

Chapter 11. Community Analysis-Based Methods

Yiping Cao¹, Cindy H Wu², Gary L Andersen², Patricia A Holden³

¹ Southern California Coastal Water Research Project, 3535 Harbor Blvd, Suite 110, Costa Mesa, CA 92626; Email: yipingc@sccwrp.org, Tel: 714-755-3241

² Lawrence Berkeley National Laboratory, One Cyclotron Road, Mail Stop 70A-3317, Berkeley, CA 94720; Email: CHWu@lbl.gov, GLAndersen@lbl.gov

³ Donald Bren School of Environmental Science & Management, University of California, Santa Barbara, CA 93106; Email: holden@bren.ucsb.edu

Abstract

Microbial communities are each a composite of populations whose presence and relative abundance in water or other environmental samples are a direct manifestation of environmental conditions, including the introduction of microbe-rich fecal material and factors promoting persistence of the microbes therein. As shown by culture-independent methods, different animal-host fecal microbial communities appear distinctive, suggesting that their community profiles can be used to differentiate fecal samples and to potentially reveal the presence of host fecal material in environmental waters. Cross-comparisons of microbial communities from different hosts also reveal relative abundances of genetic groups that can be used to distinguish sources. In increasing order of their information richness, several community analysis methods hold promise for MST applications: phospholipid fatty acid (PLFA) analysis, denaturing gradient gel electrophoresis (DGGE), terminal restriction fragment length polymorphism (TRFLP), cloning/sequencing, and PhyloChip. Specific case studies involving TRFLP and PhyloChip approaches demonstrate the ability of community-based analyses of contaminated waters to confirm a diagnosis of water quality based on host-specific marker(s). The success of community-based MST for comprehensively confirming fecal sources relies extensively upon using appropriate multivariate statistical approaches. While community-based MST is still under evaluation and development as a primary diagnostic tool, results presented herein demonstrate its promise. Coupled with its inherently comprehensive ability to capture an unprecedented amount of microbiological data that is relevant to water quality, the tools for microbial community analysis are increasingly accessible, and community-based approaches have unparalleled potential for translation into rapid, perhaps real-time, monitoring platforms.

11.1 Introduction

11.1.1 Challenges in water quality diagnosis

Microbiological water quality is a serious public health concern for drinking water, recreational swimming, shellfish consumption, and agricultural food production. The microbial pollutants of concern are pathogens, discharged with various fecal sources including human sewage and septage, domestic pet waste, and livestock manure. Pathogen-related coastal water quality problems are recognized in a National Research Council (NRC) report (1993) where “pathogens and toxins that affect human health” and the “introduction of nonindigenous species” are listed among major U.S. coastal environmental issues.

Historically, routine monitoring of environmental waters for all known bacterial and viral pathogens has been impractical; monitoring for all possible pathogens also has been impossible given that most of the human gut microbiome, from which fecal material originates, has not been identified (Eckburg et al. 2010). Instead, fecal indicator bacteria (FIB, Fig 11.1) are used to infer the presence of fecal contamination (Eaton et al. 1998), the presence of pathogens, and the risk of disease (Fig 11.2). While FIB, and thus the “chain of inference” for human health risks (Fig 11.2), are accepted for safely monitoring drinking water, the scientific and regulatory communities increasingly regard FIB as flawed for monitoring fecal material in the environment. The historical basis and advantages for using FIB in environmental water quality monitoring has been reviewed elsewhere (Wade et al. 2006). Early epidemiological studies showed a direct connection between swimmer illness and FIB (Fig 11.2), but such studies were performed in

