UNIVERSITY OF CALIFORNIA
RIVERSIDE

Adapting Polony Technology to Oligonucleotide Fingerprinting of Ribosomal rRNA Genes for Microbial Community Analysis

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Genetics, Genomics and Bioinformatics

by

Paul Michael Ruegger

March 2011

Dissertation Committee:
Dr. James Borneman, Chairperson
Dr. Thomas Girke
Dr. Tao Jiang

The Dissertation of Paul Michael Ruegger is approved:

_____

_____

_____
Committee Chairperson

University of California, Riverside

ACKNOWLEDGMENTS

I wish to thank my advisor, Dr. James Borneman, for all his help, support and guidance. I also wish to thank my parents, family and close friends, whose blessing and encouragement sustained me innumerable times.

ABSTRACT OF THE DISSERTATION

Adapting Polony Technology to Oligonucleotide Fingerprinting of Ribosomal rRNA Genes for Microbial Community Analysis

by

Paul Michael Ruegger

Doctor of Philosophy, Graduate Program in
Genetics, Genomics and Bioinformatics
University of California, Riverside, March 2011
Dr. James Borneman, Chairperson

Bacteria are present in nearly all terrestrial environments and play varied and important roles. Understanding their impacts on the environments and hosts where they reside is greatly aided by an accurate estimation of the number and types present. We have adapted polony technology to Oligonucleotide Fingerprinting of Ribosomal rRNA Genes (OFRG), a hybridization-based method for clustering similar 16S rDNA sequences. We present a new OFRG probe set design method that utilizes the available taxonomic information of training sequences to improve the clustering of fingerprints into biologically meaningful groups. A software tool is presented that quickly and accurately identifies randomly placed polonies in microarray images. The polony OFRG method is applied to DNA from a mock bacterial community created from a clone library, as well as to PCR amplicons made from the same mock community to examine PCR bias. We also examine several natural bacterial communities, making

polonies starting directly from genomic DNA templates. The method successfully clusters the known bacterial community and reveals the presence of artifacts in template from the mock community PCR. Natural bacterial communities are differentiated using a weighted UniFrac analysis. Due to the initial spatial separation of sample DNA strands, polonies are essentially free of the PCR bias and chimeric sequence formation that occurs in mixed-template PCR reactions. An additional benefit of the polony format is that sequences of near full-length rDNA can be obtained when desired – a feature not possible with current high-throughput sequencing methods. We anticipate polony OFRG may be an invaluable tool for microbial population studies where these two characteristics are required.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

Microbes often exist in complex and dynamic communities that can have profound effects on the environments or hosts in which they live. A better understanding of these interactions and the impacts microbes have on their hosts is needed and can begin by an assessment of which microbes are present. An even better understanding of these interactions is made possible by frequent sampling, such that the changes in population levels themselves can be scrutinized for clues regarding the interplay between microbe and host.

Many methods currently exist to study microbial communities. These methods range from inexpensive, coarse-grained tools such as culturing, to methods that detect various characteristic differences in microbial rRNA genes such as denaturing gradient gel electrophoresis (DGGE)(Muyzer 1999) and terminal restriction fragment length polymorphism (T-RFLP) (Schütte et al. 2008), to the significantly more expensive and more accurate “gold-standard” of sequencing near full-length rRNA genes (Frank et al. 2007). Recently, strategies for using high-throughput sequencing machines for microbial community analysis have been developed as well (Wu et al. 2010)(Caporaso, Lauber, et al. 2010).

The coarse-grained methods are useful for examining large changes in microbial communities but the low resolution is inadequate for many studies. Sequencing near full-length 16S rRNA genes provides the highest available taxonomic resolution when an accurate “snapshot” of a microbial community is

required. However, though costs are dropping, multi-sample longitudinal studies that employ full-length sequencing are often still too expensive for many labs. High-throughput sequencing currently provides the best compromise between accuracy and throughput but due to the short read-lengths these are still limited in describing the taxonomic makeup of a microbial community. Currently, taxonomic assignments can be confidently made only at the order level; assignments at the genus level can also be made but with less confidence. (Wu et al. 2010)(Caporaso, Lauber, et al. 2010).

The focus of this research is on improving an alternative method for detecting changes in microbial communities termed oligonucleotide fingerprinting of ribosomal rRNA genes (OFRG) (Valinsky, G. Della Vedova, T. Jiang, et al. 2002). OFRG may be useful for multi-sample studies requiring low cost and high taxonomic resolution. In addition, the new OFRG method that this research focuses on has two important advantages over current sequencing methods. First, the pre-sequencing PCR step known to bias results is skipped (Suzuki and Giovannoni 1996). Second, near full-length rRNA genes are available for sequencing, when desired. The former allows for a truer depiction of the microbes present and the latter provides a way to more confidently assess the identity of any microbe or group of microbes present in a sample.

To estimate the proportions of microbial species present in an environment the OFRG method uses a set of 40 computer-designed DNA probes chosen from a set of training sequences and hybridizes them against an array of

sample-derived microbial rRNA gene clones. The hybridization affinity of each probe/clone combination is quantified and then processed in one of two ways. Traditionally these data are transformed into a 40-digit binary “fingerprint” for each clone, where a 0 denotes no hybridization and a 1 denotes a successful hybridization event. These fingerprints are then clustered based on their similarity to the fingerprints of other clones in the array. Alternatively, hierarchical clustering can be performed on the data without a binary transformation of the hybridization intensities. These clusters provide a low-cost estimate of the relative proportions of the various microbial taxa present in an environment since similar fingerprints arise from similar rRNA genes.

OFRG originally employed a printed macroarray format. In a labor intensive procedure, the DNA for each spot on the array was printed from the PCR products of a clone library of microbial rRNA genes originating from environmental samples. The capacity of the method was 9,600 clones per experiment and has been used successfully in several studies (E. Bent et al. 2006)(Lee et al. 2008)(McGuire et al. 2010). The new OFRG method will replace the labor-intensive macroarray with a low-cost microarray, termed a “polony” microarray, with a current capacity of 1K-5K clones per sample and a theoretical capacity of perhaps millions.

Polonies, or “polymerase colonies,” are analogous to and replace the spots of the macroarray. Rather than being printed, each polony is grown in place; each polony consists of many thousands of localized copies of an

individual DNA molecule generated through solid-phase PCR in a polyacrylamide hydrogel. Because diffusion of the PCR amplicons is inhibited somewhat by the gel the process results in spots of DNA molecules – the polonies – randomly placed in the gel, one for each original DNA molecule.

Polony technology has been employed in several applications such as DNA sequencing, SNP detection, gene expression studies, genotyping, haplotyping and alternative pre-mRNA splicing (Shendure et al. 2005)(Butz et al. 2004)(Rieger et al. 2007)(Robi D Mitra et al. 2003)(Zhang et al. 2006)(Jun Zhu et al. 2003). There are several important characteristics of a polony microarray that also make it a useful tool for OFRG. These are, 1) polony DNA is anchored to the gel and can be made single-stranded, making it durable and available for probe hybridizations, 2) sample microbial DNA can be spatially isolated, thus eliminating the formation of chimeric amplicons during PCR and 3) microbial genomic DNA can be used directly, without an intermediate PCR step, thus reducing PCR bias of the true makeup of a microbial community.

Despite these advantageous characteristics, however, polony technology is not without its challenges. The most difficult of these challenges is the random placement of polonies in the hydrogel microarray. As a consequence, some polonies overlap with other polonies to varying degrees. Another challenge is that polonies vary in diameter and can contain different amounts of DNA. These characteristics make detection and/or quantification of polony intensities difficult, and most existing microarray software is unable to properly handle these issues

as they are designed for ordered arrays. Lastly, the OFRG paradigm of an "inverted" array (where the spots are sample-DNA and a small set of probes are sequentially hybridized to the array) necessitates that the probes be carefully selected in order to maximize the information gleaned from them. Part of this research involves the development of bioinformatics tools to address the challenging aspects of polonies. Finally, several samples of bacterial DNA were analyzed with the polony OFRG method and the results are reported herein.

This research involves strategies to overcome these challenges. Chapter two explains a method for OFRG probe set design that considers the taxonomic information of available training sequences. Chapter three deals with an approach to handle the inherent difficulties involved in gathering data from polony hybridization images. Chapter four presents the working polony OFRG method with experiments that show its ability to distinguish a range of bacterial communities.

BIBLIOGRAPHY

Bent, E., B. Yin, A. Figueroa, J. Ye, Q. Fu, Z. Liu, V. McDonald, D. Jeske, T. Jiang, and J. Borneman. 2006. Development of a 9600-clone procedure for oligonucleotide fingerprinting of rRNA genes: Utilization to identify soil bacterial rRNA genes that correlate in abundance with the development of avocado root rot. *J Microbiol Methods* 67, no. 1: 171-80.

Butz, James A, Hai Yan, Venugopal Mikkilineni, and Jeremy S Edwards. 2004. Detection of allelic variations of human gene expression by polymerase colonies. *BMC Genetics* 5 (February 16): 3. doi:10.1186/1471-2156-5-3.

Caporaso, J Gregory, Christian L Lauber, William A Walters, Donna Berg-Lyons, Catherine A Lozupone, Peter J Turnbaugh, Noah Fierer, and Rob Knight. 2010. Microbes and Health Sackler Colloquium: 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* (June 3). doi:10.1073/pnas.1000080107. http://www.ncbi.nlm.nih.gov/pubmed/20534432.

Frank, Daniel N, Allison L St Amand, Robert A Feldman, Edgar C Boedeker, Noam Harpaz, and Norman R Pace. 2007. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proceedings of the National Academy of Sciences of the United States of America* 104, no. 34 (August 21): 13780-13785. doi:10.1073/pnas.0706625104.

McGuire, Krista L, Elizabeth Bent, James Borneman, Arundhati Majumder, Steven D Allison, and Kathleen K Tresederi. 2010. Functional diversity in resource use by fungi. *Ecology* 91, no. 8 (August): 2324-2332.

Mitra, Robi D, Vincent L Butty, Jay Shendure, Benjamin R Williams, David E Housman, and George M Church. 2003. Digital genotyping and haplotyping with polymerase colonies. *Proceedings of the National Academy of Sciences of the United States of America* 100, no. 10 (May 13): 5926-5931. doi:10.1073/pnas.0936399100.

Muyzer, G. 1999. DGGE/TGGE a method for identifying genes from natural ecosystems. *Current Opinion in Microbiology* 2, no. 3 (6): 317-322. doi:10.1016/S1369-5274(99)80055-1.

Rieger, C, R Poppino, R Sheridan, K Moley, R Mitra, and D Gottlieb. 2007. Polony analysis of gene expression in ES cells and blastocysts. *Nucleic Acids Research* 35, no. 22: e151. doi:10.1093/nar/gkm1076.

Schütte, Ursel M E, Zaid Abdo, Stephen J Bent, Conrad Shyu, Christopher J Williams, Jacob D Pierson, and Larry J Forney. 2008. Advances in the use of terminal restriction fragment length polymorphism (T-RFLP) analysis of 16S rRNA genes to characterize microbial communities. *Applied Microbiology and Biotechnology* 80, no. 3 (September): 365-380. doi:10.1007/s00253-008-1565-4.

Shendure, Jay, Gregory J Porreca, Nikos B Reppas, Xiaoxia Lin, John P McCutcheon, Abraham M Rosenbaum, Michael D Wang, Kun Zhang, Robi D Mitra, and George M Church. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science (New York, N.Y.)* 309, no. 5741 (September 9): 1728-1732. doi:10.1126/science.1117389.

Suzuki, M T, and S J Giovannoni. 1996. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Applied and Environmental Microbiology* 62, no. 2 (February): 625-630.

Valinsky, L., G. Della Vedova, T. Jiang, and J. Borneman. 2002. Oligonucleotide fingerprinting of rRNA genes for analysis of fungal community composition. *Appl Environ Microbiol* 68, no. 12: 5999-6004.

Wu, Gary D, James D Lewis, Christian Hoffmann, Ying-Yu Chen, Rob Knight, Kyle Bittinger, Jennifer Hwang, et al. 2010. Sampling and pyrosequencing methods for characterizing bacterial communities in the human gut using 16S sequence tags. *BMC Microbiology* 10, no. 1: 206. doi:10.1186/1471-2180-10-206.

Ye, Jingxiao, Jimmy W Lee, Laura L Presley, Elizabeth Bent, Bo Wei, Jonathan Braun, Neal L Schiller, Daniel S Straus, and James Borneman. 2008. Bacteria and bacterial rRNA genes associated with the development of colitis in IL-10(-/-) mice. *Inflammatory Bowel Diseases* 14, no. 8 (August): 1041-1050. doi:10.1002/ibd.20442.

Zhang, Kun, Jun Zhu, Jay Shendure, Gregory J Porreca, John D Aach, Robi D Mitra, and George M Church. 2006. Long-range polony haplotyping of individual human chromosome molecules. *Nature Genetics* 38, no. 3 (March): 382-387. doi:10.1038/ng1741.

Zhu, Jun, Jay Shendure, Robi D Mitra, and George M Church. 2003. Single molecule profiling of alternative pre-mRNA splicing. *Science (New York, N.Y.)* 301, no. 5634 (August 8): 836-838. doi:10.1126/science.1085792.

Chapter 2: Probe Set Design

2.1 Introduction

Microbes often exist in complex and dynamic communities that can have profound effects on the environments or hosts in which they live. A better understanding of these interactions and the impacts microbes have on their hosts is needed and must begin by an assessment of which microbes are present. An even better understanding of these interactions would be made possible by frequent sampling, such that the changes in population levels themselves could be scrutinized for clues regarding the interplay between microbe and host.

Many methods currently exist to study microbial communities. These methods range from inexpensive, coarse-grained tools such as culturing, denaturing gradient gel electrophoresis (DGGE) (Muyzer 1999) and terminal restriction fragment length polymorphism (T-RFLP) (Schütte et al. 2008), to the significantly more expensive and more accurate “gold-standard” of sequencing near full-length rRNA genes (Frank et al. 2007). .

The coarse-grained methods are useful for examining large changes in microbial communities but the low resolution is inadequate for some types of studies. Sequencing near full-length 16S rRNA genes provides the highest available taxonomic resolution when an accurate “snapshot” of a microbial community is required. However, though costs are dropping, multi-sample longitudinal studies that employ full-length sequencing are often still too expensive for many labs. High-throughput sequencing currently provides the

best compromise between accuracy and throughput but due to the short read-lengths these are limited to describing a microbial community confidently only at the order level and some confidence at the genus level (Wu et al. 2010) (Caporaso, Lauber, et al. 2010).

This study focuses on improving an alternative method for analyzing microbial communities, termed oligonucleotide fingerprinting of ribosomal rRNA genes (OFRG) (Valinsky, G. Della Vedova, T. Jiang, et al. 2002), which can be made both accurate and inexpensive, and may be useful for studies that require many samples at higher taxonomic resolution than current high-throughput sequencing methods provide.

To estimate the proportions of microbial phylotypes present in an environment the OFRG method uses a set of 40 computer-designed DNA probes chosen from a set of training sequences to hybridize against an array of sample-derived microbial rRNA gene clones (J Borneman et al. 2001). The hybridization affinity of each probe/clone combination is quantified and transformed into a 40-digit binary “fingerprint” for each clone. These experimentally-derived fingerprints can be clustered based on their similarity to the fingerprints of other clones in the array. Because similar fingerprints arise from similar rRNA genes, and because the new OFRG arrays are inexpensive to produce and contain many thousands of clones, these clusters can provide a low-cost estimate of the relative proportions of the various microbial species present in an environment.

One challenge with the ORFG method is choosing an optimal set of probes. Previous work to create a probe set for OFRG built upon the work of Drmanac and Maier (R Drmanac and S Drmanac 1999)(S Drmanac and R Drmanac 1994) (Meier-Ewert et al. 1998) which investigated strategies to screen cDNA and BAC clone libraries with carefully chosen sets of probes. This concept was adapted to microbial community analysis by Borneman (J Borneman et al. 2001) that used available 16S rRNA gene sequences as training data. The optimal probe set of Borneman most pertinent to this work is termed the Maximum Distinguishing Probe Set (MDPS).

As the name implies, the MDPS attempts to create a probe set that produces a distinct binary fingerprint for all training sequences – maximizing the ability of the probe set to distinguish all sequences. Neither sequence similarity nor taxonomy is taken into account, however. By chance, fingerprints from similar DNA sequences do tend be similar or identical to each other, and fingerprints coming from dissimilar DNA sequences tend to be dissimilar to each other – but this is not always the case.

Although MDPS has been used successfully in several studies (Valinsky, G. Della Vedova, Scupham, et al. 2002) (Yin et al. 2003) (Scupham et al. 2006) (E. Bent et al. 2006)(Jingxiao Ye et al. 2008), the limitation of the MDPS from a biological perspective is that it considers all undistinguished clones (those having the same fingerprint) equally undesirable. More specifically, it makes no attempt to produce a probe set based on the taxonomy of sequences having the same

fingerprints – very divergent sequences having the same fingerprint are considered no worse than very similar sequences having the same fingerprint. In the present study we address this shortcoming of the MDPS with a new formulation for heuristic probe selection termed the Maximum Fidelity Probe Set (MFPS) and a new processing pipeline for preparing the training data used by the MFPS.

2.2 Methods

The new method involves a change to the cost function within the simulated annealing algorithm used by Borneman et al (2001) (J Borneman et al. 2001). In addition, a processing pipeline was developed to prepare the training data. Within the simulated annealing algorithm the Maximum Fidelity Probe Set (MFPS) scores each transient probe set using multiple penalty levels corresponding to the taxonomic levels of the training sequences. After many iterations of (random) probe substitution/probe set evaluation a final probe set is output. Below we describe the new pipeline and cost function, highlighting the elements contributing to improved performance.

Data processing pipeline. The processing pipeline prepares the training data for the cost functions to operate on. The three most important differences between the new and original processing pipelines are that in the new pipeline the sequences have their hypervariable regions removed, are clustered into species-like operational taxonomic units (OTUs) and are labeled with their OTU and higher-level taxonomic information.

