR method for streamlining Illumina library preparation protocol and automating PCR clean-ups to speed up the recovery of PCR products.

The shortcomings of conventional library preparation protocols for the Illumina platform has motivated the development of a high-throughput method for creating libraries of environmental DNA extracts to facilitate large phylogenetic studies of the human microbiome and other microbial ecosystems. By enabling size selection of sheared genomic DNA for overlapping reads, the Illumina platform can find new application in metagenomics studies. In addition, techniques for addressing PCR bias and artifacts in PCR enrichment of 16S regions will be addressed.

2. *Materials & Methods*

2.1 *Preparation of Genomic DNA*

Extraction of DNA from lake water (Upper Mystic Lake, Middlesex County, MA) and human saliva samples made use of the Qiagen DNeasy Blood & Tissue Kit and proceeded as follows. Filters (25 mm diameter, 0.22 μm pore size; Millipore Express Plus membrane) were placed inside filter holders (Swinnex Filter Holder SX0002500, 25 mm) and autoclaved for 30 min on wet-cycle following assembly. A lysis buffer was made by mixing 20 mM Tris HCl, 2 mM EDTA, and 1.2% Triton-X100 with the final pH adjusted to 8.0 with NaOH. Following preparation, this mixture was autoclaved. Enzymatically-active lysis solution was prepared fresh daily by adding the necessary amount of dry lysozyme (Roche) to the lysis buffer to obtain a 40 mg/mL solution. One sterile 20 mL syringe was needed for each sample processed. The syringe is attached to the top of the filter holder by twist-locking into the female Luer-Lok®. The male Luer slip at the bottom of the filter holder connects to a female Luer slip attached to tubing that directs the fluid a waste vessel. A vacuum pump was used to provide sufficient pressure to pull the full sample volume through the filter. A large vacuum flask with an arm to connect the vacuum pump was used to collect the waste liquid after it passed through the filters.

After filtration the filter holders are opened using sterile technique and the filters with collected biomass are transferred to a sterile Petri dish using tweezers. Using a flame-sterilized razor blade, the filter was cut into 9 pieces (tic-tac-toe cut). Flame-sterilized tweezers are used to transfer the filter pieces to 2 mL zirconium bead-beating tubes and 180 μL of the lysis buffer (without lysozyme) was added. The tube was sealed tightly and inserted into a Mini Beadbeater-1 (BioSpec Products, Inc.) shaker for 1 min at room-temperature at the maximum speed setting 48. Following the cell-disruption procedure, 180 μL of lysis buffer containing 40 mg/mL lysozyme was added and after briefly vortexing the sample to mix, the tube was incubated on a shaking heat block for 1 h at 37°C and 450 rpm. After the 1 h incubation period, 50 μL of Proteinase K was added, then 400 μL of Buffer AL (without ethanol) from the Qiagen DNeasy Kit was added. The sample was gently vortexed to mix and centrifuged briefly to collect the material at the bottom of the tube. The tube was then incubated for 45-60 min at 56°C.

To heat-inactivate the Proteinase K, following the second incubation, the sample was incubated for 5 min at 95°C. The tube was then centrifuged for 1 min at full speed (16.1 rcf) to separate the liquids from the solids. The liquid was then collected (about 800 μL) and transferred to a fresh microcentrifuge tube. Next, 400 μL of molecular-grade ethanol (100-proof) was added, the mixture was vortexed thoroughly at 3000 rpm to mix, and then an aliquot of 500 μL was applied to the DNA collection column

from the DNeasy Kit. The column was centrifuged at 8000 rcf for 30 s. The flow-through was discarded, then the remainder of the DNA solution is loaded onto the column and centrifuged at 8000 rcf for 1 min. Following the second centrifugation, the column was transferred to a new collection tube and 500 µL of AW1 solution (wash 1) was added. The tube was centrifuged for 1 min at 8000 rcf, the flow-through was discarded and the column was transferred to another fresh collection tube. Then, 500 µL of AW2 (wash 2) was added, the tube was centrifuged for 3 min at 20000 rcf, the flow-through and used-collection tube are discarded, and the column was transferred to a fresh 1.5 mL microcentrifuge tube. The column was air-dried for 1 min in a laminar flow hood before 100 µL of AE solution was applied. The column was incubated for 7 min before it was centrifuged for 30 s at 8000 rcf, rotated 180°, then centrifuged again for 1 min at 8000 rcf. This flow-through contained the genomic DNA from the sample, and the material was saved, labeled and stored at -20°C until needed.

Extraction of DNA from human fecal samples made use of the QIAamp DNA Stool Mini Kit (Qiagen) and proceeded as follows. Samples were collected by using a 1 mL sterile pipette tip to core out a small volume of fecal matter from a fresh bowel movement, which was then placed into a sterile 15 mL conical tube and immediately frozen at -80°C until DNA extraction. Prior to DNA extraction, stool samples were thawed to room temperature and weighed. The average weight of the 15 mL tube and 1 mL pipette tip was subtracted from the measured weight to determine the weight of the stool sample inside. This measurement is used to calculate the volume of ASL solution needed (700 µL ASL/100 mg stool). After the appropriate volume of ASL was added to the samples, the contents of one 2 mL tube of garnet beads (0.70 mm, MoBio Laboratories, Inc.) was added to each 15 mL tube. The contents were vortexed vigorously at 3000 rpm (setting 10) for 10 s, then centrifuged at 3000 rcf for 1 min to collect the material at the bottom of the tube. Flame-sterilized tweezers were used to extract the 1 mL pipette tip from each conical tube, and then each sample was resuspended in the ASL by vortexing vigorously for 10 s at 3000 rpm. The tubes were then briefly centrifuged at 3000 rcf for 10-20 s to pellet non-homogenized material.

Next, 1.6 mL of the homogenized stool/ASL suspension was transferred to a 2 mL bead-beating tube containing zirconium/glass beads (0.1 mm, MoBio Laboratories, Inc.). The bead-beating tubes were placed on a Vortex Genie II (Scientific Industries, Inc.) with a horizontal microtube holder and shaken for 10 min at 3000 rpm. Then each tube was placed on a heat block and incubated at 95°C for 5-6 min. The tubes were placed back onto the horizontal tube holder and vortexed for 15 s at setting 7, and then centrifuged for 1 min at 20000 rcf and 20°C. Following centrifugation, 1.2 mL of the supernatant from each centrifuged sample was transferred to a clean 2 mL microcentrifuge tube, and

one Inhibitex tablet was added to each. The sample and tablet were vortexed for 1 min at 3000 rpm to completely dissolve the tablet, incubated at room temperature for 1 min, then centrifuged for 3 min at 20000 rcf and 20°C. All of the supernatant was then transferred to a fresh 1.5 mL microcentrifuge tube, and centrifuged again for 3 min at 20000 rcf and 20°C.

200 µL of supernatant from this centrifuge step was transferred to a fresh 1.5 mL microcentrifuge tube for each sample and 15 µL of Proteinase K was then added. Following addition of Proteinase K, 200 µL of AL buffer is added to each sample. The tubes were vortexed briefly to mix, then centrifuged to collect drops in the bottom of the tube. Then, all of the tubes were placed on a heat block at 70°C for 10 min. After the 10 min incubation, the tubes were quickly centrifuged to collect drops, then 200 µL of molecular-grade ethanol (100-proof) was added and mixed in by vortexing briefly. This solution was transferred onto a QIAamp spin column, taking care not to wet the top rim of the column, then centrifuged for 1 min at 20000 rcf and 20°C. Both the flow-through and tube were discarded, each column was placed into a new 2 mL tube, and 500 µL of AW1 was added. The tubes were centrifuged for 3 min at 20000 rcf and 20°C. Both the flow-through and tubes were discarded again, the columns were placed into new 2 mL tubes, and 500 µL of AW2 was added. The tubes were centrifuged for 30 s at 20000 rcf, then rotated 180° in the rotor, and spun again for 1 min at 20000 rcf and 20°C. The columns were placed into new 1.5 mL microcentrifuge tubes and air-dried for 1 min in a laminar flow hood. After the drying step, 200 µL of AE was added onto the membrane and allowed to incubate for 7 min at room temperature. Finally, the DNA was harvested by repeating the 2-step centrifugation procedure (30 s at 20000 rcf, rotate 180°, 1 min at 20000 rcf). The samples were labeled and stored at -20°C until needed.

2.2 *Procedure for Shearing of Genomic DNA Isolated from Environmental/Clinical Samples*

Genomic DNA (gDNA) was sheared via ultrasonication using a Bioruptor® sonicator (UCD-200, Diagenode) to simultaneously process up to 6 genomic samples in 1.5 mL microcentrifuge tubes. DNA samples were sheared using 18-24 cycles of alternating 30 s ultrasonic bursts and 30 s pauses in a 4°C water bath. The size distribution of the resulting fragments ranged from ~100 to ~800 bp, as determined by Agilent Bioanalyzer DNA 1000 assays (Figure 6 and Figure 7). For a detailed description of the Bioanalyzer capillary electrophoresis system, please refer to Panaro NJ, et al (63).

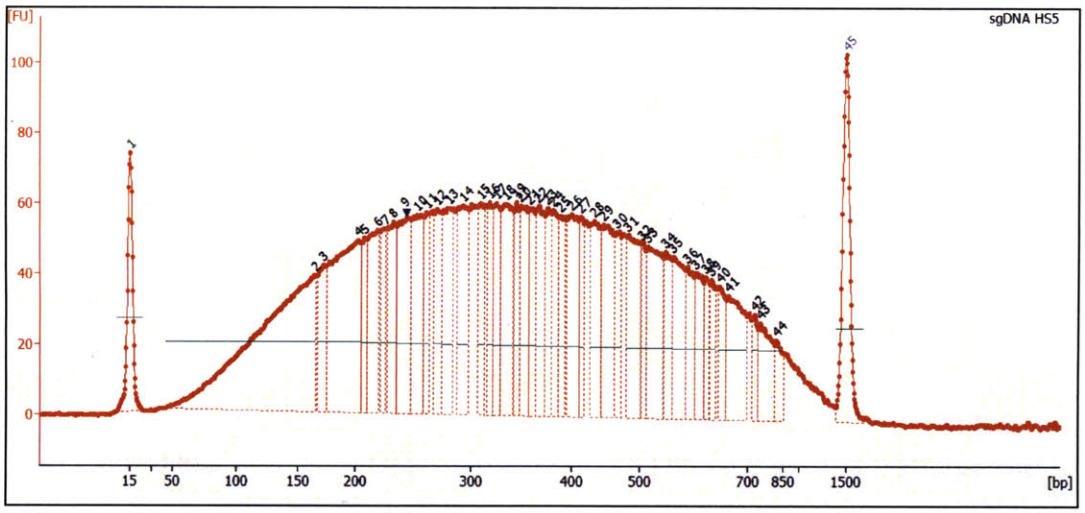

Figure 6. Example of electropherogram output from Bioanalyzer DNA 1000 Assay of sheared genomic DNA (sgDNA). The y-axis indicates arbitrary fluorescence units, and the x-axis indicates DNA fragment length in base pairs. The two sharp peaks on either side of the sample readout are markers used to calculate both the fragment length and concentration of DNA, as compared to a DNA ladder control run with each assay.

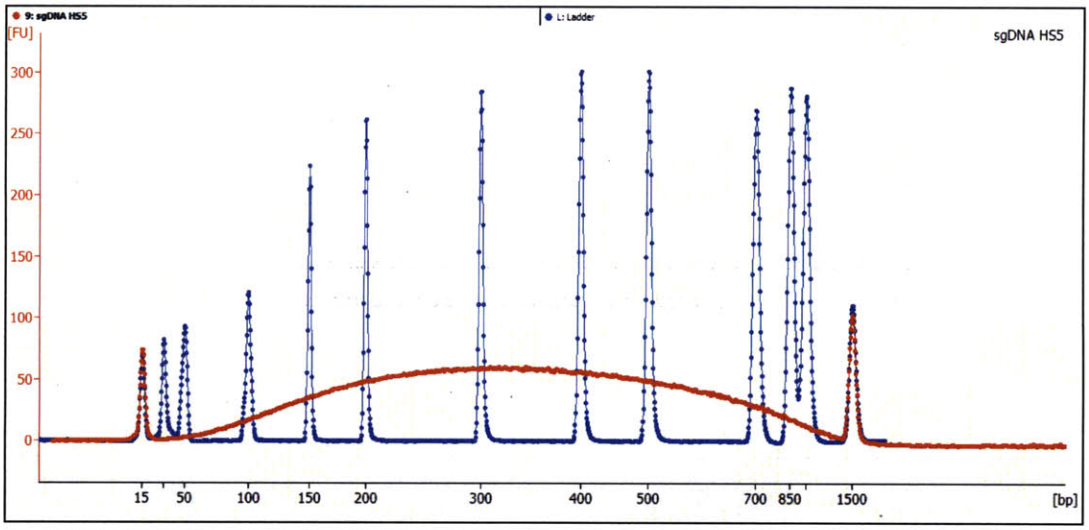

Figure 7. Overlay of electropherogram outputs of sheared genomic sample with the ladder control. The ladder is used in determining fragment lengths and DNA concentration. By normalizing the fluorescence of the sgDNA sample against the ladder, the concentration is determined. DNA fragment length is found by correlating the time for specific ladder peaks to be visualized.

2.3 *SPRI DNA Fragment Size Selection*

2.3.1 *Single-SPRI (sSPRI) Reaction Clean-up*

The sSPRI clean-up procedure was developed to provide a method for primer/dimer removal from PCR reactions, and buffer exchange. To begin, AMPure XP magnetic beads (Agencourt) were removed from 4°C storage and vortexed at 1600 rpm for 15-30 s to resuspend and provide a homogeneous mixture for aliquots. The volume of PCR to be cleaned was then adjusted to 50 µL by

adding Buffer EB (Qiagen). For a 50 µL PCR, 45 µL of SPRI beads are required (larger PCR volumes can be used by maintaining the 50 µL DNA: 45 µL SPRI solution ratio). One 45 µL aliquot of SPRI beads per PCR sample was placed in a 1.5 mL microcentrifuge tube and allowed to equilibrate to room temperature. Then, 50 µL PCR mixture was added to the tube containing the SPRI beads, vortexed briefly at 1600 rpm to mix, and incubated at room temperature for 5-7 min to bind the DNA.

Next, each tube was placed on a DynaMag™ (Invitrogen) or similar magnetic separator for 2 min to form a SPRI pellet, and the supernatant was removed and discarded while on the magnet. While still on the magnet, the pellet was gently washed twice with 70% ethanol in water, allowing the pellet to stay submerged in the wash solution for 15-30 s during each wash. Once all of the ethanol from the second wash was removed, the pellet was left to dry for 15 min, while on the magnet. For DNA elution, the tube was removed from the magnet, 20 µL of Buffer EB or sterile deionized water was added, and the tube was vortexed at 2000 rpm until the pellet was completely resuspended. After incubating in the elution liquid for at least 1 min, the tube was placed back onto the magnet for 2 min until a pellet formed again. The supernatant (EB or water) was then carefully collected and transferred to a fresh 1.5 mL microcentrifuge tube.

2.3.2 *Double-SPRI (dSPRI) DNA Fragment Size Selection*

The dSPRI size selection procedure is virtually the same as the sSPRI procedure, except after the first separation on the magnet, the supernatant is saved rather than discarded. This supernatant contains smaller fragments of DNA which were competitively inhibited from binding to the original 45 µL of SPRI beads by larger DNA fragments. DNA is negatively-charged and preferentially binds to the polymer on the surface of the magnetic beads in the presence of the buffer. Longer DNA fragments carry larger charges and despite being slower to diffuse to the beads due to their size, their electrostatic interaction with the beads is more enthalpically favorable, allowing them to displace shorter, less charge-dense DNA fragments. To isolate different size fragments from the DNA contained in the supernatant, a varying volume of fresh SPRI beads (45-100 µL) is added, vortexed to mix, and incubated for 5 minutes, before placing on the magnet for 2 min to form a pellet. The supernatant from this step is collected and discarded, the pellet is washed twice with 70% ethanol in water (as in sSPRI), allowed to dry for 15 min, then eluted in water or Buffer EB. After a pelleting on the magnet for 2 min, the supernatant is collected and transferred to a fresh 1.5 mL microcentrifuge tube.

2.4 *Primer-Clipping Library Preparation (16S-specific)*

Amplification of 16S rDNA hypervariable regions is carried out using 5'-biotinylated universal primers featuring T→U substitutions at the 3' termini. PCR is carried out using primers V4-U515-3'dU-B-F and V5-U926-3'dU-B-R, which are complementary to the universally-conserved regions at positions 515 and 926, respectively. A 1× master mix (25 µL reaction volume) contained 16.25 µL H_2O, 2.5 µL Turbo Cx Buffer (10x), 0.5 µL dNTP (10 mM), 2.5 µL of both the forward and reverse primers (5 µM), 0.25 µL undiluted template DNA (genomic DNA), and 0.5 µL of Pfu Turbo Cx Hotstart DNA polymerase (Agilent Technologies). For larger reaction mixtures (typically an 8× master mix was prepared), the total volume was aliquoted into 25 µL volumes in a PCR tube strip. The thermal cycling scheme involved a denaturation step of 95°C for 40 s, an annealing step of 52°C for 30 s, and an elongation step of 72°C for 1 min. This was repeated for 30 cycles total.