coastal waters continuously receiving sewage (Cabelli et al. 1982). Now, most wastewaters in the U.S. are treated to either primary or secondary standards (National Research Council 1993), and, except for acute short-term sewage spills, sources of human waste are diffuse, e.g. from leaking sewer pipes, cross-connections to storm drains, or failed septic systems. The disadvantages in relying on FIB for diffuse, or non-point, source monitoring arise primarily from the fact that FIB are nonspecific as to source, e.g. to human waste which is most relevant as a carrier for human pathogens. While total maximum daily load (TMDL) compliance is a major driver for microbial source tracking (MST; (U.S. Environmental Protection Agency 2005)), another is the non-specificity of FIB to human or other wastes. Other disadvantages of FIB, including the method's reliance on cultivation when FIB can become nonculturable in the environment (Leadbetter 1997), partly motivate the subject of this chapter: i.e. culture-independent microbial community analysis as a MST approach. Lastly, and related to the non-specificity of FIB to source, some FIB can become indigenous, e.g. to plant surfaces (Hazen 1988) and to beach sands (Ishii et al. 2007), which results in false positive indications of an increased risk of pathogen presence. Each of these disadvantages contributes to the rate of error associated with FIB monitoring for predicting pathogen presence and protecting public health. Thus, while one MST objective may be discovering the sources of FIB, which is particularly pertinent to TMDL applications (U.S. Environmental Protection Agency 2005), another view, particularly in light of the fundamental problems with interpreting FIB environmental data, is that MST should be mainly motivated by discovering 1) if fecal material is present, 2) the host source of fecal material, and 3) the geographical origin of fecal material.

11.1.2 The MST toolkit, including microbial community analysis

Because of the ubiquity of routine FIB monitoring, MST is often employed as a follow-up to determine if FIB signal the presence of fecal material. Frequently used in such follow-up studies, a stepwise MST study approach has been successful in some circumstances (Vogel et al. 2007), including a study revealing a leaking sanitary sewer line as an acute source of FIB near a beach (Boehm et al. 2003), and another discovering that aged sanitary sewers contribute diffuse contamination to storm drains discharging directly to coastal waters (Sercu et al. 2009). In such cases, study steps included: analyzing historical FIB time-course data in a spatial context, performing existing data and field system reconnaissance, nominating hypothetical sources based on sewer infrastructure and nearby features, designing a field sampling and sample analysis program to test the hypotheses, and employing one or more source-specific markers in sample analysis. This has been previously referred to as a “tiered approach” (Boehm et al. 2003), but the differences in decay rates between abundances of FIB and source-specific markers (Dick et al. 2010) means that FIB data may not provide a solid foundation upon which a tiered source diagnosis can be built. FIB data are useful for generally circumscribing the spatial emphases of MST, but not for reliably tracking sources of waste.

MST has been advanced significantly by the discovery of microbial host-specific waste markers, most notably DNA sequence encoding the 16S ribosomal RNA (rRNA) of a *Bacteroidetes* specific to human feces, as discovered by Bernhard and Field (Bernhard and Field 2000b). The advantages of testing for this DNA sequence include its strong association with human waste and its culture-independence, coupled with quantitative polymerase chain reaction (qPCR) protocols (Seurinck et al. 2005) that enable quantifying waste markers relative to possible sources such as

sewage (Seurinck et al. 2005, Sercu et al. 2009). Host specific markers are reviewed more comprehensively elsewhere in this book (Chapter 3, 4, 5, 6 and 7), and are continuing to be an important, emerging component of the MST toolkit. There is still much work to be done in order to understand the true specificity of host-specific waste markers, the behavior of individual DNA-based waste markers under environmental conditions including their rates of decay relative to pathogens (Field and Samadpour 2007), and the potential for waste markers to become indigenous to the environment (a problem also noted for FIB). For these reasons, individual markers will generally not be used alone but will instead be employed within a suite of MST tools. Further, while a combination of bacterial host-specific markers (Chapter 3 and 4), coupled with chemical markers (Chapter 8), host-related protozoa markers (Chapter 7), and assays for viral pathogens or host-specific bacteriophages indicators (Chapter 5 and 6), can comprise an effective suite of MST tools, a universally-applicable and optimal tool combination has yet to be defined (Chapter 9). Meanwhile, the orientation of MST towards culture-independent waste markers suggests a potentially powerful use of DNA extracted from environmental samples: analysis of entire microbial communities which inevitably include pathogens, FIB and other marker organisms (Fig 11.3). This type of analysis could trace various environmental conditions, including biotic and abiotic perturbations.

