Published online January 10, 2006

Nucleic Acids Research, 2006, Vol. 34, No. 1 ***e5***
doi:10.1093/nar/gnj007

Rapid quantitative profiling of complex microbial populations

Chana Palmer, Elisabeth M. Bik[1,4], Michael B. Eisen[5,6], Paul B. Eckburg[1,2,4], Theodore R. Sana[7], Paul K. Wolber[7], David A. Relman[1,2,4] and Patrick O. Brown[3,8,*]

Department of Genetics, [1]Department of Microbiology and Immunology, [2]Division of Infectious Diseases and Geographic Medicine and [3]Department of Biochemistry, Stanford University School of Medicine, Stanford, CA, USA, [4]Veterans Affairs Palo Alto Health Care System, Palo Alto, CA, USA, [5]Lawrence Berkeley National Laboratory, Berkeley, CA, USA, [6]Department of Molecular and Cell Biology, University of California, Berkeley, CA, USA, [7]Agilent Technologies, Santa Clara, CA, USA and [8]Howard Hughes Medical Institute, Chevy Chase, MD, USA

Received October 14, 2005; Revised and Accepted December 13, 2005

ABSTRACT

Diverse and complex microbial ecosystems are found in virtually every environment on earth, yet we know very little about their composition and ecology. Comprehensive identification and quantification of the constituents of these microbial communities—a 'census'—is an essential foundation for understanding their biology. To address this problem, we developed, tested and optimized a DNA oligonucleotide microarray composed of 10 462 small subunit (SSU) ribosomal DNA (rDNA) probes (7167 unique sequences) selected to provide quantitative information on the taxonomic composition of diverse microbial populations. Using our optimized experimental approach, this microarray enabled detection and quantification of individual bacterial species present at fractional abundances of <0.1% in complex synthetic mixtures. The estimates of bacterial species abundance obtained using this microarray are similar to those obtained by phylogenetic analysis of SSU rDNA sequences from the same samples—the current 'gold standard' method for profiling microbial communities. Furthermore, probes designed to represent higher order taxonomic groups of bacterial species reliably detected microbes for which there were no species-specific probes. This simple, rapid microarray procedure can be used to explore and systematically characterize complex microbial communities, such as those found within the human body.

INTRODUCTION

Microorganisms are the 'unseen majority' in almost every ecosystem on our planet; they far surpass plants and animals in abundance and diversity (1). The human body is no exception—the number of bacterial cells in and on the human body is estimated to be of an order of magnitude greater than the number of human cells (2). Commensal microbes have been shown to play numerous important roles in host physiology, including protection against intestinal epithelium injury (3), nutrient absorption, development of the neonatal gut epithelium (4) and regulation of host fat storage (5). Still, many fundamental questions about the human microbial flora remain unanswered because of their technical intractability. Which organisms occupy the diverse specific niches and microenvironments of the human body; how do these populations vary from individual to individual and over time; how are they affected by disease, geography, hygiene, diet and genotype; how might they influence human physiology and disease risk?

Ribosomal RNA gene sequence analysis is a powerful method for identifying and quantifying members of microbial communities (6,7). This gene is present in all living organisms, contains diverse species-specific domains and can be used to infer phylogenetic relationships reliably at multiple taxonomic levels. Existing techniques for surveying bacterial populations—small subunit (SSU) ribosomal DNA (rDNA) sequence analysis (8,9), temperature/denaturing gel electrophoresis (D/TGGE) (10–12) and terminal-restriction fragment length polymorphism (T-RFLP) (13)—are useful for identifying the dominant members of a population and for discovering new rDNA species, but inadequate for detection and quantification of rare species, while currently available quantitative techniques, such as fluorescence *in situ* hybridization (FISH) (14), dot-blot hybridization (15) and real-time

*To whom correspondence should be addressed. Tel: +1 650 723 0005; Fax: +1 650 723 1399; Email: pbrown@pmgm.stanford.edu

PCR (16), are not easily applied to large numbers of taxonomic groups.

DNA microarray technology, with its ability to detect and measure thousands of distinct DNA sequences simultaneously, has been recognized as a potentially valuable tool for high-throughput, quantitative, systematic and detailed studies of microbial communities. Early applications of rDNA microarrays have included small-scale microarrays for profiling specific bacterial species of interest (17–24), or for providing high-level overviews of the composition of microbial communities (25,26) and one larger-scale microarray with both species and more inclusive taxonomic level probes (27,28). Previous reports of SSU rDNA microarrays have only tested the performance of their system using small numbers bacterial species, and thus the limits of the SSU microarray approach have not yet been well defined, and it has been difficult to judge the general usefulness of this approach. Furthermore, with the exception of one recent publication (28), which demonstrated a strong correlation between relative abundance and signal intensity for five test species, there have been no reports of successful quantification of individual species in the context of complex mixtures, using SSU rDNA microarrays. In this report, we describe the development, extensive validation and application of a DNA oligonucleotide microarray with 10 462 40 nt SSU rDNA probes (7167 unique sequences) and an optimized protocol for rapid, quantitative profiling of diverse microbial populations.

MATERIALS AND METHODS

Microarray design and production

The array design was based on a database of 8681 SSU rDNA sequences representing a diverse set of bacterial, archaeal and eukaryotic species (29) (see Supplementary Data S1 and S2). We defined a set of 359 target species and nodes in the design database phylogenetic tree based on their representation of the species with which we planned to do validation experiments. For each target species and target node, we performed a local BLAST search (30) to identify 40 nt probes predicted to hybridize to 'in group' species but not to 'out group' species. The five top-scoring sequences for each target node or target species were selected, yielding a set of 8138 probes that together represented diverse taxonomic groups, with specificities ranging from species to phylum level. We also included 2324 control probes designed for systematic examination of parameters that affect hybridization. Surface-attached oligonucleotide probes were synthesized *in situ* as previously described (31) (Agilent Technologies, Palo Alto, CA). Each array had 10 462 probes (7167 unique sequences), each consisting of a 40 nt probes sequence plus a 10 nt poly(T) linker. All probes (both taxonomic and control probes) were later re-annotated and assigned a taxonomic specificity using a different algorithm that was found to better predict hybridization behavior (see Probe annotation).

SSU rDNA amplification

SSU rDNA was amplified using broad-range bacterial primer Bact-8F with either universal primer 1391R or T7-1391R (5′-GACGGGCGGTGTGTRCA-3′) (33) or T7-1391R (5′-AATTCTAATACGACTCACTATAGGGA-GACGGGCGGTGTGTRCA-3′). These primers amplify approximately 90% of the full-length prokaryotic SSU ribosomal RNA coding sequence.

Bacterial DNA test pools construction and strains

Test pools consisted of mixtures of SSU rDNA amplicons from a set of American Type Culture Collection (ATCC) bacterial strains. SSU rDNA sequences were amplified by PCR from individual lysates of 229 bacterial species using universal primers Bact-8F and T7-1391R, using 35 cycles of amplification. A common reference pool was constructed by pooling equimolar amounts of SSU rDNA amplicons from all 229 bacterial species. Artificial test pools were made by mixing varying proportions of amplified DNA from selected species (see Supplementary Data S3).