After cycling, the 25 µL aliquots were pooled to obtain a total of 200 µL of PCR product. This solution was cleaned up using the MinElute PCR Clean-Up Kit (Qiagen). For a 200 µL reaction volume, 1000 µL of Buffer PB was added, vortexed briefly to mix, and applied to a column with collection tube. The tubes were centrifuged for 1 min at 16100 rcf, and the flow-through was discarded. Then, 750 µL of PE was added to each column, the tubes were centrifuged for 1 min at 16100 rcf, and the flow-through was discarded. Next, each column was dried by centrifuging again for 30 s at 16100 rcf, rotated 180° in the rotor, and spun again for 1 min at 16100 rcf. The columns were then transferred to a fresh 1.5 mL microcentrifuge tube, and left open in a laminar flow hood to dry for 2 min. Following the air-drying step, 10 µL of EB (elution buffer) or water was added directly onto the center of the column and incubated for 7 min at room temperature. Following the incubation period, the samples were centrifuged for 30 s at 16100 rcf, rotated 180°, then centrifuged for 1 min at 16100 rcf. The flow-through from this step contains the DNA from the PCR reaction and was saved and stored at -20°C. Prior to performing the next enzymatic reaction, the DNA concentration and purity of each sample was measured using a NanoDrop 1000 (Thermo Fisher Scientific Inc.).

The next step in this library preparation procedure requires the USER™ Enzyme (Uracil-Specific Excision Reagent; New England Biolabs, Inc.) which is used to generate a single nucleotide gap at the location of a uracil. The USER Enzyme contains a mixture of two nucleases: Uracil DNA glycosylase (UDG) and DNA glycosylase-lysase Endonuclease VIII. UDG is used to catalyze the excision of a uracil base while leaving the phosphodiester backbone intact, whereas the lyase activity of the Endonuclease VIII functions to sever the backbone at the 3' and 5' sides for the abasic site to release base-free deoxyribose. USER-digest treatments were run in 50 µL volumes, where 5 µL of USER are required for

every 10 pmol of uracil. To calculate the pmol/µL of uracil in the PCR product, the average molecular weight of a DNA basepair was assumed to be 650 µg/µmol, and using the U515 and U926 primer set, a 412 bp product is expected. Since the final PCR product will have 2 uracils (one in each strand of the duplex DNA), the conversion equation becomes:

$$\text{Conc of DNA } \left(\frac{\text{ng}}{\mu\text{L}}\right) \times \frac{1\ \mu\text{g}}{1000\ \text{ng}} \times \frac{\mu\text{mol}}{650\ \mu\text{g}} \times \frac{1}{412} \times 10^6 \frac{\text{pmol}}{\mu\text{mol}} \times 2\ \text{uracil} = \frac{\text{pmol uracil}}{\mu\text{mol PCR product}}$$

A 50 µL USER master mix required a variable volume of purified PCR product (such that the amount of uracil was accurate for the required reaction chemistry), 10 µL of 5x Phusion® HF Buffer (Thermo Fisher Scientific Inc), 5 µL of USER Enzyme (New England Biolabs), and a variable volume of sterile deionized water to reach 50 µL. The reaction was run at 37°C for 1 h. For this step, it was important to use as much purified PCR product as possible, since DNA is lost in downstream reaction clean-up steps.

Following USER treatment, samples may be frozen at -20°C or carried on to the next step involving solid-phase capture of the biotinylated, uracil-clipped PCR products using Dynabeads® MyOne™ Streptavidin T1 (Invitrogen) magnetic beads. To prepare the streptavidin-coated Dynabeads, they were removed from +4°C and vortexed at 2000 rpm for 20 s. A 25 µL aliquot of the bead solution was transferred to a microcentrifuge tube and placed on a DynaMag™ (Invitrogen) magnetic separator stand for 2 min. After the 2 min incubation, a pellet of magnetic beads will have formed on the back of the tube nearest the magnetic core of the stand.

While on the magnet, the supernatant (SN) was removed and discarded. The tube was taken off the magnet, and twice the original bead volume (50 µL) of 2x concentration binding and washing buffer (B/W buffer; 10 mM Tris-HCl, 1 mM EDTA, 2 M NaCl, pH adjusted to 7.5 and autoclaved before use) was added by rinsing over the pellet, followed by vortexing briefly to resuspend the beads. The tube was then placed back on the magnet for another 2 min incubation. This process was repeated for a total of 3 washes with the B/W buffer. After the third wash and removal of SN, the tube was taken off the magnet, and 50 µL of B/W buffer (2x conc.) was added to resuspend the beads. The volume of DNA solution to be cleaned was adjusted to 50 µL with sterile deionized water so that the volume of DNA was 1:1 with the B/W buffer and beads, then the two volumes were mixed together (100 µL total volume) and vortexed briefly. The mixture was incubated for 20 min at room-temperature on a tilted shaking heat block (390 rpm), and monitored every five minutes to ensure none of the beads had settled to the bottom. At the end of the 20 min incubation, the tube was placed on the magnet for 3 min to separate

the beads. The SN was removed while the tube was on the magnet, taking care not to disturb the bead pellet.

A second series of washes was then performed to remove unbound DNA. The tube was taken off the magnet and 50 µL of B/W buffer (at 1x, made by dilution of 2x buffer with sterile deionized water) was added by gently rinsing over the pellet. The tube was then placed back on the magnet for 2 min to reform a pellet. The SN was removed and discarded. This was repeated for a total of 3 washes. After the third wash and removal of SN, 40 µL of elution buffer (Buffer EB, Qiagen; 10 mM Tris-Cl, pH 8.5) was added, and the tube was gently inverted to suspend the beads. For denaturation of the DNA, the tube was placed on a tilted shaking heat block at 300 rpm and 59°C for 40 min. After the 40 min incubation, the tube was quickly removed from the heat block and placed on the magnet for 45-50 s to form a pellet before carefully collecting all of the supernatant and transferring it to a fresh microcentrifuge tube. This short magnetic separation time was necessary to prevent re-annealing of the DNA to the small DNA fragment still attached to the streptavidin beads.

Although the magnetic streptavidin-coated bead isolation procedure is capable of buffer exchange, effectively 'cleaning' the USER reaction products, a sSPRI clean-up step was included to further purify DNA products. Prior to performing the next enzymatic reaction, the DNA concentration and purity of each sample was measured using a NanoDrop 1000 (Thermo Fisher Scientific Inc).

The final enzymatic treatment in the primer-clipping library preparation system involved treatment using S1 nuclease (Invitrogen) to chew back single-stranded DNA, producing blunt ends for Illumina adaptor ligation. Prior to use, the stock S1 nuclease (at ~1000 units/µL) is diluted to a working solution of 5 units/µL using the dilution buffer included with the enzyme. The necessary ratio of enzyme to DNA was 2 µL nuclease per µg of DNA (10 units/µg DNA, where 1 µL S1 nuclease = 5 units). S1-digests were run in 50 µL reactions (similar to USER treatments), which required 5 µL of 10x S1 Nuclease Buffer, 5 µL of NaCl (3M), a variable amount of DNA and enzyme, followed by enough sterile deionized water to reach 50 µL total volume. The digest was run at 30°C for 1 h, then stored at -20°C. The final step in the primer-clipping protocol involves using a minElute PCR Clean-Up Kit (Qiagen), which was described earlier. For a 50 µL reaction, 250 µL of PB were used. The product was eluted in 10 µL of EB or water to ensure a high concentration of DNA.

After primer-clipping, the final amplicon is $412 \text{ bp} - 17 \text{ bp} - 14 \text{ bp} = 381 \text{ bp}$ long (Figure 8). In order to complete library preparation for Illumina sequencing, adapter sequences must be ligated to both ends of the 381 bp product, then enriched using amplification primers.

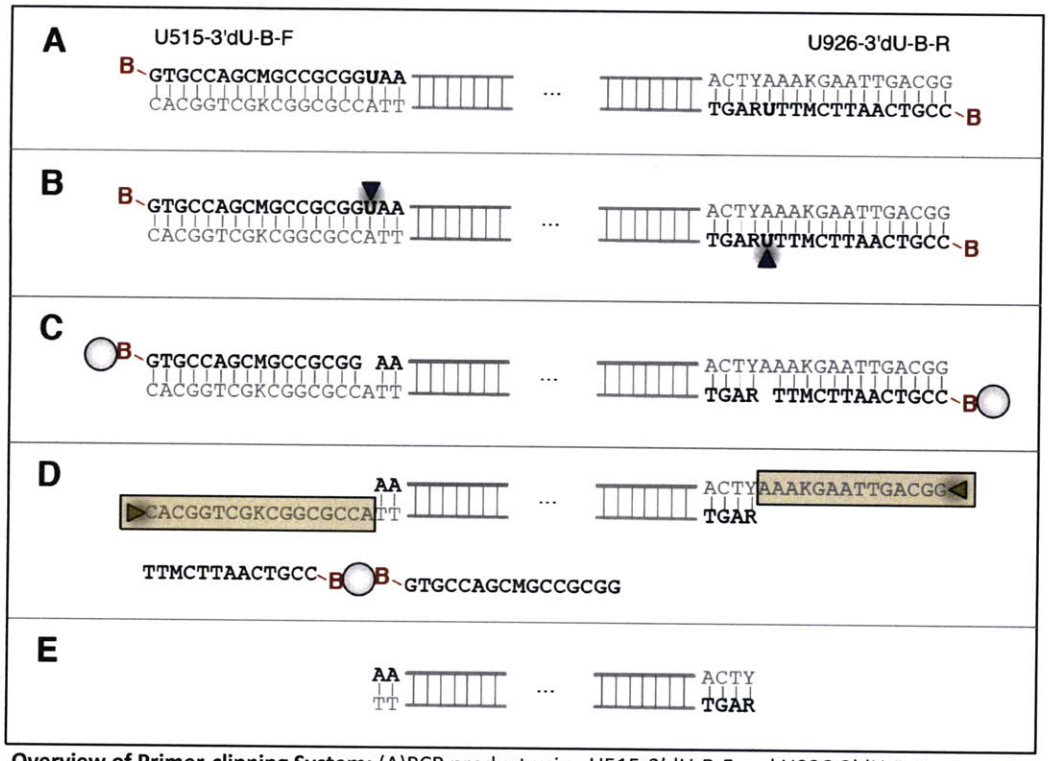

Figure 8. Overview of Primer-clipping System; (A)PCR product using U515-3'dU-B-F and U926-3'dU-B-R primer set where B indicates the location of biotin. (B)3'-uracil base is recognized and removed by the activity of USER enzyme mix (uracil-specific excision reagent). (C)PCR product is immobilized on streptavidin-coated magnetic beads through the strong noncovalent interaction of 5'-biotin and streptavidin (K_d = 4×10-14 M) (64). (D)After heat-denaturing the DNA to separate the parts of the PCR product containing the original primer sequence from the rest of the DNA, the overhang is treated with S1 exonuclease to produce blunt-ended product. (E) The final product following primer-clipping no longer contains the universal primer sequence.

2.5 *Two-step PCR Primer-Skipping Library Preparation (16S-specific)*

In the first step, amplification of 16S rDNA hypervariable regions is carried out using primers V4-U515-IAlm3.1-F (me16S-F) and V5-U926-R (me16S-R), which are complementary to the universally-conserved regions at positions 515 and 926, respectively (Figure 9). The forward primer (U515-IAlm3.1-F) contains some non-complementary sequence at the 5'-end which is used in the second amplification step to improve step-2 primer binding.

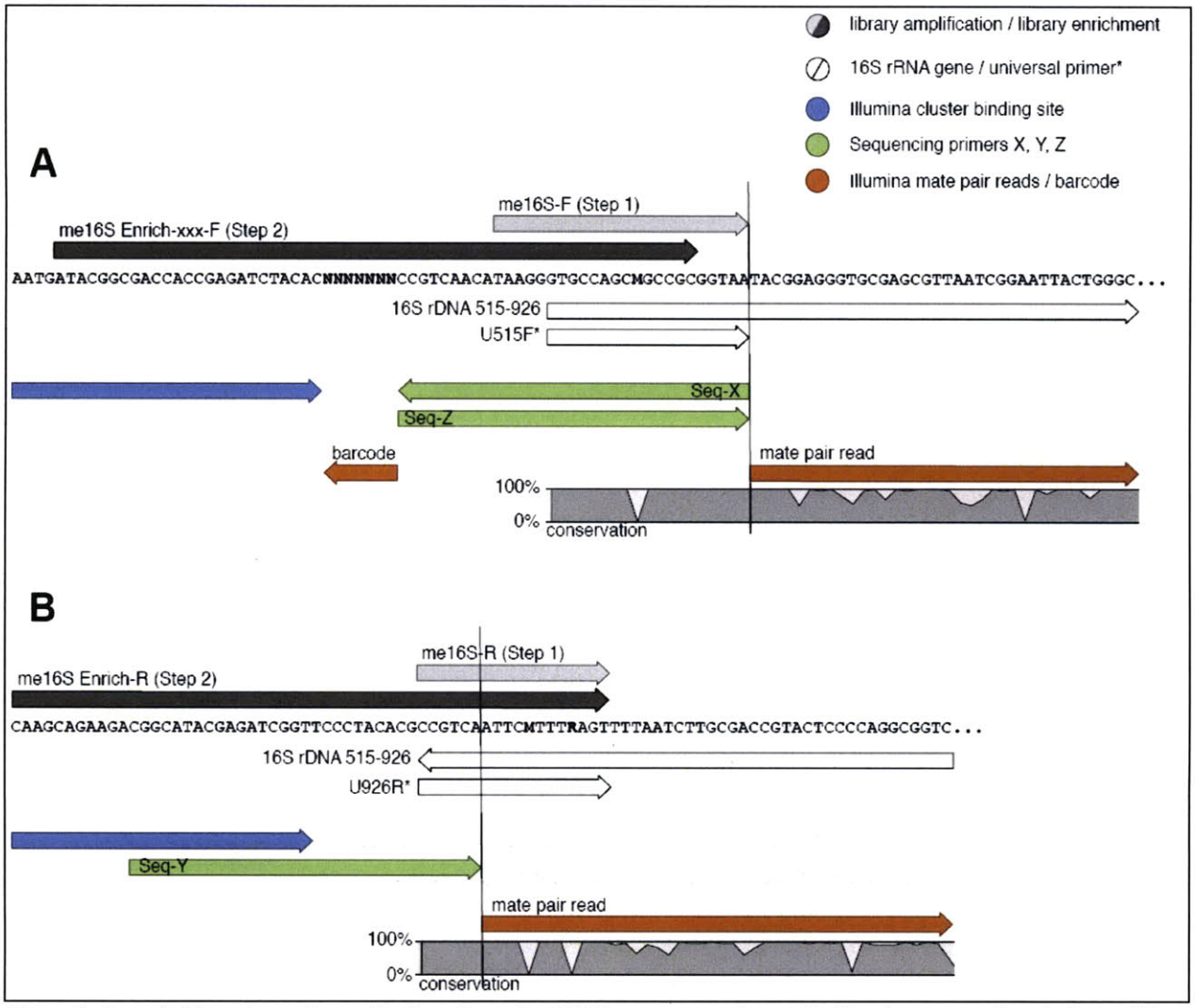

Figure 9. Barcoded two-step, primer-skipping scheme for 16S library preparation on the Illumina sequencing platform. This figure depicts the sequence of the primers used in library preparation (Step 1/Step 2), the Illumina adapter sequence for clustering and the binding sites of the sequencing primers. The mate-pair read and barcode are data obtained from sequencing.

Step 1 reactions were made in 100 µL total volume per template, which was then partitioned into four 25 µL volumes on a PCR-strip for thermal cycling. A 1x mastermix (25 µL, total volume) contained 12.25 µL of sterile deionized water, 5 µL of 5x Phusion® HF Buffer, 0.5 µL of dNTP mix (10 mM), 2.5 µL of each primer (3 µM), 2 µL of template (diluted to ~10 ng/µL), and 0.25 µL of Phusion® DNA Polymerase (Finnzymes, distributed by NEB). The Phusion® polymerase was chosen for its high-fidelity, with an error rate 50-fold lower than Taq polymerase. The thermal cycling scheme involved an initial denaturation step of 98°C for 30 s, followed by cycles containing a denaturation step of 98°C for 30 s, an annealing step of 52°C for 30 s, and an elongation step of 72°C for 15 s. This was repeated for 25 cycles total.

Immediately after cycling, the volumes from each of the four 25 µL reactions were pooled into a single microcentrifuge tube and a sSPRI clean-up procedure was used to obtain purified PCR product.

In the second step, amplification of each purified product from the Step 1 reaction (each originating from a unique genomic template) was carried out using a unique forward primer containing a 7-nucleotide barcode, PE-PCRxxx-1 (where -xxx is a unique numerical tag from -000 to -127), and a common reverse primer PE-PCR-2 (in both primers, PE stands for 'primer-enrichment'). Both the forward and reverse Step 2 primers contain sequence that allows them to hybridize with immobilized primers on the Illumina flowcell, allowing them to participate in 'clustering' prior to sequencing (Figure 10 and Figure 11). They also contain sequence that enables specially-designed sequencing primers X & Y to bind for obtaining sequence information in the hypervariable 16S region of the DNA insert, while sequencing primer Z is used to obtain the sequence of the 7-bp barcode in a third sequencing run. By using these primers, the universal primer sequence is not included in the sequencing readout, effectively increasing the informative read-length.