Here, we use community analysis to refer to culture-independent characterization of the microbial community in a sample, i.e., what microbes, in what abundance, comprise the community in a sample. Conceptually, fecal input can alter microbial communities in the receiving environmental waters either directly or indirectly, i.e., by fecal source communities acting as inoculants, or by altering the environmental conditions of receiving water such as through changing water chemistry. Community analysis in MST has strong potential to be

useful for waste source assessment for the following reasons. First, gut microbial communities of various hosts vary significantly by host species and diet (Ley et al. 2008), therefore microbial communities in feces would differ by host animal. Second, microbial communities are sensitive to perturbations and respond rapidly to environmental changes (Hwang et al. 2009), suggesting that their composition may reflect recent or ongoing contamination events. Third, the culture-independent analysis of microbial communities, by now a mature research endeavor (Liu and Jansson 2010), has successfully captured changes of microbial communities in the environment along gradients such as carbon availability (LaMontagne, Schimel and Holden 2003) and proximity to hydrocarbon seeps (LaMontagne et al. 2004), and changes in microbial community due to perturbations such as inundation (Córdova-Kreylos et al. 2006), and vegetation and pollutant variations (Cao et al. 2006, Cao, Green and Holden 2008) in estuaries. Furthermore, community analysis detects many microbial signals and pathogens simultaneously to provide what is potentially a more robust and relevant approach to MST when compared to single indicators (Fig 11.3).

This Chapter examines the use of culture-independent microbial community analysis in MST. Briefly, this Chapter reviews technical methods and data analysis approaches, provides two case studies, a critical evaluation of advantages and needs, and a view of community analysis in the future of MST.

11.2 Community analysis methods

11.2.1 Description of community analysis techniques

A variety of holistic microbial community analysis techniques have been used in microbial ecology (for reviews, see (Osborn and Smith 2005, Nocker, Burr and Camper 2007)). Culture-independent community analysis techniques differ on the analysis target and target profiling techniques (Fig 11.4). The analysis targets divide techniques into phenotypic and genotypic methods. The most common phenotypic method, phospholipid fatty acid (PLFA) analysis, targets phospholipids contained in the cell membranes of all living cells. Most genotypic methods are DNA-based, where total DNA extracted from an environmental sample is analyzed directly through techniques such as “shotgun cloning”, “shotgun sequencing”, or via DNA probe hybridization. Alternatively, polymerase chain reaction (PCR) is used to amplify specific genetic markers from community DNA and the PCR products are subsequently analyzed by sequencing, hybridization, DNA melting behavior, or length polymorphism (Nocker et al. 2007).

Here we describe methods that are most widely used. In particular, we focus on techniques that have the greatest potential to be used in MST studies. For each method, brief descriptions are provided regarding the method background, sample processing, and basic method-specific data processing. Multivariate data analysis is discussed in section 11.3.

11.2.1.1 Phospholipid fatty acid (PLFA) analysis

Phospholipids are essential membrane components that control cell permeability. Microbial fatty acids are largely linked to phospholipids and, in most cases, specific types of fatty acids predominate in a given taxon and are commonly associated with groups metabolizing similar substrates (Zelles 1999). Microorganisms also change membrane fatty acids composition in

adaptation to environmental conditions including stressors (Loffhagen, Härtig and Babel 2004). Because PLFAs decompose quickly upon cell death, PLFA community analysis is typically regarded as reflecting viable or recently-living cells (White 1994).