Colon biopsy samples

Colonic tissue biopsies were collected from the cecum and transverse colon of three human subjects, aged 43, 50 and 50 years, who were healthy controls from a population-based case–control study of inflammatory bowel disease in Manitoba, Canada. All participants in the study provided their signed informed consent. The use of these subjects was approved by the Stanford University Administrative Panel on Human Subjects in Medical Research. The tissue samples were obtained at the University of Manitoba, placed immediately on dry ice and shipped to Palo Alto, CA for analysis, where they were stored in their original tubes at −80°C. DNA was extracted from intestinal tissue using the QIAamp® DNA Mini Kit (Qiagen, Inc., Valencia, CA), eluted in a final volume of 200 µl elution buffer and stored at −20°C. For microarray hybridization experiments, the SSU rDNA gene was amplified from the extracted DNA using primers Bact-8F and T7-1391R, using a 20-cycle PCR.

Construction and phylogenetic analysis of SSU ribosomal DNA clone libraries

An SSU ribosomal DNA clone library was constructed from the biopsy samples. Briefly, the SSU rDNA gene was amplified from the biopsy DNA using primers Bact-8F and 1391R. Purified PCR products were cloned and sequenced, and the sequences were taxonomically classified as described previously (34).

Direct labeling of double-stranded DNA (method 1)

Individual or pooled, gel-purified SSU rDNA sequences, amplified using primers Bact-8F and 1391R, were used as a template for random-octomer-primed synthesis of Cy-dye labeled double-stranded DNA (dsDNA), using a modification of Invitrogen's BioPrime DNA labeling system. Cy-labeled DNA was purified using the QIAquick PCR Purification Kit (Qiagen) and quantified by ultraviolet spectrophotometry.

Indirect labeling of single stranded RNA (method 2)

Individual or pooled, gel-purified SSU rDNA sequences, amplified using primers Bact-8F and 1391R, were used as a template for *in vitro* transcription-based synthesis of amino-allyl labeled single stranded RNA (ssRNA) using the MEGAScript T7 *In Vitro* Transcription Kit (Ambion, Austin,

TX). RNA was fragmented to 50–200 nt and stored at −20°C (Ambion Fragmentation Reagents). Immediately before hybridization, 1–2 μg of sample were coupled to Cy3 or Cy5 NHS-esters.

Hybridization

The hybridization mix typically contained 40–500 ng of Cy5-labeled test sample (smaller amounts used in low complexity tests) mixed with 230 ng of a Cy3-labeled reference pool. The Cy3- and Cy5-labeled samples were mixed together in a final volume of 100 μl, heated to 95°C for 5 min, and cooled on ice. We then added 30 μl of 10× control targets (Agilent, Palo Alto, CA), 150 μl of 2× hybridization buffer (Agilent Life Sciences *In Situ* Hybridization Kit) and 20 μl of water to each 100 μl sample. Of this mixture, 200 μl was applied to the slide, sealed (22 K hybridization chambers; Agilent Technologies), and hybridized in a rotisserie rotating oven for ∼16 h at 60°C. Slides were washed in 6× SSC, 0.005% Triton X-102 for 10 min at room temperature, ice-cold 0.1× SSC, 0.005% Triton X-102 for 5 min and scanned immediately using an Agilent DNA Microarray Scanner. Washing and scanning were performed in a low ozone environment (35).

Probe annotation

A set of BLAST parameters was empirically derived in order to maximize the correlation between signal intensity and BLAST score. We determined an alignment score below which appreciable signal was virtually never observed (28 out of a maximum of 40) where the score for a given alignment was the number of matched bases minus the number of internal mismatches (Supplementary Figure S1). With this set of parameters, BLAST was used to predict the hybridization of each taxonomic and control probe (10 462 probes, 7167 unique sequences) to the RDP type strains database (4370 sequences, downloaded June 2004). Each probe was annotated according to the most specific taxonomic group encompassing all of the species that scored >28 out of 40 (matches–mismatches) (Supplementary Data S4 and S5).

Microarray data analysis

All experiments involved co-hybridization of a Cy5-labeled test sample and a Cy3-labeled reference of known composition. The relative abundance of each bacterial species was expected to be proportional to the mean of the Cy5/Cy3 ratios of the corresponding species-specific probes for that species. The relative abundance of cognate species for more inclusive (>1 target species) probes ('abundance score') was estimated by multiplying the observed Cy5/Cy3 ratio for each probe by an 'expected reference binding score', reflecting the proportion of sequences in the reference pool that would be expected to hybridize to that probe. We determined the 'expected reference binding score' for each probe for our reference mixture by using BLAST to predict hybridization of the common reference pool to the probe sequence. We obtained abundance estimates for taxonomic groups using 'composite probe sets'. For each taxonomic group with at least one representative probe, we defined a 'composite probe set' by identifying the set of probes that captured as many of the species in that group as possible (using relationships defined by taxonomy) without representing any species more than once (Supplementary Data S6). The relative abundance of each taxonomic group was estimated by summing the 'abundance scores' (Cy5/Cy3 ratio multiplied by probe-specific reference binding factor) across the corresponding composite probe set. These probe sets are not necessarily comprehensive and thus provide only lower-bound estimates for each more inclusive taxonomic group.

Data filtering and normalization

Data were extracted from microarray images using the most current version of the Agilent Feature Extraction software (Versions 5.1.1–7.1.1). Cy5/Cy3 ratios were computed directly from Cy5 and Cy3 raw background subtracted data. The data from each array were normalized by applying a common scaling factor to each probe—the ratio of the Cy5/Cy3 ratio of universal probes to the known Cy5/Cy3 sample mass ratio. Data from colon biopsies were filtered for probes that satisfy both of the following criteria: (i) reference channel (Cy3) signal above background in at least 50% of samples, where background is defined as 90th percentile Cy3 signal intensity for a set of 290 negative control (antisense) sequences and (ii) at least one bacterial species in the common reference pool was predicted to match the probe with a BLAST score (matches minus mismatches) of least 25 out of 40. This filter limited further analysis to the 7343 probes (4620 unique sequences) with sequence homology to the 16S rDNA gene of one or more species in the reference pool.

Validation of selected microarray results

To confirm the presence of species detected by the microarray but not by sequencing, we performed a specific PCR with one universal primer and one specific primer (either the microarray 40mer in question, or a newly designed 20mer primer) on the original colon biopsy DNA as well as on a set of positive and negative control bacterial lysates. PCR conditions were identical to those for the original universal SSU rDNA PCRs, with 35 cycles. Amplified rDNA was cloned and sequenced. Amplified sequences were taxonomically annotated by online BLAST searching.

Detailed experimental and data analysis procedures are available as Supplementary Data (S7).