Similar to Step 1 reactions, Step 2 reactions were carried out in 100 µL total volume per template, which was then partitioned into four 25 µL volumes on a PCR-strip for thermal cycling. A 1x mastermix (25 µL, total volume) contained 10.65 µL of sterile deionized water, 5 µL of 5x Phusion® HF Buffer, 0.5 µL of dNTP mix (10 mM), 3.3 µL of each primer (3 µM), 2 µL of template (purified PCR product from Step 1), and 0.25 µL of Phusion® DNA Polymerase. The thermal cycling scheme involved an initial denaturation step of 98°C for 30 s, followed by cycles containing a denaturation step of 98°C for 30 s, an annealing step of 83°C for 30 s, and an elongation step of 72°C for 15 s. This was repeated for 15 cycles total. Immediately after cycling, the volumes from each of the four 25 µL reactions were pooled into a single microcentrifuge tube and a sSPRI clean-up procedure was used to obtain purified sequence-ready product.

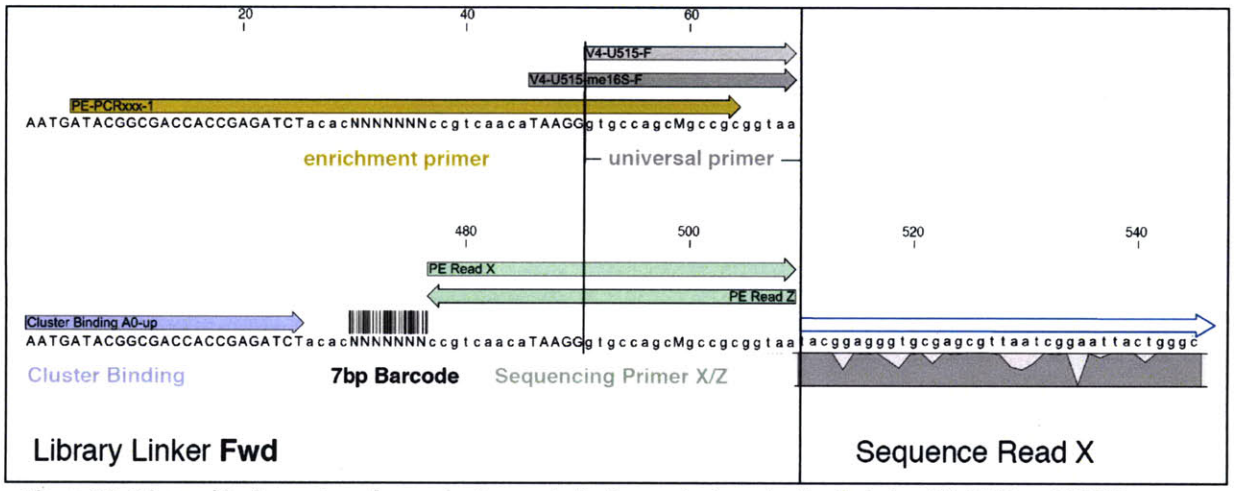

Figure 10. Primer-skipping system, forward primer set; the forward primer for the first step (V4-U515me16S-F) has some sequence not complementary to the universal region, but allows the barcoded forward primer for the second step (PE-PCRxxx-1) to hybridize. This barcoded primer also includes the sequence necessary for binding to the primer lawn on the Illumina flowcell. Sequence Read X collects sequence data in the V4 hypervariable region, whereas Sequence Read Z only collects the 7 nucleotide barcode.

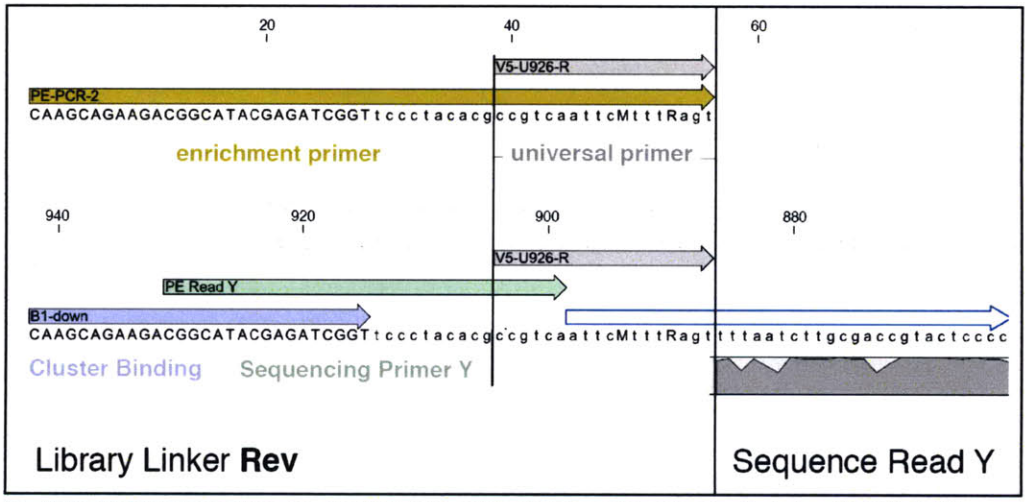

Figure 11. Primer-skipping system, reverse primer set; the reverse primer for the first step (V5-U926-R) has no added sequence for attachment of the second step primer (PE-PCR-2). PE-PCR-2 does contain the sequence necessary for binding to the primer lawn on the Illumina flowcell. Sequence Read Y collects sequence data in the V5 hypervariable region.

2.6 *Cycle Optimization to Reduce Chimera Formation and PCR Bias*

It is well known that PCR amplification of highly conserved genes from mixed templates can lead to artifacts that diminish informative reads during sequencing and disrupt our understanding of microbial ecosystems. The amount of PCR-generated chimeras should theoretically decrease with cycle number, since there are less opportunities for recombination events to occur. To incorporate this knowledge into our library preparation protocol, we developed a technique referred to as cycle-

optimized PCR. Prior to preparing large sets of 16S libraries using universal primers, a subset of samples was selected and their concentrations normalized to 10 ng/μL by dilution with sterile water. The diluted samples were then subjected to a 50 cycle quantitative-PCR using SYBR Green as the fluorescent indicator. The optimal-cycle was arbitrarily chosen to be the cycle number reached when the majority of the samples approached the mid-point of the PCR linear phase. This technique complements the statements made previously (40, 45) about the effect of decreasing cycle number on PCR bias and artifacts.

Quantitative real-time PCR was carried out in triplicate for each template examined to determine the optimal-cycle number. A 1x master mix (25 μL, total volume) for Step1/Step2 contained 12.125/10.525 μL sterile deionized water, 5 μL of 5x Phusion® HF Buffer, 0.5 μL of dNTP mix (10 mM), 2.5/3.3 μL of each primer (3 μM), 2 μL of template (or purified PCR product from Step 1), 0.125 μL SYBR Green I (at 1/10,000 dilution in 1x TE, TAE or TBE buffer; TE = 10 mM Tris-HCl, 1 mM EDTA, pH 8.0; TBE = 89 mM Tris base, 89 mM boric acid, 1 mM EDTA, pH 8.0; TAE = 40 mM Tris-acetate, 1 mM EDTA, pH 8.0) and 0.25 μL of Phusion® DNA Polymerase. The reaction was aliquoted into PCR tubes or 96-well plate with optical adhesive sleeve and run on an Opticon™ real-time PCR detector (MJ Research). The thermal cycling protocol began with an initial denaturation step of 98°C for 20 s, followed by cycles containing a denaturation step of 98°C for 30 s, an annealing step of 52°C or 83°C (Step 1 and Step 2, respectively) for 30 s, and an elongation step of 72°C for 15 s. This was repeated for 50 cycles total. At least 3 replicates per sample and 3 replicates of a non-template (water) control were used for each experiment.

2.7 *Emulsion PCR to Reduce Chimera Formation and PCR Bias*

An emulsion PCR protocol was developed by incorporating reagents and techniques from two references (53, 54). Briefly, our procedure involves separately preparing an oil/surfactant phase and an aqueous/PCR reaction phase, producing micelles by vigorously vortexing these phases together on a horizontal microtube holder, aliquoting the resulting emulsion into a PCR tube strip, thermal cycling, breaking the emulsion with a high salt solution via osmotic pressure, and using ethanol precipitation to harvest the PCR products.

The oil phase was prepared by adding 9 mL of mineral oil to a 15 mL Falcon tube, then adding 450 μL of Span 80 surfactant (Fluka), 40 μL of Tween 80 surfactant (Sigma-Aldrich), and 5 μL of Triton X-100 surfactant (Fisher Scientific). More mineral oil was then added to obtain a final volume of 10 mL. The oil phase was mixed by vortexing while inverted at 1600 rpm for 2 min, filter-sterilized and incubated overnight at +4°C to allow air bubbles to surface before using.

The aqueous phase was prepared in the same manner as other PCR mastermixes described previously, except 50 µL total volume per template was used (2x mastermix). For Step 1 PCR, this equates to 24.5 µL of sterile deionized water, 10 µL of 5x Phusion® HF Buffer, 1 µL of dNTP mix (10 mM), 5 µL of each primer (3 µM), 4 µL of template, and 0.5 µL of Phusion® DNA Polymerase. For Step 2 PCR, this equates to 21.3 µL of sterile deionized water, 10 µL of 5x Phusion® HF Buffer, 1 µL of dNTP mix (10 mM), 6.6 µL of each primer (3 µM), 4 µL of template (purified PCR product from Step 1), and 0.5 µL of Phusion® DNA Polymerase.

For emulsification, 200 µL of the oil phase was added to each 50 µL PCR in a 1.5 mL microcentrifuge tube. The oil and aqueous phases were blended on Vortex Genie II (Scientific Industries, Inc.) with horizontal bead-beating adapter at 3000 rpm (or maximum speed) for 2 min. Shorter emulsification times were found to produce micelles that were too large, while longer vortexing times produced smaller micelles. The emulsion PCR mix was then separated into 4 PCR tubes, each containing approximately 50 µL (the emulsion mixture was very difficult to pipette accurately since the oil phase was extremely viscous). The thermal cycling protocol began with an initial denaturation step of 98°C for 20 s, followed by cycles containing a denaturation step of 98°C for 30 s, an annealing step of 52°C or 83°C (Step 1 and Step 2, respectively) for 30 s, and an elongation step of 72°C for 15 s. This was repeated for 45 cycles total. The high cycle number was chosen to give each of the compartmentalized reactions sufficient opportunities to amplify.

Emulsion breaking required pooling the emulsion volumes from the 4 PCR tubes back into a 1.5 mL microcentrifuge tube. Accounting for liquid entrainment on the walls of the PCR tubes and pipette tip used for pooling, only about 185 µL of the original 200 µL is collected. After pooling, ammonium acetate (10 M) is added the emulsion to a final concentration of 2 M (for 185 µL of emulsion, 46.25 µL of 10 M ammonium acetate is needed). The ammonium acetate and emulsion is briefly vortexed at 3000 rpm for 5 s. Next, one volume of isopropanol is added, such that there is a 1:1 ratio of isopropanol to emulsion/ammonium acetate (185 µL emulsion + 46.25 µL ammonium acetate = 231.25 µL of isopropanol needed). This mixture is vortexed thoroughly at full speed (3000 rpm) for 15-30 s, then incubated on ice for 15 min. Following the ice incubation, the samples are centrifuged at 20,000 rcf for 10 min at 4°C. The supernatant is removed and discarded, then 150 µL isopropanol is added and vortexed briefly to mix. The sample is again centrifuged at 20,000 rcf for 10 min at 4°C, then the supernatant is removed and discarded, and 150 µL of 70% ethanol in water is added. The sample is then centrifuged a third time at 20,000 rcf for 10 min at 4°C, the supernatant is removed and discarded, and

each sample is left in a laminar-flow hood for 15 min to air-dry. The DNA is finally eluted by resuspending in 20 µL of Buffer EB or sterile deionized water.

2.8 *Visualization of Emulsions*

To determine the correct template dilution and vortexing conditions for preparing emulsion PCRs, a variety of conditions were tested and prepared. These samples were then imaged on a Nikon Eclipse TE2000-E to gather rough estimates on micelle size distributions, and the frequency of successful amplifications by counting fluorescent droplets stained with fluorescein/uranine (0.034 mg/mL, TCI America) or Quant-iT™ PicoGreen (Invitrogen) following thermal cycling.

Emulsion samples were placed on glass slides or polystyrene dishes, with glass coverslips to produce monolayers of emulsion droplets for imaging. In some cases, emulsions were diluted with mineral oil to improve image quality and fluorescent detection. For fluorescein staining, the aqueous PCR phase was replaced with a dilute solution of fluorescein in water. For PicoGreen staining, each 50 µL ePCR was run for 15-35 cycles, then stained with 0 to 2.5 µL of PicoGreen (undiluted), incubated at room-temperature for 2 min in the dark, then imaged.

2.9 *Automation of Library Preparation*

Automation of the two step Primer-skipping library preparation and SPRI reaction clean-up protocols was enabled by training requisite labware and developing command scripts for a Freedom EVO 150 (Tecan) robotic liquid-handling workstation. The EVO 150 is equipped with a four-channel liquid handling arm (LiHa) and 96-channel pipette arm (MCA96) that can accurately handle volumes between 1 and 500 µL. The platform also includes two thermal control (4°C and higher) plate racks, and two ambient plate racks. To increase the throughput of SPRI clean-up, a 96-well SPRIPlate Super Magnet Plate (Agencourt) was obtained for preparing 96 clean-ups in parallel.

Robotic commands were compiled into program scripts for individual steps of library preparation. Each piece of labware (tip-boxes, troughs, skirted and unskirted 96-well plates, SPRIPlate Super Magnet, microcentrifuge tube rack, and labware-combinations) on the working deck was meticulously trained. The x- and y- coordinates, and the associated z-travel, z-start, z-dispense, and z-max coordinates for each individual well on each type of labware were selected to prevent the creation of obstacles for the robotic liquid-dispensing arms. The only human involvement in library preparation

involves physically moving robotically-prepared 96-well plates to a thermal cycler, and moving 96-well SPRI reactions on and off of the SPRIPlate Super Magnet when prompted by the program.

3. *Results & Discussion*

3.1 *Bioanalyzer Results from Development of SPRI Protocols*

A technique for enabling gel-less size selection of sheared genomic DNA fragments, as well as a method for replacing column-based reaction clean-ups, enabling PCR product purification to be automated, was developed by extensive testing with AMPure XP (Agencourt) SPRI beads. The dSPRI protocol for efficient capture of DNA fragments is tunable with respect to fragment lengths by varying the ratio of SPRI beads to DNA solution in two consecutive DNA-binding reactions. Fragment lengths from each elution were characterized using Bioanalyzer DNA 1000 or High Sensitivity DNA Assays (Agilent). Integration under each curve was used to quantify the concentration and molarity of the DNA. An extensive set of experiments were conducted with sheared genomic DNA to determine the necessary parameters for isolating the desired range of DNA fragment lengths (Table 2 and Figure 12, Figure 15, Figure 16, and Figure 17).

Table 2. Matrix of experiments performed to develop the dSPRI protocol.