In performing PLFA analysis, PLFAs are first recovered from a sample by organic solvent extraction or solid phase extraction (Zelles 1999). The extracted fatty acids and their derivatives are analyzed by gas chromatography to generate a PLFA profile which contains a list of fatty acids and their molar abundances. For data processing, the total mass of all PLFAs is often calculated as an indicator of total biomass, and the total mass of specific groups of PLFAs or mass ratios of certain PLFAs are also calculated as indicators of target groups of microorganisms or of environmental stresses (Zelles 1999). For example, several PLFAs are considered as biomarkers for members of the sulfate-reducer functional group; branched PLFAs are used as biomarkers for Gram-positive bacteria, and certain FAs can indicate heavy metal pollution (Pennanen et al. 1996, Córdova-Kreylos et al. 2006). Depending on downstream data analysis, the PLFA mass (i.e. absolute abundance) is also converted to percentage composition (i.e. relative abundance) so that biomass influences can be down-weighted and sample comparisons will be mainly based on community composition (Cao et al. 2006).

11.2.1.2 Terminal Restriction Fragment Length Polymorphism (TRFLP)

TRFLP produces genotypic fingerprints, or profiles, of the microbial community based on length polymorphism of PCR products of a specific marker gene, most frequently the gene encoding 16S rRNA which contains regions that are conserved and regions that are variable among microorganisms. Additionally, other functional genes such as nitrite reductase genes (*nirS*,

nirK) have also been used as marker genes for community analysis by TRFLP (Braker et al. 2001).

In TRFLP analysis (Liu et al. 1997), DNA is extracted from an environmental sample and genes from total community DNA (usually 16S rRNA gene) are amplified via PCR. One or both of the two (forward and reverse) primers are labeled with a fluorescent dye. The resulting terminally-labeled PCR products (amplicons) are digested with restriction enzymes that recognize specific cutting sites. These sites are located at different positions on the amplicons due to differences in gene sequence among microorganisms. Thus the enzyme digestion generates fluorescently-labeled terminal restriction fragments (TRFs) of various sizes (or lengths in base pairs). The sizes and abundances of the TRFs are determined using an automated DNA sequencer, where each TRF is represented as a peak on the electropherogram. The electropherogram is the graphical representation of the TRFLP profile, and the profile itself is a list of TRFs (a.k.a. peaks) in order of their base pair length and their abundance in relative fluorescent units. Both TRF height and area, for each peak, are provided in the profile dataset. Because the estimated sizes of TRFs from the same phylotype differ slightly due to run-to-run variability, TRFLP profiles are aligned across runs before further data analysis (Dunbar, Ticknor and Kuske 2001). Absolute abundances of TRFs are often converted to relative abundances, expressed as percentages of the total peak height or peak area for all TRFs in a sample. Normalization in this way is necessary to adjust for slight variations in the amount of DNA loaded onto the sequencer. If PCR bias is a concern, TRFLP data can be converted, prior to further analysis, to a simple binary format based on TRF presence or absence (Cao et al. 2006).

11.2.1.3 Denaturing Gradient Gel Electrophoresis (DGGE)

DGGE also generates community fingerprints based on profiling PCR products of a specific marker gene. However, the profiling is achieved based on separating the PCR products in a gel formulated with a chemical denaturant such as urea (Muyzer and Smalla 1998). Specially-designed primers with a GC clamp are used in PCR so that the GC clamp can hold together the two separated strands of amplicon during denaturation. Briefly, as the PCR products are separated by electrophoresis, they denature in the gel due to exposure to the chemical denaturant. Denaturation converts easily-electrophoresed double-stranded DNA into DNA whose migration is sterically-hindered by its denatured content. Because sequence differences cause different melting behaviors, the timing and extent of denaturation of the PCR products differ along the gradient, which in turns determines how fast the various PCR products migrate in the gel. A community profile results from the pattern of separately-migrated bands where each band represents a different organism or group. A similar technique is known as temperature gradient gel electrophoresis (TGGE) where a temperature gradient is used for denaturation.

After electrophoresis, the gel is stained with a nucleic acid-binding dye and photographed, resulting in an image of the banding pattern that is often imported into computer software for band density analysis (Esseili, Kassen and Sigler 2008). Depending on the crispness of band separation, bands can also be excised, PCR-amplified, and sequenced for phylogenetic identification.