RESULTS

The goal of this work was to develop a fast, efficient and quantitative procedure that enables comprehensive profiling of complex microbial populations in natural environments. The essential features of our experimental method are (i) isolation of genomic DNA from microbial populations; (ii) PCR amplification of nearly full-length SSU rDNA sequences, using phylogenetically conserved primers; (iii) preparation of fluorescently labeled copies of the resulting amplified sequences; and (iv) quantitative determination of the species and taxonomic groups represented in the sample, by comparative fluorescent hybridization to a microarray of SSU rDNA sequences. (For a schematic diagram, see Supplementary Figure S2.) The microarray was designed by using a BLAST-based (30) algorithm to simulate hybridization specificity and identify SSU rDNA sequences capable of

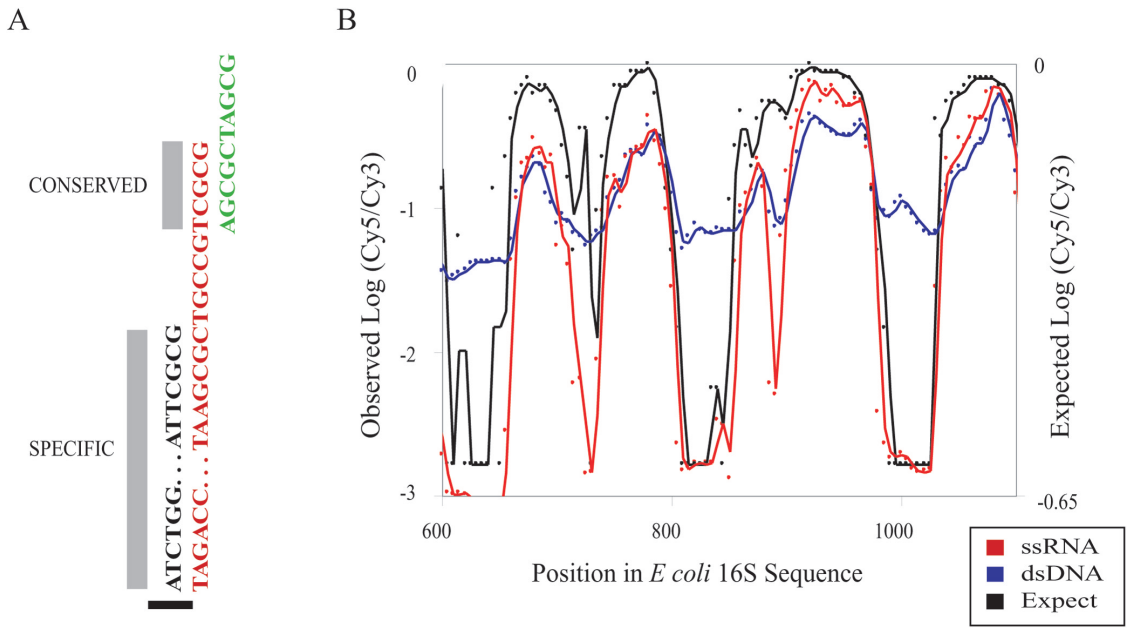

Figure 1. Comparison of dsDNA and ssRNA Labeling Methods. (**A**) 'Probe Hitchhiking' model for non-independence of Cy5 and Cy3 signal. (**B**) Signal:noise comparison for hybridization using dsDNA and ssRNA methods. *B.subtilis* and *E.coli* SSU rRNA genes were amplified and labeled with Cy5 and Cy3 respectively, using either method 1 (dsDNA) or method 2 (ssRNA), and were co-hybridized to microarrays. Cy5/Cy3 ratios are shown for tiled *E.coli* SSU rDNA sequences. Expected ratio for each probe is the ratio of BLAST scores of *B.subtilis* and *E.coli* SSU rDNA sequence to that probe. Red, ssRNA; blue, dsDNA; and black, expected.

distinguishing specific bacterial species and taxonomic groups from all other species.

Protocol optimization

Initially, we hybridized Cy-dye labeled dsDNA produced by direct incorporation of Cy-dyes during random-octomer-primed DNA synthesis from an rDNA amplicon template. With this protocol, we observed several instances in which a probe hybridized to both Cy5 and Cy3-labeled rDNA species when only one (either Cy5 or Cy3) labeled species had sequence homology to the probe (Supplementary Figure S3). We hypothesized that the conserved sequences that inevitably flank the phylogenetically specific sequences in rDNA were enabling indirect hybridization of non-specific rDNA sequences (Figure 1A), and that this 'hitchhiking' would be eliminated by hybridizing labeled nucleic acid of a single complementarity. Indeed, we found that our specificity increased greatly when we modified our protocol and hybridized labeled ssRNA instead of dsDNA (Figure 1B). The key difference between the results obtained with the two protocols involved the Cy5/Cy3 ratios of the probes that were not homologous to the Cy5-labeled sample—these were 30-fold lower with the ssRNA protocol than with the dsDNA protocol (mean log (Cy5/Cy3) of −1.9 and −0.5, respectively). Therefore, we used the 'ssRNA' protocol (method 2 in Materials and Methods) for all subsequent experiments.

Sources of hybridization variation

By analysing sets of probes that covered, in an overlapping 'tiling', the entire SSU rRNA gene sequence of two species (*Bacillus subtilis* and *Escherichia coli*), we discovered that the raw signal intensity obtained from the hybridization of

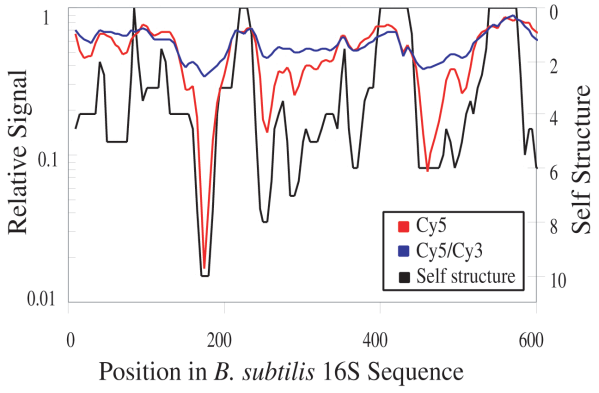

Figure 2. Variation in signal intensity across *B.subtilis* SSU Sequence. Hybridization of identical Cy5 and Cy3-labeled *B.subtilis* rRNA to overlapping probes tiling along the *B.subtilis* SSU rDNA sequence illustrates variation in signal intensity across perfectly matched sequences. Cy5 and Cy5/Cy3 ratio are both normalized to the maximum across the entire tiled region. Self-structure is measured as the length of the longest hairpin. Red, Cy5; blue, Cy5/Cy3; and black, self structure.

cognate RNA to different 40mers derived from the same 16S rDNA gene varied over a 67-fold range. We found that potential self-structure was a strong predictor of variation in signal intensity; sequences with high potential to form stable intra-molecular duplexes [measured as the length of the longest hairpin using Vienna RNA fold (36)] had the lowest signal intensity upon hybridization of their cognate sequence ($r = -0.52$ and $r = -0.36$ for *B.subtilis* and *E.coli*, respectively). GC content had a negligible effect on hybridization intensity ($r = 0.07$ and $r = 0.17$ for *B.subtilis* and *E.coli*, respectively; data not shown). We were able to reduce this variation considerably by performing comparative hybridization of each unknown sample (Cy5-labeled) with a defined