	Separation 1		Separation 2			
Method	Volume Ratio SPRI beads: DNA	Incubation time (min)	Volume Ratio SPRI beads: Sep1 vol	Incubation time (min)	Average fragment length (bp)	Fragment length CV (%)
A	65:50 (µL); 1.3	20	100:115 (µL); 0.87	15	134	25.5
B	65:50 (µL); 1.3	20	60:115 (µL); 0.52	15	139	23.3
C	45:50 (µL); 0.9	20	55:95 (µL); 0.58	7	187	26.8
D	45:50 (µL); 0.9	20	40:95 (µL); 0.42	7	195	24.8
E	45:50 (µL); 0.9	20	30:95 (µL); 0.32	7	205	23.2
F	45:50 (µL); 0.9	20	20:95 (µL); 0.21	7	218	22.1
G	45:50 (µL); 0.9	20	10:95 (µL); 0.11	7	243	18
H	40:50 (µL); 0.8	5	100:85 (µL); 1.18	5	207	32.4

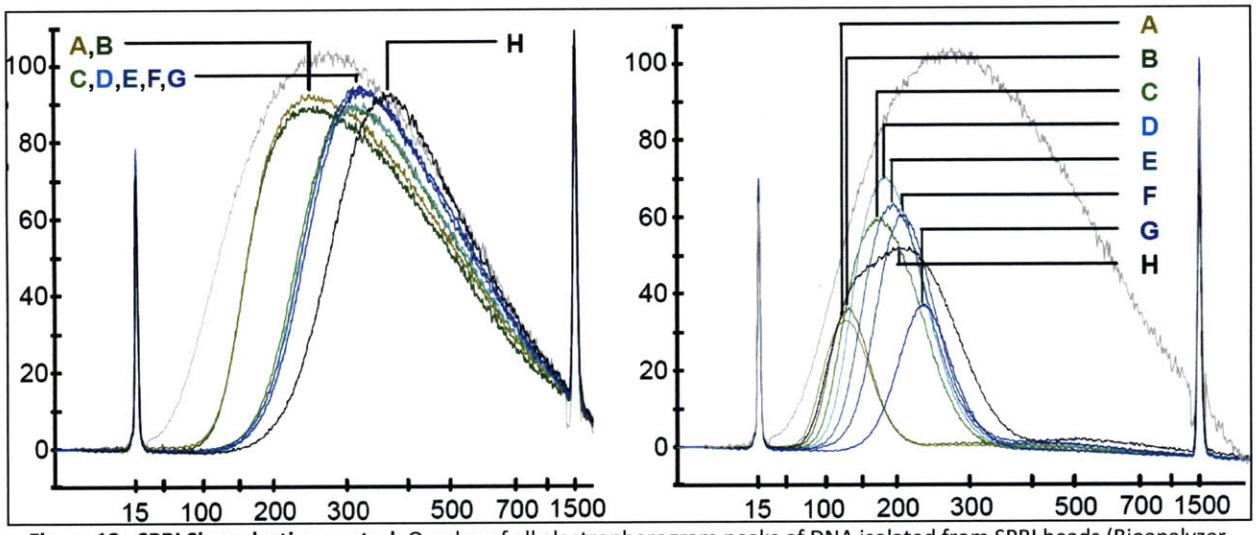

Figure 12. SPRI Size selection control; Overlay of all electropherogram peaks of DNA isolated from SPRI beads (Bioanalyzer DNA 1000 assay) in the first separation (left) or second separation (right); x-axis = DNA fragment length (bp), y-axis = arbitrary fluorescence units for the experiments (A-H) listed in Table 2

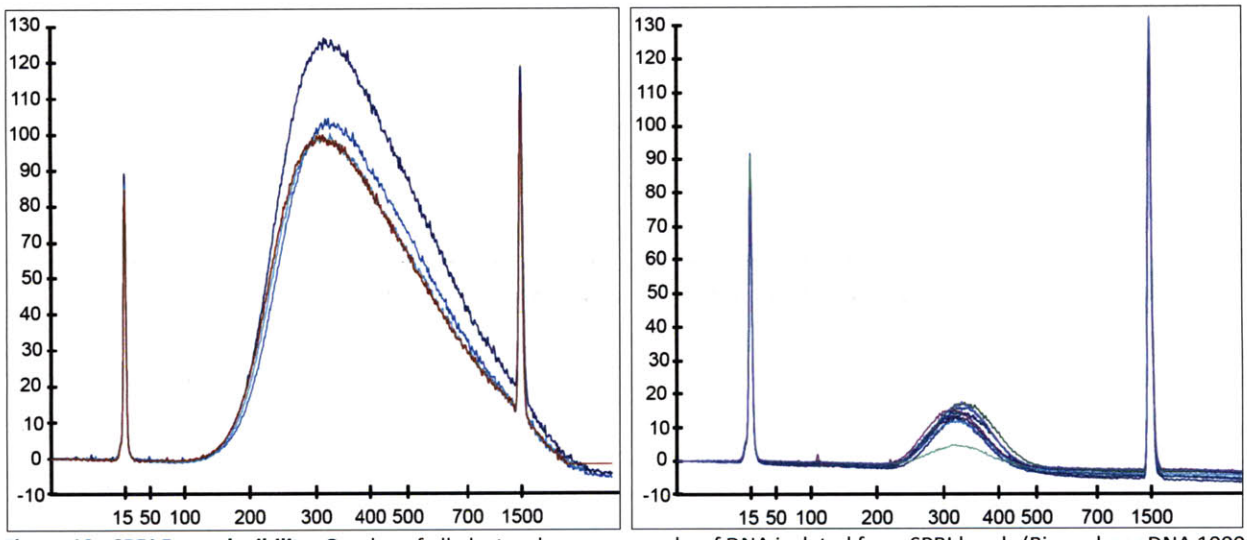

Figure 13. SPRI Reproducibility; Overlay of all electropherogram peaks of DNA isolated from SPRI beads (Bioanalyzer DNA 1000 assay, first separation on right, second separation on left) to demonstrate reproducibility of DNA fragment isolation across 4-8 independent experiments; x-axis = DNA fragment length (bp), y-axis = arbitrary fluorescence units

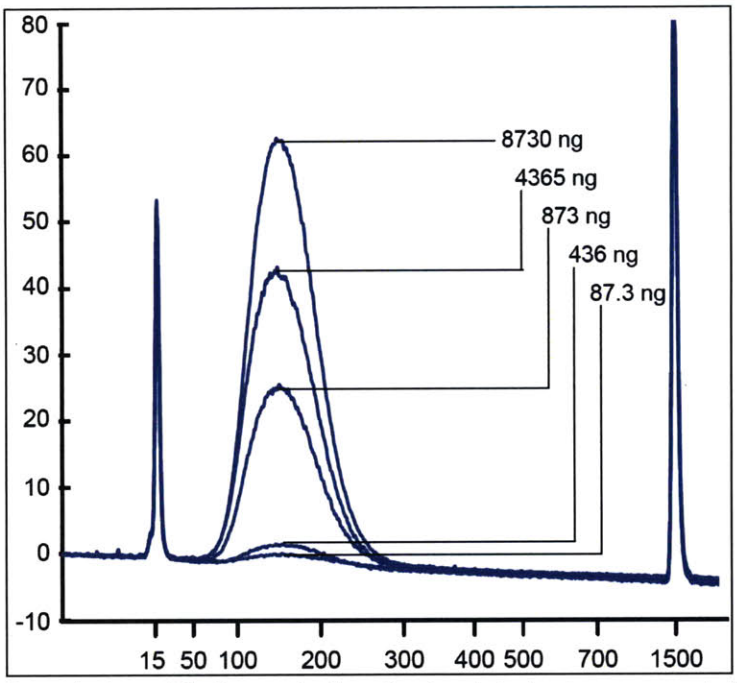

Figure 14. DNA recovery dependent on original concentration; Overlay of electropherogram peaks DNA distributions isolated from SPRI beads after the second separation, when using decreasing amounts of sheared genomic DNA; x-axis = DNA fragment length (bp), y-axis = arbitrary fluorescence units

There are several trends that are immediately noticeable using the SPRI beads for DNA size selection:

1. The concentration of DNA in the original solution has no effect on fragment size enrichment, it only controls the amount of DNA of a particular size fraction that can be isolated (e.g. higher amounts of DNA will be isolated if the original pool contains more DNA, Figure 14)

2. Decreasing the volume of SPRI bead solution during the first separation shifts and narrows the distribution towards larger fragment sizes.

3. Decreasing the volume of SPRI bead solution during the second separation also shifts, but widens, the distribution towards larger fragment sizes.

4. DNA binding to SPRI beads reaches equilibrium within 5 min. Longer incubation times do not have a substantial effect on the distribution of sizes that bind during each separation.

5. Isolation of DNA from either the first or the second SPRI separation using the same experimental conditions is highly reproducible across independent experiments (Figure 13).

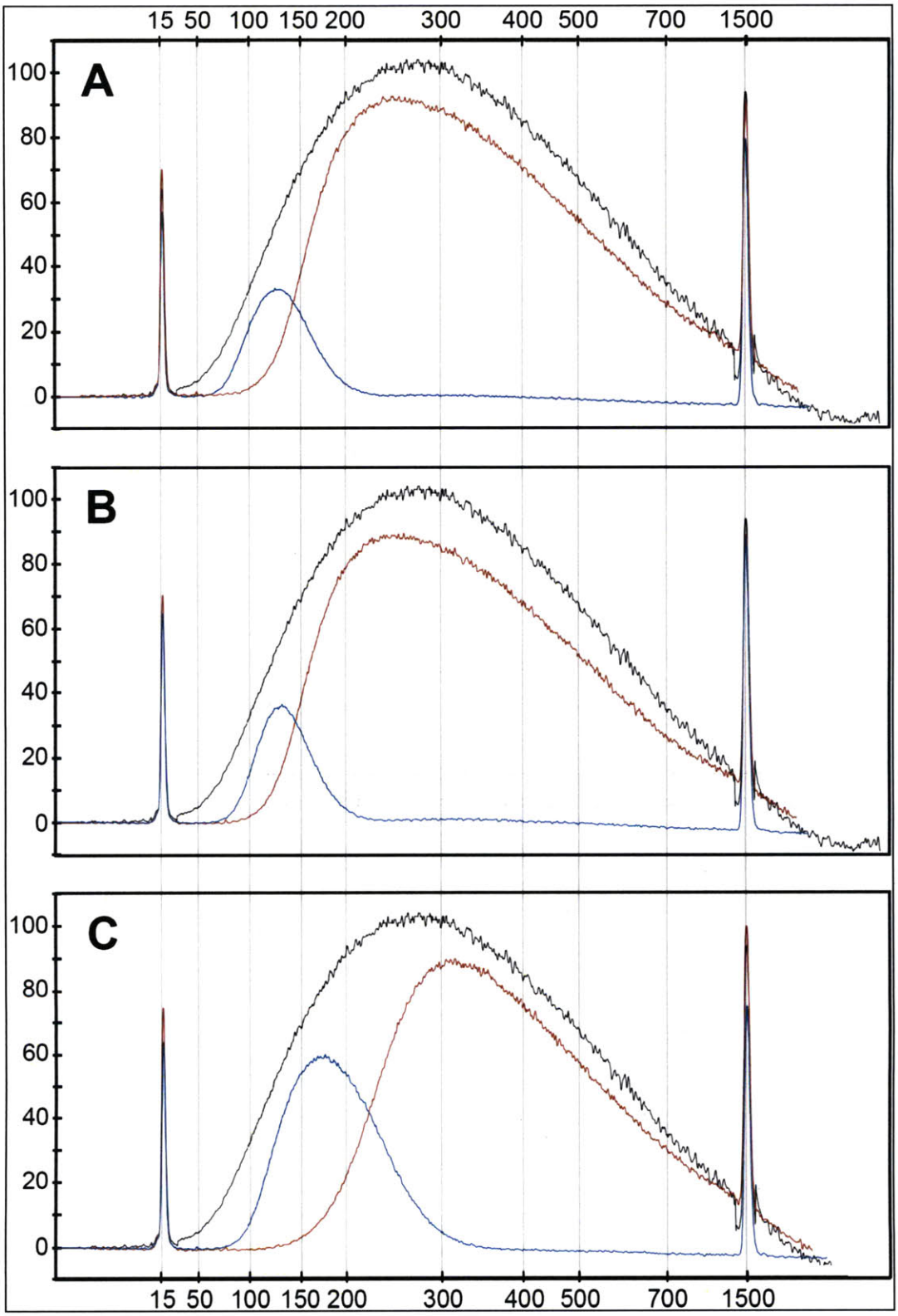

Figure 15. SPRI DNA Size Selection Control-1; Overlay of fluorescence peaks of DNA isolated from SPRI beads in the first separation (red) or second separation (blue), with original sheared genomic DNA length distribution (black). Refer to **Table 2** for experimental details pertaining to each plot (A, B and C); x-axis = DNA fragment length (bp), y-axis = arbitrary fluorescence units

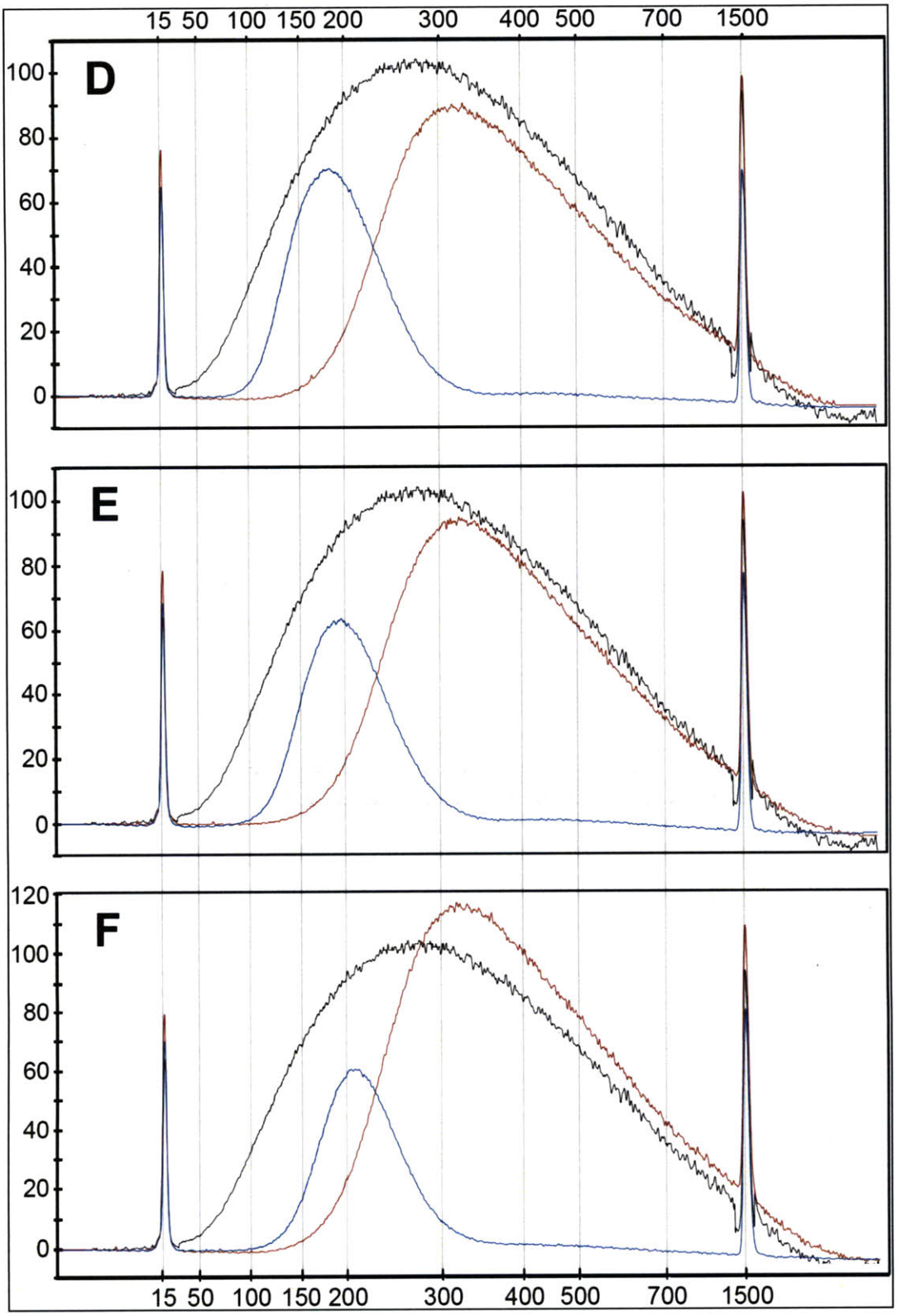

Figure 16. SPRI DNA Size Selection Control-2; Overlay of fluorescence peaks of DNA isolated from SPRI beads in the first separation (red) or second separation (blue), with original sheared genomic DNA length distribution (black). Refer to **Table 2** for experimental details pertaining to each plot (D, E and F); x-axis = DNA fragment length (bp), y-axis = arbitrary fluorescence units

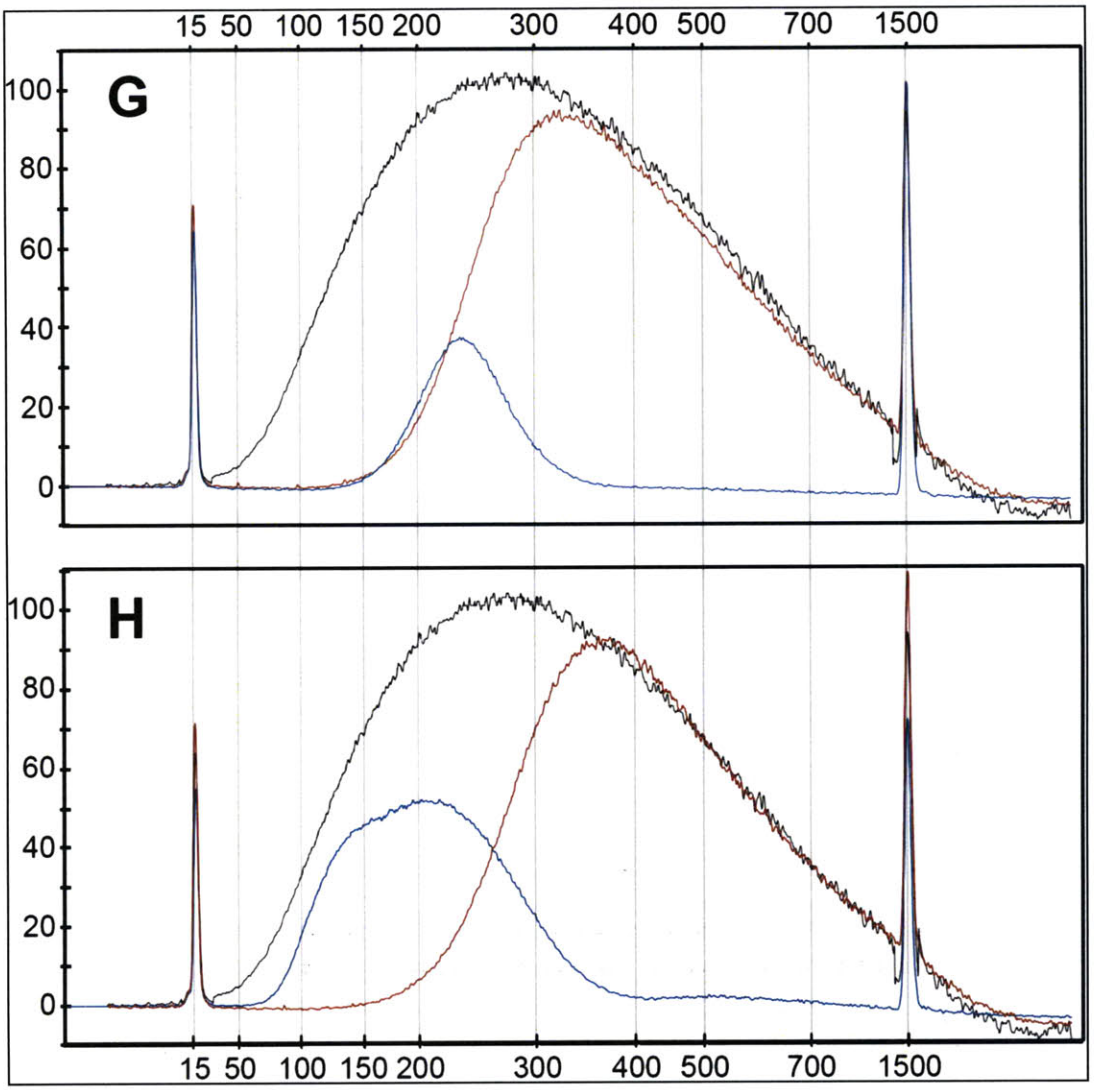

Figure 17. SPRI DNA Size Selection Control-3; Overlay of fluorescence peaks of DNA isolated from SPRI beads in the first separation (red) or second separation(blue), with original sheared genomic DNA length distribution (black). Refer to **Table 2** for experimental details pertaining to each plot (G and H); x-axis = DNA fragment length (bp), y-axis = arbitrary fluorescence units

3.2 *Paired-End Overlapping Reads for Metagenomics*

The Illumina platform is capable of performing paired-end (mate-paired) reads in which sequencing is conducted from both ends of the target DNA. To date, the Illumina method is capable of sequencing ~140 bp from each end of an insert, although the quality of the reads past 100 bp drops significantly (Figure 18). Using only 200-250 bp fragments of DNA for sequencing, the paired-end read functionality of the Illumina method can enable full sequencing of these fragments, greatly increasing the total read-length. By allowing for a small amount of sequencing overlap, the quality of the composite read increases dramatically (Figure 19). The selective isolation of these DNA fragments

requires application of the dSPRI protocol that was developed. The experiment detailed in Rodrigue et al. (17) describes the successful application of the dSPRI technique to isolate genomic DNA fragments of the necessary size to be used for creating overlapped, mate-pair sequencing reads to extend the overall length of sequencing reads, to improve confidence in phylogenic assignments and gene-function predictions.