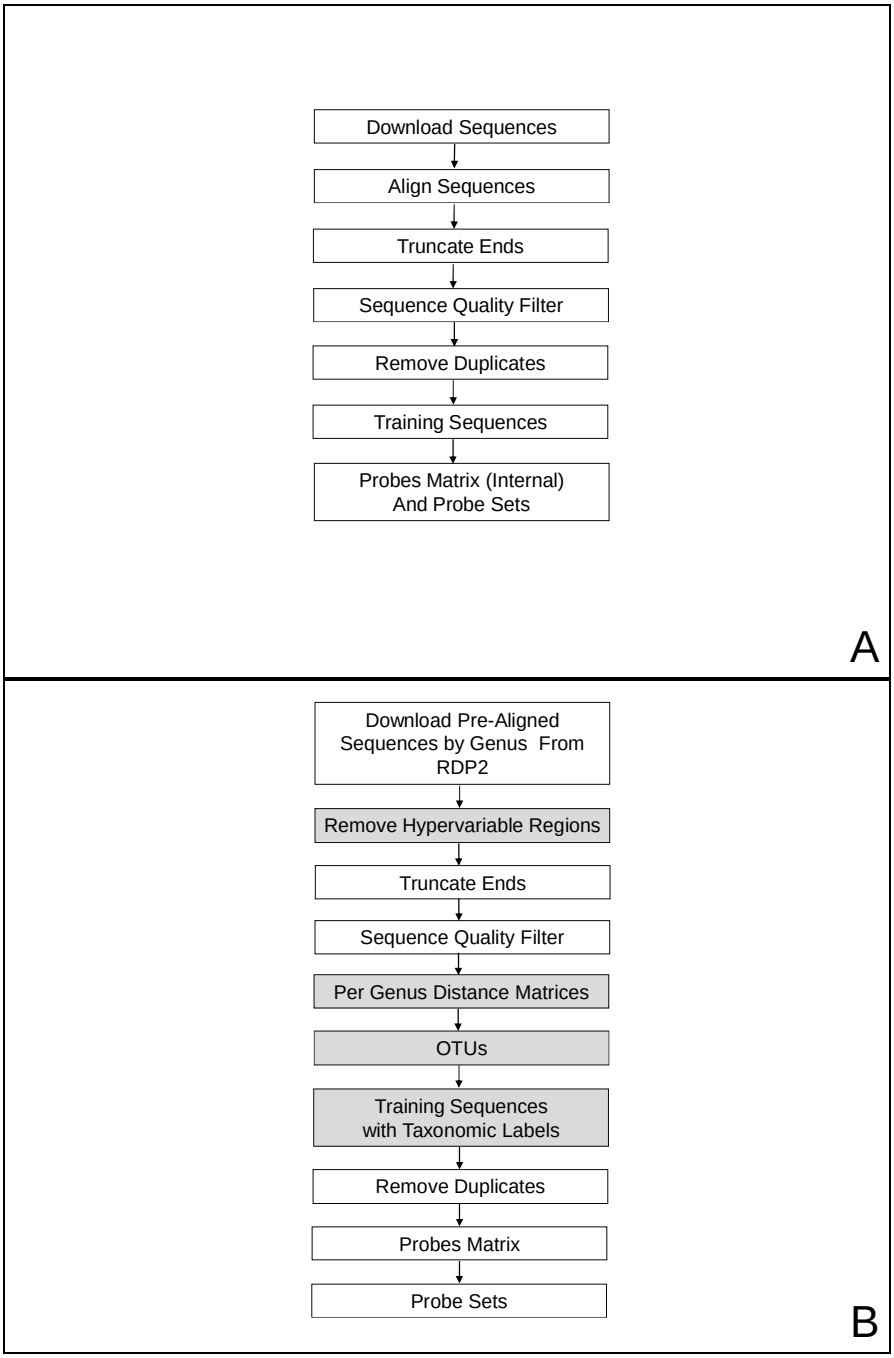

Figure 2.1. Diagrams of the New and Original Processing Pipelines. (**A**) the original processing pipeline and (**B**) the new processing pipeline for training sequences. The three main differences (shaded boxes) in the new are 1) sequences have their hypervariable regions removed, 2) distance matrices allow grouping (<= 1% sequence difference) into Operational Taxonomic Units (OTUs), and 3) sequences are labeled with their taxonomic designations.

Figures 2.1, A and B show the new and original processing pipelines, respectively. Both pipelines start with downloading rRNA gene sequences. However, the new processing pipeline gathers pre-aligned sequences and a "mask" sequence denoting the location of hypervariable regions within the alignment (see first shaded box, Figure 2.1B); these are used in combination to remove the hypervariable regions in the sequences, as any probes designed to bind in those regions would hybridize to only a few taxonomic groups and thus provide little to no help in distinguishing most other taxonomic groups.

In addition, the pre-aligned sequences simplify the creation of distance matrices used to create OTUs and the task of truncating the ends of the sequences. It is useful to truncate the ends to create more consistent training data. To do so we truncated ten nucleotide positions "inward" of the locations of two highly conserved primer regions (27F – AGAGTTTGATCMTGGCTCAG and 1392R – ACGGGCGGTGTGTRC), thus leaving only the portions of the 16S molecule intended as the target for probes. For both pipelines a sequence was considered too short and rejected if there was an end gap in the alignment after truncation and the truncated section from that end contained only gaps.

Using these aligned sequences, we then use the program MOTHUR (Schloss et al. 2009) to make distance matrices and OTUs on a per genera basis (middle two shaded boxes, Figure 2.1B). OTUs were made with a minimum sequence similarity of 99%. The OTU, genera and phyla information was then

concatenated to the corresponding DNA sequences. The last step in preparing the training sequences is checking for and removing any duplicate sequences.

Both processing pipelines then create a probe matrix from the training sequences. The matrices are comprised of a list of candidate probes and their putative binding ability to each of the training sequences and include the taxonomic information of each sequence (bottom shaded box, Figure 2.1B). Making a matrix once and saving it allows the cost functions to operate more efficiently. Constructing the probe matrix begins by creating a list of all 10mers that occur at least once in the training sequence data. This list can grow to over 750,000 probes depending on the size of the data set and must be reduced due to practical considerations of computational time and memory limitations. The size reduction is accomplished by a filtering step to keep only 1000 of the most highly conserved probes (based on how many OTUs a probe is found in). For each probe/sequence combination in the probes matrix, a 1 or 0 denotes whether the probe sequence was found in or not found in the training sequence, respectively. Taxonomic data are converted to numbers and added to the probes matrix so it is accessible to the MFPS. The original MDPS cost function uses the same probe and binding information but the taxonomic information is ignored.

To compare the two pipelines we made training sequences and probe matrices with both. The training data from the original pipeline differs from the new in that the hypervariable regions were not removed from the sequences prior

to making the probe matrix, and the list of candidate probes in the two matrices are not identical because of this. To examine just the pipeline's effect on probe sets, apart from any added benefit of using taxonomic information, we employed only the original MDPS cost function, making probe sets of sizes 20, 30, 40, 60 and 80 probes per probe set.

Maximum Fidelity Probe Set (MFPS). By employing a heuristic strategy, the MFPS scores a randomly chosen set of probes using multi-level penalties corresponding to the taxonomic levels of the training sequences. By doing so, it addresses the main weakness of the original MDPS cost function, which attempts to choose a probe set that creates a distinct binary fingerprint for each training sequence without regard to sequence similarity or taxonomy.

To adequately explain the MFPS, we first define several terms. A *simulated fingerprint* is a binary vector of *k* digits representing the putative hybridization pattern of *k* DNA probes on a DNA sequence of interest. For our purposes, the sequences we are interested in are bacterial 16S rRNA genes and the DNA probes are 10 bases long. If the sequence of a probe occurs exactly in the sequence of a gene, we assume it would hybridize to the gene in a real hybridization experiment and if it does not occur exactly we assume it would not hybridize. Therefore, we place a 1 or 0 into each of the *k* characters of the simulated fingerprint of a gene sequence to denote a putatively successful or unsuccessful hybridization event for each of the *k* probes of a probe set.

A *distinct fingerprint* is a single representative of a group of identical simulated fingerprints produced by a probe set $\mathcal{P}$ in a set of sequences $\mathcal{S}$. It is useful in determining a probe set's quality score - its *fidelity*.

The term *fidelity* is explained as follows. If a distinct fingerprint *f* is produced by probe set $\mathcal{P}$ on one or more sequences in taxonomic group γ in a set of sequences $\mathcal{S}$, and *f* is *not* produced in any other taxonomic group at the same level as γ, then *f* is said to have high fidelity – a desirable trait. Conversely, if fingerprint *f* is produced on one or more sequences outside of taxonomic group γ in $\mathcal{S}$, then *f* is said to have low fidelity. Additionally, the more groups outside of γ where fingerprint *f* is produced, the lower its fidelity is said to be.

Note that fidelity is always associated with a taxonomic level. For instance, a distinct fingerprint *f* may have low fidelity at the OTU level (if it occurs in the sequences of two or more OTUs) yet have high fidelity at the genus level (if it occurs in the sequences of only one genus). The goal is to choose a set of probes that together produce high-fidelity distinct fingerprints at the taxonomic level(s) desired. If this can be achieved, distinct fingerprints arise within biologically meaningful taxonomic groupings and can be used as proxies for them. To that end, probe sets are evaluated in the MFPS by the cost function,

$$C = \frac{1}{2}\sum_{f=1}^{N}\sum_{i=1}^{3} P_i \gamma_{i,f}\left(\gamma_{i,f} - 1\right)$$

where *C* is the total cost, N is the number of distinct fingerprints produced by the probe set on the training sequences, *i* is one of three taxonomic levels (we used OTU, genus and phyla but others could be used), *f* is an individual, distinct fingerprint, $\gamma_{i,f}$ is the number of taxonomic groups where *f* occurs at taxonomic level *i*, and P_i is the penalty (for low-fidelity fingerprints) at taxonomic level *i*. Note that if a distinct fingerprint is found in only one taxonomic group ($\gamma_{i,f}$=1) then no penalty will accrue to the probe set from that fingerprint. This cost function of our MFPS replaces the cost function in the simulated annealing algorithm used by Borneman (J Borneman et al. 2001).

Note that the cost function allows one to vary the penalty level for up to three taxonomic levels simultaneously. Experiments to find optimal penalty settings were conducted by systematically varying them and comparing the results. These experiments were conducted with probe sets containing 20, 30, 40, 60 and 80 probes. For each experiment at each penalty level and probe set size, one hundred probe sets were created for the MFPS and MDPS cost functions.

When cross-validation was performed, we used a variation of 5-fold cross-validation. Instead of the traditional 80% training/20% validation, we chose to use a 20% training/100% validation strategy. Due to the nature of one of our evaluation metrics, this strategy allowed us to better compare the results of other tests where we used 100% of the training data to make and evaluate probe sets.

The 20%/100% also provides a more stringent test of probe set design than 80%/20%. All cross-validation data shown are an average of 5-fold results.

Evaluation Metrics. Two evaluation metrics are used to compare the two pipelines and cost functions. The first metric is termed the high fidelity ratio (HFR), which is the ratio of distinct high-fidelity fingerprints produced by probe set $\mathcal{P}$ (on validation data) and the total number of distinct fingerprints produced by $\mathcal{P}$ on the same data. In essence, the HFR is a measure of how closely the simulated fingerprints arising from a probe set on the training sequences are representing real OTUs and genera. Importantly, the HFR metric is comparable across probe sets; because the raw scores of the cost functions are dependent upon the penalty levels chosen, as well as the number of probes in a probe set, they cannot be used to compare probe sets made with different penalty levels or different numbers of probes. Note that a probe set can have one HFR for each taxonomic level evaluated. In our experiments we examine OTU and genera HFRs only as phyla HFR automatically improves when lower-level fidelity improves.

The second metric is the average pairwise sequence distances of each distinct low-fidelity fingerprint in a probe set. Rather than a single number, this metric is shown as a line graph and is constructed as follows. For each low-fidelity distinct fingerprint *f* in probe set $\mathcal{P}$, we take all sequences having *f* and compute their average pairwise sequence distance. Bin each average into bin sizes of 1% difference. Continue this for as many probe sets as were made for

the experiment (usually 100) and graph the overall averages for each bin. Note that it is not necessary to examine the high-fidelity distinct fingerprints as they cannot, by definition, exceed the OTU cutoff threshold of 1% sequence difference.

Both new and original processing pipeline scripts were written in Perl, version 5.8. Sequences and taxonomic information were downloaded from Ribosomal Database II (Maidak et al. 1999).

2.3 Results and Discussion

Comparison of Data Processing Pipelines. We compared the new and original processing pipelines using the High Fidelity Ratio (HFR) metric and the Maximum Distinguishing Probe Set (MDPS) cost function of Borneman (J Borneman et al. 2001); the MDPS does not use taxonomic information so any differences in the results can be attributed solely to the pipelines.

The new processing pipeline shows an improved OTU HFR over the original pipeline in probe sets ranging in size from 20 – 80 probes (Figure 2.2A). The improvement is approximately the same across the range of probe set sizes. The poorer performance of the original pipeline is likely due to the increased number of OTUs created by it, as having more OTUs will tend to lower the odds of successfully distinguishing them. There were 203218 sequences distributed in 34701 OTUs using the new pipeline and 216414 sequences distributed in 52983 OTUs with the original. The average OTU sizes for the new and original pipelines are 5.86 and 4.08 sequences, respectively. The increased numbers of

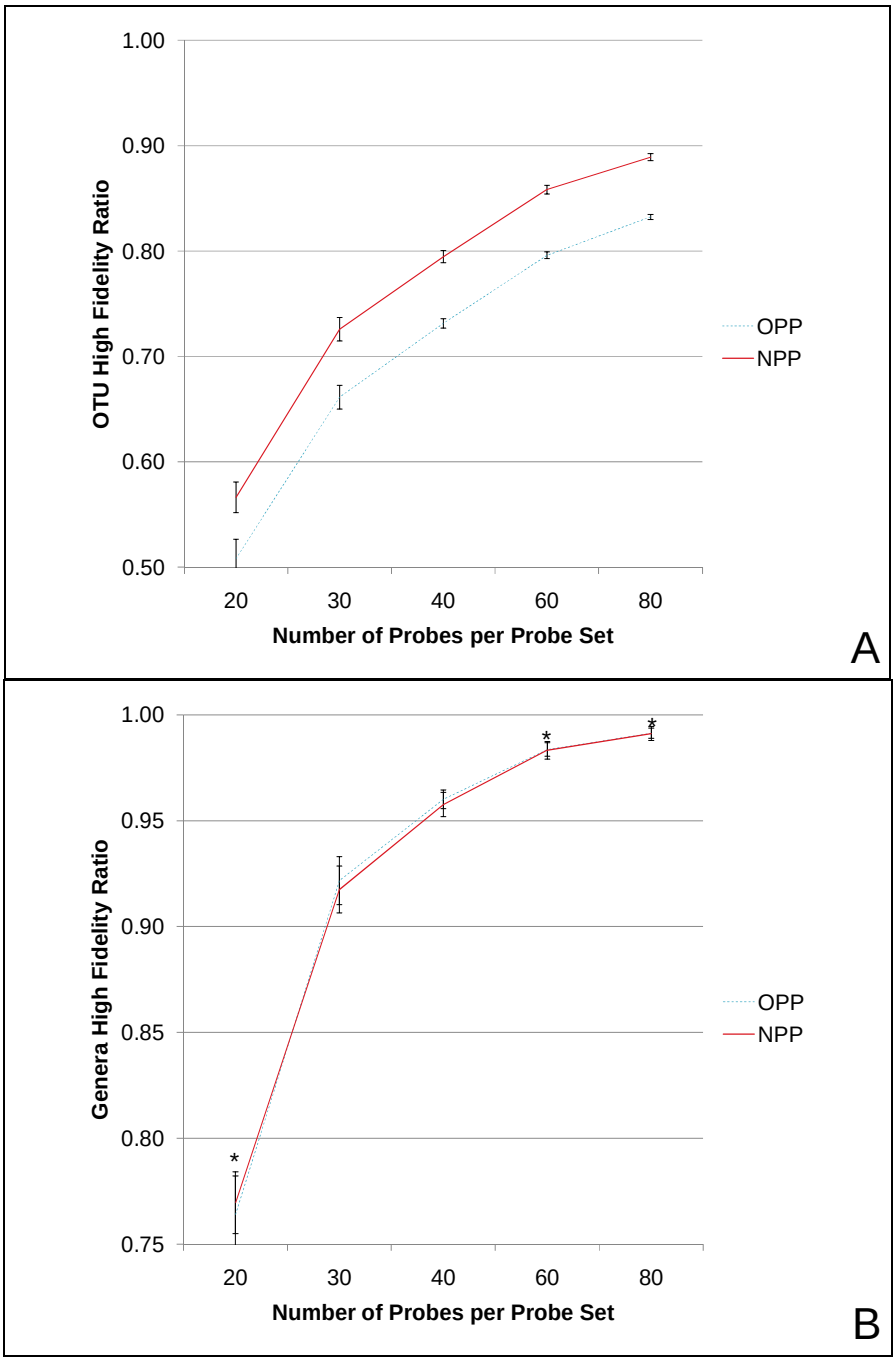

Figure 2.2. Processing pipeline's effect on fidelity. The effect on fidelity of the new (NPP) and original processing pipelines (OPP) in a range of probe set sizes using only the MDPS cost function with the 20%/100% training/validation data. (**A**) OTU fidelity appears higher in the new. (**B**) Genera fidelity shows little difference between the two pipelines. Error bars are standard deviations of 25 probe sets per data point. All unstarred comparative points are significant using the two-tailed Student's t-test assuming unequal variance ($P < 0.005$). Stars indicate no significant difference.

OTUs, in turn, is due to both the greater number of sequences allowed into the training set by the original pipeline and the presence of the hypervariable regions, which often makes the average pairwise sequence distances greater and thus leads to more and smaller OTUs. The genera-level HFRs were very similar to each other, however, with a slightly better score seen in the original pipeline with probe sets of size 30 and 40 (Figure 2.2B). The high overall similarity of HFR scores at the genera level is reflective of the fact that the number of genera represented in the data from both pipelines is the same; genera designations are made by the RDP2 database, unlike OTU designations that are made by the processing pipelines. The slightly better genera-level HFR in the original pipeline is thus either due to the presence of hypervariable regions or the increased numbers of training sequences per genera.

Regarding the hypervariable regions, the rationale for removing them in the new pipeline is that candidate probes arising from these areas may target only a narrow range of taxa and may thus be less informative than more conserved probes – yet they may be common enough in the training data (where some taxa may be overrepresented) to be chosen for a final probe set.

By removing the hypervariable regions the average pairwise sequence similarities will tend to increase – a situation that leads to larger and fewer OTUs that can potentially contain sequences exceeding the maximum pairwise identity of 97% that traditionally defines a species (Stackebrandt and Goebel 1994). To

compensate for this, we increased the stringency for inclusion into an OTU to 99% sequence similarity.

The new pipeline's contribution to better probe sets is supportive and indirect. It enriches the pool of more informative candidate probes and attaches the taxonomic information of the sequences for the MFPS cost function to operate on. In addition, the new pipeline facilitates updating an OFRG probe set with the latest sequence information. With relatively minor modifications the pipeline can be adapted for use on ribosomal (or other) genes of different microorganisms.