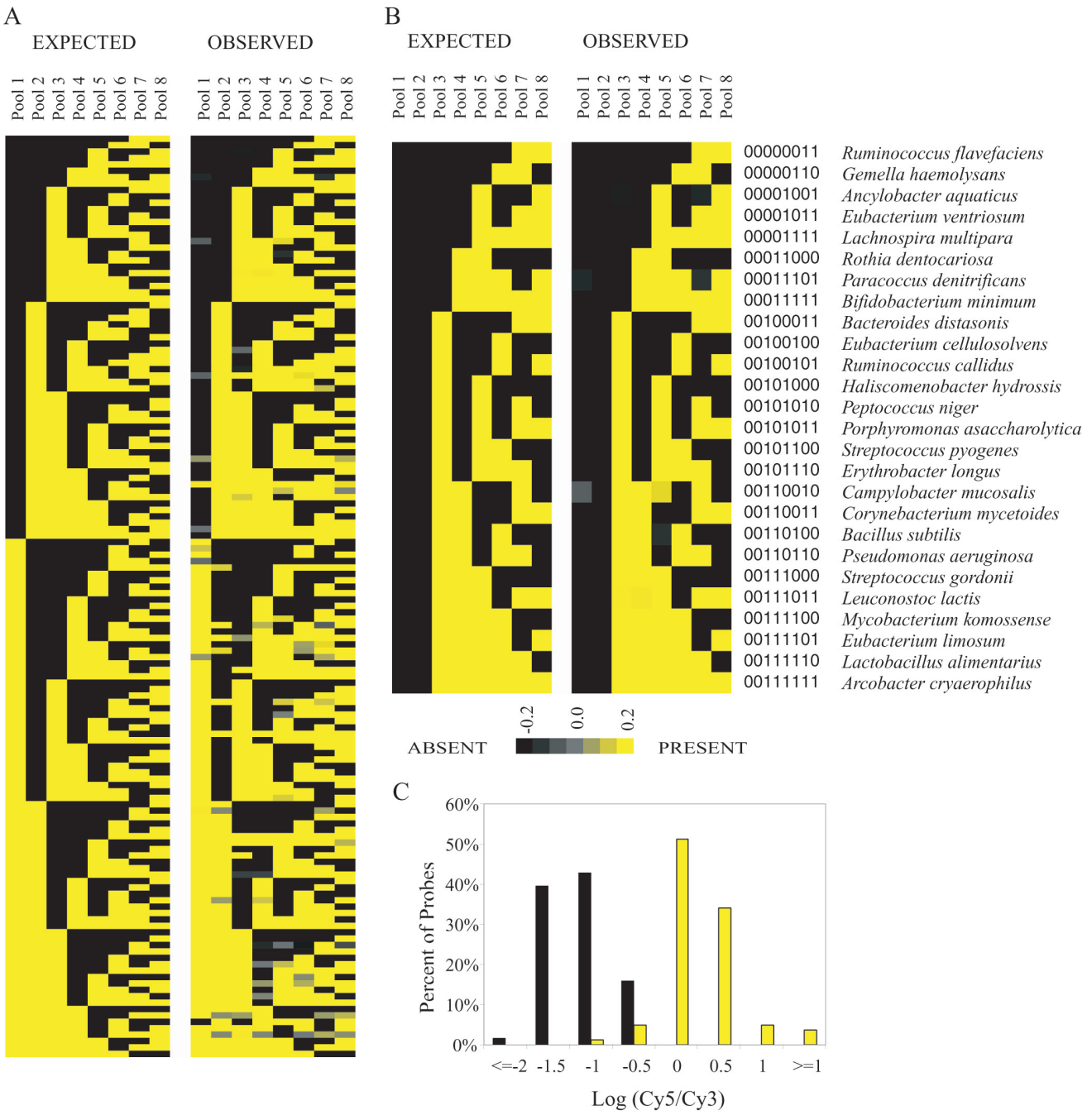

Figure 3. Identification of species in complex mixtures ('Binary Pools') (**A**) Overview of observed versus expected results for all 145 species with 1 or more species probes. Expected values are given as present (yellow = 1) or absent (black = −1). Observed values are deviations in the log (Cy5/Cy3) ratios from 0.7 such that log ratios >0.7 appear yellow and those <0.7 appear black. When multiple species-specific probes were present for a single species, we averaged the log (Cy5/Cy3) ratios for all available probes (2–5 probes per species). (**B**) Expanded view of observed versus expected data for an arbitrary subset of species. (**C**) Distribution of log (Cy5/Cy3) for species probes according to presence or absence of the cognate target (Pool 1). Black = Absent; Yellow = Present.

common reference mixture (Cy3-labeled), and interpreting the Cy5/Cy3 fluorescence ratios for each probe. This comparative hybridization approach provided a probe-by-probe correction for variation in inherent hybridization efficiency (Figure 2).

Detection of species in complex mixtures

In order to test the ability of species-specific probes to detect their cognate sequence at fractional concentrations of <1%, we constructed complex pools of SSU rDNA amplicons. Each of eight such pools contained an equimolar mix of a subset of 115–128 SSU rDNA amplicons drawn from a common set of 229 bacterial species. Each 'binary pool' was labeled with Cy5 and co-hybridized with a Cy3-labeled common reference pool consisting of an equimolar mix of SSU rDNA amplicons from all 229 bacterial species. This series of experiments demonstrated that, with few exceptions, the hybridization signal of a species-specific probe

correlated strongly with the presence/absence of its cognate species, even in the presence of an excess (>100) of diverse species (median $r = 0.99$) (Figure 3).

Quantification of low abundance species

We determined the limits of detection and quantification of species-specific probes by pooling SSU rDNA amplicons from diverse species in defined proportions, ranging over five orders of magnitude. We constructed six 'dilution pools', each containing different proportions of a common set of 190 bacterial species. Each pool had 31–32 bacterial species at each of six levels of abundance: 3% (~4 ng), 0.3%, 0.03%, 0.003%, 0.0003% and 0%. We co-hybridized each Cy5-labeled rRNA pool with an equal amount of the Cy3-labeled common reference rDNA pool, consisting of an equimolar mix of amplified SSU rDNA sequences from the same 192 species. We found that hybridization signals [log (Cy5/Cy3) ratios] were distinguishable from background for probes whose cognate species were present at relative abundances of 0.03% or greater (Figure 4; *t*-test, $P < 10^{-6}$). Furthermore, the fluorescence ratios measured for each probe correlated strongly with relative abundance of its cognate species across samples [median $r = 0.97$ between observed and expected log (Cy5/Cy3) ratios, using log (0.003%/0.5%) as the expected value for relative abundance $\leqslant$0.003%].