11.2.1.4 Cloning and sequencing

A higher degree of phylogenetic resolution for community analysis is achieved by obtaining the actual sequence of a marker gene (or a metagenome, Fig 11.4) through cloning followed by sequencing (Nocker et al. 2007) or direct sequencing (Shendure and Ji 2008). Here, a

metagenome refers to the entire collection of genetic material recovered directly from environmental samples. During the cloning process, PCR products (or fragmented community DNA, i.e., the pieces of the metagenome) are inserted into plasmid vectors that are then transformed into *E. coli*. The plasmids harboring specific PCR products (from specific organisms, presumably) are then multiplied through growing the transformed *E. coli* cells. The insertion is harvested by plasmid preparation and subsequently sequenced. After quality check and trimming of the vector sequence (i.e. alignment), the marker gene sequences can be compared to sequence databases (Maidak et al. 2001) for phylogenetic identification. Computer programs used for sequence quality check, alignment, and comparison are widely available. Phylogenetic trees can be constructed to reveal similarities between sequences (<http://rdp.cme.msu.edu/treebuilder/treeing.spr>).

11.2.1.5 *PhyloChip* microarray

Microarrays are high-throughput devices that allow for simultaneous detection of multiple DNA fragments (Bodrossy and Sessitsch 2004, Andersen et al. 2010). Detection is based on strand complementation and hybridization of the fluor-labeled target DNA with the probes representing known DNA sequences fixed on the array. After washing away the un-bound target DNA fragments, the array is scanned at defined excitation wavelengths to image the bound, fluorophore-labeled DNA fragments. The locations of the probes with hybridized target DNA on the array indicate the presence of specific nucleic acid sequences in the query sample. The fluorescence intensity can also be used to quantify the relative abundance of the target when compared to the same probe on a separate array. However, different probes cannot be directly compared to each other due to differences in GC content and hybridization efficiency

The PhyloChip (G2, i.e. the 2nd generation) is an Affymetrix (Santa Clara, CA)-platform microarray designed by researchers at the Lawrence Berkeley National Laboratory (Berkeley, CA). This high-density custom microarray encodes all known Archaea and Bacteria 16S rRNA sequences found in the 2004 public databases, and can identify up to 8,741 operational taxonomic units (OTU) (Brodie et al. 2006, DeSantis et al. 2006, DeSantis et al. 2007). Currently, a G3 PhyloChip is being developed, and the probes are designed based on all 16SrRNA sequences available in the 2007 public databases. The G3 PhyloChip can detect up to ~30,000 OTUs and select pathogen specific genes. Additionally, the probes on the G2 and G3 PhyloChips are constantly being dynamically re-annotated based on the most current database information.

Sample processing includes extraction of the nucleic acids from an environmental microbial community, and amplification of the genomic DNA using universal bacterial and archaeal primers targeting the 16S rRNA gene (Fig 11.5). The amplicons are then fragmented, biotin-labeled, denatured and hybridized onto the microarray. Following overnight hybridization, the microarray is washed, stained with streptavidin-labeled fluorophore conjugate and an image is acquired using a confocal laser scanner. The response of individual oligonucleotide probes that make up a probe-set is calculated to determine both the presence and relative concentration of defined bacterial and archaeal taxa. The data is subsequently normalized and ready for downstream processing with uni-, multi-variate, and phylogenetic analyses. Key features that set the PhyloChip apart from other similar technologies are the use of multiple oligonucleotide probes for every known category of Bacterial and Archaeal organisms for high confidence of detection and the pairing of a mismatch probe for every perfectly matched probe to minimize the effect of non-specific hybridization. A strong linear correlation has been confirmed between

microarray probe set intensity and concentration of OTU specific 16S rRNA gene copy number, allowing quantification over a wide dynamic range (Brodie et al. 2007, La Duc et al. 2009).