Optimizing Penalty Levels of the MFPS Cost Function. The new cost function (MFPS) has three penalty settings that correspond to the three levels of taxonomic information supplied in the training data (OTU, genus and phylum) and a series of optimizations of these settings was performed before comparing it to the MDPS.

Figure 2.3 shows how the HFR metric is affected as the genus penalty increases relative to the OTU penalty. In each panel (A and B) two results are shown. The dashed lines show the average HFR scores of 100, 5x cross-validation probe sets per point, and the solid lines show the average scores of 100 probe sets per point but using 100% of the data for training and validation.

Notice in Figure 2.3A that there is a slight increase in the OTU HFR before beginning a downward trend. This effect is seen in both 100% and 20% cross-validation probe sets, with the 20% cross-validation reaching a maximum at a

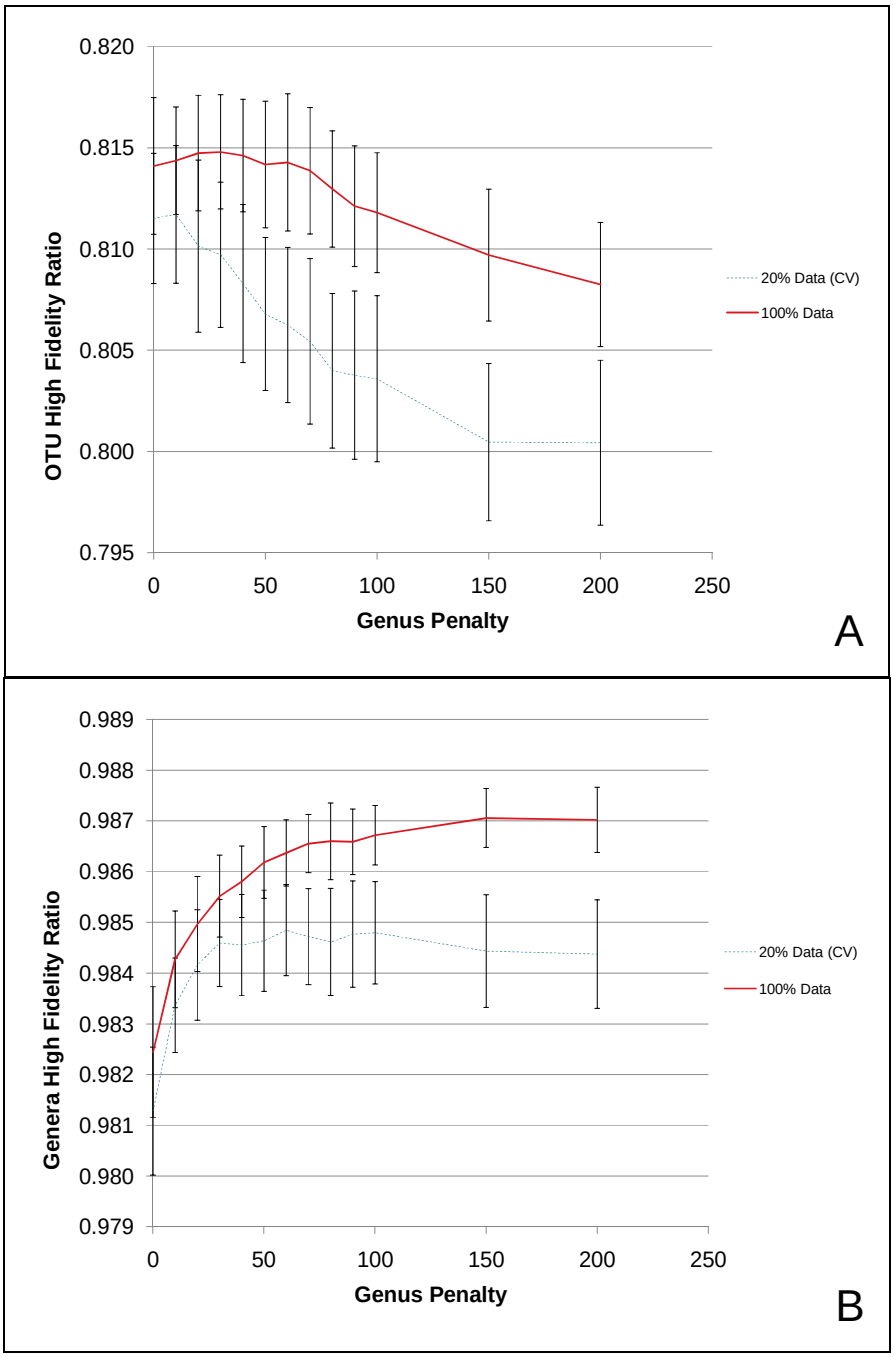

Figure 2.3. Optimizing the genus penalty for the MFPS cost function. The genus penalty was varied from 0 to 200 while holding the OTU penalty at 1. (**A**) OTU fidelity rises slightly as the genus penalty increases from zero then declines. (**B**) Genera fidelity rises sharply then plateaus. A penalty of 10 achieves the highest OTU fidelity in the 20% cross-validation data (CV) and OTU fidelity is highest in the 100% data set at a penalty of 30.

genus penalty of 10 and the 100% sets reaching a maximum at a genus penalty of 30. Figure 2.3B shows how genera-level HFR is affected as the genus penalty increases. This number rises and eventually plateaus, with more variation and a lower plateau seen in the 20% cross-validation data.

An OTU penalty of 1 and a genus penalty of 30 for the MFPS were chosen as optimal for a comparison to the MDPS. Our rationale for choosing a genus penalty of 30 was as follows. The initial rise in OTU fidelity makes intuitive sense because the increasing genus penalty improves the chances a distinct fingerprint will occur in only one genus – but if more distinct fingerprints are occurring in only one genus it becomes more likely some will also occur in only one OTU within that genus. However, as the genus penalty increases further and the total penalty score for a candidate probe set becomes dominated by any mistakes in genera classification, the MFPS begins to sacrifice OTU fidelity for better genera fidelity. Finally, the peak OTU fidelity occurs at a lower genus penalty level in the smaller 20% cross-validation data than in the 100% data set (10 and 30, respectively), suggesting that the size and/or makeup of the training data influences the optimal genera penalty level.

This led us to conclude that the larger the data set the farther to the right the OTU maximum might appear. And, since we planned to order a set of probes for laboratory use on environmental samples we should design them with a large data set in mind. Nevertheless, choosing a genus penalty above 30 would be an extrapolation.

The risk of overfitting may be higher when using the full data set, but since it is impossible to predict what bacteria a sample will contain, it is not clear how we can know we have or have not over-fit the data. Also, based on the severe tests of removing whole phyla (see Figure 2.6 and discussion) and using only 20% cross-validation data evaluated on 100%, the solution-space appears to be broad, and good solutions abundant, even if an optimal one is elusive.

Comparison of MFPS and MDPS Cost Functions. For this comparison we prepared training sequences using the new pipeline only. Figure 2.4 shows the performance of the MFPS and MDPS cost functions, using the HFR metric, with probe sets containing between 20 and 80 probes. In both the OTU and genera HFRs the MFPS scores higher than the MDPS in all probe set sizes examined. The difference is most pronounced in probe sets of size 20 and gradually narrows up to probe sets of size of 80. For OTU HFRs, the scores at n=80 are nearly identical, but for genera HFR the MFPS still shows a slightly improved performance over the MDPS.

As a control, probe sets were created randomly from one of two differently-sized probe matrices – either 1000 probes (the same one used to compare the cost functions) or 4000 probes, and are also included in Figure 2.4. The HFRs of the MFPS and MDPS cost functions are indeed higher than the random probe sets from the 1000 probe matrix. Interestingly, the HFRs of random probe sets from the 4000 probe matrix were much lower than the probe sets made from the 1000 probe matrix.

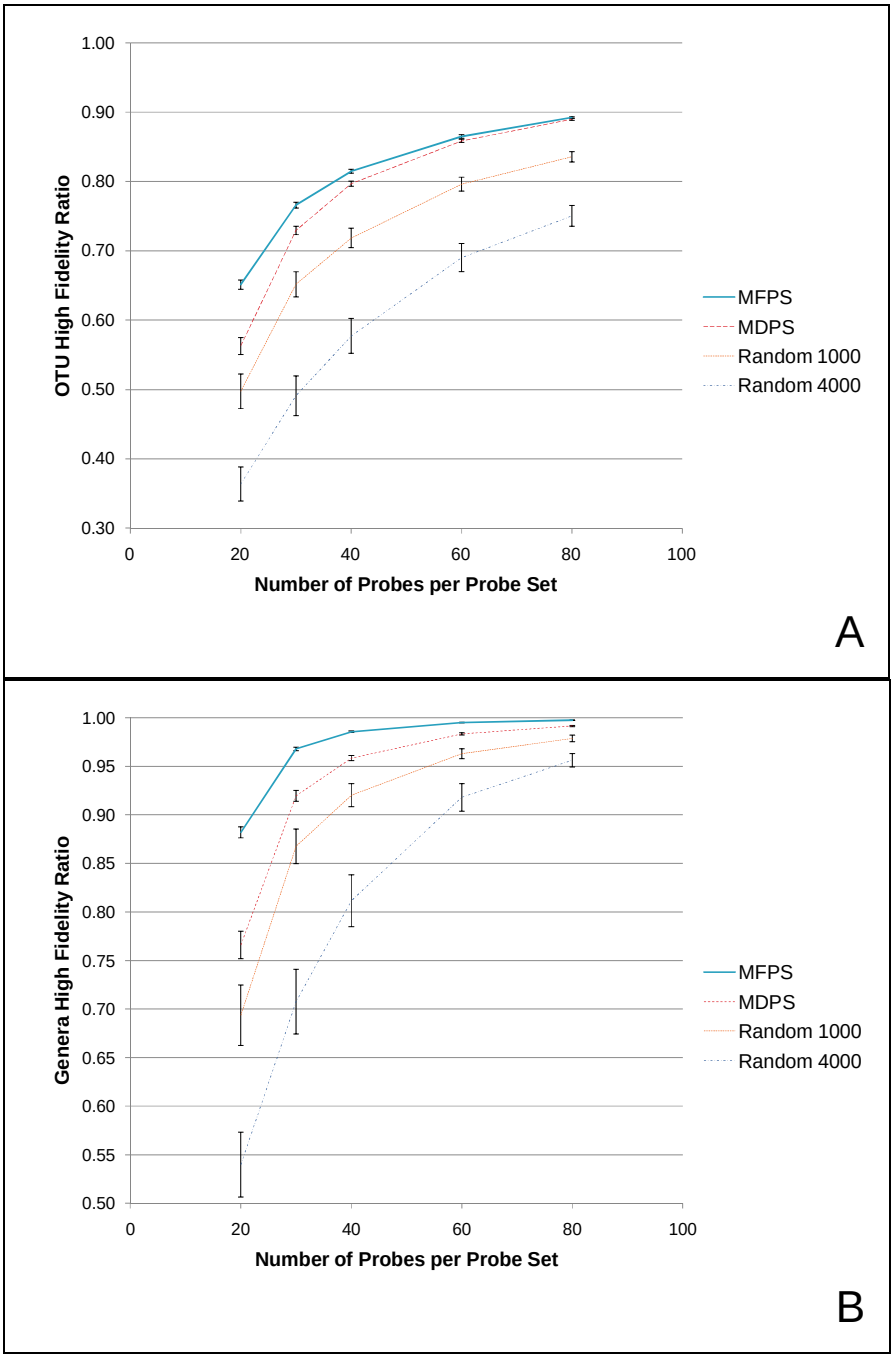

Figure 2.4. Comparison of the MFPS and MDPS cost functions using the High Fidelity Ratio (HFR). In both (**A**) OTU and (**B**) genera HFRs the MFPS scores higher than the MDPS in all probe set sizes examined but the difference narrows from n=20 to n=80 probes. Randomly chosen probe sets perform less well. The Random 1000 probe sets selected probes from the top 1000 most conserved probes and the Random 4000 from the top 4000. The MFPS and MDPS also selected probes from the 1000 most conserved probes. All four values within each probe set size are significantly different from each other using the two-tailed Student's t-test assuming unequal variance ($P << 0.01$) with 100 probe sets per data point.

To explain this difference, recall that the random 1000 probe sets contain probes from the top 1000 most conserved probes and the random 4000 from the top 4000. The higher HFR scores observed from the smaller probe matrix therefore suggests these are somehow more informative taxonomically. More work could be done to see how far this could be taken, and at what point some type of bias could be introduced, if any.

Our laboratory experiments will be done with a set of 40 probes as this is a practical maximum and provides very high (theoretical) fidelity. Genera-level HFR is over 98% and OTU-level HFR is over 81%. It is also worth noting that with 40 probes the majority of low-fidelity distinct fingerprints (~55%) occur in only two OTUs, but within the same genus.

Average Pairwise Sequence Distances. The average pairwise sequence distances results are shown in Figure 2.5. Unlike the High Fidelity Ratio, which is a measure of the taxonomic accuracy of a probe set, this metric focuses on the inaccuracy of a probe set's low-fidelity fingerprints, measuring the dissimilarity of the underlying DNA sequences from which they arose. Figure 2.5 reveals a considerable overall improvement of the MFPS over the MDPS cost function, as well as the effects different penalty settings have in the MFPS. To evaluate the two cost functions for this metric, we compared their results using two different penalty schemes.

Compared to the MDPS line, MFPS A (OTU and genus penalties set to 1 and 0, respectively) is superior except for having a few more sequences from 0% to 1%. The improved scores beyond 1% difference reflect the tendency of all distinct fingerprints (high and low fidelity) to more closely pattern real taxonomic groups; even if they do occur in more than one OTU they tend to occur in more

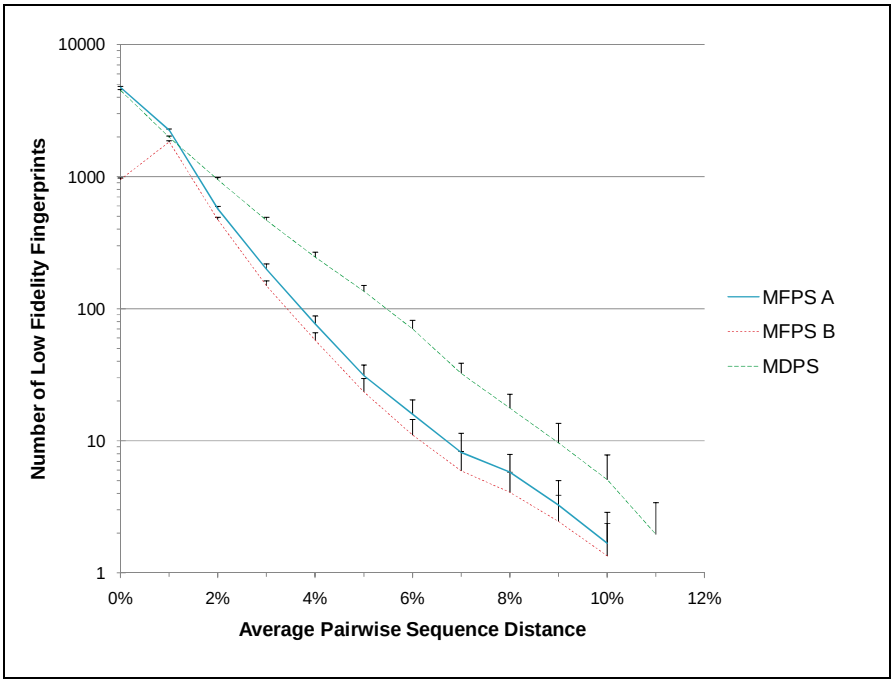

Figure 2.5. Average pairwise sequence distance metric. This metric focuses on the inaccuracy of a probe set's low-fidelity fingerprints, measuring the dissimilarity of the underlying DNA sequences from which they arose. MFPS A (OTU and genus penalties set to 1 and 0, respectively) is superior to MDPS except for having a few more sequences from 0% to 1%; scores in this range are from highly similar sequences but from OTUs in different genera. MFPS B (OTU and genus penalty levels set to 1 and 30, respectively) shows further improvement in distances greater than 1%, but unlike MFPS A or MDPS, has markedly fewer low-fidelity distinct fingerprints with distances less than 1%. The improvement in distances greater than 1% is the same windfall seen in HFR scores when the genus-level penalty was set to 30 (see figure 2.3). Error bars are standard deviations of 100 probe sets per data point.

similar sequences. Likely for the same reason, the MFPS A performs more poorly from 0% to 1%. These scores are from highly similar sequences in different OTUs but presumably from different genera (otherwise they would have

been grouped into the same OTU). This phenomenon is consistent with the fact that there was no genus-level penalty imposed in MFPS A.

MFPS B (OTU and genus penalty levels set to 1 and 30, respectively) shows further improvement in distances greater than 1%, but unlike MFPS A or MDPS, has markedly fewer low-fidelity distinct fingerprints with distances less than 1%. The latter is clearly an effect stemming from the genus-level penalty imposed during probe set creation; now, probe sets are shepherded away from these "near-misses." The improvement in distances greater than 1% is the same windfall seen in HFR scores when the genus-level penalty was set to 30 (see Figure 2.3).

Effect of Removing Whole Phyla. To examine how the fidelity of probe sets might behave if sequences from unknown phyla are encountered, MFPS and MDPS probe sets were made after sequentially removing several of the largest phyla, each ranging in size from approximately 10% to 33% of all training sequences.

Evaluations of the probe sets were performed with all phyla included. The results shown in Figure 2.6 indicate that although both MFPS and MDPS are negatively affected generally, the effect is relatively minor, and the MFPS outperforms the MDPS.