Profiling microflora of the human colon: comparison with SSU rDNA sequencing

We tested the performance of our microarray in profiling complex, natural microbial communities, of the type for which it was intended, using cecum and transverse colon mucosa biopsies from three healthy individuals. We analysed each of the six samples using our SSU oligonucleotide microarray and compared the results with those obtained by SSU rDNA sequencing (34). This comparison allowed us to assess the performance of our DNA microarray under 'field conditions' and directly compare it with the sequencing method. SSU rDNA was amplified from independent aliquots of these samples either in triplicate (patient 2) or in duplicate (patients 1 and 3). One set of amplifications was analysed by cloning and sequencing (461–641 sequences per sample), and the other PCR(s) were analysed using our microarray. We found that the microbial profiles of the cecum and transverse colon samples from the same individual were as similar to each other ($r = 0.98$–1.00) as were replicate PCR amplifications from the same sample ($r = 0.98$–1.00) as measured by pairwise correlations of Cy5/Cy3 ratios (see Supplementary Figure S4 for graphical comparison of samples).

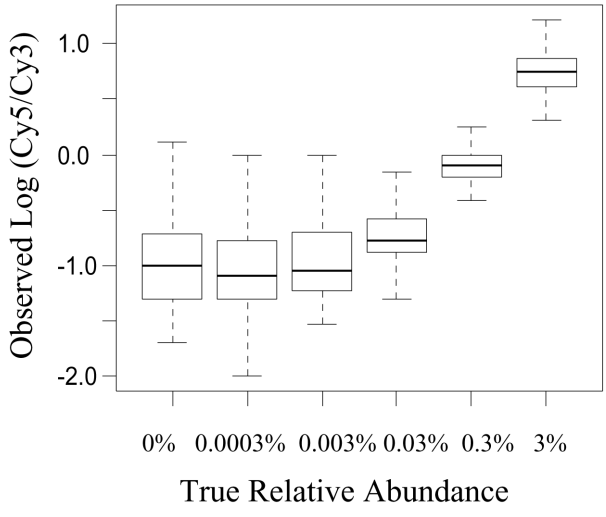

Figure 4. Quantification of species in complex mixtures ('Dilution Pools'). True relative abundance is compared with observed log (Cy5/Cy3) ratio for the 101 species with 1 or more species-specific probes. The *x*-axis corresponds to the six different relative abundance levels used in this experiment. The *y*-axis shows the distribution of observed log (Cy5/Cy3) ratios for species at the specified relative abundance level. When more than one probe was available for a given species, we averaged the log (Cy5/Cy3) ratios across all available probes (2–5 probes per species). Box-whisker plot format: box spans the 25% quantile to the 75% quantile surrounding the median; 'whiskers,' extend to span the full dataset excluding outliers; outliers are defined as points beyond 3/2 the interquantile range from the edge of the box.

The quantitative population profiles obtained with these two techniques were very similar at each taxonomic level. Figure 5 illustrates the similarities and differences between samples and between methods in estimates of relative abundances for all taxonomic groups measured by the microarray (see Supplementary Data S8 for raw data). Both approaches identified members of the genera *Bacteroides, Clostridium, Eubacterium, Ruminococcus* and *Faecalibacterium* as the major constituents of the colonic flora. Furthermore, both methods suggested that the mucosa-associated microbial flora patterns are similar between distinct anatomic sites in the colon within each individual, but differ between individuals. There were, however, several discrepancies between the results obtained with the two techniques. While the microarray data never had a difference of >2-fold in estimates of relative abundance between paired samples from different anatomical sites in the same individual, there were several examples of species detected by sequencing in only one of two such sample pairs. We hypothesized that the detection of a sequence in only one of the two paired samples in the sequence profiles, but in both of the microarray samples, reflected incomplete coverage of the clone library. We tested this hypothesis in three cases (*Eubacterium ventriosum*, *Bacteroides thetaiotaomicron* and genus *Streptococcus*—data not shown), and in each case we were able to confirm that the species or genus detected by the microarray and missed by sequencing was indeed present in the sample in question. Since the microarray data in Figure 5 provides only lower-bound estimates for each more inclusive taxonomic group (see Microarray data analysis methods), additional artifactual differences can occur when sequencing detected members of a taxonomic group that was only partially covered by the microarray probes.

By using a simple model based on BLAST to predict hybridization results from rDNA sequence data, we were able to compare the observed microarray hybridization signals with the signals predicted from the sequencing data, in a quantitative way, on a probe-by-probe basis. Briefly, for each of the six colon biopsies, we used BLAST to simulate the hybridization of the corresponding set of rDNA clone sequences to each probe on the microarray. This comparison revealed a strong correlation between microarray-based and sequencing-based estimates of relative abundance, as shown in Figure 6 for one of the six samples (mean $r = 0.88$) (see Supplementary Figure S5 for all six samples).