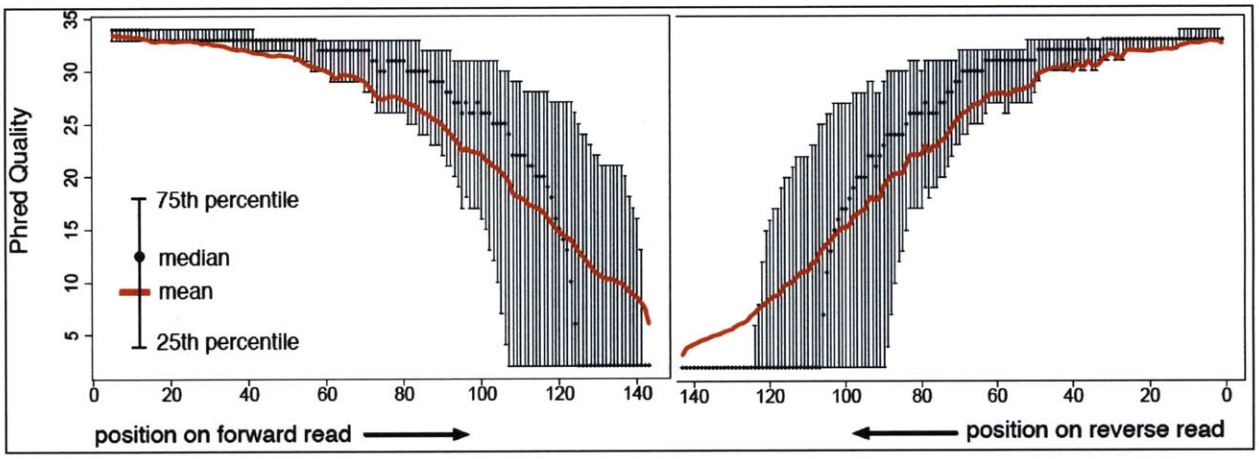

Figure 18. Phred quality score data for original Illumina paired end reads; Rodrigue et al. (17) displaying decrease in quality with read length out to 140 bp.

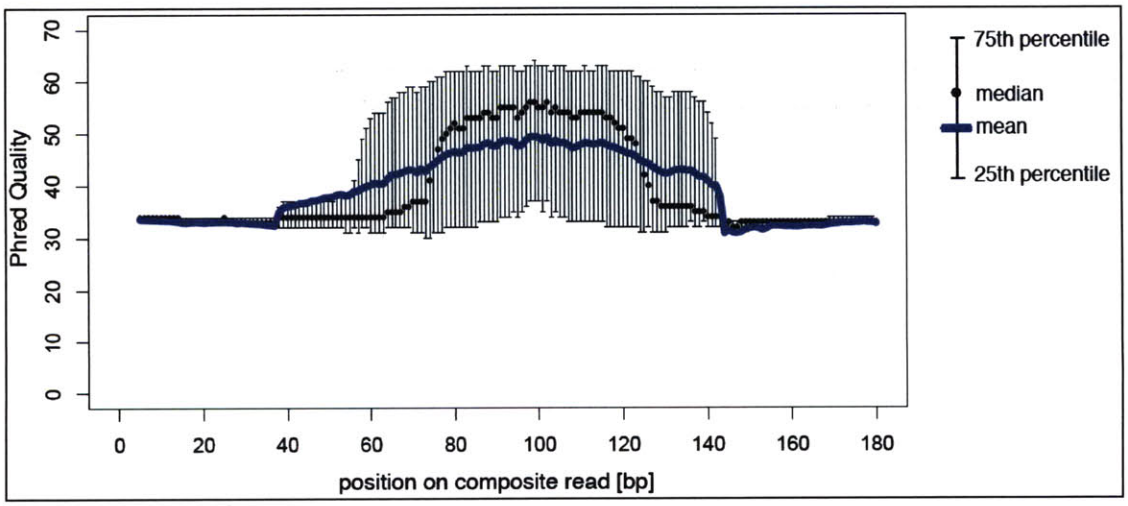

Figure 19. Phred quality score data for composite reads; displaying increase in quality in region of overlap (Rodrigue et al. (17)).

3.3 *Validation of Primer-Clipping Technique*

A DNA 1000 Assay was run on the Bioanalyzer containing samples from each of the steps in the primer-clipping protocol. Taking into account the sizing resolution of the instrument (for DNA 1000 assays which accept 25-1000 bp samples, the resolution for 100-500 bp fragments is 5%, or accurate to

within 5-25 bp, respectively) the results compare extremely well with the expected fragment lengths (Figure 20 and Figure 21). These results represent the successful application of a new technique for removal of PCR primers from PCR-generated libraries, increasing the informative read-length of the library insert by removing sequence that is already known.

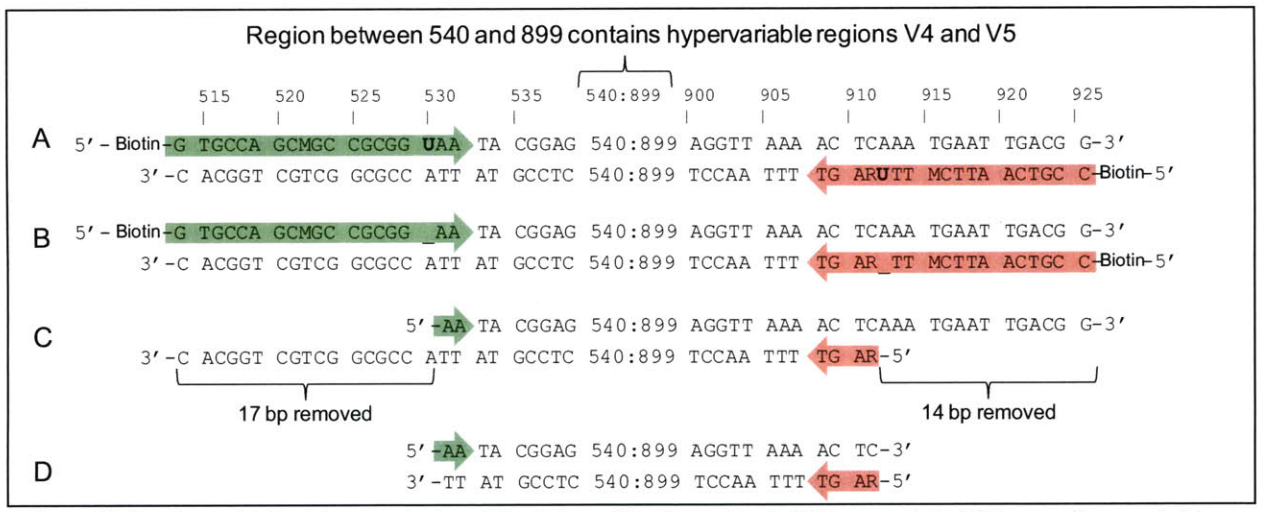

Figure 20. Products obtained from each step of the Primer-Clipping method; (A) the complete 412 bp amplicon containing hypervariable regions V4 and V5, along with 5′-biotin and 3′-uracil, (B) removal of uracil with USER enzyme, (C) product formed after immobilization of biotinylated segments on streptavidin-coated magnetic beads and thermal denaturation, (D) final 381 bp product formed after treatment with 3′-overhangs with S1 nuclease

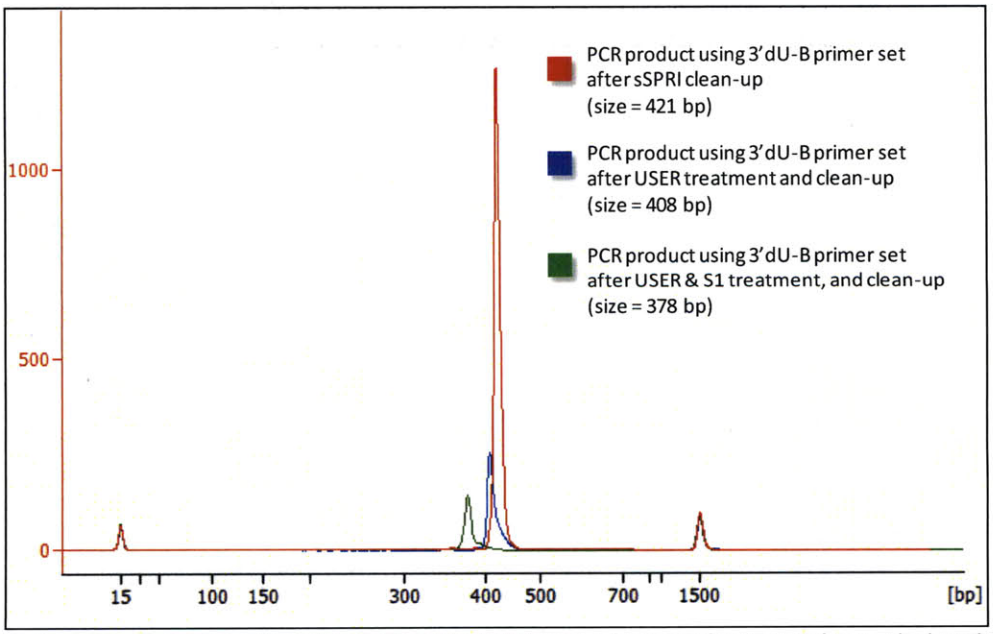

Figure 21. Electropherogram output from Bioanalyzer DNA 1000 assay of primer-clipping products. The lengths of each of the products correlate well with the expected product lengths when taking into account machine error. The expected length of the final product is 381 bp, which is very close to the 378 bp fragment detected (x-axis = bp, y-axis = arbitrary fluorescence units)

3.4 *Characterization and Optimization of ePCR*

In vitro compartmentalization of individual polymerase chain reactions within oil-phase vesicles was considered as an approach for reducing PCR bias and the frequency of chimera formation during amplification of 16S rDNA for microbial diversity studies. Initially, work was done to understand the formation of the emulsions, which involved testing a variety of different conditions (Figure 22 and Table 3). We found that the ratio of oil phase to aqueous phase (water stained with fluorescein) after 2 min of vortexing was the most important variable influencing the size distribution of vesicles formed. We settled on using a ratio of 50:200 (aqueous phase to oil phase).

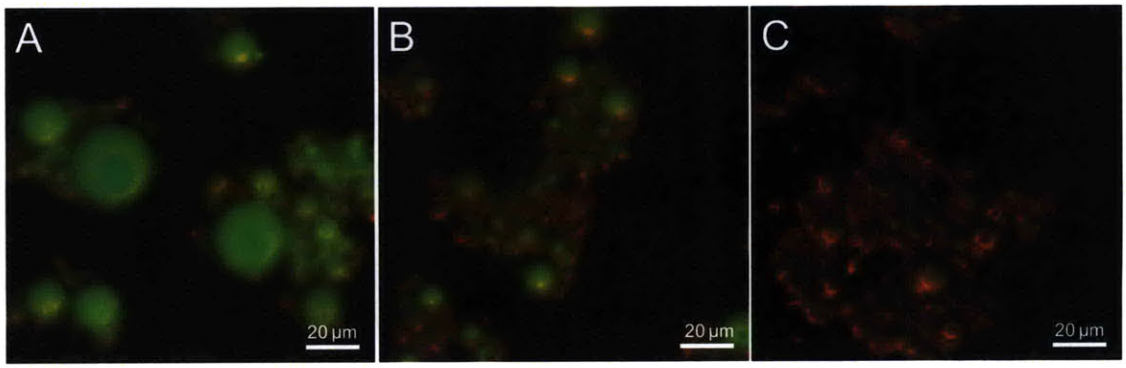

Figure 22. Effect of changing ratio of aqueous phase to oil phase. (A) aqueous : oil = 50 µL: 100 µL (A/O = 0.5), (B) aqueous : oil = 50 µL: 200 µL (A/O = 0.25), (C) aqueous : oil = 50 µL: 300 µL (A/O = 0.17). All images taken at 40x magnification, diluted 1:5 in mineral oil with coverslip applied. An un-cycled, aqueous solution of fluorescein was used to determine the effects of mixture composition on emulsion vesicle sizes.

Table 3. Average diameter and volume of emulsion vesicles formed; using un-cycled samples containing fluorescein (n = 50 to 100)

Ratio Aqueous : Oil	Vesicle Diameter	Vesicle Volume
50 : 100 (0.50)	11.55 µm ± 6.56 µm	7.68×10^{-7} µL ± 2.4×10^{-7} µL (0.768 pL ± 0.24 pL)
50 : 200 (0.25)	7.78 µm ± 5.46 µm	2.39×10^{-7} µL ± 4.15×10^{-8} µL (0.239 pL ± 0.0415 pL)
50 : 300 (0.17)	5.68 µm ± 2.84 µm	4.91×10^{-8} µL ± 1.43×10^{-8} µL (0.0491 pL ± 0.0143 pL)

The number of fluorescein-containing vesicles was used to provide a rough estimate for the theoretical maximum number of independent chambers in which separate PCR reactions could occur (Figure 23). We then replaced the fluorescein solution with enzymatically active PCR mixture, containing the necessary buffering agents, primers, dNTPs, template and polymerase (Phusion), and prepared emulsions following the 50:200 ratio and 2 min vortexing step. The 250 µL ePCR was then placed into PCR tubes, and run through 45 thermal cycles. This mix was then mixed and incubated with PicoGreen so that vesicles that participated in PCR could be detected. During imaging of these samples, the entire

viewing field would contain at most 2-5 fluorescent vesicles that successfully amplified template (Figure 24). This was a valuable indicator that the template was sufficiently diluted into the emulsion phase such that the majority of the vesicles did not contain a copy. If the template had not been diluted sufficiently, ePCR images would have approached the appearance of the fluorescein images (template saturation).

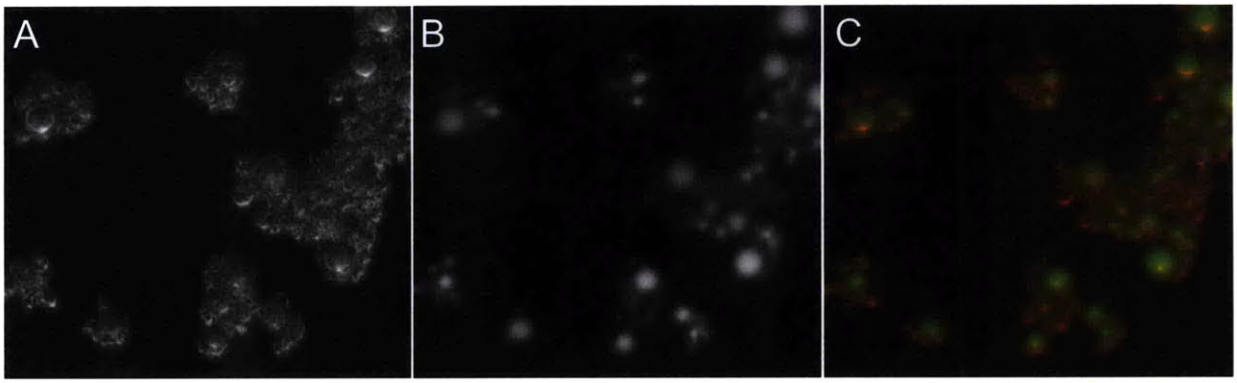

Figure 23. Emulsion vesicles containing aqueous fluorescein solution without cycling to determine theoretical maximum number of PCR vesicles. Emulsion was prepared by using a ratio of aqueous phase to oil phase was 50 µL:200 µL and vortexing at 3000 rpm for 2 min. Transmission light image, 5 ms exposure (A), fluorescence image when excited with 494-nm wavelength light, 300 ms exposure (B), image overlay (C); 40x magnification, diluted 1:5 in mineral oil with coverslip.