Interestingly, OTU HFR went up in the MFPS and MDPS when the phyla Proteobacteria and Actinobacteria were removed, respectively. When looking at the genera-level HFRs for these phyla, removing Proteobacteria does not

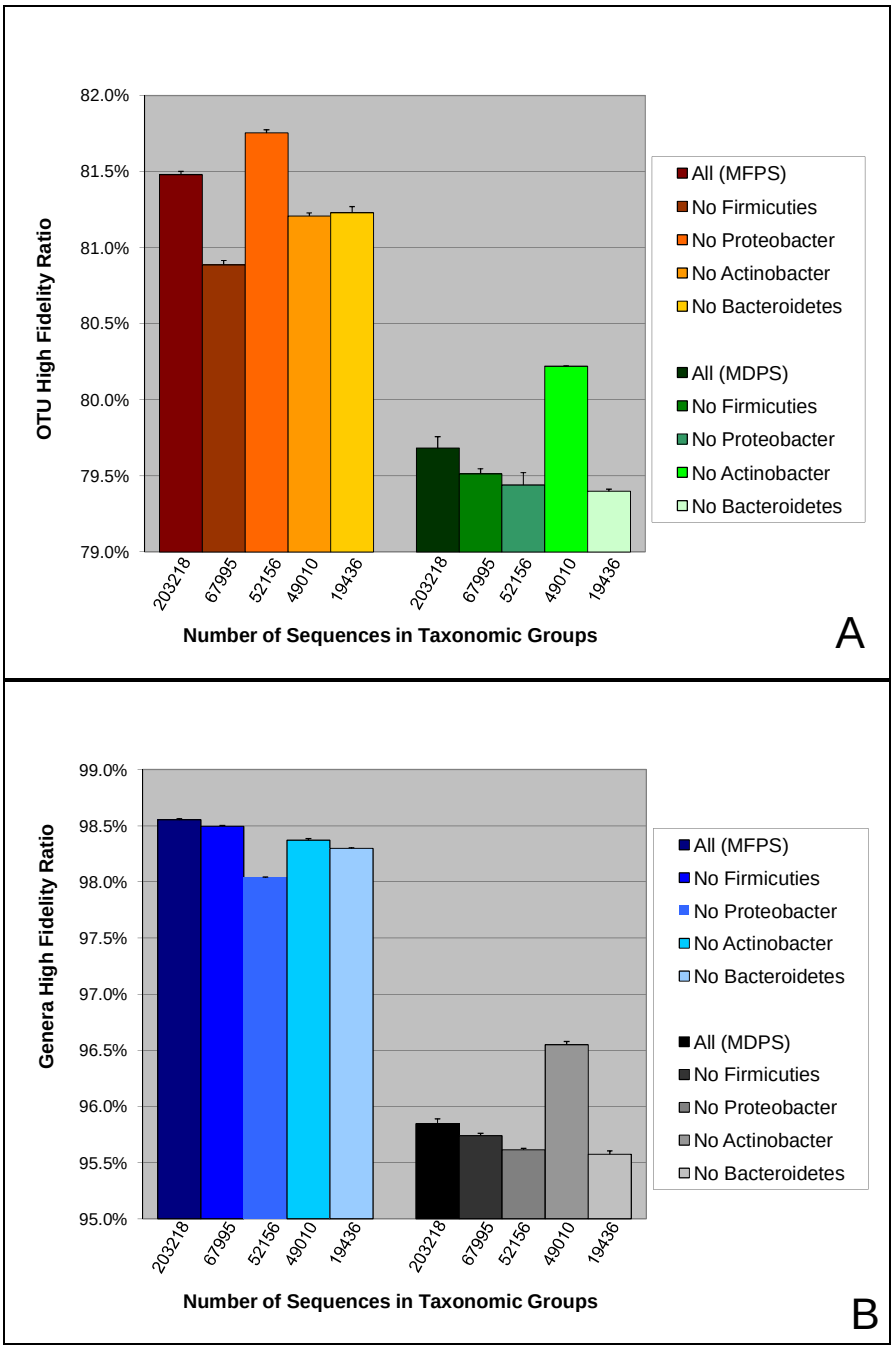

Figure 2.6. Fidelity Effect When Removing Whole Phyla From Training Data. (**A**) OTU and (**B**) genera fidelity. Although both MFPS and MDPS are negatively affected generally, the effect is relatively minor, and the MFPS outperforms the MDPS. OTU HFR goes up in the MFPS and MDPS when the phyla Proteobacteria and Actinobacteria are removed, respectively. At the genera-level HFRs for these phyla, removing Proteobacteria does not improve in MFPS, yet removal of Actinobacteria improves HFR in the MDPS.

improve in MFPS, yet HFR still improves in the MDPS when removing Actinobacteria. It is not clear why an increase of HFR scores would occur when removing a phylum before making probe sets, other than that something in these phyla are causing the cost functions to become confused, perhaps trapping them in a local minimum.

Positional Bias of Probes in MFPS and MDPS. We were curious if the

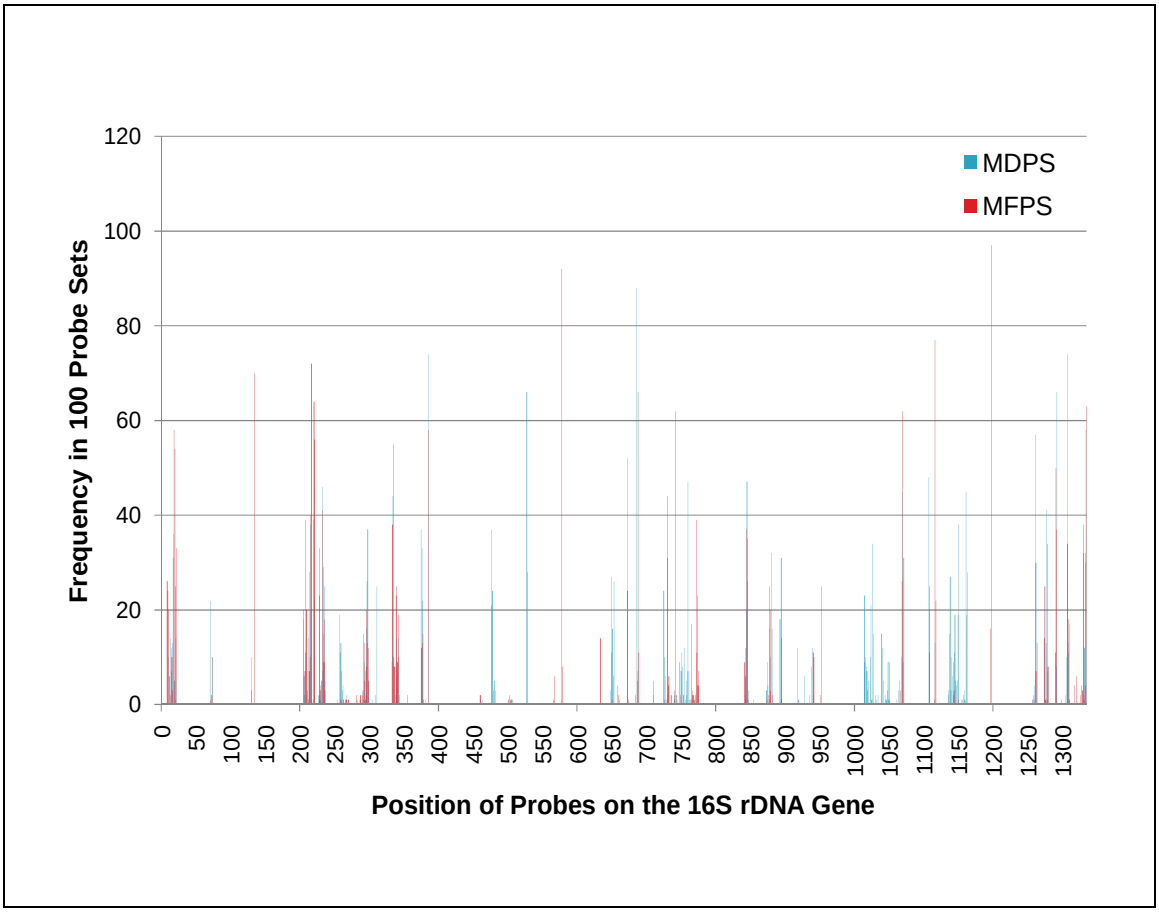

Figure 2.7. Positional Bias of Probes from MFPS and MDPS. This graph was constructed by finding the starting positions of all probes in 100 probe sets of size n=40 probes and plotting the frequency where they occurred at each position for both cost functions. Although probes arising from some positions appear to be chosen by both cost functions there are several positions that appear to be favored by the MFPS or MDPS, sometimes exclusively.

probes chosen by the two cost functions would show any positional bias on the 16S rDNA molecule. Figure 2.7 was constructed by finding the starting positions of all probes in 100 probe sets of size 40 and plotting the frequency they occurred at each position for both cost functions. Although probes arising from some positions appear to be chosen by both cost functions there are several positions that appear to be favored by the MFPS or MDPS, sometimes exclusively.

The regions favored by the MFPS suggest these may tend to be more conserved within taxonomic groups, whereas the regions favored by the MDPS may tend to be less conserved within the same groups. Alternatively, because probes in a probe sets are chosen to work together to provide information about the sequences there may be some kind of complex within-group conservation between the regions being favored. More investigation would need to be performed to determine if there was some underlying biological significance to these patterns.

2.4 Conclusion

With its multi-level penalty scheme the MFPS improves the quality of probe sets as measured by two biologically relevant metrics: fidelity and sequence distances. By pre-clustering training sequences into biologically meaningful groups and then choosing probe sets based on how closely their resultant fingerprints represent those groupings we increase the chances that the

underlying sequences of those fingerprints are more similar to each other than they would be in the original MDPS.

Attempts have been made to assign taxonomy using the relatively small sequencing reads produced by high-throughput sequencing technology such as 454 and Illumina (Wu et al. 2010)(Caporaso, Kuczynski, et al. 2010). With the Illumina platform in Caporaso et al, 16S rRNA gene reads were placed on a pre-built guide tree and the taxonomy reported was the highest level that can confidently be predicted; these sequences can be from any portion of the rRNA gene. Using the 454 platform and reads of ~450 nucleotides, Wu et al. compared taxonomic predictions from the sequencing results of different hypervariable regions. Both methods were able to achieve genus level taxonomic predictions in most cases, though not always. With 10-mer probes and 40 probes per probe set, the MFPS is essentially interrogating 400 nucleotides but is not restricted to a contiguous portion of the molecule, as is the case in 454. And the probes act together to produce a fingerprint for each clone, unlike Illumina reads which are disconnected from each other and so can be utilized only in isolation.

One future improvement in the MFPS would be to take into account more complex interactions between the probe and DNA strand. It is known, for instance, that in real hybridization experiments a probe can hybridize with varying degrees of affinity depending on several factors. These factors include being able to hybridize at a detectable level even when there is a single nucleotide

mismatch between the probe and DNA, or that a probe may bind to more than one location on a DNA strand simultaneously. In a small bioinformatics trial (data not shown), allowing an A:G mismatch at the 10^{th} position of the probe to count as a successful hybridization event actually enhanced the HFR scores of probe sets. In a post-hoc analysis these events could be detected and characterized, and this information incorporated into the cost function itself, potentially leading to a much higher fidelity probe set

BIBLIOGRAPHY

Bent, E, B Yin, A Figueroa, J Ye, Q Fu, Z Liu, V Mcdonald, D Jeske, T Jiang, and J Borneman. 2006. Development of a 9600-clone procedure for oligonucleotide fingerprinting of rRNA genes: Utilization to identify soil bacterial rRNA genes that correlate in abundance with the development of avocado root rot. *Journal of Microbiological Methods* 67, no. 1 (10): 171-180. doi:10.1016/j.mimet.2006.03.023.

Borneman, J, M Chrobak, G Della Vedova, A Figueroa, and T Jiang. 2001. Probe selection algorithms with applications in the analysis of microbial communities. *Bioinformatics (Oxford, England)* 17 Suppl 1: S39-48.

Caporaso, J Gregory, Justin Kuczynski, et al. 2010. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* 7, no. 5 (4): 335-336. doi:10.1038/nmeth.f.303.

Caporaso, J Gregory, Christian L Lauber, William A Walters, Donna Berg-Lyons, Catherine A Lozupone, Peter J Turnbaugh, Noah Fierer, and Rob Knight. 2010. Microbes and Health Sackler Colloquium: 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* (June 3). doi:10.1073/pnas.1000080107. http://www.ncbi.nlm.nih.gov/pubmed/20534432.

Drmanac, R, and S Drmanac. 1999. cDNA screening by array hybridization. *Methods in Enzymology* 303: 165-178.

Drmanac, S, and R Drmanac. 1994. Processing of cDNA and genomic kilobase-size clones for massive screening, mapping and sequencing by hybridization. *BioTechniques* 17, no. 2 (August): 328-329, 332-336.

Frank, Daniel N, Allison L St Amand, Robert A Feldman, Edgar C Boedeker, Noam Harpaz, and Norman R Pace. 2007. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proceedings of the National Academy of Sciences of the United States of America* 104, no. 34 (August 21): 13780-13785. doi:10.1073/pnas.0706625104.

Maidak, B L, J R Cole, C T Parker, G M Garrity, N Larsen, B Li, T G Lilburn, et al. 1999. A new version of the RDP (Ribosomal Database Project). *Nucleic Acids Research* 27, no. 1 (January 1): 171-173.

Meier-Ewert, S, J Lange, H Gerst, R Herwig, A Schmitt, J Freund, T Elge, R Mott, B Herrmann, and H Lehrach. 1998. Comparative gene expression profiling by oligonucleotide fingerprinting. *Nucleic Acids Research* 26, no. 9 (May 1): 2216-2223.

Muyzer, G. 1999. DGGE/TGGE a method for identifying genes from natural ecosystems. *Current Opinion in Microbiology* 2, no. 3 (6): 317-322. doi:10.1016/S1369-5274(99)80055-1.

Schloss, P. D., S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, et al. 2009. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Applied and Environmental Microbiology* 75, no. 23 (10): 7537-7541. doi:10.1128/AEM.01541-09.

Schütte, Ursel M E, Zaid Abdo, Stephen J Bent, Conrad Shyu, Christopher J Williams, Jacob D Pierson, and Larry J Forney. 2008. Advances in the use of terminal restriction fragment length polymorphism (T-RFLP) analysis of 16S rRNA genes to characterize microbial communities. *Applied Microbiology and Biotechnology* 80, no. 3 (September): 365-380. doi:10.1007/s00253-008-1565-4.

Scupham, A. J., L. L. Presley, B. Wei, E. Bent, N. Griffith, M. McPherson, F. Zhu, et al. 2006. Abundant and diverse fungal microbiota in the murine intestine. *Appl Environ Microbiol* 72, no. 1: 793-801.

Stackebrandt, E., and B. M. Goebel. 1994. Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. *International Journal of Systematic Bacteriology* 44, no. 4 (10): 846-849. doi:10.1099/00207713-44-4-846.

Valinsky, L., G. Della Vedova, T. Jiang, and J. Borneman. 2002. Oligonucleotide fingerprinting of rRNA genes for analysis of fungal community composition. *Appl Environ Microbiol* 68, no. 12: 5999-6004.

Valinsky, L., G. Della Vedova, A. J. Scupham, et al. 2002. Analysis of bacterial community composition by oligonucleotide fingerprinting of rRNA genes. *Appl Environ Microbiol* 68, no. 7: 3243-50.

Wu, Gary D, James D Lewis, Christian Hoffmann, Ying-Yu Chen, Rob Knight, Kyle Bittinger, Jennifer Hwang, et al. 2010. Sampling and pyrosequencing methods for characterizing bacterial communities in the human gut using 16S sequence tags. *BMC Microbiology* 10, no. 1: 206. doi:10.1186/1471-2180-10-206.

Ye, Jingxiao, Jimmy W. Lee, Laura L. Presley, Elizabeth Bent, Bo Wei, Jonathan Braun, Neal L. Schiller, Daniel S. Straus, and James Borneman. 2008. Bacteria and bacterial rRNA genes associated with the development of colitis in IL-10 Mice. *Inflammatory Bowel Diseases* 14, no. 8 (8): 1041-1050. doi:10.1002/ibd.20442.

Yin, B., L. Valinsky, X. Gao, J. O. Becker, and J. Borneman. 2003. Bacterial rRNA genes associated with soil suppressiveness against the plant-parasitic nematode Heterodera schachtii. *Appl Environ Microbiol* 69, no. 3: 1573-80.

Chapter 3: Polony Image Analysis

3.1 Introduction

Polony technology has been employed in several applications, such as DNA sequencing, SNP detection, gene expression studies, genotyping, haplotyping and alternative pre-mRNA splicing (Shendure et al. 2005)(Butz et al. 2004)(Rieger et al. 2007)(Robi D Mitra et al. 2003)(Zhang et al. 2006)(Jun Zhu et al. 2003), and has now been adapted to oligonucleotide fingerprinting of ribosomal rRNA genes (OFRG)(Valinsky, G. Della Vedova, T. Jiang, et al. 2002). The raw data produced by polony technology are scanned images, analogous to a microarray but different in important aspects. The data from the polony array must therefore be collected with software tailored to its particular characteristics.

Unlike a traditional array, where unbound sample DNA is hybridized to thousands of anchored probes, the polony microarray is inverted: sample DNA is anchored and unbound probes are hybridized to them. Polony microarrays are created by adding sample DNA into a polyacrylamide gel mixture and casting the mixture in a thin layer on a microscope slide. After the gel has fully polymerized a chamber is placed over the gel and PCR reagents are added, the chamber is sealed and the slide is subjected to thermocycling. During thermocycling, the DNA strands are duplicated again and again, diffusing outward from each original strand. Because diffusion of the amplicons is inhibited somewhat by the gel the process results in randomly placed spots of DNA molecules – the polonies – one for each original DNA molecule. Once made, the polonies can be visualized with

Sybr or ethidium bromide staining or interrogated, either by single-base extension using fluorescently labeled nucleotides or by hybridization using fluorescently labeled DNA probes.

There are several advantages of polonies over traditional arrays that make them an attractive tool for certain applications. Polonies are inexpensive and easy to make; more importantly, they can be used to accurately quantify and characterize a sample of DNA, such as in haplotyping, since each polony originates from a single, isolated DNA molecule (Zhang et al. 2006). The ability to amplify spatially isolated DNA molecules is an important property that makes polony technology a useful tool for analyzing bacterial communities via 16S rRNA gene analysis. Other molecular approaches for bacterial community analysis necessarily involve a "mixed-template" PCR step that can bias the original ratios of the species or create chimeric amplicons of two or more species (Suzuki and Giovannoni 1996)(Lahr and Katz 2009a).

We have adapted polony technology to OFRG, a hybridization-based method for grouping similar rRNA genes. With this method, 42 different DNA probes are sequentially hybridized to a polony microarray. Then, each probe's affinity for the DNA of each polony is measured and together become that polony's hybridization "fingerprint." This fingerprint is used to cluster together polonies with similar fingerprints and the clusters tentatively represent real taxonomic groups.

There are also characteristics that make the use of polonies difficult, such as their random placement, the problem of overlapping polonies, and diameters that can vary substantially. These characteristics make detection and quantification of polonies a challenge, especially at higher densities. Most existing software for microarray analysis is designed for organized grids and fails when applied to the randomly ordered polony array. Also, existing software packages are not designed to deal with overlapping spots on arrays, such as by flagging them or by selecting a suitable non-overlapping portion from which to measure.

The Illumina high-throughput sequencing platform is most similar to polonies because template DNA is also randomly placed before being grown into "clusters" of DNA via bridge PCR amplification. To locate clusters, both Illumina's "Firecrest" and an open source program called, "Swift" (Whiteford et al. 2009), employ thresholding approaches. Although these packages may be adaptable to polony images they are tailored to the specific needs and characteristics of the sequencing platform.