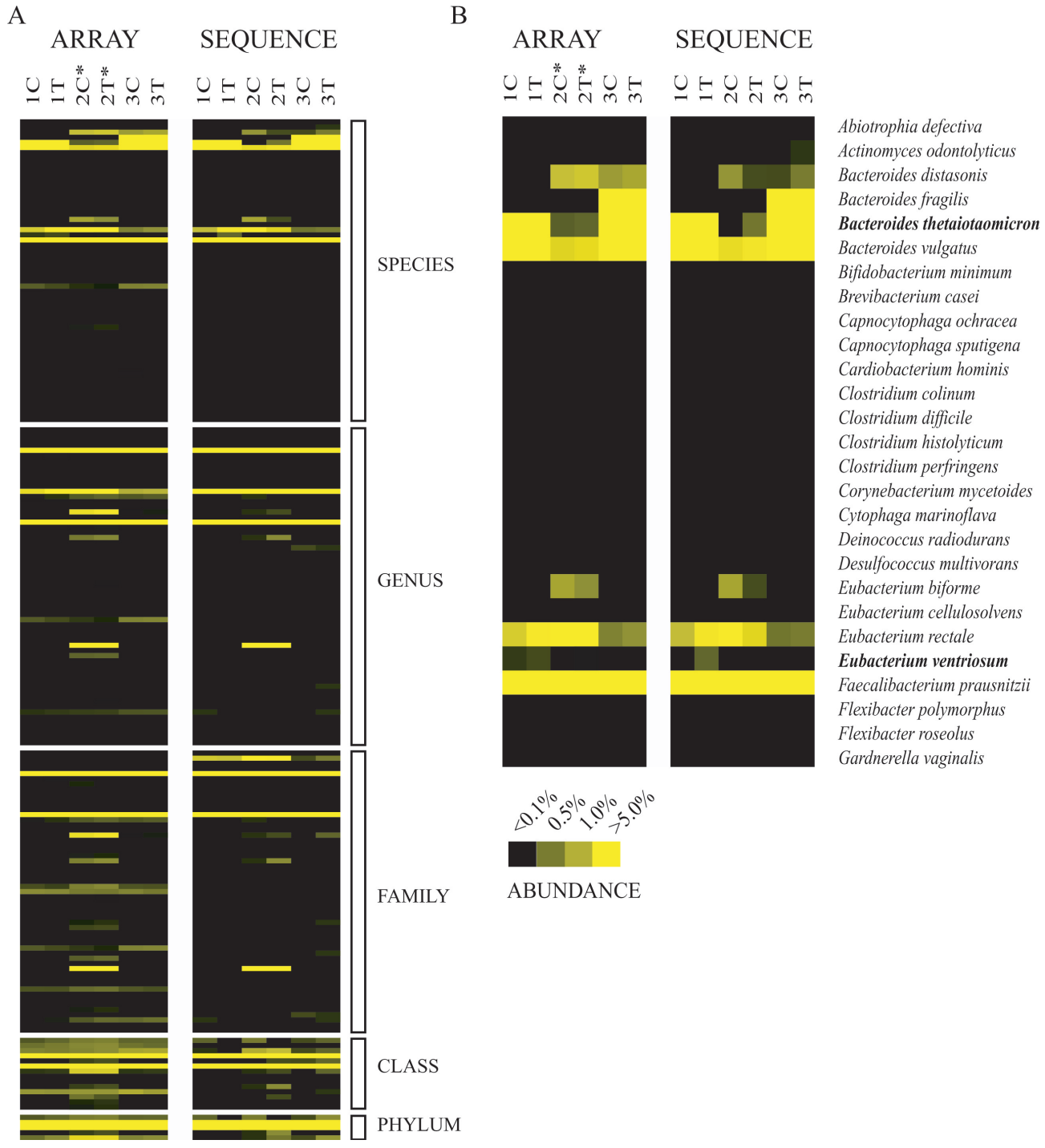

Figure 5. Comparison of taxonomic profiles from microarray and sequencing data. Each column represents one patient sample; each row represents a taxonomic group. Samples are labeled by subject (subjects 1–3) and by anatomical site (C: cecum; T: transverse colon). Both microarray data (Cy5/Cy3 ratios) and sequencing data (clone counts) have been converted to a fractional abundance scale. (**A**) Relative abundance estimates are shown for all taxonomic groups represented by at least 1 probe with well-measured reference signal. (**B**) Expanded view of an arbitrary subset of species-probe data. Bold font indicates species whose presence was confirmed by PCR. Asterisk indicates averaged microarray values from two replicate PCRs.

DISCUSSION

In the last several years, rDNA microarrays are emerging as a sensitive and efficient way to screen samples systematically for bacterial species of interest. We set out to extend and optimize the microarray approach, with the goal of designing a microarray and experimental protocol that could provide a robust, reliable quantitative census of diverse and complex microbial populations in a wide range of microenvironments. We did this by identifying a set of rDNA sequence probes that was able to represent specifically a large and diverse number of taxonomic groups, ranging in scope from species to phylum

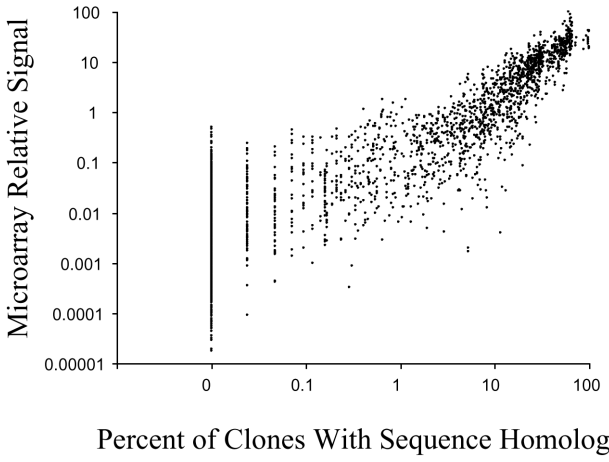

Figure 6. Quantitative comparison of microarray and sequencing results. Probe-by-probe comparison of sequence-based and microarray-based abundance estimates for cecum biopsy from subject 1 (4605 unique probes). Relative abundance estimates for sequence data (*x*-axis) are weighted sums of the number of clone sequences that matched the probe as determined by BLAST. Relative abundance estimates for microarray data (*y*-axis) were derived from Cy5/Cy3 fluorescence ratios and the known composition of the Cy3-labeled common reference pool as described in Materials and Methods.

level, and by concurrently developing a molecular protocol and analysis approach that enabled us to obtain quantitative measurements at multiple taxonomic levels.

Hybridization experiments directed at discriminating and quantifying rDNAs from different taxa are more complex than experiments directed at analysing mRNAs from the diverse genes in a genome. Two major complicating factors are the presence of phylogenetically conserved sequences flanking most phylogenetically specific sequences, and the propensity of the rRNA molecule for self-structure. We were able to overcome these obstacles by performing two-color hybridizations using labeled ssrRNA [as in (19,20)]. Using this approach, we were able to detect and quantify individual rRNA species in complex mixtures of >100 strains, at relative abundances of <0.1%.

This microarray design performed very well in our first test of 'real world' biological samples. We were able to characterize the bacterial composition of six colonic mucosal endoscopic biopsies, both broadly and specifically, and to systematically compare the samples with each other. SSU rDNA sequence analysis of the same samples gave qualitatively and quantitatively similar results (Figures 5 and 6). Both methods were consistent with previous studies in terms of the dominant species and taxa (37,38), and the greater extent of inter-individual differences as compared with differences between anatomic sites in the same individual (39).

SSU rDNA clone library sequencing and microarray analysis of rDNAs each offers distinct advantages for profiling microbial populations: the microarray approach is substantially more rapid (several days versus weeks to months) and reproducible, and can detect bacterial species missed by sequencing of >600 clones. The main limitations of the microarray approach are that (i) it can only measure species and taxonomic groups for which probes were both successfully designed and printed and (ii) it cannot directly discover novel species. We can minimize these limitations by including many higher-level taxonomic probes, which ensure that any species, novel or known, will hybridize to the array. Indeed, we have already designed a more comprehensive 'next-generation' microbial SSU rDNA microarray, which aims to represent most known bacterial species at multiple taxonomic levels. Still, sequencing of rDNA populations remains invaluable for its ability to discover new rDNA species and thereby infer new microbial species.

The microarray-based method described here is also subject to several biases that are inherent in the use of amplified rDNA sequences for identification and quantification of bacterial species. These include biases introduced by each of the steps in the preparation of labeled sample: DNA extraction, PCR amplification and *in vitro* transcription, [reviewed in (40)], as well as by interspecies variation in rRNA gene copy number (41). We have tried to minimize these biases by using rigorous lysis methods, highly conserved PCR primers and a minimal number (20) of PCR cycles. Several studies using rDNA microarrays have avoided amplification biases by labeling and hybridizing rRNA directly isolated from microbial samples, but these studies have not included thorough testing of this method with complex communities (19–21,26). Our future studies will explore this approach as a complement to amplified rDNA-based community profiling. An additional caveat that this method shares with other 16S rDNA-based censusing methods is that different strains of the same species may have the same 16S rDNA sequences yet differ at other significant loci; such microheterogeneity will not be revealed by methods that rely exclusively on 16S rDNA as a taxonomic identifier.