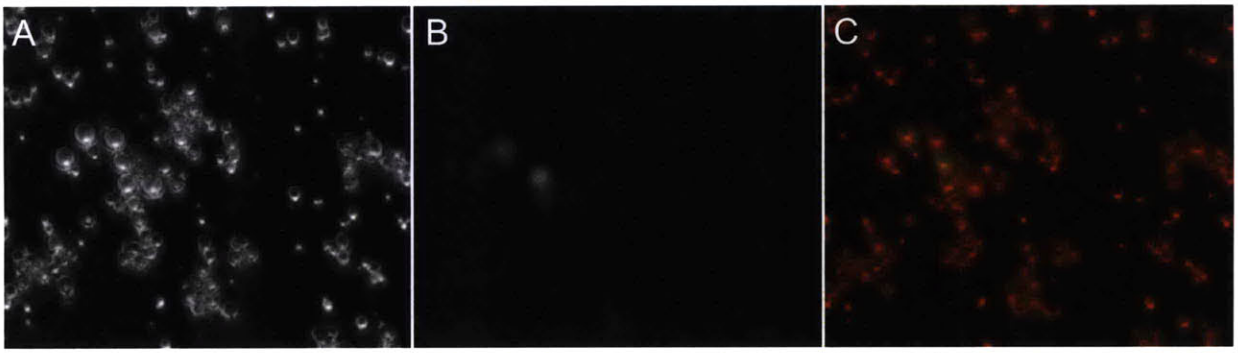

Figure 24. Emulsion vesicles containing PCR aqueous phase after 35 cycles and staining with PicoGreen. Transmission light image, 5 ms exposure (A), fluorescence image when excited with 480-nm wavelength light, 500 ms exposure (B), image overlay (C); 40x magnification, diluted 1:5 in mineral oil without coverslip.

Once it was determined that PCR was functioning in our emulsions, we prepared more samples with templates, developed the emulsion-breaking method, and proceeded to collect and analyze the DNA from our ePCR libraries. A High Sensitivity DNA Assay was run on the Bioanalyzer containing samples of DNA from broken ePCRs, as well as samples prepared using normal PCR conditions. While the regularly-prepared PCR sample yielded significantly more DNA (Figure 25, the markers present in the ladder are reduced to small sharp peaks due to the large peak of the DNA sample), it also contained a large amount of non-specific products, evidenced by the large 'skirt' trailing after the peak indicating the desired PCR product (around 412 bp). The benefit of the ePCR method can be seen in the Bioanalyzer

output for one of the broken emulsions, where the correct product peak is present and all the larger, hypothetically chimeric, products are absent (Figure 26). The ePCR technique was applied in the preparation of an actual sequencing experiment (called the 'Matrix', discussed below), in which 4 mock bacterial communities were used to determine the sensitivity of our 16S rDNA amplification method and whether the ePCR reduced the total amount of chimeric products as determined by sequencing reads.

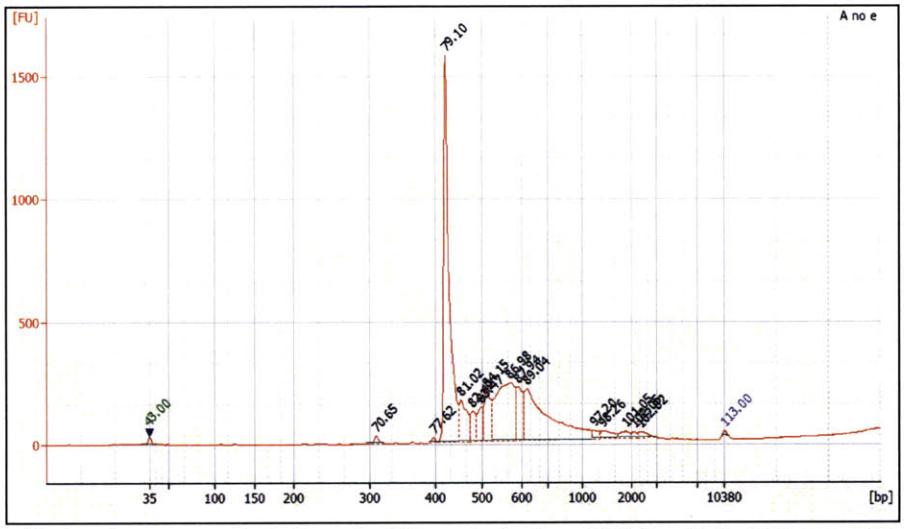

Figure 25. Electropherogram output of regular PCR product using a High Sensitivity DNA assay (Bioanalyzer)

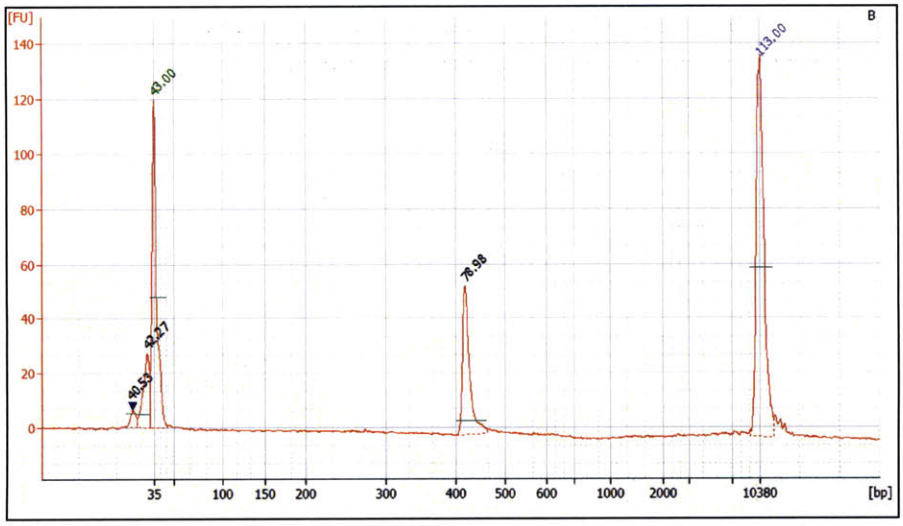

Figure 26. Electropherogram output of ePCR product using a High Sensitivity DNA assay (Bioanalyzer)

3.5 *Cycle-Optimization as an Alternative to ePCR*

During the preparation of the 'Matrix' experiment detailed below, different PCR chemistries were compared for their ability to limit PCR bias. A set of 4 different polymerases, and their associated buffers were used to prepare normal PCR and ePCR 16S libraries. It was known from the ePCR optimization studies that the Phusion polymerase reaction chemistry was stable in the oil phase, but the KAPA2G™ Robust HotStart (KAPA Biosystems) reaction chemistry and the SequalPrep™ Long PCR Kit (Invitrogen) reaction chemistry had never been tested for compatibility with the oil/surfactant mix. When this was attempted, it was found that after thermal cycling the emulsion phase had collapsed, so that the ePCR had become a bulk-phase PCR, with no compartmentalization of individual amplification reactions (Figure 27). Due to concern for using these samples for sequencing, another method, called cycle-optimization, was developed for reducing PCR bias.

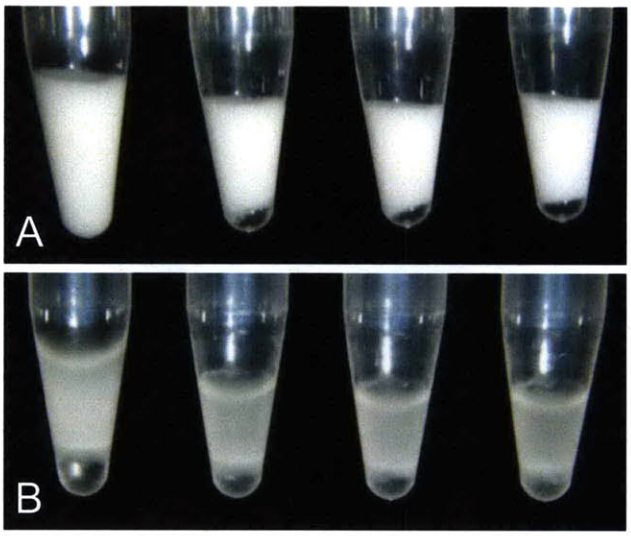

Figure 27. Premature emulsion breaking; pictures of ePCR before (A) and after (B) running through a thermal cycling program when the KAPA or SequalPrep reaction kits were used to prepare the aqueous phase of 16S libraries

Cycle-optimization involved taking a set of DNA extracts, and using DNA concentration measurements from a NanoDrop 1000 spectrophotometer, the appropriate dilution was made to obtain concentrations of ~10 ng/μL for each template. These samples were run in triplicate for quantitative PCR, from which the 'optimal' cycle number was determined by choosing the number of cycles required to reach the midpoint of the linear phase of PCR for the majority of the concentration-normalized templates. This point in the PCR was chosen for the endpoint since the probability of approaching the 'plateau-effect' in which limiting concentrations of dNTPs and primer favor the formation of homologous hybridization products was minimized (Figure 28).

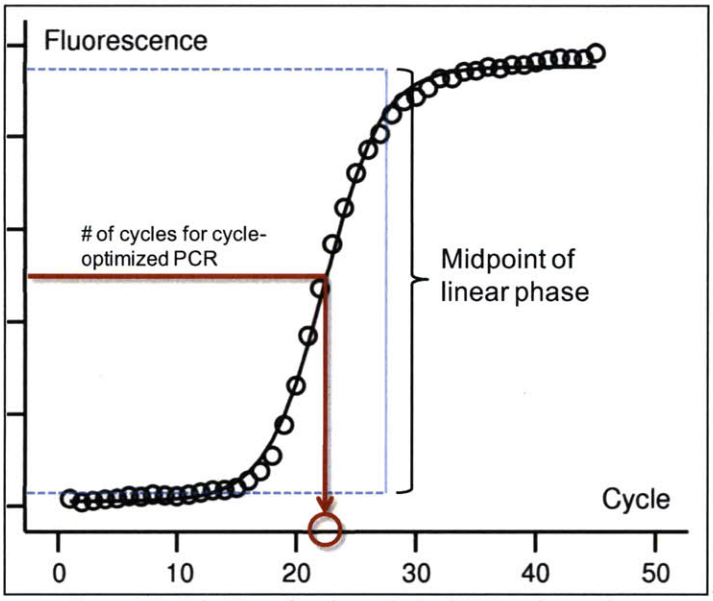

Figure 28. Selection of cycle-optimized PCR cycle number

3.6 *Validation of Primer-Skipping Technique - Sequencing Data from 'Matrix' Experiment*

To better understand the effect of PCR bias and chimera formation during 16S library preparation using the primer-skipping method, an experiment was constructed to compare the effect of different DNA polymerases as well as compare the results of emulsion PCR (ePCR) and normal, cycle-optimized PCR (Table 4). A set of mock bacterial communities were prepared by mixing pure DNA from 10 different species at different concentrations and mixing the mock community DNAs with DNA obtained from a human stool sample at pre-determined ratios. These DNA mixtures were used for library preparation to understand the limit of detection of less-abundant species in a mixed-template PCR, as well as learn whether or not our changes to library preparation (polymerase used, ePCR or cycle-optimized PCR) were capable of decreasing PCR bias and the formation of PCR artifacts.

Table 4. The 'Matrix' of experiments used to construct a barcoded, 16S library. The polymerase used for each sample is listed in colored font. The template used is either pure-mock DNA, or a mixture of the mock DNA and DNA extracted from a healthy human gut microbial community at the described ratios. Each of the experiments was prepared in triplicate.

Normal PCR			Emusion PCR		
Phusion; DNA-M1	KAPA2G-robust; DNA-M1	SequalPrep; DNA-M1	Phusion; DNA-M1	KAPA2G-robust; DNA-M1	SequalPrep; DNA-M1
Phusion; P/DNA-M1 = 1:3			Phusion; P/DNA-M1 = 1:3		
Phusion; P/DNA-M1 = 3:1			Phusion; P/DNA-M1 = 3:1		
Phusion; P/DNA-M1 = 1:1			Phusion; P/DNA-M1 = 1:1		
Phusion; P/DNA-M1 = 9:1	KAPA2G-robust; P/DNA-M1 = 9:1	SequalPrep; P/DNA-M1 = 9:1	Phusion; P/DNA-M1 = 9:1	KAPA2G-robust; P/DNA-M1 = 9:1	SequalPrep; P/DNA-M1 = 9:1
Phusion; P/DNA-M1 = 99:1			Phusion; P/DNA-M1 = 99:1		
Phusion; P/DNA-M2 = 9:1			Phusion; P/DNA-M2 = 9:1		
Phusion; P/DNA-M3 = 9:1			Phusion; P/DNA-M3 = 9:1		
Phusion; P/DNA-M4 = 9:1			Phusion; P/DNA-M4 = 9:1		

Table 5. Pipetting scheme for two-step barcoded 16S library for normal PCR (emulsions used 45 cycles for each step)

	Step 1 PCR			Step 2 PCR		
Polymerase Used	1× master mix (25 µL)		Thermal cycling scheme	1× master mix (25 µL)		Thermal cycling scheme
Phusion	Reagent	Volume		Reagent	Volume	
	1) H_2O	12.25 µL	T_D/t_D = 98°C/30 s	1) H_2O	10.65 µL	T_D/t_D = 98°C/30 s
	2) HF Buffer (5x)	5 µL	T_A/t_A = 52°C/30 s	2) HF Buffer (5x)	5 µL	T_A/t_A = 83°C/30 s
	3) dNTP (10 mM)	0.5 µL	T_E/t_E = 72°C/15 s	3) dNTP (10 mM)	0.5 µL	T_E/t_E = 72°C/15 s
	4) primer (3 µM)		(25 cycles)	4) primer (3 µM)		(15 cycles)
	i) V4-U515-Alm3.1-F	2.5 µL	D = denaturation	i) PE-PCRxxx-1	3.3 µL	
	ii) V5-U926-R	2.5 µL	A = annealing	ii) PE-PCR-2	3.3 µL	
	5) template	2 µL	E = elongation	5) template	2 µL	
	6) Phusion polymerase	0.25 µL		6) Phusion polymerase	0.25 µL	
KAPA2G™ Robust HotStart	1) H_2O	7.4 µL	95°C, 3 min	1) H_2O	5.8 µL	95°C, 3 min
	2) Buffer A (5x)	5 µL	T_D/t_D = 95°C/30 s	2) Buffer A (5x)	5 µL	T_D/t_D = 95°C/30 s
	3) Enhancer (5x)	5 µL	T_A/t_A = 52°C/30 s	3) Enhancer (5x)	5 µL	T_A/t_A = 52°C/30 s
	4) dNTP (10 mM)	0.5 µL	T_E/t_E = 72°C/20 s	4) dNTP (10 mM)	0.5 µL	T_E/t_E = 72°C/20 s
	5) primer (3 µM)		(25 cycles)	5) primer (3 µM)		(15 cycles)
	i) V4-U515-Alm3.1-F	2.5 µL		i) PE-PCRxxx-1	3.3 µL	
	ii) V5-U926-R	2.5 µL		ii) PE-PCR-2	3.3 µL	
	6) template	2 µL		6) template	2 µL	
	7) KAPA2G polymerase	0.10 µL		7) KAPA2G polymerase	0.10 µL	
SequalPrep™ Long PCR	1) H_2O	13.3 µL	94°C, 2 min	1) H_2O	11.7 µL	94°C, 2 min
	2) Rxn Buffer (10x)	2.5 µL	T_D/t_D = 94°C/10 s	2) Rxn Buffer (10x)	2.5 µL	T_D/t_D = 94°C/10 s
	3) DMSO (5x)	0.5 µL	T_A/t_A = 52°C/30 s	3) DMSO (5x)	0.5 µL	T_A/t_A = 83°C/30 s
	4) Enhancer A (10x)	1.25 µL	T_E/t_E = 68°C/35 s	4) Enhancer A (10x)	1.25 µL	T_E/t_E = 68°C/40 s
	5) primer (3 µM)		(10 cycles)	5) primer (3 µM)		(10 cycles)
	i) V4-U515-Alm3.1-F	2.5 µL	T_D/t_D = 94°C/10 s	i) PE-PCRxxx-1	3.3 µL	T_D/t_D = 94°C/10 s
	ii) V5-U926-R	2.5 µL	T_A/t_A = 52°C/30 s	ii) PE-PCR-2	3.3 µL	T_A/t_A = 83°C/30 s
	6) template	2 µL	T_E/t_E = 68°C/35 s	6) template	2 µL	T_E/t_E = 68°C/40 s
	7) SeqPrep polymerase	0.45 µL	+ 20 s/cycle	7) SeqPrep polymerase	0.45 µL	+ 20 s/cycle
			(15 cycles)			(5 cycles)
			72°C, 3 min			72°C, 3 min

For the sequencing run, the samples were roughly split into two different lanes, since the large PCR products that were seen in the electropherogram traces of normal PCR could cause severe over-clustering on the flowcell. The normal PCR products had been through 25 cycles in Step 1 and 15 cycles in Step 2 (Table 5). Despite limiting the number of cycles for each step, there were still PCR products larger than expected, so these products were separated using gel electrophoresis and the desired size range (400-700 bp) was extracted for sequencing. This extra step gave the normal PCR a considerably lower chance of including chimeras, providing an unfair comparison against the ePCR libraries. Once the sequencing run was complete, it was found that the lane containing these larger products had not been diluted sufficiently, and many of the reads from that lane were not usable. The results from the other lane, which mostly contained libraries made from ePCR, are given below. A total of 13,097,730 reads (144bp mate-paired) were obtained from this single lane. The original number of reads for each sample is given below, along with the number of reads that were kept after filtering for quality and removal of hypothetical chimeric sequences. Although 144 bp were sequenced from each end, only the first 92 in the forward read and the first 80 in the reverse read were usable for phylogenetic analysis.