This paper presents a straightforward algorithm for finding polonies using a simple approach based on finding local maximum pixels and harnessing the additive information contained within the multiple images of OFRG hybridizations of the same polony microarray. Because of the nature of the hybridization patterns across the 42 scanned images – overlapped polonies can occasionally "appear" isolated when a probe binds to only one polony of an overlapping group

of polonies – an opportunity to precisely locate these polonies presents itself. By taking advantage of this phenomenon we can ameliorate the problem of overlap in the polony microarray, allowing for higher polony densities and more accurate quantification.

We compare the ability of our approach with a new approach developed recently that combines the Expectation-Maximization (EM) algorithm with a model of polony intensities using multiple Gaussian distributions (Wei Li et al. 2010). Both algorithms were originally developed to find polonies using a single image but have been extended to use the multiple images produced in a polony OFRG microarray.

3.2 Methods

Polony microarrays are created from randomly-placed DNA molecules that are subsequently amplified via PCR, and these newly created strands of DNA are fixed to the acrylamide gel in which they were grown (see Figure 3.1). Hybridization/wash cycles of 42 different probes to the polony array are performed sequentially two probes at a time. The probes are designed to bind to different taxonomic subsets of bacterial DNA. Thus, each of the 42 images has a unique pattern of visible polonies. Figure 3.2 shows just six hybridization scans of a small section of a polony microarray. Polony arrays are imaged with a two-laser scanner to gather the intensities of each probe/polony pair. Techniques developed to accurately locate and measure polony intensities within the raw images are the subject of this study. All image processing with our algorithm is

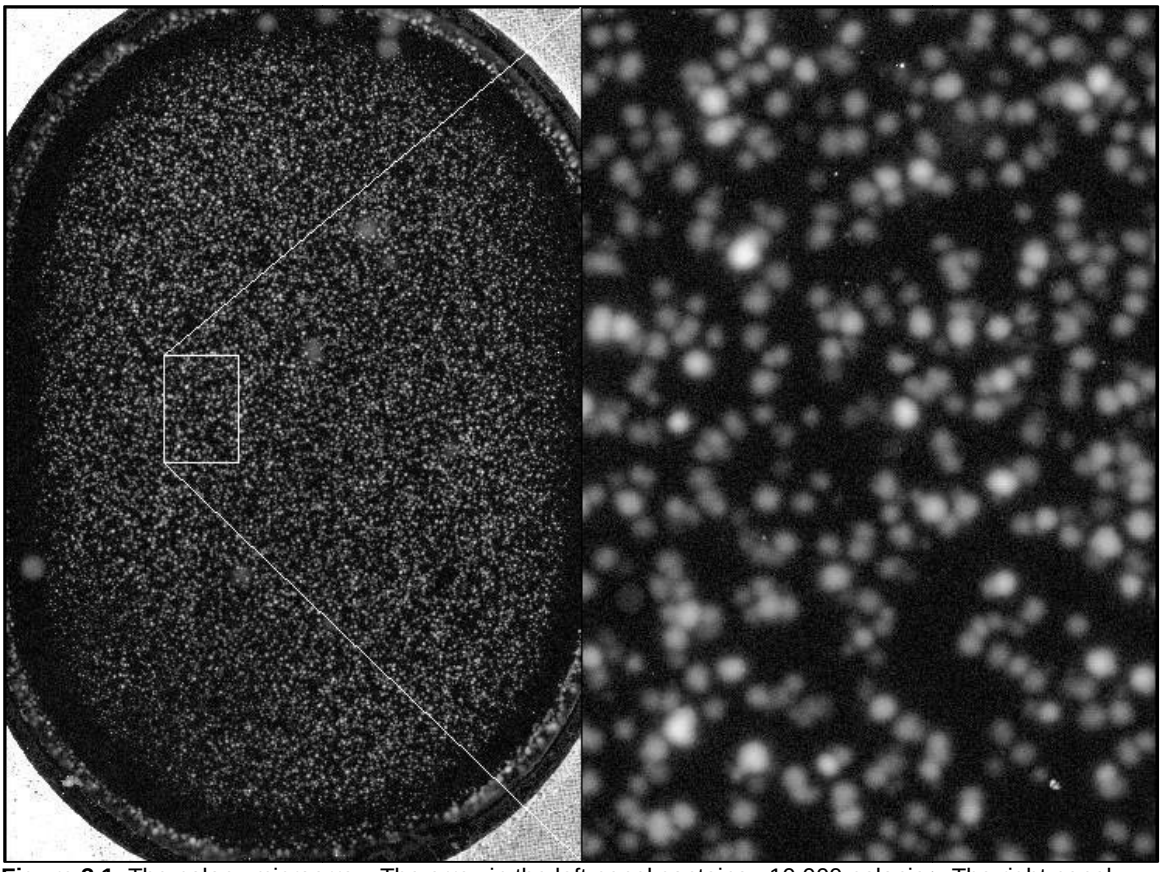

Figure 3.1. The polony microarray. The array in the left panel contains ~10,000 polonies. The right panel shows a close up of a small area. Maximum polony width is ~22 pixels (110um). The contrast of both images has been increased for display purposes.

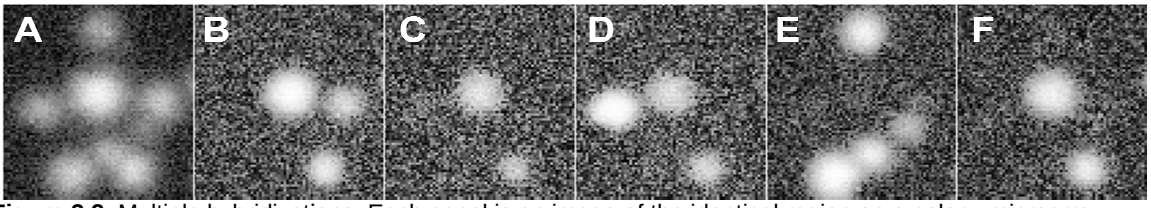

Figure 3.2. Multiple hybridizations. Each panel is an image of the identical region on a polony microarray, hybridized and scanned with one of six fluorescently labeled probes (**A**: GYACACACCGCCCG, **B**: ATACCGCATA, **C**: GCCTAACACA, **D**: GCTAGTTGGT, **E**: CAATGGGCGA, **F**: GACTGAGACA). The leftmost panel (**A**) is the reference probe, showing all eight polonies. Though physically still present, some polonies do not appear in subsequent images because the indicated probe did not hybridize to their DNA. The contrast in all images has been increased for display purposes.

performed using a plugin written for the National Center for Biotechnology Information's ImageJ program (Rasband 1997). Post data-acquisition filtering is performed in R.

Image analysis. Polony gels are affixed to a microscope slide and a 2700 x 4000 pixel image (16 bit grey-scale TIFF) is sufficient to encompass all polonies. The resolution of each pixel is currently 5um and the largest polonies are ~22 pixels in diameter, and images contain ~1,000-10,000 polonies. The images are processed in ImageJ in a single "stack" of images; each image in a stack is referred to as a "slice." Once all hybridization images have been acquired and loaded as a stack the processing can begin and involves several main steps: *i*) align the slices, *ii*) determine putative polony locations and background regions in each slice, *iii*) select the polonies most likely to be real, *iv*) determine areas of overlap, *v*) measure polony and background intensities. Table I shows the sizes and polony densities of images used to compare the EM and LM algorithms

Aligning images. A convenient way to measure polony intensities throughout all slices in an ImageJ stack is to define a region of interest (ROI) for each polony. An ROI's size, shape and position can be defined for each polony once they are known. To get accurate measurements, however, the images in the stack must be aligned first so that polonies remain in the same x-y position as their corresponding ROIs. Although the 42 images are of the same polony array, they may be out of alignment with each other due to positional variations that can

occur when placing the slides in the laser scanner, or due to the slight offset between its two lasers. The first image in a stack is of the reference probe and it is used as the standard to which all other images are aligned. By checking in a +/- 5 pixel offset range, alignment is performed by comparing the local maximum pixels in the reference image to local maximums of the other images. The x-y offsets where the most local maximum pixels align are chosen to shift an image into alignment. Local maximum pixels are defined as those whose intensity values are brighter than their eight neighbor pixels and are 200 or more above the average intensity value of the whole image but less than the saturation value of 65535.

Determining polony locations and measurement areas. In images, polonies are located by exploiting three simple and common physical characteristics: they are brightest in the center, their intensities taper off gradually when moving out from the center and they are circular. Polony centers are located first by finding local maximums across an image. A local maximum is defined as a pixel whose intensity is greater than its eight neighbor pixels. However, due to the high amount of variation in the pixel intensities of raw images, using this criterion alone would result in finding many false positives. Therefore, local maximums in a slice are determined only after a smoothing step. The search for local maximums is repeated for each slice in a stack and results in a list of putative polony-center locations.

Next, three successive perimeters surrounding each local maximum are found. The perimeters are used to define ROIs for intensity and size measurements. The innermost perimeter defines the pixels to be used for measuring a polony's intensity. They are found by searching outward from the local maximum in eight directions until encountering pixels whose intensity is close to but not less than 85% of the maximum intensity. The second set of points delineate where (the DNA boundary of) a polony ends. Each of these points are found when, starting from the center and moving outwards, any one of the following are true: *i*) a rise in intensity is detected, indicating another polony may be nearby, *ii*) little to no change in intensity is detected, indicating the DNA boundary of the polony has been reached, *iii*) a preset maximum distance from the center has been reached; this can occur when polonies closely overlap and a long, gentle slope of pixel intensities are encountered that will not trigger *i* and *ii*. The third and outermost set of points is the background-measurement area. These points are just beyond the polony DNA boundary and are found by simply extending outward from it by five pixels. Each local maximum represents, together with their three corresponding perimeters, the putative locations and measurement areas of real polonies. Several validation checks must then be performed before a final ROI list is made.

Selecting the highest quality polonies. There are three main types of polonies in the initial list of putative polonies: real, pseudo and duplicates. Within the group of real polonies are two subtypes: isolated and overlapping.

Detecting each main type involves assigning a confidence level that is based on the observation that real polonies almost always have a circular form. After pseudo and duplicate polonies are removed, overlapping cases can be determined.

Duplicate polonies arise from any real polony that is bright enough to have its local maximum detected in more than one slice in a stack. Recall that local maximum pixels represent putative polony centers and a search for them is performed in every slice of a stack. We say that a local maximum pixel in one slice having the same x-y coordinates (or within a distance of 3 pixels to allow for alignment variations) as local maximum pixels in other slices represent the same physical polony. Therefore only one of these duplicates need be kept as the ROI to measure it. To choose the best duplicate we select the one we are most confident represents the size and shape of the physical polony at that position on the gel. Typically this will be the largest and most circular of the duplicates. The circularity of each duplicate is gauged by finding the standard deviation of the distances from its center to its eight center-measurement boundary points. The most circular is kept and the rest are discarded.

A pseudo polony is defined as a local maximum that does not represent the true center of any physical polony. Pseudo polonies arise as a result of overlapping polonies that fluoresce simultaneously or from debris. Overlapping polonies often appear in an image as a single, misshapen polony with no clear

boundary between them. A local maximum in this case will often be found at a point somewhere in between the two (or more) real polony centers.

Detecting overlapping pseudo polonies. To detect pseudo polonies caused by overlapping we use the circularity confidence criteria. A pseudo polony will usually (but not always) have a non-circular perimeter. Polonies whose circularity falls below a user-defined threshold are removed from the list of polonies. Any pseudo polonies that are not removed at this step can usually be detected in measurement data using the standard deviations of their center-measurement area's pixel values; if a pseudo polony is centered between two or more physical polonies, measurements will be of their downward sloping sides and they will have considerable intensity changes across the measurement area. In contrast, real polony ROIs are correctly centered and the variation in measured pixel intensities are much smaller.

Detecting debris pseudo polonies. Debris on a gel is similar enough to real polonies to be detected as local maximums and the algorithm will find three perimeters for them. Detection of debris is done in several ways. The initial filter for debris uses the circularity criterion; most debris ROIs will not be circular and so will be removed. However, for debris that is circular, it is convenient to detect and remove them using an R script (R Development Core Team 2010) after measurements are made. To do so, three detection criteria are used. When debris is large it will often be very bright and the image will be saturated. These are filtered out by removing polonies whose average intensity is above a

threshold. When debris is small it will usually have large standard deviations in pixel intensities and can be filtered out accordingly. The last method to detect debris is by examining the ratio of its center-measurement area to its DNA area. Real polonies have a small center to DNA area ratio. Debris will have a ratio close to one, however, because they have an almost instantaneous drop from bright to background; this causes their DNA areas to be only slightly larger than their center-measurement areas.

Determining overlaps and optimal measurement areas. After duplicate and pseudo polonies are filtered out it is necessary to detect which real polonies are overlapping each other. Measurements are made in two of the three areas defined for each polony – its center area and the background area surrounding the polony. (Two of the three areas defined for each polony are where intensity measurements are made – the polony's center pixels and the background area surrounding the polony). Overlaps in these areas are detected and handled by creating a "mask" image (Figure 3.3). The mask is created by superimposing the three areas of all polonies found. Each polony's three areas are plotted in the mask using white and three shades of grey. White, medium grey and dark grey denote center-measurement, DNA and background-measurement areas, respectively. Light grey pixels denote the overlap of one polony into another polony's center-measurement area. Overlap areas are reported for that polony during the measurement step. This information can be used to eliminate potentially bad data if desired. Overlaps in background-measurement areas are

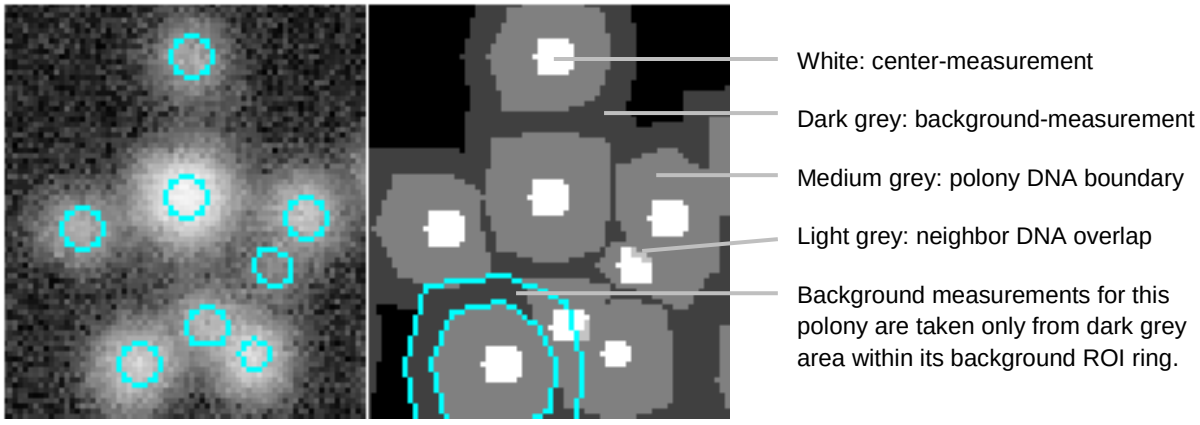

Figure 3.3. Measurement areas and mask. The left panel shows polonies with their center-measurement Region Of Interest (ROI) areas displayed. The right panel is the measurement areas mask of the same section. White pixels in the mask indicate center-measurement areas. Dark grey pixels indicate allowed background-measurement areas. Medium grey pixels designate polony DNA boundary areas (not measured). Light grey pixels denote the amount of neighbor DNA areas that overlap with a center-measurement area. The two large rings in the lower left denote one polony's total background-measurement area. When measurements are taken, however, only values corresponding to the dark grey pixels within this boundary are used. The contrast of the image in the left panel has been increased for display purposes.

handled in a different manner. Rather than reporting how many pixels from a neighbor polony are intruding into the background area, during the measurement step the program refers to the mask and only measures the portion of each polony's background area where no overlap occurs.

Measuring polony intensities. Once a final list of ROIs has been determined the intensities of each polony in each slice can be measured using a modified ImageJ ROI Manager. The data is saved and is ready for further processing with an R script that performs multiple quality control checks (such as detecting any remaining pseudo polonies), transforms the data and clusters the polonies according to the similarity of their transformed hybridization patterns.

3.3 Results

We compare the accuracy and relative speeds of the local maximum (LM) and Expectation Maximization (EM) algorithms in (Wei Li et al. 2010) in single and multiple slices. Concerning accuracy, Li et al. (2010) found that for a single slice the algorithms performed similarly when polony density (defined as the number of polonies divided by the total number of pixels in an image) was low (less than 0.6×10^{-3}) but the EM algorithm outperformed the others when polony density was high. For multiple slices, however, we find the LM begins to have an advantage in both speed and accuracy.

Accuracy comparison. We use a Precision-Recall curve (PR) and an Area Under the PR curve (AUPR) to compare the accuracy of each algorithm in identifying polony locations. The PR curves are constructed from their precision and recall performances on a polony image or series of images (slices). If N real polonies exist and an algorithm locates M polonies, K of which are correct, then *precision* = *K/M* and *recall* = *K/N*. The x-y position of the center pixel of a polony is defined as its location. These locations were manually annotated in real images and known precisely in simulated images. If one of an algorithm's x-y coordinates for a polony are less than 3 pixels in distance from a known polony location it is deemed to have found that polony.

Single real image accuracy. Figure 3.4 shows the precision/recall values for the LM and EM algorithms in a high density real polony image (see Table 3.1, Image 1). We can see that EM has correctly identified more polonies than LM

(recall of 0.56 versus 0.33) but has incorrectly identified some as well (precision of 0.86 versus 1.0 for LM). Using the AUPR to compare we find the EM has outperformed the LM by 0.72 to 0.67.

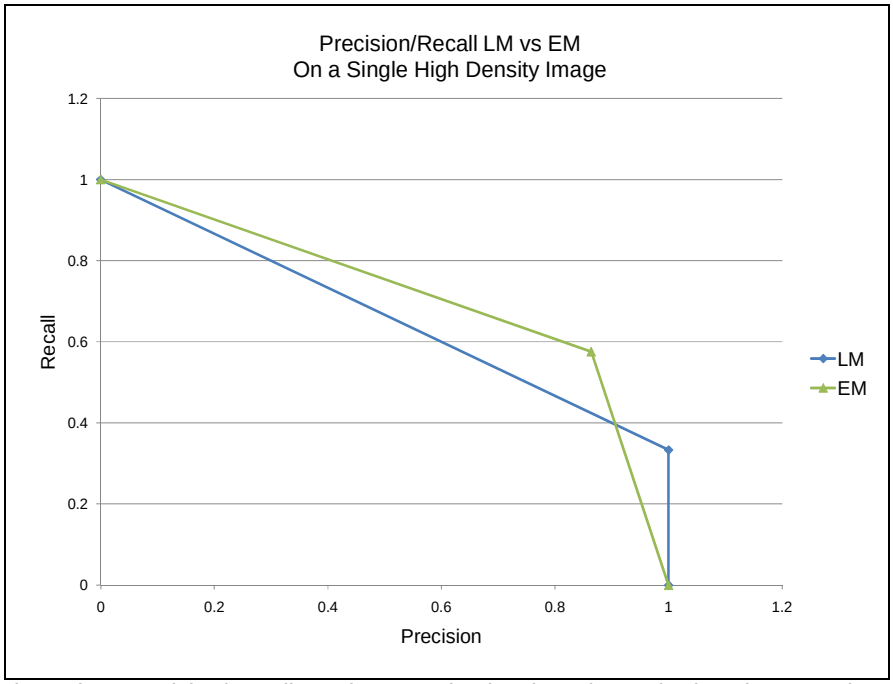

Figure 3.4. Shown is a Precision/Recall graph comparing local maximum (LM) and expectation-maximization (EM) algorithms. The image used is a small, high density section of a single image of a real polony microarray (see Table 1, Image 1).