Comprehensive identification and quantitative profiling of the members of microbial communities is a challenging problem, and important not only for understanding the critical roles microbes play in shaping our environment, but also in defining their rich symbiotic relationships with our own bodies. Microbial flora has been found to vary in composition between hosts (12) and over time (42), while individuals have been shown to vary in their responses to diverse microbial stimuli (43). Moreover, alterations in the intestinal flora have been associated with diverse disorders ranging from autism (44) to ankylosing spondylitis (45) and inflammatory bowel disease (46,47). Using the microarray approach described here, we can now begin large-scale systematic, quantitative, comparative studies of bacterial populations and their relationships with their human hosts and other environments, which are likely to reveal new and unexpected principles of human biology and microbial ecology.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank Rey Cypress for donation of ATCC strains, Steve R. Gill and Karen E. Nelson at TIGR for sequencing, Charles N. Bernstein for colon biopsy samples, as well as Stephen Popper and Jerel Davis for helpful discussions. This work was supported by the Stanford Genome Training Grant, the Horn Foundation, NIH grant AI051259 (D.A.R.),

Ellison Medical Foundation Senior Scholar Award [ID-SS-0103 (D.A.R.)] and the Howard Hughes Medical Institute (P.O.B.). P.O.B. is an investigator of the Howard Hughes Medical Institute. Funding to pay the Open Access publication charges for this article was provided by the Howard Hughes Medical Institute.

Conflict of interest statement. Paul K. Wolber works for and holds stock in Agilent Technologies who provided the microarrays that were used in the study.

REFERENCES

1. Whitman,W.B., Coleman,D.C. and Wiebe,W.J. (1998) Prokaryotes: the unseen majority. *Proc. Natl Acad. Sci. USA*, **95**, 6578–6583.
2. Savage,D.C. (1977) Microbial ecology of the gastrointestinal tract. *Annu. Rev. Microbiol.*, **31**, 107–133.
3. Rakoff-Nahoum,S., Paglino,J., Eslami-Varzaneh,F., Edberg,S. and Medzhitov,R. (2004) Recognition of commensal microflora by toll-like receptors is required for intestinal homeostasis. *Cell*, **118**, 229–241.
4. Hooper,L.V., Wong,M.H., Thelin,A., Hansson,L., Falk,P.G. and Gordon,J.I. (2001) Molecular analysis of commensal host-microbial relationships in the intestine. *Science*, **291**, 881–884.
5. Backhed,F., Ding,H., Wang,T., Hooper,L.V., Koh,G.Y., Nagy,A., Semenkovich,C.F. and Gordon,J.I. (2004) The gut microbiota as an environmental factor that regulates fat storage. *Proc. Natl Acad. Sci. USA*, **101**, 15718–15723.
6. Woese,C.R. (1987) Bacterial evolution. *Microbiol. Rev.*, **51**, 221–271.
7. Olsen,G.J., Lane,D.J., Giovannoni,S.J., Pace,N.R. and Stahl,D.A. (1986) Microbial ecology and evolution: a ribosomal RNA approach. *Annu. Rev. Microbiol.*, **40**, 337–365.
8. Schmidt,T.M., DeLong,E.F. and Pace,N.R. (1991) Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. *J. Bacteriol.*, **173**, 4371–4378.
9. Giovannoni,S.J., Britschgi,T.B., Moyer,C.L. and Field,K.G. (1990) Genetic diversity in Sargasso Sea bacterioplankton. *Nature*, **345**, 60–63.
10. Felske,A., Wolterink,A., Van Lis,R. and Akkermans,A.D. (1998) Phylogeny of the main bacterial 16S rRNA sequences in Drentse A grassland soils (The Netherlands). *Appl. Environ. Microbiol.*, **64**, 871–879.
11. Muyzer,G., de Waal,E.C. and Uitterlinden,A.G. (1993) Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Appl. Environ. Microbiol.*, **59**, 695–700.
12. Zoetendal,E.G., Akkermans,A.D. and De Vos,W.M. (1998) Temperature gradient gel electrophoresis analysis of 16S rRNA from human fecal samples reveals stable and host-specific communities of active bacteria. *Appl. Environ. Microbiol.*, **64**, 3854–3859.
13. Liu,W.T., Marsh,T.L., Cheng,H. and Forney,L.J. (1997) Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. *Appl. Environ. Microbiol.*, **63**, 4516–4522.
14. DeLong,E.F., Wickham,G.S. and Pace,N.R. (1989) Phylogenetic stains: ribosomal RNA-based probes for the identification of single cells. *Science*, **243**, 1360–1363.
15. Stahl,D.A., Flesher,B., Mansfield,H.R. and Montgomery,L. (1988) Use of phylogenetically based hybridization probes for studies of ruminal microbial ecology. *Appl. Environ. Microbiol.*, **54**, 1079–1084.
16. Tajima,K., Aminov,R.I., Nagamine,T., Matsui,H., Nakamura,M. and Benno,Y. (2001) Diet-dependent shifts in the bacterial population of the rumen revealed with real-time PCR. *Appl. Environ. Microbiol.*, **67**, 2766–2774.
17. Castiglioni,B., Rizzi,E., Frosini,A., Sivonen,K., Rajaniemi,P., Rantala,A., Mugnai,M.A., Ventura,S., Wilmotte,A., Boutte,C. *et al.* (2004) Development of a universal microarray based on the ligation detection reaction and 16S rRNA gene polymorphism to target diversity of *Cyanobacteria*. *Appl. Environ. Microbiol.*, **70**, 7161–7172.
18. Loy,A., Lehner,A., Lee,N., Adamczyk,J., Meier,H., Ernst,J., Schleifer,K.H. and Wagner,M. (2002) Oligonucleotide microarray for 16S rRNA gene-based detection of all recognized lineages of sulfate-reducing prokaryotes in the environment. *Appl. Environ. Microbiol.*, **68**, 5064–5081.
19. Koizumi,Y., Kelly,J.J., Nakagawa,T., Urakawa,H., El-Fantroussi,S., Al-Muzaini,S., Fukui,M., Urushigawa,Y. and Stahl,D.A. (2002) Parallel characterization of anaerobic toluene- and ethylbenzene-degrading microbial consortia by PCR-denaturing gradient gel electrophoresis, RNA-DNA membrane hybridization, and DNA microarray technology. *Appl. Environ. Microbiol.*, **68**, 3215–3225.
20. Guschin,D., Mobarry,B., Proudnikov,D., Stahl,D., Rittmann,B. and Mirzabekov,A. (1997) Oligonucleotide microchips as genosensors for determinative and environmental studies in microbiology. *Appl. Environ. Microbiol.*, **63**, 2397–2402.
21. Small,J., Call,D.R., Brockman,F.J., Straub,T.M. and Chandler,D.P. (2001) Direct detection of 16S rRNA in soil extracts by using oligonucleotide microarrays. *Appl. Environ. Microbiol.*, **67**, 4708–4716.
22. Liu,W.T., Mirzabekov,A.D. and Stahl,D.A. (2001) Optimization of an oligonucleotide microchip for microbial identification studies: a non-equilibrium dissociation approach. *Environ. Microbiol.*, **3**, 619–29.
23. Wang,R.F., Beggs,M.L., Erickson,B.D. and Cerniglia,C.E. (2004) DNA microarray analysis of predominant human intestinal bacteria in fecal samples. *Mol. Cell Probes*, **18**, 223–234.
24. Wang,R.F., Beggs,M.L., Robertson,L.H. and Cerniglia,C.E. (2002) Design and evaluation of oligonucleotide-microarray method for the detection of human intestinal bacteria in fecal samples. *FEMS Microbiol. Lett.*, **213**, 175–182.
25. Busti,E., Bordoni,R., Castiglioni,B., Monciardini,P., Sosio,M., Donadio,S., Consolandi,C., Rossi Bernardi,L., Battaglia,C. and De Bellis,G. (2002) Bacterial discrimination by means of a universal array approach mediated by LDR (ligase detection reaction). *BMC Microbiol.*, **2**, 27.
26. El Fantroussi,S., Urakawa,H., Bernhard,A.E., Kelly,J.J., Noble,P.A., Smidt,H., Yershov,G.M. and Stahl,D.A. (2003) Direct profiling of environmental microbial populations by thermal dissociation analysis of native rRNAs hybridized to oligonucleotide microarrays. *Appl. Environ. Microbiol.*, **69**, 2377–2382.
27. Wilson,K.H., Wilson,W.J., Radosevich,J.L., DeSantis,T.Z., Viswanathan,V.S., Kuczmarski,T.A. and Andersen,G.L. (2002) High-density microarray of small-subunit ribosomal DNA probes. *Appl. Environ. Microbiol.*, **68**, 2535–2541.
28. DeSantis,T.Z., Stone,C.E., Murray,S.R., Moberg,J.P. and Andersen,G.L. (2005) Rapid quantification and taxonomic classification of environmental DNA from both prokaryotic and eukaryotic origins using a microarray. *FEMS Microbiol. Lett.*, **245**, 271–278.
29. Hugenholtz,P. (2002) Exploring prokaryotic diversity in the genomic era. *Genome Biol.*, **3**, reviews0003.1–reviews0003.8.
30. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–10.
31. Blanchard,A.P., Kaiser,R.J. and Hood,L.E. (1996) High-density oligonucleotide arrays. *Biosens. Bioelectron.*, **11**, 687–690.
32. Edwards,U., Rogall,T., Blocker,H., Emde,M. and Bottger,E.C. (1989) Isolation and direct complete nucleotide determination of entire genes. Characterization of a gene coding for 16S ribosomal RNA. *Nucleic Acids Res.*, **17**, 7843–7853.
33. Lane,D.J., Pace,B., Olsen,G.J., Stahl,D.A., Sogin,M.L. and Pace,N.R. (1985) Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc. Natl Acad. Sci. USA*, **82**, 6955–6959.
34. Eckburg,P.B., Bik,E.M., Bernstein,C.N., Purdom,E., Dethlefsen,L., Sargent,M., Gill,S.R., Nelson,K.E. and Relman,D.A. (2005) Diversity of the human intestinal microbial flora. *Science*, **308**, 1635–1638.
35. Fare,T.L., Coffey,E.M., Dai,H., He,Y.D., Kessler,D.A., Kilian,K.A., Koch,J.E., LeProust,E., Marton,M.J., Meyer,M.R. *et al.* (2003) Effects of atmospheric ozone on microarray data quality. *Anal. Chem.*, **75**, 4672–4675.
36. Hofacker,I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429–3431.
37. Hold,G.L., Pryde,S.E., Russell,V.J., Furrie,E. and Flint,H.J. (2002) Assessment of microbial diversity in human colonic samples by 16S rDNA sequence analysis. *FEMS Microbiol. Ecol.*, **39**, 33–39.
38. Wang,X., Heazlewood,S.P., Krause,D.O. and Florin,T.H. (2003) Molecular characterization of the microbial species that colonize human