For a sequence to be included in a library it had to fulfill the following criteria: the bar code sequence had to match a used bar code exactly (no ambiguities or short sequences); none of the first 92 bases in the forward read or of the first 80 bp in the reverse read could have ambiguous base calls (i.e. no N's) or have Illumina's Read Segment Quality Score Indicator (B), which designates the quality of that base unreliable and that base unusable in downstream analysis. If both the forward and the reverse reads fulfilled the criteria listed above, the reverse complement sequence of the reverse read was concatenated to the end of the forward read for a total of 172 unambiguous, usable bases. The concatenated sequences were used as the input for UCLUST 2.0.5, a program commonly used to group sequences into operational taxonomic units (OTUs) for large datasets. Sequences were grouped into sequences clusters at 95% identity. One representative sequence from each cluster was used in ChimeraSlayer (available through the Broad Institute) with the reference alignment truncated to match the same 172 bases of the 16S sequence that were retained in this analysis.

Table 6. Results of multiplexed sequencing run containing 16S libraries prepared from different mock communities

16S rDNA source, Polymerase Used	Barcode	Preparation type	Total no. reads that match each barcode	Total no. reads kept after filtering
DNA-M1, Phusion	AACTCGG	ePCR	163486	**131309**
DNA-M1, Phusion	CAGTCGT	ePCR	169519	**133807**
DNA-M1, Phusion	GCATCTA	ePCR	391544	**309546**
P/DNA-M1 (9:1), Phusion	CTCTCCT	ePCR	418298	**339180**
P/DNA-M1 (9:1), Phusion	GATTCGA	ePCR	377286	**304115**
P/DNA-M2 (9:1), Phusion	GTGTCCA	ePCR	479685	**390469**
P/DNA-M2 (9:1), Phusion	TTTTCCC	ePCR	329406	**268220**
P/DNA-M3 (9:1), Phusion	AGTTCAG	ePCR	391616	**319073**
P/DNA-M3 (9:1), Phusion	TGGTCAC	ePCR	334489	**273035**
P/DNA-M4 (9:1), Phusion	AATTATT	ePCR	432013	**352109**
P/DNA-M4 (9:1), Phusion	TAGTATG	ePCR	405495	**318490**
P/DNA-M4 (9:1), Phusion	GGCTCAA	ePCR	432777	**349386**
P/DNA-M1 (3:1), Phusion	CAATATA	ePCR	423628	**340494**
P/DNA-M1 (3:1), Phusion	GACTATC	ePCR	216208	**177488**
P/DNA-M1 (1:3), Phusion	ATGTAGT	ePCR	210844	**170910**
P/DNA-M1 (1:3), Phusion	TTCTAGG	ePCR	63588	**51909**
P/DNA-M1 (1:1), Phusion	CTTTAGA	ePCR	378764	**283958**
P/DNA-M1 (1:1), Phusion	GTATAGC	ePCR	562786	**446541**
P/DNA-M1 (99:1), Phusion	AGCTACT	ePCR	508441	**384284**
P/DNA-M1 (99:1), Phusion	TGATACG	ePCR	377380	**299582**
DNA-M1, Phusion	TTTTTTT	Non-emulsion	339885	**274147**
DNA-M1, SequalPrep	CTTTGAG	Non-emulsion	311750	**255722**
DNA-M1, KAPA2G-robust	CAATGCG	Non-emulsion	3912995	**3094204**
DNA-M1, KAPA2G-robust	TCTTGGA	Non-emulsion	105569	**85475**
P/DNA-M1 (9:1), KAPA2G-robust	AATTGCC	Non-emulsion	147535	**120973**
Vibrio, Phusion	GCGGTAG	ePCR	391169	**363674**
Vibrio, Phusion	TCTGTAT	ePCR	414223	**385271**

The forward and reverse complemented reverse reads were joined by 217 N's to create a gapped sequence of appropriate length for use in classification through the command line Ribosomal Database Project (RDP) classifier tool (18). RDP assigns 16S rRNA gene sequences to the new phylogenetically consistent higher-order bacterial taxonomy proposed by Garrity et al. ((18)where hierarchical taxa are based on a naïve Bayesian rRNA classifier (http://rdp.cme.msu.edu/classifier/classifier.jsp). Reference sequences for mock community members were obtained through GenBank or MicrobesOnline and the appropriate 172 base pairs were also included in UCLUST clustering, ChimeraSlayer analysis, and RDP classification. Sequences which were found in the same cluster as a mock community member, or in a cluster that was determined to be non-chimeric and classified as the same organism as a mock community member, were included as part of the mock community. This allows for some sequence variation due to differences between copies of the ribosomal DNA sequence or sequencing error. Any clusters determined to be chimeric sequences were not regarded as part of the mock community.

Expected counts were based on the volume of DNA and the stock concentration (Nanodrop reading) as a percent of total DNA added to the sample. The results of this sequence filtering are given in Table 6 and Figure 29.

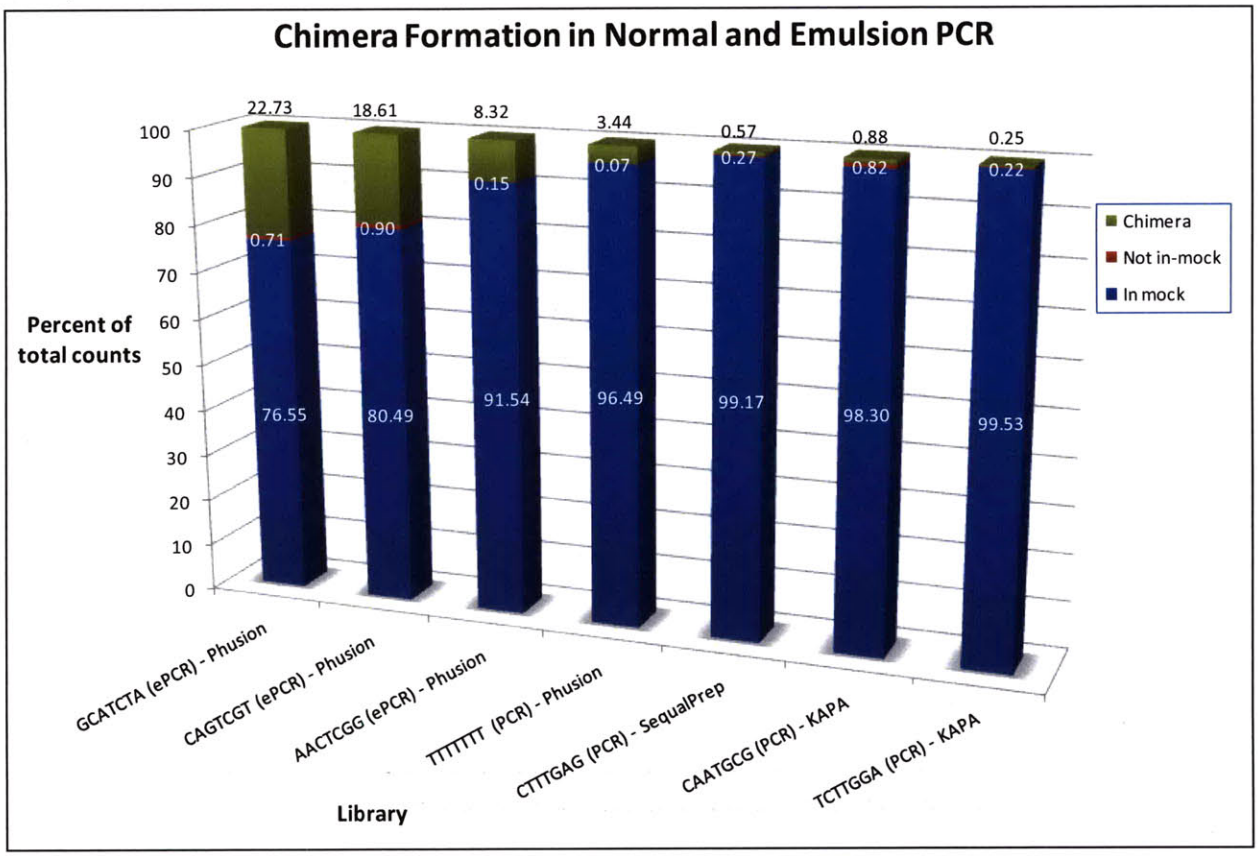

Figure 29. Chimera formation in Normal and Emulsion PCR, where chimera counts were determined by searching 172 bp, paired end sequencing reads for homology in 2 different species

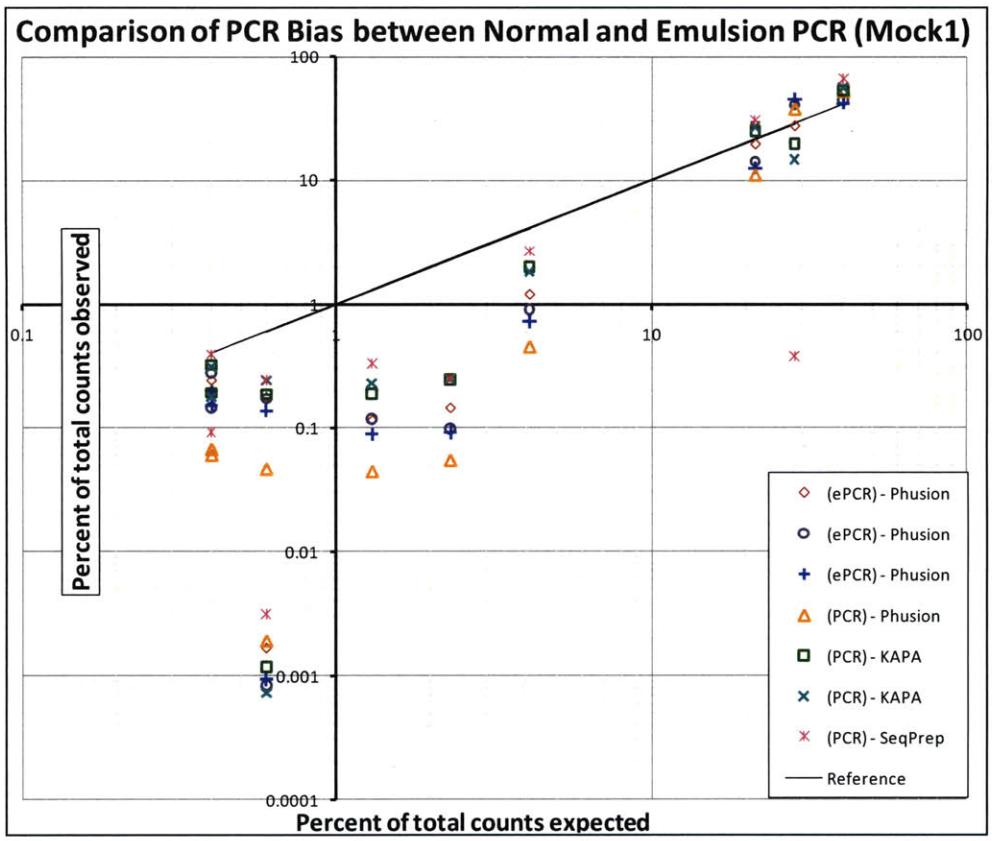

Figure 30. Log-log plot of sequence data comparing number of reads for Mock 1 species with the expected values

By comparing the observed number of sequencing reads against the expected population of the Mock 1 (DNA-M1) community, the relative accuracy of each sample preparation method could be compared. Most of the sample preparation methods underestimate the abundance of most of the species in the mock community (as compared with the reference line, Figure 30).

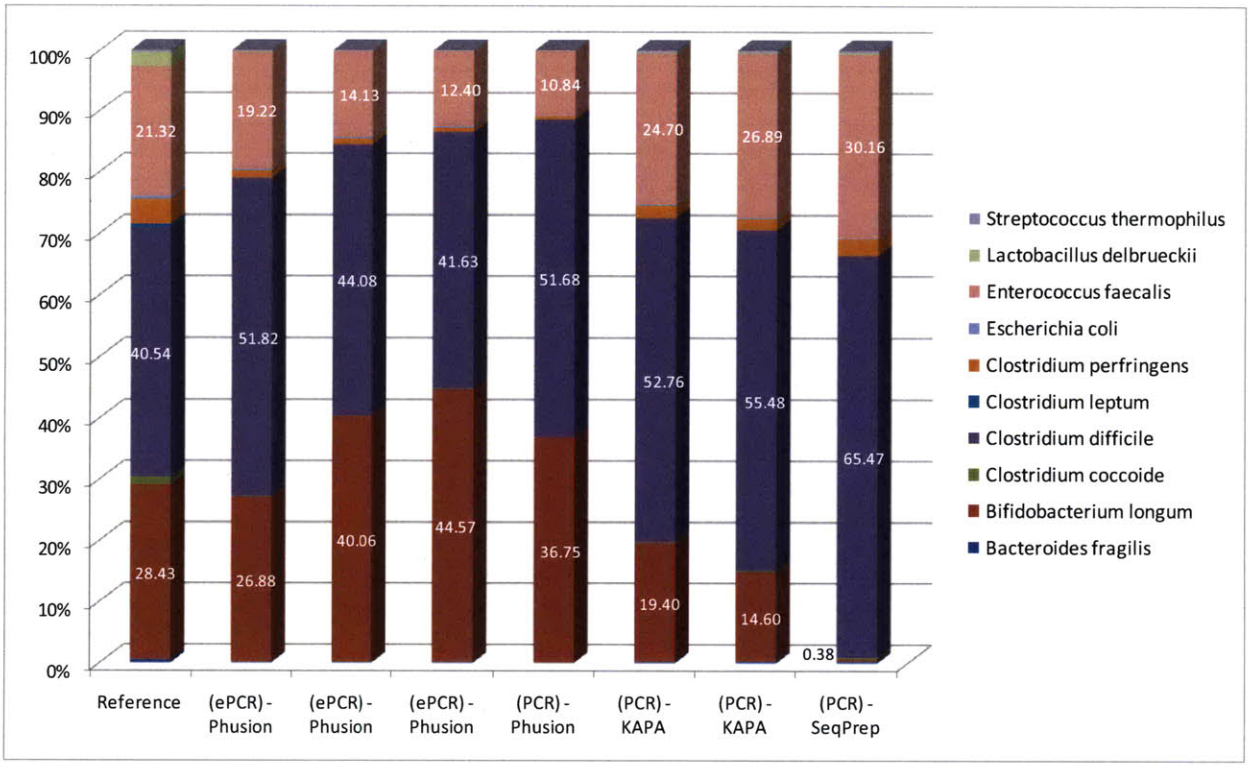

Figure 31. Percentage of reads associated with each species in the Mock 1 community; compared with the Reference community many of the species are under-represented in the sequencing reads

Table 7. Percentage of reads associated with each species in the Mock 1 community from various library preparations

Mock 1 Species	Ref	AACTCGG ePCR-Phusion	GCATCTA ePCR-Phusion	CAGTCGT ePCR-Phusion	TTTTTTT PCR-Phusion	TCTTGGA PCR-KAPA	CAATGCG PCR-KAPA	CTTTGAG PCR-SeqPrep
Bacteroides fragilis	0.60	0.18	0.17	0.14	0.05	0.19	0.24	0.24
Bifidobacterium longum	28.43	26.88	40.06	44.57	36.75	19.40	14.60	0.38
Clostridium coccoide	1.30	0.11	0.12	0.09	0.04	0.19	0.23	0.33
Clostridium difficile	40.54	51.82	44.08	41.63	51.68	52.76	55.48	65.47
Clostridium leptum	0.60	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Clostridium perfringens	4.10	1.20	0.91	0.73	0.45	2.00	1.82	2.67
Escherichia coli	0.40	0.20	0.28	0.19	0.06	0.19	0.18	0.09
Enterococcus faecalis	21.32	19.22	14.13	12.40	10.84	24.70	26.89	30.16
Lactobacillus delbrueckii	2.30	0.14	0.10	0.09	0.05	0.25	0.25	0.25
Streptococcus thermophilus	0.40	0.24	0.15	0.15	0.07	0.32	0.31	0.39

Both of the normal PCRs using KAPA2G-Robust polymerase have very little variation between the two libraries (Figure 31). Samples that used template DNA consisting of a mixture of human gut community

and mock community (P/DNA-M) at a defined ratio have not been included in the PCR bias analysis since the pure gut community is in the process of being sequenced. This data is needed in order to determine the obfuscating effect of the gut community in detecting the mock communities.

The ePCR libraries were expected to have lower amount of chimeric sequences due to the template diluting and PCR compartmentalization, however in Figure 29 the normal PCR libraries appear to have a lower amount of detected chimeras. This is mostly likely due to the fact that the normal libraries were run on a gel and size-selected prior to sequencing. In addition, the ePCR libraries were expected to have reduced PCR bias, since all templates had an equally likely chance of successfully amplifying in an individual emulsion vesicle. The *Bacteroides, Clostridium, E coli, Lactobacillus and Streptococcus* were not as successfully amplified in ePCR libraries (Figure 31 and Table 7). The KAPA-2G robust and SequalPrep polymerases were more successful in capturing the presence of *Clostridium perfringens*. There are no clear trends in the ability for ePCR to improve PCR bias.