Multiple real images accuracy. Figure 3.5 shows the AUPR for LM and EM algorithms in the same high density polony microarray (see Table 3.1, Image 1) but where multiple images (slices) are used. The EM shows an initial increase when a second image is used but its performance mostly decreases or levels out with additional images. However, the performance of LM continues to improve with additional images, leveling off after 20 images.

Multiple simulated images accuracy. Because no very high density real images were available we simulated a highly dense polony microarray and

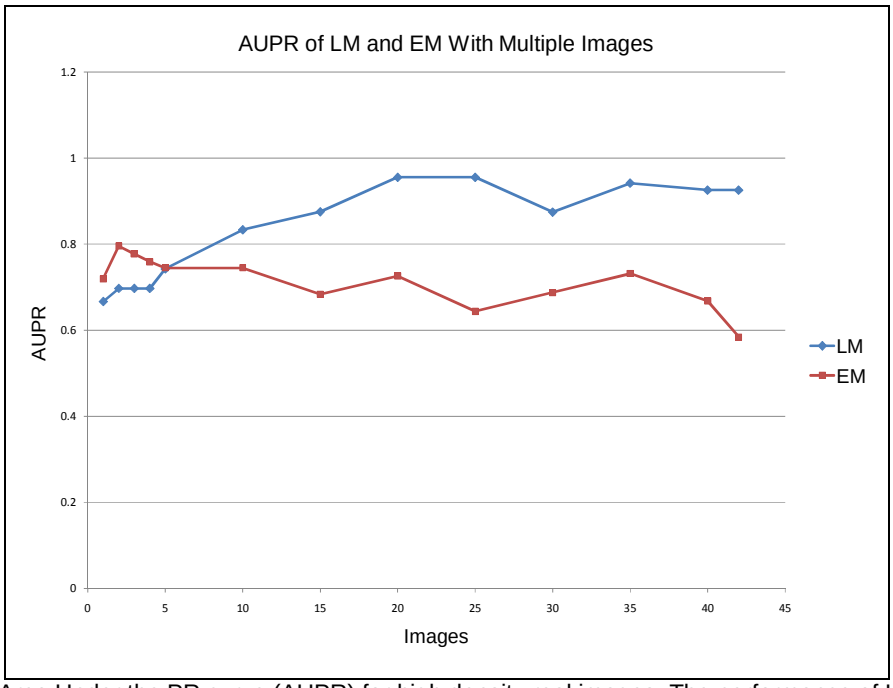

Figure 3.5. Area Under the PR curve (AUPR) for high density real images. The performance of LM is worse than EM when processing 1 to 5 images but its performance improves over EM after being supplied with 10 or more images. The image used is a small, high density section of a single image of a real polony microarray (see Table 1, Image 1).

Figure 3.6. Area Under the PR curve (AUPR) for very high density simulated images. In this simulated set of images the performance of LM is slightly better than EM when processing only one image. Its performance improves rapidly over EM after being supplied with more images. The image used is a very high density simulated polony microarray (see Table 1, Image 4).

compared the LM and EM algorithms. Figure 3.6 shows the AUPR of both algorithms in images with a polony density of 3.00×10^{-3} (see Table 3.1, Image 4). Surprisingly, in these simulated images the AUPR of the LM is higher than that of the EM in all cases, even when using only one image.

Relative Speeds. Runtimes of both algorithms were found on images using a full, 42-slice stack and reported on a time-per-slice basis. In Table 3.2, the results on Image 2 using the LM and EM algorithms shows the average time per slice of the LM algorithm is over 200 times less than the EM. The EM is not practical for large images but the time per slice of the LM on full-sized polony microarray images was about 4 seconds (Image 3).

Image ID	Image Source	Pixels	Polonies	Density
1	Real	150 x 150	33	1.47×10^{-3}
2	Real	500 x 500	247	9.88×10^{-4}
3	Real	2968 x 4400	8020*	1.03×10^{-3}
4	Simulated	500 x 500	750	3.00×10^{-3}

Table 3.1. Source, size and density information on images used. * Estimated.

Image ID	Algorithm	Pixels	Slices	Time per Slice	Density
2	LM	500 x 500	42	0.36s	9.88×10^{-4}
2	EM	500 x 500	42	77.95s	9.88×10^{-4}
3	LM	2968 x 4400	42	4.02s	1.03×10^{-3}

Table 3.2. Processing times of Local Maximum (LM) and Expectation Maximum (EM) algorithms. All images were of real polony microarrays.

3.4 Discussion

As with any microarray, accurate data collection is of primary importance. Polony microarrays offer the researcher a low cost platform with the unique ability to interrogate thousands of genomic DNA molecules simultaneously by spatial isolation of the molecules and subsequent PCR amplification. The random placement of the DNA molecules leads to an inherent difficulty associated with polonies – namely, the problem of overlap.

To address this weakness we have developed a simple yet effective algorithm to locate polonies in scanned images. We have also developed a method to locate suitable areas for background measurements, as well as estimate overlap amount by neighbor polonies. Our approach achieves higher precision-recall rates than other published polony finding algorithms at a fraction of the speed. The higher precision-recall rates are achieved by taking advantage of the fact that in the multiple hybridization images produced by polony OFRG, polonies will often appear isolated as their overlapping neighbors go dark, revealing their truer, more circular shapes. Metrics for post data acquisition quality control have been developed that enable false polonies or highly overlapped polonies to be detected and removed from analysis.

A straightforward method for polony image alignment has been developed that relies on the same local maximum information used to find polonies. By using only these data points instead of correlating all pixels in an image, as is done in Swift (Whiteford et al. 2009), we also achieve some improvement in

speed and accuracy (data not shown) since the algorithm is made to operate on only a small subset of the most relevant pixels in polony images.

The advantage the EM algorithm has over the LM is most pronounced when only one high density image is used. The reason for this is that the EM models whole intensity profiles of overlapping polonies well, even during overlap events, and can often tease out their locations better than the LM. In contrast, the LM focuses on a single bright pixel to make its determination, which during overlap events is often not a true polony center.

For multiple images, however, the LM algorithm has the advantage in both accuracy and speed. The results of the very high density simulated image (Figure 3.6) shows a continual increase in the AUPR value when more images are used, reaching a high of nearly 0.74. The EM AUPR values peak at 0.56 after 20 images are used and then mostly plateaus, even declining at some points. The steady uptick of the LM demonstrates that it is effectually using the additional information provided by more images but that the EM unable to do so.

A slightly different situation is seen in Figure 3.5, the comparison on a real image. Here, the LM AUPR values also decline occasionally with more images, though both LM and EM achieve higher overall AUPR values. The higher AUPR values can probably be best explained by the lower polony density, which makes the job of finding polonies easier for both algorithms. The occasional decline of LM values may be due to the greater variation in real images – including the range of polony intensity variation, the presence of debris and the slightly

different polony shapes, which is an artifact of the offset between the two lasers in the microarray scanner.

We have created a software package, written as an ImageJ plugin, for accurately and efficiently finding polonies and measuring their intensities across multiple images. The measured intensity data and other quality control information about each polony can be saved in a comma delimited file for post data acquisition processing. The software is open source and can easily be used or adapted for use in other methods that employ polony technology.

BIBLIOGRAPHY

Butz, James A, Hai Yan, Venugopal Mikkilineni, and Jeremy S Edwards. 2004. Detection of allelic variations of human gene expression by polymerase colonies. *BMC Genetics* 5 (February 16): 3. doi:10.1186/1471-2156-5-3.

Lahr, Daniel J G, and Laura A Katz. 2009. Reducing the impact of PCR-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity DNA polymerase. *BioTechniques* 47, no. 4 (October): 857-866. doi:10.2144/000113219.

Li, Wei, Paul M. Ruegger, James Borneman, and Tao Jiang. 2010. Polony Identification Using the EM Algorithm Based on a Gaussian Mixture Model. In *2010 IEEE International Conference on BioInformatics and BioEngineering*, 220-225. Philadelphia, PA, USA. doi:10.1109/BIBE.2010.43. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5521685.

Mitra, Robi D, Vincent L Butty, Jay Shendure, Benjamin R Williams, David E Housman, and George M Church. 2003. Digital genotyping and haplotyping with polymerase colonies. *Proceedings of the National Academy of Sciences of the United States of America* 100, no. 10 (May 13): 5926-5931. doi:10.1073/pnas.0936399100.

R Development Core Team. 2010. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. http://www.R-project.org.

Rasband, W. S. 1997. *ImageJ*. Bethesda, MD, USA. http://rsb.info.nih.gov/ij/.

Rieger, C, R Poppino, R Sheridan, K Moley, R Mitra, and D Gottlieb. 2007. Polony analysis of gene expression in ES cells and blastocysts. *Nucleic Acids Research* 35, no. 22: e151. doi:10.1093/nar/gkm1076.

Shendure, Jay, Gregory J Porreca, Nikos B Reppas, Xiaoxia Lin, John P McCutcheon, Abraham M Rosenbaum, Michael D Wang, Kun Zhang, Robi D Mitra, and George M Church. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science (New York, N.Y.)* 309, no. 5741 (September 9): 1728-1732. doi:10.1126/science.1117389.

Suzuki, M T, and S J Giovannoni. 1996. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Applied and Environmental Microbiology* 62, no. 2 (February): 625-630.

Valinsky, L., G. Della Vedova, T. Jiang, and J. Borneman. 2002. Oligonucleotide fingerprinting of rRNA genes for analysis of fungal community composition. *Appl Environ Microbiol* 68, no. 12: 5999-6004.

Whiteford, N., T. Skelly, C. Curtis, M. E. Ritchie, A. Lohr, A. W. Zaranek, I. Abnizova, and C. Brown. 2009. Swift: primary data analysis for the Illumina Solexa sequencing platform. *Bioinformatics* 25, no. 17 (6): 2194-2199. doi:10.1093/bioinformatics/btp383.

Zhang, Kun, Jun Zhu, Jay Shendure, Gregory J Porreca, John D Aach, Robi D Mitra, and George M Church. 2006. Long-range polony haplotyping of individual human chromosome molecules. *Nature Genetics* 38, no. 3 (March): 382-387. doi:10.1038/ng1741.

Zhu, Jun, Jay Shendure, Robi D Mitra, and George M Church. 2003. Single molecule profiling of alternative pre-mRNA splicing. *Science (New York, N.Y.)* 301, no. 5634 (August 8): 836-838. doi:10.1126/science.1085792.

Chapter 4: Polony OFRG

4.1 Introduction

Microbes often exist in complex and dynamic communities that can have profound effects on the environments or hosts in which they live. A better understanding of these interactions and the impacts microbes have on their hosts is needed and can begin by an assessment of which microbes are present. An better understanding of these interactions can be made possible by frequent sampling, such that the changes in population levels themselves can be scrutinized for clues regarding the interplay between microbe and host.

Many methods currently exist to study microbial communities. These methods range from inexpensive, coarse-grained tools such as culturing, to methods that detect various characteristic differences in microbial rRNA genes such as denaturing gradient gel electrophoresis (DGGE) (Muyzer 1999) and terminal restriction fragment length polymorphism (T-RFLP) (Schütte et al. 2008), to the significantly more expensive and more accurate “gold-standard” of sequencing near full-length rRNA genes (Frank et al. 2007). Recently, strategies for using high-throughput sequencing machines for microbial community analysis have been developed as well (Wu et al. 2010)(Caporaso, Lauber, et al. 2010).

The coarse-grained methods are useful for examining large changes in microbial communities but the low resolution is inadequate for many studies. Sequencing near full-length 16S rRNA genes provides the highest available taxonomic resolution when an accurate “snapshot” of a microbial community is

required. However, though costs are dropping, multi-sample longitudinal studies that employ full-length sequencing are often still too expensive for many labs. High-throughput sequencing currently provides the best compromise between accuracy and throughput but due to the short read-lengths these are still limited in describing the taxonomic makeup of a microbial community. Currently, taxonomic assignments can be confidently made only at the order level; assignments at the genus level can also be made but with less confidence. (Wu et al. 2010)(Caporaso, Lauber, et al. 2010). Single molecule sequencing offers perhaps the most promising eventual solution to microbial community analysis but the read lengths of this technology are still too short currently.

The focus of this research is on improving an alternative method for detecting changes in microbial communities termed oligonucleotide fingerprinting of ribosomal rRNA genes (OFRG) (Valinsky, G. Della Vedova, T. Jiang, et al. 2002). OFRG may be useful for multi-sample studies requiring low cost and high taxonomic resolution. In addition, the new OFRG method that this research focuses on has two important advantages over current sequencing methods. First, the pre-sequencing PCR step known to bias results is skipped (Suzuki and Giovannoni 1996). Second, near full-length rRNA genes are available for sequencing, when desired. The former allows for a truer depiction of the microbes present and the latter provides a way to more confidently assess the identity of any microbe or group of microbes present in a sample.

To estimate the proportions of microbial species present in an environment the OFRG method uses a set of 40 computer-designed DNA probes chosen from a set of training sequences and hybridizes them against an array of sample-derived microbial rRNA gene clones. The hybridization affinity of each probe/clone combination is quantified and then processed in one of two ways. Traditionally these data are transformed into a 40-digit binary “fingerprint” for each clone, where a 0 denotes no hybridization and a 1 denotes a successful hybridization event. These fingerprints are then clustered based on their similarity to the fingerprints of other clones in the array. Alternatively, hierarchical clustering can be performed on the data without a binary transformation of the hybridization intensities. These clusters provide a low-cost estimate of the relative proportions of the various microbial taxa present in an environment since similar fingerprints arise from similar rRNA genes.

OFRG originally employed a printed macroarray format. In a labor intensive procedure, the DNA for each spot on the array was printed from the PCR products of a clone library of microbial rRNA genes originating from environmental samples. The capacity of the method was 9,600 clones per experiment and has been used successfully in several studies (E. Bent et al. 2006)(Lee et al. 2008)(McGuire et al. 2010). The new OFRG method will replace the labor-intensive macroarray with a low-cost microarray, termed a “polony” microarray, with a current capacity of 1K-5K clones per sample and a theoretical capacity of perhaps millions.

Polonies, or “polymerase colonies,” are analogous to and replace the spots of the macroarray. Rather than being printed, each polony is grown in place; each polony consists of many thousands of localized copies of an individual DNA molecule generated through solid-phase PCR in a polyacrylamide hydrogel. Because diffusion of the PCR amplicons is inhibited somewhat by the gel the process results in spots of DNA molecules – the polonies – randomly placed in the gel, one for each original DNA molecule.

Polony technology has been employed in several applications such as DNA sequencing, SNP detection, gene expression studies, genotyping, haplotyping and alternative pre-mRNA splicing (Shendure et al. 2005)(Butz et al. 2004)(Rieger et al. 2007)(Robi D Mitra et al. 2003)(Zhang et al. 2006)(Jun Zhu et al. 2003). There are several important characteristics of a polony microarray that also make it a useful tool for OFRG. These are, 1) polony DNA is anchored to the gel and can be made single-stranded, making it durable and available for probe hybridizations, 2) sample microbial DNA can be spatially isolated, thus eliminating the formation of chimeric amplicons during PCR and 3) microbial genomic DNA can be used directly, without an intermediate PCR step, thus reducing PCR bias of the true makeup of a microbial community.

Despite these advantageous characteristics, however, polony technology is not without its challenges. The most difficult of these challenges is the random placement of polonies in the hydrogel microarray. As a consequence, some polonies overlap with other polonies to varying degrees. Another challenge is

that polonies vary in diameter and can contain different amounts of DNA. These characteristics make detection and/or quantification of polony intensities difficult, and most existing microarray software is unable to properly handle these issues as they are designed for ordered arrays. Lastly, the OFRG paradigm of an "inverted" array (where the spots are sample-DNA and a small set of probes are sequentially hybridized to the array) necessitates that the probes be carefully selected in order to maximize the information gleaned from them.

This work presents the polony OFRG method with several experiments that show its ability to distinguish a range of bacterial communities.

4.2 Methods

Probes. Probes were designed computationally using the Maximum Fidelity Probe Set algorithm developed as part of this work and were purchased from Exicon, Woburn, MA. To increase annealing temperatures of the short (10-mer) probes we added three Locked Nucleic Acid™ nucleotides on the 5' ends. Two spacers and a fluorophore were also added at the 5' ends. Fluorophores were either TYE™ 563 or TYE™ 665.

Polony slides. Polony slides are created by adding a sample of DNA to an acrylamide gel mixture, casting the mixture onto a microscope slide and then performing solid-phase PCR on the cured gel.

Bind Silane. Teflon-coated slides (Thermo Fisher, Waltham, MA) were cleaned by submersing them in 2M HCL for 2 hours with gentle stirring then triple-rinsing in DI H2O. The slides were then treated with Bind Silane (GE

Healthcare Biosciences, Pittsburgh, PA) by submersing them in a treating solution (193mL EtOH, 6mL DI H2O, 600ul glacial acetic acid, 500ul Bind Silane) overnight. Slides were washed 2x in 100% EtOH, allowed to air dry and stored in a desiccator.

Template. For the artificial community experiment, ten clones of 16S rDNA genes, each from different phyla, were selected from a clone library. Phylum designations were determined using Ribosomal Database Project's Naïve Bayesian Classifier (Qiong Wang et al. 2007) with the full sequences of each clone. Plasmids from broth cultures of these clones were extracted with a Miniprep Kit (Qiagen Sciences, Germantown, MD). Their concentrations were determined with a NanoDrop spectrophotometer, mixed in equimolar amounts and diluted.

Using the mixture of plasmids as template, the the 35 cycle template was prepared by PCR (94C for 2 min; 35 cycles of 94C 20 s, 50C 30 s, 72C 60 s; 72C 4 min) using M13 forward and reverse primers in 10ul reactions x 4 replicates. Replicates were pooled and gel purified using a Gel Extraction Kit (Qiagen, Sciences) without the use of ethidium bromide staining or UV light. Polonies were made with either the 0-cycle plasmid DNA or the purified 35-cycle DNA.