ileal and colonic mucosa by using 16S rDNA sequence analysis. *J. Appl. Microbiol.*, **95**, 508–520.
39. Zoetendal,E.G., von Wright,A., Vilpponen-Salmela,T., Ben-Amor,K., Akkermans,A.D.L. and de Vos,W.M. (2002) Mucosa-associated bacteria in the human gastrointestinal tract are uniformly distributed along the colon and differ from the community recovered from feces. *Appl. Environ. Microbiol.*, **68**, 3401–3407.
40. von Wintzingerode,F., Gobel,U.B. and Stackebrandt,E. (1997) Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. *FEMS Microbiol. Rev.*, **21**, 213–229.
41. Klappenbach,J.A., Saxman,P.R., Cole,J.R. and Schmidt,T.M. (2001) rrndb: the Ribosomal RNA Operon Copy Number Database. *Nucleic Acids Res.*, **29**, 181–184.
42. Favier,C.F., Vaughan,E.E., De Vos,W.M. and Akkermans,A.D. (2002) Molecular monitoring of succession of bacterial communities in human neonates. *Appl. Environ. Microbiol.*, **68**, 219–226.
43. Boldrick,J.C., Alizadeh,A.A., Diehn,M., Dudoit,S., Liu,C.L., Belcher,C.E., Botstein,D., Staudt,L.M., Brown,P.O. and Relman,D.A. (2002) Stereotyped and specific gene expression programs in human innate immune responses to bacteria. *Proc. Natl Acad. Sci. USA*, **99**, 972–977.
44. Finegold,S.M., Molitoris,D., Song,Y., Liu,C., Vaisanen,M.L., Bolte,E., McTeague,M., Sandler,R., Wexler,H., Marlowe,E.M. *et al.* (2002) Gastrointestinal microflora studies in late-onset autism. *Clin. Infect. Dis.*, **35**, S6–S16.
45. Stebbings,S., Munro,K., Simon,M.A., Tannock,G., Highton,J., Harmsen,H., Welling,G., Seksik,P., Dore,J., Grame,G. *et al.* (2002) Comparison of the faecal microflora of patients with ankylosing spondylitis and controls using molecular methods of analysis. *Rheumatology*, **41**, 1395–13401.
46. Ott,S.J., Musfeldt,M., Wenderoth,D.F., Hampe,J., Brant,O., Folsch,U.R., Timmis,K.N. and Schreiber,S. (2004) Reduction in diversity of the colonic mucosa associated bacterial microflora in patients with active inflammatory bowel disease. *Gut*, **53**, 685–693.
47. Seksik,P., Rigottier-Gois,L., Gramet,G., Sutren,M., Pochart,P., Marteau,P., Jian,R. and Dore,J. (2003) Alterations of the dominant faecal bacterial groups in patients with Crohn's disease of the colon. *Gut*, **52**, 237–242.