11.2.2 Applicability and demonstration of community analysis approaches for MST

Applicability of the various community analysis methods to specific types of questions differ because of their technological differences, state of method development and evaluation, and logistics. The demonstrated and potential usages of the methods for MST also vary.

PLFA analysis has been widely used to study changes in microbial community structure along environmental gradients or in response to environmental perturbations (Frostegard, Tunlid and Baath 1993, Macalady et al. 2000, Kaur et al. 2005). Because phenotypic adaptation reflects the microorganisms' particular habitat, including host intestine and other sources, PLFA is potentially a useful tool for MST. As it mostly reflects living or recently living cells, PLFA may have the added benefit of detecting recent pollution sources, but not older ones. However, extensive knowledge about fatty acid patterns is generally required for interpreting the significance of specific fatty acids or fatty acid groups and for most efficient usage of PLFA data. Accessible databases for relating taxa or environmental stresses with fatty acid patterns are not available, and such interpretation often relies on a researcher's experience or familiarity with the PLFA literature. Furthermore, PLFA extraction methods influence the types of fatty acids recovered from a sample, and some extraction protocols may liberate fatty acids from non-living organic matter, in which case PLFA composition could reflect more than the living microbial community (Zelles 1999). Nonetheless, these potentially confounding issues may be alleviated

by commercial service laboratories that offer standardized PLFA analysis. Total PLFA abundance and PLFA profiles can be useful to track overall biomass and microbial community changes in MST studies without specifically focusing on individual fatty acids or fatty acid groups. For example, similarity and dissimilarity of the PLFA profiles from different samples were utilized to rule out kelp, but imply beach sand, as sources of fecal contamination to beach water (Izbicki et al. 2009).

For genotypic methods based on genes encoding the 16S rRNA, abundant sequence data has been generated and is accessible in large databases such as the Ribosomal Database Project (RDP) (Maidak et al. 2001). Ever-growing databases are increasingly accessible because of developments in computational biology and bioinformatics that provide new and better tools for data handling. Performing most community analysis methods does require specialized equipment and expertise, but many genomic facilities can provide such services. However, it is important to be aware that, like culture-based or phenotypic methods, genotypic methods have their share of technical shortcomings arising from variations in DNA extraction efficiency, PCR bias, and/or sequencing accuracy and comprehensiveness.

Application of genotypic community analysis methods in MST has included 1) identification of source-specific species as candidates for developing source-specific single indicators, and 2) differentiation and/or tracking source of pollution based on the similarity of microbial community profiles from potential sources and sinks. Numerous studies have characterized microbial communities associated with human or animal feces using community analysis methods (Zoetendal et al. 2004). Although MST was not the objective, these studies provided abundant information regarding the host-specificity of microbial communities and factors that

affect such specificity (Ley et al. 2008); such studies also demonstrated the potential of community analysis for MST (Li, Hullar and Lampe 2007). The following paragraphs discuss the application of TRFLP, cloning and sequencing, and PhyloChip in MST.

Because it is a high-throughput, sensitive and reproducible approach whose data are readily amenable to quantitative statistical comparisons, TRFLP has been frequently used to analyze communities from a wide range of environments, including feces and digestive tracts of insects and mammals, and to characterize microbial community responses to environmental changes (for review, see (Thies 2007, Schütte et al. 2008)). Furthermore, because of its popularity in microbial ecology, an abundant literature and many automated data processing software applications are available for adapting TRFLP for MST studies (Kent et al. 2003, Shyu et al. 2007).