For the multi-environment experiment, three human gut samples (CD, UC and healthy) and one ocean sample were selected. Genomic DNA was extracted with a FastDNA Spin kit (MP Biomedicals, Solon, OH) following manufacturer's protocol and a Fast Prep instrument setting of 5.5 for 30 s.

Pacific Ocean water (one gallon, collected at lat. 33.193104, long. -117.386267, in Oceanside, CA), was first passed through a 0.22um filter (Millipore Co., Billerica, MA). The filter membrane was removed, air-dried and cut into small pieces with a clean razor blade. One fourth of the membrane was placed in the Fast Prep tube for DNA extraction. The DNA was gel purified with the same technique used for the 35 cycle template described above. Polonies were made using these genomic DNA as template.

Primers. To increase the efficiency of solid-phase PCR on full-length bacterial 16S sequences we used a modified 27F primer, adding an M13R sequence to the 5' end (ACAGGAAACAGCTATGACCATGAGRRTTTGATYHTGGYTCAG). The reverse primer is the universal bacterial primer 1492R but with an acrydite moiety at the 5' end that covalently attaches to the acrylamide gel during polymerization ([5Acrd]GBTACCTTGTTACGACTT).

ABD mix. A stock of ABD mix was made with 450ul of 40% Acrylamide, 50ul of Bis:Acrylamide and 10mg of DATD.

Acrylamide Gels. The acrylamide gels are prepared first then a PCR master mix is applied to the cured gel for thermocycling. Eighteen microliters of gel mixture (2uM reverse acrydite primer, 20% ABD mix, 2mg/mL BSA, 0.1% TEMED, 0.1% APS) plus template were placed on a coverglass (Fisherbrand 12-543-A, 22x40-2) and a 17x40mm, single-well Teflon-coated slide (Thermo Fisher, Waltham, MA) was carefully brought upside down to the droplet until capillary

action pulled the coverglass off the bench and onto the slide; this helps prevent the formation of bubbles in the gel. The slides were allowed to polymerize for 45 minutes at RT. Coverglasses were removed and the slides were washed in H2O for 15 minutes to remove unbound acrylamide. The gels were dried in air completely before adding 45ul of PCR master mix.

PCR master mix (50mM Tris (pH 8.3), 2.5mM MgCl2, 250uM each dNTP, 400nM of forward primer and 0.24U/ul ThermoSequenase DNA polymerase (USB Corp. Cleavland, OH) and 2mg/mL BSA) was spread onto the gel and allowed to soak in for at least 2 minutes. A 20x35mm coverglass was carefully lowered onto the mix to allow bubbles to escape. A 22x40mm Secure Seal hybridization chamber (SA-500, Grace Bio Labs, Bend, OR) was affixed over the coverglass and filled with mineral oil to prevent evaporation. The slides were cycled in a PTC-100 thermocycler (Biorad Laboratories, Hercules, CA) as follows: 94C for 2 min; 8 cycles of 95C 10 s, 48C 30 s, 72C 4 min; 62 cycles of 94C 20 s, 60C 30 s, 72C 2 min; 72C 5 min.

Post PCR. Slides were placed in a glass hexane-filled Copeland jar for ~10 minutes to loosen the chambers and dissolve the mineral oil. The slides were separated from the chambers and placed in another Copeland jar with fresh hexane for a few moments to remove residual mineral oil. The coverglasses were then removed and four slides each were placed into a LockMailer™ Microscope Slide Jar (Electron Microscopy Sciences, Hatfield PA) for subsequent processing.

Stripping unbound DNA strands. Seven mL of 70% formamide solution were added to each mailer tube and the tubes were placed in a water bath at 75C for 15 minutes to strip the free DNA strands away from the acrydite anchored strands. To cool the solution the tubes were placed in RT H2O for 5 minutes. The formamide solution was removed and replaced with DI H2O then gently shaken for 3 min. This was repeated two more times to remove residual formamide.

Hybridizations. For each hybridization, two fluorescently labeled 10-mer probes were added to a hybridization solution (6x SSPE) at a final concentration of 10nM and 7mL of the solution was added to each mailer tube containing four slides. The tubes were placed in an 87C waterbath for 4 minutes for a short denature step then transferred to a RT waterbath for 30 minutes.

Washes. Hybridization solution was replaced with 3x SSC and the tubes were gently shaken for 5 to 30 minutes, depending on which probes were being hybridized. At the end of the wash time the tubes were immediately placed on ice to inhibit further washing during transport to the scanner.

Scanning. Slides were scanned with a GenePix 4000b Microarray Scanner (Molecular Devices, Sunnyvale, CA) using both the 635nm and 532nm wavelengths at 100% power. The gain settings for each laser were adjusted for each probe-pair to achieve the maximum brightness possible without saturation of the polonies. All subsequent slides in an experiment were scanned at the

same optimized gain setting. Each scan for each probe was saved as a 2969 x 4400 pixel image in 16-bit tiff format.

Removing probes. The 3x SSC solution was replaced with DI H2O and the tubes were placed in a 65C waterbath for 7 minutes. The slides were then either hybridized with the next set of probes or stored overnight in Wash 1E at 4C.

Data acquisition and processing. Raw intensities for each polony and hybridization were measured using ImageJ (National Center for Biotechnology Information, Bethesda, MD) and the PolonyFinder plugin. The intensities were processed using a script written in the R programming language. The raw data for each polony was kept as a vector of 42 values, one value for each of the 42 probes, and was processed as follows. All intensities were background subtracted. To adjust for polony size the values were converted into a percentage of the total intensity measured for all probes for each polony. A distance matrix was created between all polonies from all slides in the experiment using a Euclidian distance metric and then hierarchical clustering was performed to create a dendrogram (not shown).

To better visualize 0-cycle and 35-cycle data, the distance matrix was converted to two separate 2D plots using R's "cmdscale" function. This function attempts to make the distances between points on the plot as close as possible to the distances found in the distance matrix from which it was made. Note that a single distance matrix was made with combined data (three replicates per

treatment); the two scatter plots (0-cycle and 35-cycle) were plotted separately but were taken from this combined data.

The four environmental samples with replicates were compared using a community-level comparison method. UPGMA trees with jackknife support were created where each leaf in a tree represents a whole bacterial community/polony microarray. Distances between bacterial communities were calculated using the Weighted UniFrac beta diversity metric (C. A. Lozupone et al. 2007) via Chiime (Caporaso, Kuczynski, et al. 2010).

Chiime is a processing pipeline made for sequence data. In order to use it with polony intensity data it was necessary to start "in the middle" of the pipeline. At the point of entry, Chiime requires two input files: an OTU table designating the OTUs and the OTU membership counts from each environment, and a Newick-formatted tree file containing the distances between those OTUs. R scripts were written to create these two files from the dendrogram and polony hybridization vectors. A Euclidean distance matrix of the polony intensity vector averages from each cluster was created and the Newick-formatted tree file was created from it using the "ctc" package in R (Lucas and Gautier 2005).

Rather than compare the environmental samples from clusters formed at a single dendrogram "cut level," three pairs of OTU tables and trees were created from the full dendrogram, creating subtrees having 10, 100 and 200 clusters. Weighted UniFrac analyses of the different environments were performed using the OTU tables and trees at these cut levels.

Polony sequences were obtained as follows. Starting with a probe-free microarray, the gel was placed in DI H2O for 3 minutes to remove buffer salts. The array was then removed and gently shaken to remove excess water. 200ul of 1x Sybr Gold was placed on the gel for 5 minutes to stain the DNA. The slide was rinsed and placed in clean DI H2O. Individual polonies were manually excised from a moist gel under UV illumination and magnification using a Leica MZ FLII stereo microscope (Leica Microsystems, Heerbrugg, Switzerland) with a no. 15 stainless steel needle (Minucie Sphinx, Czech Republic) held by a pair of clamping forceps. The needle and gel fragment containing the polony was placed into 6ul of DNA elution buffer (10mM Tris-HCl pH 8.3, 50mM KCl, 1.5mM MgCl2, 0.1% Triton-X-100) in a 200ul microcentrifuge tube, heated for 60 minutes at 95C (Sanguinetti, Dias Neto, and Simpson 1994) and subjected to thermocycling (94C 2min, 35 cycles of 94C 20s, 50C 30s, 72C 60s, then 72C 5min) using universal bacterial primers modified for the USER vector (Invitrogen) 27F (GGAGACAUAGRRTTTGATYHTGGYTCAG) and 1392R (GGGAAAGUACGGGCGGTGTGTRC), and sequenced.

4.3 Results and Discussion

To see if polony OFRG could differentiate the clones of a known bacterial community, and to see if any affect from 35 cycles of PCR on the same community could be detected (versus 0 cycles of PCR), we constructed an artificial bacterial community using clone library plasmids having inserts from 10

different bacterial phylotypes and used this as our 0-cycle template. Our 35-cycle template was the 35-cycle PCR products of the 0-cycle template.

The polony counts for each replicate microarray in the artificial bacterial community are shown in Table 4.1. A dendrogram containing this data was constructed but was too large for display. For easier visualization the same data is displayed as two scatter plots in Figure 4.1. Figure 4.1A, showing polonies made from plasmid DNA (0-cycles of PCR) clearly show the ten bacterial phylotypes as ten relatively tight clusters of points, with some overlap seen in clusters near the center. In Figure 4.1B, showing polonies made from the 35-cycle PCR products of the same plasmid DNA used in Figure 4.1A, one can see a similar pattern of clustering, yet the groups are much more diffuse than in Figure 4.1A. The clusters near the center are essentially merged and a new cluster appears to have formed in the lower-left quadrant (see arrow in Figure 4.1B).

There are several possible explanations for this pattern but they all likely involve a change to the DNA relative to their plasmid originals. PCR artifacts are known to occur in mixed-template reactions. One of those artifacts is the formation of chimeric sequences, where a partially duplicated amplicon from one

	Artificial Community		**Multi-Environment**			
Replicate	**0-Cycle**	**35-Cycles**	**Ocean**	**CD**	**UC**	**HC**
1	1771	2050	1502	4051	1727	2428
2	2054	2287	1049	4685	1595	2239
3	1219	2419	1052	4457	1959	2174
Total	5044	6756	3603	13193	5281	6841

Table 4.1. Polony counts for each microarray in our artificial community (10-clone) and multi-environment samples.

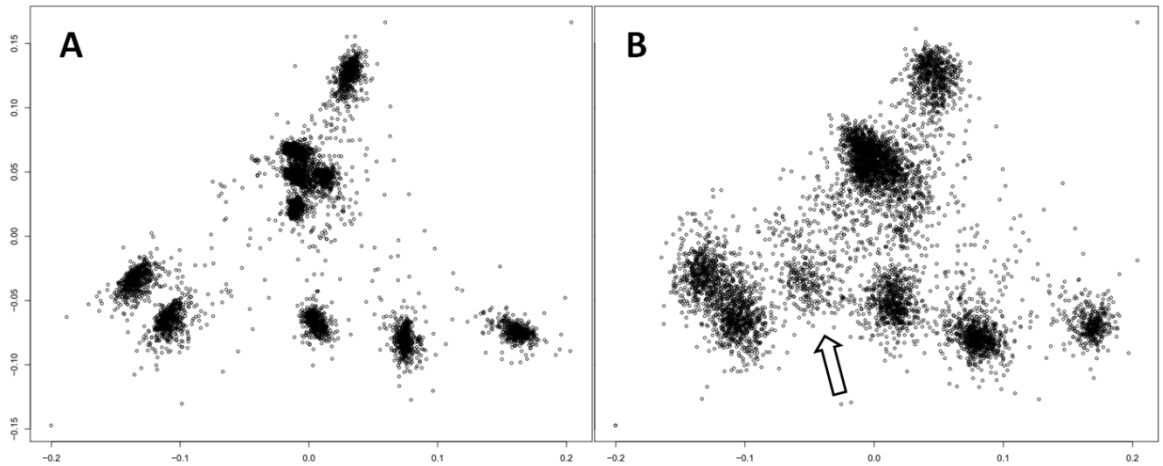

Figure 4.1. Two dimensional scatter plots of the (**A**) 0-cycle and (**B**) 35-cycle templates. Polonies made from 35-PCR cycle templates form much more diffuse groupings as their genomic template counterparts do in **A**, indicating changes to the DNA have been introduced. An apparently new group in **B** (see arrow) suggests a certain error is occurring with some regularity under the PCR conditions used.

species is fully extended from the template of one or more different species (Suzuki and Giovannoni 1996)(Lahr and Katz 2009b). If this occurred in the 35-cycle templates the hybridization patterns from them would also be chimeric. Another source of DNA change is the errors introduced by the polymerase itself during the 35 cycles of PCR. In this case, a polony arising from a mutated amplicon would have the same mutations or more, potentially causing one or more probes to hybridize differently than the original DNA would have. Whatever artifacts may have occurred in the 35-cycle templates, the new cluster in Figure 4.1B (see arrow) may indicate a certain type has occurred with some regularity.

Though the ten bacteria are from ten different phyla and they cluster into ten readily distinguishable groups (Figure 4.1A), a few of those groups are rather close together. It is not known how close two species in the same genus would cluster using polony OFRG. We do know from past experiments using the 9600-clone macroarray version of OFRG that clones with the same 40-digit binary

fingerprint have 97% or more sequence similar to each other (E. Bent et al. 2006). We also know similar species have similar computationally predicted (binary) fingerprints, though about 20% of the time they are identical, suggesting polony OFRG will not be able to distinguish some sequences at the species level (though it is able to distinguish above 98% at the genus level).

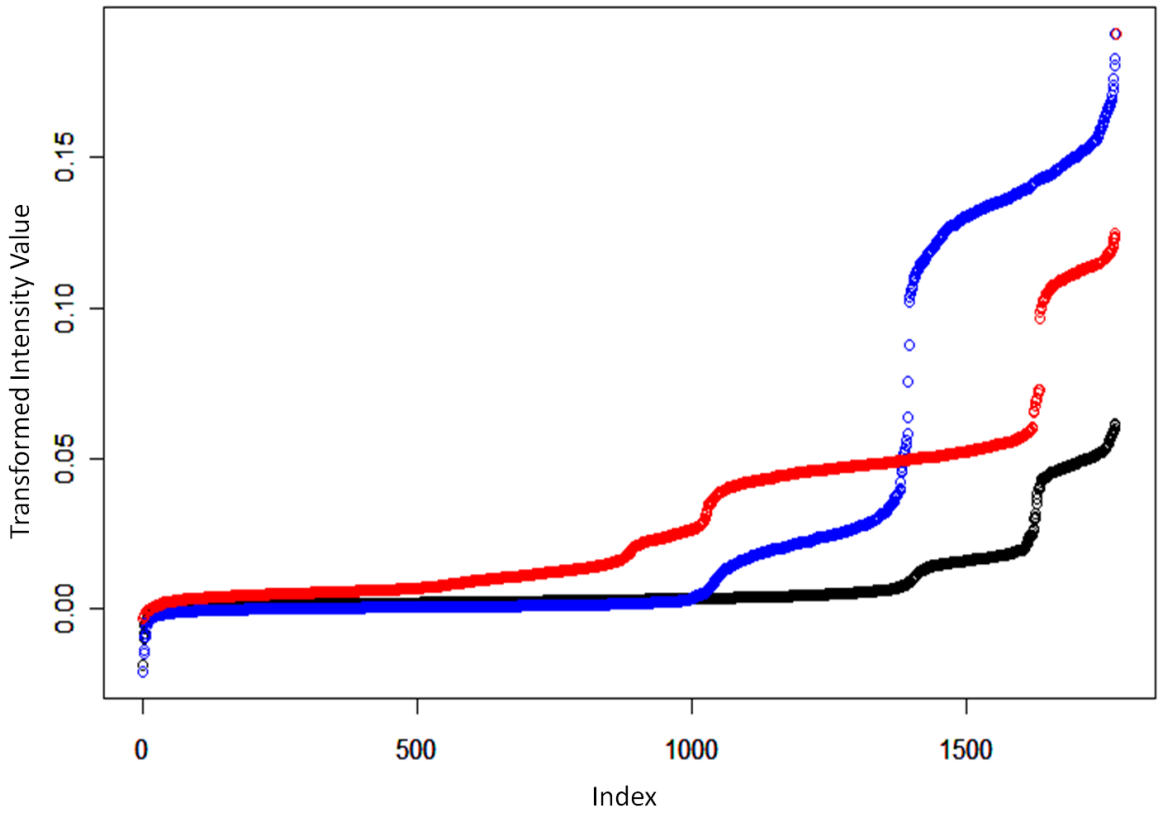

Figure 4.2. Hybridization data provides more than binary information. Each of the three colors represents the transformed and sorted hybridization intensities of a different probe to the same polony microarray, revealing multiple levels of hybridization affinity exist and can be detected. This additional information may allow a higher taxonomic resolution than predicted in the probe design phase, which was based on binary classification of hybridization intensities.

However, actual hybridization behavior is not always as predicted, and there is reason to believe that the real, more complex behavior, can lead to higher resolution. The reason for such optimism is due to the fact that although each probe was designed to provide only one bit of information about a strand of

DNA (1 or 0), indicating whether the probe can bind perfectly or not – it often provides more information, and it may be possible to leverage that information into higher resolution. Specifically, partial match hybridization events may be detectable. Figure 4.2 shows sorted hybridization values for three different probes on the same polony microarray. It is readily apparent from the figure that hybridization data contains intermediate values between a perfect match and a non-perfect match (i.e., where at least one base is a mismatch). This additional information may contribute to a higher resolution than predicted during the probe design phase.

The polony counts for each replicate slide in the multi-environment experiment are shown in Table 4.1. We tested the ability of polony OFRG to differentiate bacterial communities in different environments using a Weighted UniFrac analysis. Because in this experiment the true number of bacterial OTUs is unknown we performed the analysis using several cutoff levels to give differing numbers of clusters. Figures 4.3A-C show the results for 10, 100 and 200 clusters, respectively. We also included one of the 10-clone, 0-PCR cycle replicates in the data for comparison (topmost leaf in all trees). One can see in all three figures that the replicate polony microarrays cluster together much more closely than the different microbial communities themselves, indicating a degree of consistency can be achieved with the method. Many more optimizations could be done, such as automating the hybridization and scanning steps, which would lead to even better repeatability and with less noise.