3.7 *Development of Liquid Class and Program Scripts for Tecan Robotic Liquid Handler*

Several scripts for preparing primer-skipping 16S libraries were developed for the Tecan Freedom EVO 150 robotic liquid-handling work station. Libraries prepared in this manner also utilize the cycle-optimized method rather than ePCR for reducing PCR bias and artifacts. The process of developing a program for each step of the library preparation was described in 2.9 *Automation of Library Preparation* of Materials & Methods. The main operations are described in Table 8.

Table 8. Robotic Library Preparation Workflow

Workflow Operation	Description
Template Plating	Takes 24 DNA extracts (diluted to 10 ng/µL) from microcentrifuge tubes and dispenses them in four 2 µL aliquots on a 96-well plate.
Step 1 PCR	Takes a human-prepared, bulk master mix from a deep 96-well plate and dispenses 23 µL (total volume per well = 25 µL) onto the prepared template plates, then mixes. The plates are then transferred to a thermal cycler by hand.
Sample Pooling	Recombines the four 25 µL aliquots back into one well on a 96-well plate, creating a Step 1 PCR product plate holding 96 distinct samples.
SPRI Cleanup	Removes primers and dimers from PCR mixture using the sSPRI cleanup procedure, yielding 96 distinct Step 1 PCR products (eluting in 40 µL water or EB).
Sample Aliquoting	Redistributes each 40 µL sample into four 4 µL aliquots on a 96-well plate (24 samples per plate), the rest of the Step 1 PCR product is saved.
Barcoded Primer Plating	Takes 96 barcoded primers (at working concentration) from microcentrifuge tubes and dispenses them into a deep 96-well plate.
Step 2 PCR	Takes a human-prepared, bulk master mix from a deep 96-well plate and dispenses 17.7 µL onto the prepared Step 1 PCR product plates (total volume per well = 21.7 µL). Then adds 3.3 µL of unique barcoded-primer solution (total volume per well = 25 µL) and mixes. The plates are then transferred to a thermal cycler by hand.
Sample Pooling	Recombines the four 25 µL aliquots back into one well on a 96-well plate, creating a Step 2 PCR product plate holding 96 distinct samples.
SPRI Cleanup	Removes primers and dimers from PCR mixture using the sSPRI cleanup procedure, yielding 96 distinct Step 2 PCR products (eluting in 40 µL water or EB).
Samples are frozen at -20°C until needed for sequencing.	

4. *Conclusions & Future Directions*

A set of engineering challenges, which focused on developing techniques for increasing the scale and quality of genomic libraries from environmental samples, were introduced for the Illumina sequencing platform. The first of these involved increasing the read-length of DNA inserts to 200 bp to improve the applicability of the Illumina system to metagenomics studies. A method for selecting DNA fragments of the appropriate size range using SPRI beads was developed. These fragments were then sequenced as mate-paired reads with a small amount of overlap in the center to dramatically enhance the quality of the ends of each read, successfully producing sequenced DNA inserts with a length twice that of conventional Illumina reads that could be used for more confident assignments of gene function and phylogenetic classification.

The second engineering challenge was designing universal primers with some degenerate nucleotides to specifically amplify the V4 and V5 hypervariable regions of the bacterial 16S ribosomal RNA gene. In order to improve the informative read-length of sequencing reads, one set of universal 16S primers was designed to include 5'-biotin and 3'-uracil, so that the known universal sequence could be 'clipped' from the amplified V4 and V5 region. This primer set was successfully used to produce sequences that could be used in the standard Illumina library preparation protocol. Additionally, a method of streamlining the traditional Illumina library preparation protocol by removing time-consuming gel electrophoresis and enzymatic steps was developed. This system utilized a two-step PCR library preparation method, that amplifies the hypervariable regions in the first step, then appends a unique molecular barcode along with the necessary sequences for clustering on the Illumina flowcell. Sequencing primers were designed to 'skip' the known universal sequence and target the informative nucleotides in the hypervariable regions. The SPRI bead technique was used for PCR clean-ups. Libraries were successfully prepared, multiplexed and sequenced using this method, demonstrating compatibility with Illumina chemistry and reliable segregation of reads based on barcodes. By replacing gel electrophoresis and column-based reaction clean-ups with SPRI beads, the two-step 'primer-skipping' library preparation method could increase throughput. A set of program scripts for operating a robotic liquid-handling workstation were prepared and are in the process of being validated and tested for contamination.

Finally, a method for reducing PCR bias and artifacts (chimeras and heteroduplexes) by preparing libraries in emulsions to compartmentalize individual templates for PCR was developed. A method for breaking emulsion vesicles using osmotic pressure was also developed to harvest the PCR

products. Preliminary data from a sequencing experiment that examined the effect of different polymerases in normal and emulsion PCRs is inconclusive regarding the improvement of PCR bias and artifact formation.

Future work involves preparing 16S libraries applying the automated, primer-skipping protocol from over 800 human saliva and stool DNA extracts (HuGE Project, Human Gut Ecology; 2 human subjects, sampling daily for a year), preparing libraries from DNA extracts of water samples collected across the full depth of Mystic Lake, and preparing libraries to understand the chicken gut microbiome for the food industry.

The experiment involving the comparison of ePCR and normal PCR for minimizing PCR bias and artifact formation should be repeated, taking care to prepare both libraries as similarly as possible. In the original experiment, the Step 1 PCR was run for 45 cycles in ePCR and only 25 cycles in the normal PCR. By reducing the number of cycles in the ePCR, there should be a dramatic change in the amount of chimeras detected. Also, the normal PCR libraries were size-selected via gel electrophoresis, which may have removed a number of chimeric sequences, whereas the ePCR libraries were cleaned using SPRI bead selection. These discrepancies may explain why there were more chimeras found in the ePCR samples than was expected. Since the KAPA2G-Robust and SequalPrep enzymes performed very well in normal PCR, a concerted effort should be made to prepare an oil phase that is stable enough to maintain the emulsion in the presence of their respective chemistries throughout thermal cycling. Both of these enzymes come with 'enhancers' which may contain high levels of salt that break the emulsion during cycling.

As a suggestion for future experiments to improve upon the ePCR library preparation, a more controlled method for encapsulating templates within individual emulsion vesicles should be investigated. The haphazard, batch mode of emulsion preparation works to some degree, but the size distribution of vesicles and ratio of template to vesicles cannot be fine-tuned with the method presented. RainDance Technologies has demonstrated almost complete control over these variables (51, 52, 65), and have commercial products available. A comparison of our ePCR preparations against theirs would prove useful for determining how well our bulk-preparation method performs.

Other suggestions for improving the study of PCR bias using a mock community involve sequencing each individual species in the mock using the same two-step PCR preparation procedure, taking care to accurately quantify the volume and concentration of template DNA used to obtain accurate estimates on the rrn copy number. Samples could also be prepared using the touch-down PCR technique by incorporating it into the cycle-optimized PCR procedure to see if there was any significant

effect on PCR bias and artifact formation. Finally, to better capture the less abundant species in the 'rare biosphere', special care should be taken to make sure there is no inhibition of PCR amplification. Von Wintzingerode et al. (35) suggested using bovine serum albumin and T4 gene 32 protein (gp32) in PCR preparations to prevent inhibition by contaminants from environmental DNA extracts.

5. *References*

1. Chistoserdova L (2010) Recent progress and new challenges in metagenomics for biotechnology. *Biotechnology Letters*.

2. Tringe SG, Rubin EM (2005) Metagenomics: DNA sequencing of environmental samples. *Nature Reviews Genetics* 6:805-14.

3. Wooley JC, Godzik A, Friedberg I (2010) A primer on metagenomics. *PLoS Computational Biology* 6:e1000667.

4. Kunin V, Copeland A, Lapidus A, Mavromatis K, Hugenholtz P (2008) A bioinformatician's guide to metagenomics. *Microbiology and Molecular Biology Reviews* 72:557-78.

5. Venter JC et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304:66-74.

6. Denef VJ, Mueller RS, Banfield JF (2010) AMD biofilms: using model communities to study microbial evolution and ecological complexity in nature. *The ISME Journal* 4:599-610.

7. Lazarevic V et al. (2009) Metagenomic study of the oral microbiota by Illumina high-throughput sequencing. *Journal of Microbiological Methods* 79:266-71.

8. Ley RE, Lozupone CA, Hamady M, Knight R, Gordon JI (2008) Worlds within worlds: evolution of the vertebrate gut microbiota. *Nature Reviews Microbiology* 6:776-88.

9. Dethlefsen L, Huse S, Sogin ML, Relman DA (2008) The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. *PLoS Biology* 6:e280.

10. Bäckhed F, Ley RE, Sonnenburg JL, Peterson DA, Gordon JI (2005) Host-bacterial mutualism in the human intestine. *Science* 307:1915-20.

11. Eckburg PB et al. (2005) Diversity of the human intestinal microbial flora. *Science* 308:1635-8.

12. Wommack KE, Bhavsar J, Ravel J (2008) Metagenomics: read length matters. *Applied and Environmental Microbiology* 74:1453-63.

13. Brady A, Salzberg SL (2009) Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nature Methods* 6:673-6.

14. Hoff KJ et al. (2008) Gene prediction in metagenomic fragments: a large scale machine learning approach. *BMC Bioinformatics* 9:217.

15. Sorber K et al. (2008) The Long March: a sample preparation technique that enhances contig length and coverage by high-throughput short-read sequencing. *PloS ONE* 3:e3495.

16. Mitra S, Schubach M, Huson DH (2010) Short clones or long clones? A simulation study on the use of paired reads in metagenomics. *BMC Bioinformatics* 11 Suppl 1:S12.

17. Rodrigue S et al. (2010) Unlocking Short Read Sequencing for Metagenomics. *PLoS ONE* 5:e11840.

18. Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology* 73:5261-7.

19. Liu Z, Lozupone C, Hamady M, Bushman FD, Knight R (2007) Short pyrosequencing reads suffice for accurate microbial community analysis. *Nucleic Acids Research* 35:e120.

20. Ewing B, Hillier L, Wendl MC, Green P (1998) Base-Calling of Automated Sequencer Traces Using Phred I Accuracy Assessment. *Genome Research*:175-185.

21. Gray MW, Sankoff D, Cedergren RJ (1984) On the evolutionary descent of organisms and organelles: a global phylogeny based on a highly conserved structural core in small subunit ribosomal RNA. *Nucleic Acids Research* 12:5837-5852.

22. Woese CR, Fox GE (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences of the United States of America* 74:5088-90.

23. Olsen GJ, Lane DJ, Giovannoni SJ, Pace NR, Stahl DA (1986) Microbial ecology and evolution: a ribosomal RNA approach. *Annual Reviews in Microbiology* 40:337-65.

24. Ward D, Weller R, Bateson M (1990) 16s rRNA sequences reveal numerous uncultured microorganisms in a natural community. *Nature* 345:63-65.

25. Baker G, Smith J, Cowan D (2003) Review and re-analysis of domain-specific 16S primers. *Journal of Microbiological Methods* 55:541-555.

26. Metzker ML (2010) Sequencing technologies - the next generation. *Nature Reviews Genetics* 11:31-46.

27. MacLean D, Jones JD, Studholme DJ (2009) Application of 'next-generation' sequencing technologies to microbial genetics. *Nature Reviews Microbiology* 7:287-296.

28. Shendure J, Ji H (2008) Next-generation DNA sequencing. *Nature Biotechnology* 26:1135-45.

29. Ansorge WJ (2009) Next-generation DNA sequencing techniques. *New Biotechnology* 25:195-203.

30. Metzker ML (2005) Emerging technologies in DNA sequencing. *Genome Research* 15:1767-76.

31. Margulies M et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376-80.

32. Fedurco M, Romieu A, Williams S, Lawrence I, Turcatti G (2006) BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Research* 34:e22.

33. Turcatti G, Romieu A, Fedurco M, Tairi A (2008) A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic Acids Research* 36:e25.

34. Mayer P, Farinelli L, Kawashima E (1998) Method of nucleic acid amplification - WO9844151A1.

35. von Wintzingerode F, Göbel UB, Stackebrandt E (1997) Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. *FEMS Microbiology Reviews* 21:213-29.

36. Frostegård A et al. (1999) Quantification of bias related to the extraction of DNA directly from soils. *Applied and Environmental Microbiology* 65:5409-20.

37. Kauffmann IM, Schmitt J, Schmid RD (2004) DNA isolation from soil samples for cloning in different hosts. *Applied Microbiology and Biotechnology* 64:665-70.

38. Watanabe K, Kodama Y, Harayama S (2001) Design and evaluation of PCR primers to amplify bacterial 16S ribosomal DNA fragments used for community fingerprinting. *Journal of Microbiological Methods* 44:253-62.

39. Polz MF, Cavanaugh CM (1998) Bias in template-to-product ratios in multitemplate PCR. *Applied and Environmental Microbiology* 64:3724-30.

40. Suzuki MT, Giovannoni SJ (1996) Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR Bias Caused by Template Annealing in the Amplification of Mixtures of 16S rRNA Genes by PCR †. *Applied and Environmental Microbiology*.

41. Henke W, Herdel K, Jung K, Schnorr D, Loening SA (1997) Betaine improves the PCR amplification of GC-rich DNA sequences. *Nucleic Acids Research* 25:3957-8.

42. Farrelly V, Rainey FA, Stackebrandt E (1995) Effect of genome size and rrn gene copy number on PCR amplification of 16S rRNA genes from a mixture of bacterial species. *Applied and Environmental Microbiology* 61:2798-801.

43. Raeymaekers L (1995) A commentary on the practical applications of competitive PCR. *Genome Research* 5:91-94.

44. Qiu X, Wu L, Huang H, Donel PE (2001) Evaluation of PCR-generated chimeras, mutations, and heteroduplexes with 16S rRNA gene-based cloning. *Applied and Environmental Microbiology* 67:880-87.

45. Wang GC, Wang Y (1997) Frequency of formation of chimeric molecules as a consequence of PCR coamplification of 16S rRNA genes from mixed bacterial genomes. *Applied and Environmental Microbiology* 63:4645-50.

46. Thompson JR, Marcelino LA, Polz MF (2002) Heteroduplexes in mixed-template amplifications: formation, consequence and elimination by 'reconditioning PCR'. *Nucleic Acids Research* 30:2083-8.

47. Gomez-Alvarez V, Teal TK, Schmidt TM (2009) Systematic artifacts in metagenomes from complex microbial communities. *The ISME Journal* 3:1314-7.

48. Meyerhans A, Vartanian J, Wain-Hobson S (1990) DNA recombination during PCR. *Nucleic Acids Research* 18:1687-1691.

49. Ruano G, Kidd KK (1992) Modeling of heteroduplex formation during PCR from mixtures of DNA templates. *PCR Methods and Applications* 2:112-6.

50. Griffiths AD, Tawfik DS (2006) Miniaturising the laboratory in emulsion droplets. *Trends in Biotechnology* 24:395-402.

51. Kiss MM et al. (2008) High-throughput quantitative polymerase chain reaction in picoliter droplets. *Analytical Chemistry* 80:8975-81.

52. Leamon JH, Link DR, Egholm M, Rothberg JM (2006) Overview: methods and applications for droplet compartmentalization of biology. *Nature Methods* 3:541-3.

53. Williams R et al. (2006) Amplification of complex gene libraries by emulsion PCR. *Nature Methods* 3:545-550.

54. Cárdenas A, Castro E (2003) Breaking of multiple emulsions under osmotic pressure and the effect of W1/O relation. *Interciencia* 28:534-538.

55. Korbie DJ, Mattick JS (2008) Touchdown PCR for increased specificity and sensitivity in PCR amplification. *Nature Protocols* 3:1452-6.

56. Hiatt JB, Patwardhan RP, Turner EH, Lee C, Shendure J (2010) Parallel, tag-directed assembly of locally derived short sequence reads. *Nature Methods* 7:119-22.

57. Lennon NJ et al. (2010) A scalable, fully automated process for construction of sequence-ready barcoded libraries for 454. *Genome Biology* 11:R15.

58. Binladen J et al. (2007) The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PloS ONE* 2:e197.

59. Caporaso JG et al. (2010) Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences of the United States of America* Early Edition:1-7.

60. Hamady M, Walker JJ, Harris JK, Gold NJ, Knight R (2008) Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nature Methods* 5:235-37.

61. DeAngelis MM, Wang DG, Hawkins TL (1995) Solid-phase reversible immobilization for the isolation of PCR products. *Nucleic Acids Research* 23:4742-3.

62. Hawkins TL, O'Connor-Morin T, Roy A, Santillan C (1994) DNA purification and isolation using a solid-phase. *Nucleic Acids Research* 22:4543-4.

63. Panaro NJ et al. (2000) Evaluation of DNA fragment sizing and quantification by the Agilent 2100 Bioanalyzer. *Clinical Chemistry* 46:1851-3.

64. Holmberg A et al. (2005) The biotin-streptavidin interaction can be reversibly broken using water at elevated temperatures. *Electrophoresis* 26:501-10.

65. Tewhey R et al. (2009) Microdroplet-based PCR enrichment for large-scale targeted sequencing. *Nature Biotechnology* 27:1025-31.