TRFLP has been used successfully to develop source-specific single indicators. For example, highly reproducible, host-specific TRFLP patterns were identified in microbial communities from human and cow feces using primers specific to the *Bacteroides-Prevotella* group (Bernhard and Field 2000a), and subsequent cloning and sequencing of such source-specific TRFs led to designing human- and cow-specific single indicator (q)PCR assays for MST (Bernhard and Field 2000b, Field, Bernhard and Brodeur 2003a). More recently, TRFLP was used to find a poultry-specific *Brevibacterium* marker (Weidhaas et al. 2010). Studies also employed TRFLP to differentiate fecal sources based on overall community similarity. TRFLP was first shown to be successful in distinguishing deer fecal samples from sands while demonstrating high similarity between microbial communities in two discrete piles of deer fecal pellets (Clement et al. 1998). Using universal eubacterial primers, TRFLP analysis also clearly

differentiated microbial communities from cattle feces, dog feces and sewage (LaMontagne, Griffith and Holden 2003). In a more comprehensive study, TRFLP was employed to analyze the *Bacteroides-Prevotella* community in multiple (10~50) fecal samples from each of nine host species (cattle, chicken, deer, dog, geese, horse, humans, pig, and seagulls) from different geographical locations and times of year (Fogarty and Voytek 2005). While no single TRF was identified as exclusive to a host species, and the previously identified cow- and human-specific TRFs (Bernhard and Field 2000a) were not resolved from their respective sources (Fogarty and Voytek 2005), the *Bacteroides-Prevotella* TRFLP community profiles were highly reproducible and much more similar within host species as compared to between host species. Attempts to identify sources using single TRFs or total community similarity for mixed-source samples, however, were less successful, perhaps due to biomass dominance from one source (Field et al. 2003b) or higher redundancy in TRFs from eubacterial primers (Liu et al. 1997). It is possible that a combination of TRFs, analyzed as a subset of the consortium, would be more useful to source identification in mixed-source samples. More recent studies utilized TRFLP to conduct source tracking in defined watersheds. Potential pollutant transport pathways were identified via similarity analysis of TRFLP community profiles from different sampling locations (Ibekwe et al. 2008). A case study using this approach will be discussed in a later section of this Chapter.

DGGE has been widely used to assess diversity and to monitor dynamics of microbial communities (Ercolini 2004, Dorigo, Volatier and Humbert 2005). Although DGGE analysis is often confined to PCR products with limited lengths (<400bp), it can differentiate a single base pair difference in sequence fragments. Compared to expensive sequencers needed for TRFLP and sequencing, equipment for DGGE analysis is affordable for ordinary laboratories; however, DGGE is technically demanding. Methodological concerns such as inaccuracy and low

reproducibility of band patterns, low sensitivity and lack of reliable quantification of band intensity, and in particular the need to optimize experimental conditions may also hinder its application (Dorigo et al. 2005). Nevertheless, DGGE applied to a 126bp fragment of the β -D-glucuronidase gene (*uidA*) was successful in discriminating among *E. coli* phlotypes using DNA from cultured isolates, DNA from mixed culture-enriched *E.coli* populations, and community DNA extracted directly from environmental samples. Little difference in the DGGE patterns was observed for the latter two DNA sources, indicating the culture enrichment step may be bypassed (Farnleitner et al. 2000). More recently, DGGE based on enriched *E. coli* cultures indicated similar *E.coli* populations for samples originated from the same sampling site (Sigler and Pasutti 2006). A more comprehensive study evaluated the applicability of 15 marker genes for use with DGGE for MST, and three genes (*mdh*, *phoE* and *uidA*) were identified to provide good discrimination among horses, pigs and goats (Esseili et al. 2008). DGGE profiles from these three genes indicated greater *E. coli* population similarity (98-100%) between waste water treatment plant (WWTP) effluents and downstream water samples, and lower similarity between upstream and downstream/effluent samples, providing strong evidence for a dominant contamination source from the WWTP. However, source attribution was less successful for contamination in a pond, presumably due to mixed source from urban runoff in addition to goose feces deposition (Esseili et al. 2008).