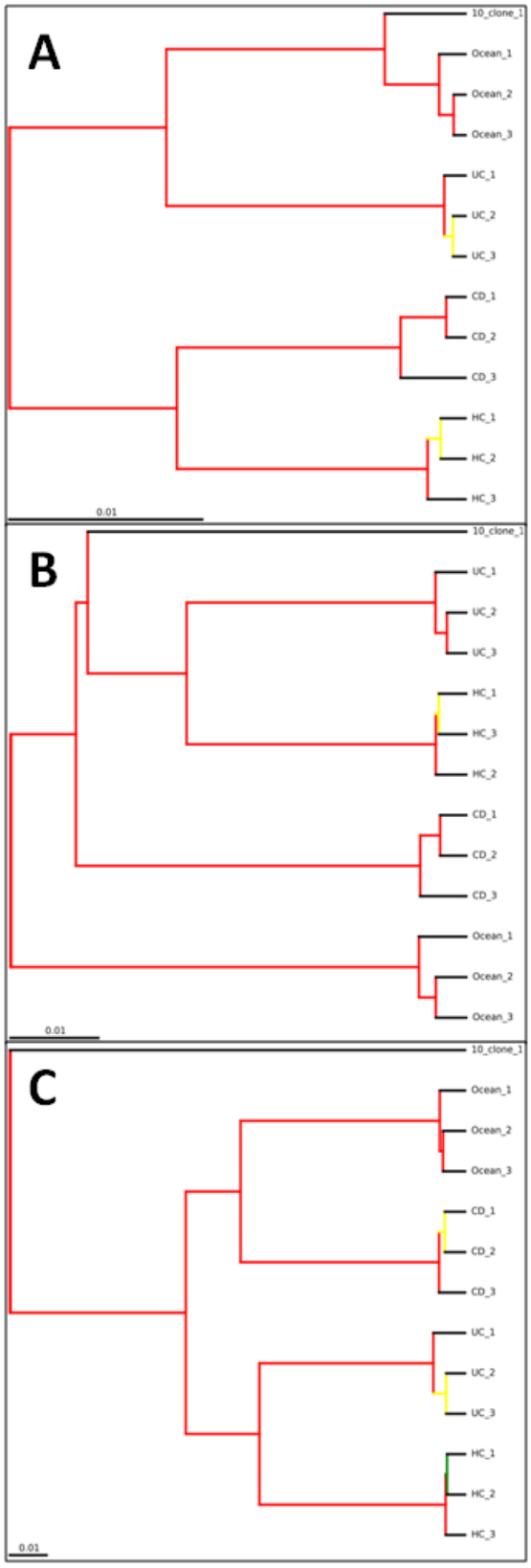

Figure 4.3. UPGMA UniFrac clustering with jackknife support comparing the similarity of several bacterial communities. Red branches indicate 75-100% jackknife support. Yellow indicates 50-75% support. Bacterial DNA used in the 13 microarrays that make up the trees came from five sources: Ocean water, Crohn's patient (CD), Ulcerative Colitis patient (UC), healthy patient (HC) and one artificial community of 10 bacterial phylotypes (10_clone). Microarray replicate numbers are indicated after label abbreviations. Each tree shown is a UPGMA tree of a Weighted UniFrac beta diversity distance matrix of the communities at three different levels of clustering. **A**) Polonies from all communities were grouped into only 10 large clusters. **B**) Polonies were grouped into 100 clusters. **C**) Polonies were grouped into 200 clusters. Grouping polonies was accomplished by selecting an appropriate "cut level" on a dendrogram, which was itself made from a Euclidean distance matrix of all polonies. Community level distance increases (see bars in lower left corners) with increasing clusters and their orders shift but replicate microarrays stay together, implying a higher degree of similarity. The distance between the 10-clone community and the others increases fastest, reflecting its simple structure in relation to the more complex natural communities.

It is interesting to observe that the artificial community (10-clone) not only clusters separately in every case (using 10, 100 and 200 clusters) but that its distance from other communities increases more than the natural communities do from themselves at higher cluster numbers. This is likely due to the fact that the simple artificial community does not contain 100 or 200 OTUs. The data does not distribute naturally into more than 10 clusters and many OTUs remain empty when dividing the data into so many clusters. The Weighted UniFrac metric detects this and reports it as a greater increase in community distance than is occurring with the natural, more complex communities.

We attempted to sequence five polonies from two clusters by excising the polonies from the gel and performing PCR, ligation, etc., as described in the methods section. Only two polonies out of the five (from different clusters) showed a strong gel band after PCR, which reduced our confidence in the sequencing results of the others. However, in Figure 4.4, using the polony sequences of the two strong bands, we predicted probe binding affinities and aligned them with tiled pictures of the group of polonies that were sequenced. The predicted binding is in good agreement with the polony intensities we observe. However, differences between what one might expect and what actually occurs can be seen. The differences, such as strong binding when 2 or more mismatches occur between probe and polony DNA, are likely due to the type and location of the mismatches; if either mismatch occurs near the middle of

the probe or on one of the LNA bases at the 5’ end it will destabilize the duplex more than if both occur near the 3’ end.

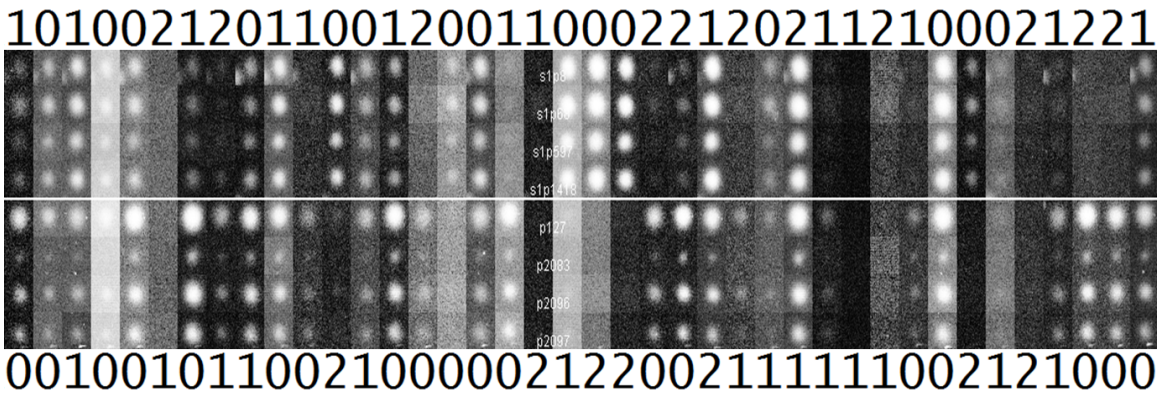

Figure 4.4. Predicted binding of differential probes versus observed binding. Images of eight polonies (rows) across 40 hybridizations (columns) are shown. The top panel shows four polonies from cluster 8 and the bottom panel shows four polonies from cluster 52. DNA sequences were obtained from the polonies in rows 2 and 7 and were used to predict three types of binding behavior. A ‘0’ means zero mismatches exists between probe and polony DNA (strong binding). A ‘1’ likewise indicates a one base mismatch (weak binding) and a ‘2’ indicates two (or more) mismatches (very weak binding).

Although we are making a crude prediction about probe binding behavior here, it is important to note that our clustering method is not dependent on predicted behavior. Rather, clustering is based on observed intensities. We included other polonies in the figure that are members of the same clusters as the sequenced polonies for an example of how our clustering method groups polonies with similar hybridization patterns. In general, other clusters have polonies with highly similar patterns as well, though there are exceptions and more work could be done to improve the method.

The cost of producing 50 polony microarrays is currently about $4400. An Illumina run with 50 libraries costs approximately $16,800. The number of Illumina reads per library in this scenario would be about 2 million, vastly outnumbering the 5000 polonies per array. Most of the polony OFRG costs are

for labor for hybridizing and scanning the arrays, however. If these steps could be automated, as the sequencing/scanning cycles for Illumina and other platforms are, the costs per array could drop dramatically; using three or four colored fluors could also decrease costs. Another way to increase throughput and lower cost is to increase the density of polonies. Polony diameters can be made smaller to allow for an estimated 5 million polonies per array (R D Mitra and G M Church 1999). Eliminating overlaps would allow even higher densities.

Polony OFRG is a new technology, and although we compare it to sequencing technologies in several ways we do not claim it is superior for most applications. Rather, the main advantages of polony OFRG – reducing PCR bias in mixed template reactions and having access to full-length 16S rDNA gene sequences – are only so for specialized applications.

BIBLIOGRAPHY

Bent, E, B Yin, A Figueroa, J Ye, Q Fu, Z Liu, V Mcdonald, D Jeske, T Jiang, and J Borneman. 2006. Development of a 9600-clone procedure for oligonucleotide fingerprinting of rRNA genes: Utilization to identify soil bacterial rRNA genes that correlate in abundance with the development of avocado root rot. *Journal of Microbiological Methods* 67, no. 1 (10): 171-180. doi:10.1016/j.mimet.2006.03.023.

Butz, James A, Hai Yan, Venugopal Mikkilineni, and Jeremy S Edwards. 2004. Detection of allelic variations of human gene expression by polymerase colonies. *BMC Genetics* 5 (February 16): 3. doi:10.1186/1471-2156-5-3.

Caporaso, J Gregory, Justin Kuczynski, et al. 2010. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* 7, no. 5 (4): 335-336. doi:10.1038/nmeth.f.303.

Caporaso, J Gregory, Christian L Lauber, William A Walters, Donna Berg-Lyons, Catherine A Lozupone, Peter J Turnbaugh, Noah Fierer, and Rob Knight. 2010. Microbes and Health Sackler Colloquium: 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* (June 3). doi:10.1073/pnas.1000080107. http://www.ncbi.nlm.nih.gov/pubmed/20534432.

Frank, Daniel N, Allison L St Amand, Robert A Feldman, Edgar C Boedeker, Noam Harpaz, and Norman R Pace. 2007. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proceedings of the National Academy of Sciences of the United States of America* 104, no. 34 (August 21): 13780-13785. doi:10.1073/pnas.0706625104.

Lahr, Daniel J G, and Laura A Katz. 2009. Reducing the impact of PCR-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity DNA polymerase. *BioTechniques* 47, no. 4 (October): 857-866. doi:10.2144/000113219.

Lozupone, C. A., M. Hamady, S. T. Kelley, and R. Knight. 2007. Quantitative and Qualitative Diversity Measures Lead to Different Insights into Factors That Structure Microbial Communities. *Applied and Environmental Microbiology* 73, no. 5 (1): 1576-1585. doi:10.1128/AEM.01996-06.

Lucas, Antoine, and Laurent Gautier. 2005. *ctc: Cluster and Tree Conversion*. http://antoinelucas.free.fr/ctc.

McGuire, Krista L, Elizabeth Bent, James Borneman, Arundhati Majumder, Steven D Allison, and Kathleen K Tresederi. 2010. Functional diversity in resource use by fungi. *Ecology* 91, no. 8 (August): 2324-2332.

Mitra, R D, and G M Church. 1999. In situ localized amplification and contact replication of many individual DNA molecules. *Nucleic Acids Research* 27, no. 24 (December 15): e34.

Mitra, Robi D, Vincent L Butty, Jay Shendure, Benjamin R Williams, David E Housman, and George M Church. 2003. Digital genotyping and haplotyping with polymerase colonies. *Proceedings of the National Academy of Sciences of the United States of America* 100, no. 10 (May 13): 5926-5931. doi:10.1073/pnas.0936399100.

Muyzer, G. 1999. DGGE/TGGE a method for identifying genes from natural ecosystems. *Current Opinion in Microbiology* 2, no. 3 (6): 317-322. doi:10.1016/S1369-5274(99)80055-1.

Rieger, C, R Poppino, R Sheridan, K Moley, R Mitra, and D Gottlieb. 2007. Polony analysis of gene expression in ES cells and blastocysts. *Nucleic Acids Research* 35, no. 22: e151. doi:10.1093/nar/gkm1076.

Sanguinetti, C J, E Dias Neto, and A J Simpson. 1994. Rapid silver staining and recovery of PCR products separated on polyacrylamide gels. *BioTechniques* 17, no. 5 (November): 914-921.

Schütte, Ursel M E, Zaid Abdo, Stephen J Bent, Conrad Shyu, Christopher J Williams, Jacob D Pierson, and Larry J Forney. 2008. Advances in the use of terminal restriction fragment length polymorphism (T-RFLP) analysis of 16S rRNA genes to characterize microbial communities. *Applied Microbiology and Biotechnology* 80, no. 3 (September): 365-380. doi:10.1007/s00253-008-1565-4.

Shendure, Jay, Gregory J Porreca, Nikos B Reppas, Xiaoxia Lin, John P McCutcheon, Abraham M Rosenbaum, Michael D Wang, Kun Zhang, Robi D Mitra, and George M Church. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science (New York, N.Y.)* 309, no. 5741 (September 9): 1728-1732. doi:10.1126/science.1117389.

Suzuki, M T, and S J Giovannoni. 1996. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Applied and Environmental Microbiology* 62, no. 2 (February): 625-630.

Valinsky, L., G. Della Vedova, T. Jiang, and J. Borneman. 2002. Oligonucleotide fingerprinting of rRNA genes for analysis of fungal community composition. *Appl Environ Microbiol* 68, no. 12: 5999-6004.

Wang, Qiong, George M Garrity, James M Tiedje, and James R Cole. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology* 73, no. 16 (August): 5261-5267. doi:10.1128/AEM.00062-07.

Wu, Gary D, James D Lewis, Christian Hoffmann, Ying-Yu Chen, Rob Knight, Kyle Bittinger, Jennifer Hwang, et al. 2010. Sampling and pyrosequencing methods for characterizing bacterial communities in the human gut using 16S sequence tags. *BMC Microbiology* 10, no. 1: 206. doi:10.1186/1471-2180-10-206.

Ye, Jingxiao, Jimmy W. Lee, Laura L. Presley, Elizabeth Bent, Bo Wei, Jonathan Braun, Neal L. Schiller, Daniel S. Straus, and James Borneman. 2008. Bacteria and bacterial rRNA genes associated with the development of colitis in IL-10 Mice. *Inflammatory Bowel Diseases* 14, no. 8 (8): 1041-1050. doi:10.1002/ibd.20442.

Zhang, Kun, Jun Zhu, Jay Shendure, Gregory J Porreca, John D Aach, Robi D Mitra, and George M Church. 2006. Long-range polony haplotyping of individual human chromosome molecules. *Nature Genetics* 38, no. 3 (March): 382-387. doi:10.1038/ng1741.

Zhu, Jun, Jay Shendure, Robi D Mitra, and George M Church. 2003. Single molecule profiling of alternative pre-mRNA splicing. *Science (New York, N.Y.)* 301, no. 5634 (August 8): 836-838. doi:10.1126/science.1085792.

Chapter 5: Conclusion

Microbes are present in nearly all terrestrial environments and play varied and complex roles. Understanding these roles – the interactions and impacts on the environments and hosts where they reside – requires gathering data on many levels. Some of the most basic facts required are to know which microbes are present and in what number. To the degree such information can be obtained accurately and inexpensively, the more varied and more thorough such studies can be performed. Finding patterns in microbe population levels can provide a basis from which to make a rational hypothesis regarding which microbes might be involved in a certain functional parameter of interest, such as inflammation level or cancer progression. Such hypotheses can form a starting point for more detailed analyses of the effect a particular microbe or microbes may be playing, helping to tease apart the myriad interactions between the microbial world and their environments.

Current high-throughput sequencing technologies are a vital tool in this endeavor, and are able to provide vast amounts of important information about the structure and function of microbial communities. Yet, their ability to accurately survey microbial communities may be negatively affected by the phenomenon of PCR bias, and their short read-lengths make species-level identification nearly impossible. Thus, for studies where such information is required, high-throughput sequencing has limited value. The “gold standard”

alternative – Sanger sequencing thousands of full-length clones – is still prohibitively expensive in many instances.

To address these issues we have improved the accuracy and throughput of OFRG using polony technology. The new probe set design method for OFRG incorporates available taxonomic information of the training sequences. It attempts to generate a set of 40 probes that, when sequentially hybridized to the small subunit rRNA genes of a polony, create a pattern of hybridization intensities that is unique for each species. These 40-digit hybridization "fingerprints" are clustered based on their similarity to other fingerprints and tentatively represent real microbial taxonomic groups. After data analysis is complete, the near full-length DNA sequence of any clusters deemed worthy of further investigation can be determined by retrieving and sequencing one or more polonies in a cluster. It may also be possible to develop a database of hybridization fingerprints and sequence information that allows tentative but rapid identification of the microbes in a microbial community prior to sequencing.

The ease of polony microarray construction is both an advantage and disadvantage. They are inexpensive mainly because the DNA is spread randomly on the microarray, which negates the need for expensive equipment that would be necessary to place them in ordered grids. Unfortunately, the random placement of DNA molecules results in randomly placed polonies, which greatly complicated the task of measuring their intensities. No software package existed that satisfactorily met the unique challenges polony OFRG images

presented. The software we developed to handle the task uses a simple yet effective strategy to identify true polony locations, overlapping polonies and suitable background measurement areas for each polony. The simple strategy of using local maximum pixel intensities to identify polony centers is extremely fast. It is also accurate when it leverages the additional information about polony locations provided in the 42 unique probe hybridization images of a polony OFRG microarray to choose the most probable location. Written as an ImageJ plugin and an R script for post-measurement processing, the software is open-source and can be used by anyone or modified for related applications if desired.

Polony OFRG replaces the previous OFRG method that used a 9600 clone macroarray. Whereas previously it was necessary to allocate the 9600 clones equally among all the samples being tested, each polony OFRG microarray contains the DNA of one and only one sample; a new microarray is made for each sample. Thus, the number of clones analyzed is independent of the number of samples and total throughput is flexible. We expect the number of clones per polony microarray can be easily raised to 10K or more with very little optimization. We may reach densities up to 50K per array with our current microarray scanner by shrinking the maximum polony sizes to 50um from a current 120um.

There are a several ways to improve polony OFRG. Hybridization conditions can be individually optimized to reduce background and increase signal intensities. New data normalization methods can be developed that use

an artificial polony control construct that has sequences that perfectly match to all probes. As mentioned earlier, a hybridization fingerprint database can be constructed that connects fingerprints to sequence information. This information would grow over time as more and more polonies are sequenced. The probe design pipeline could be improved by incorporating the knowledge gained about mismatch behavior, leading to development of a probe set specifically designed to leverage this information into higher taxonomic resolution.

Polony OFRG is a new technology with strengths and weaknesses. Although we compare it to sequencing technologies in several ways we do not claim it is superior for most applications. As a hybridization-based method it suffers from the inherent difficulties associated with these strategies, such as optimizing the hybridization conditions and normalizing data from one array to the next. Unlike sequencing, the raw hybridization data produced by OFRG is not quantitative. Nevertheless, the main strengths of polony OFRG – reducing PCR bias in mixed template reactions and allowing access to near full-length 16S rDNA gene sequences – are currently unavailable with other tools used for microbial community analysis. Polony OFRG may be a useful tool in studies where these capabilities are needed.