Cloning followed by DNA sequencing has been a widely-used tool in molecular microbial ecology from its inception. Although cloning and sequencing offers high phylogenetic resolution, the method is laborious, time consuming and costly for routine usage. Also, rarely do clone libraries provide complete coverage of entire microbial communities. However, technology advancement in automation and parallel sequencing may greatly improve its speed and lower its

cost while also improving its comprehensiveness (Shendure and Ji 2008). Because it provides actual sequence data, cloning followed by sequencing is often employed in developing and evaluating single indicator-based (q)PCR assays. For example, a library of genes encoding 16S rRNA extracted from gull feces revealed the abundance of a sequence closely related to *Catellibacter marimammali*, which was used successfully to develop a gull-specific, SYBR green (q)PCR assay (Lu et al. 2008). Libraries of *Bacteroidales* genes encoding 16S rRNA extracted from the feces of eight hosts revealed ruminant-, pig-, and horse-specific clusters of sequences while human, dog, cat, and gull *Bacteroidales* communities shared greater similarities (Dick et al. 2005). The host-specific sequences were used to design PCR assays specific to pig and horse fecal matter. The analysis of other *Bacteroidales* clone libraries comprised of genes encoding 16S rRNA extracted from gull, goose, canine, raccoon, and sewage sources revealed concerns regarding instability of source identification assays against geographic or host individual differences (Jeter et al. 2009).

In addition to developing and evaluating single indicator assays, multiple clone libraries from potential contributing sources and environmental samples have also been developed to link pollution sources with environmental sinks. *Bacteroidales* clone library analysis revealed high similarity between cattle feces and water sample clone libraries, confirming cattle fecal pollution in a small watershed (Lamendella et al. 2007). However, clone libraries constructed from a horse manure pile and water samples from up- and down-stream of the manure pile using both universal eubacterial primers and *Bacteroides* group-specific primers showed little similarity between microbial communities from the manure pile and the downstream water samples, even though the water at 5 meters downstream was visibly contaminated (Simpson, Santo Domingo and Reasoner 2004). The authors offered two explanations: 1) downstream water was

contaminated with the recently-deposited surface material of the manure pile which harbored a different microbial community than the older interior manure pile from which the clone library had been constructed; 2) universal eubacterial primers do not offer sufficient sensitivity to detect manure pollution at the dilution level in the study sites. Direct shotgun cloning and sequencing of community DNAs (i.e., metagenomic analysis) was used to characterize a viral community in human feces (Breitbart et al. 2003), however its direct application in MST has been limited.

Phylogenetic microarrays such as the PhyloChip, which targets the currently known diversity within bacteria and archaea, have been employed to determine the composition of microbial communities in a number of different environments and conditions. When the PhyloChip microarray was applied to urban aerosols, the spatio-temporal distributions of known bacterial groups, including specific pathogens, were determined to be related to meteorologically driven transport processes as well as sources (Brodie et al. 2007). This microarray has been extensively validated and successfully used on a number of complex environmental samples and the resulting findings have been confirmed by additional methods, including qPCR and 16S rRNA gene clone libraries (Brodie et al. 2006, Flanagan et al. 2007, Chivian et al. 2008, Tsiamis et al. 2008, Wrighton et al. 2008, Cruz-Martinez et al. 2009, DeAngelis et al. 2009, Sagaram et al. 2009, Sunagawa et al. 2009, Yergeau et al. 2009, Rastogi et al. 2010). Studies using split samples have confirmed that >90% of all 16S rRNA sequence types identified by the more expensive clone library method are also identified by the PhyloChip (DeSantis et al. 2007). In addition, the PhyloChip has demonstrated several-fold increases in detected microbial diversity over the clone library method. One of the reasons for this is the high sensitivity of the PhyloChip, with the ability to detect organisms present at a proportional fraction of less than 10^{-4} abundance compared to the total sample (La Duc et al. 2009). Although each sample analysis by

the PhyloChip provides detailed information on microbial composition, the highly parallel and reproducible nature of this array allows tracking community dynamics over time and treatment. With no prior knowledge, specific microbial taxa may be identified in urban watersheds that are keys to human-associated fecal influence.

The PhyloChip is ideal for characterizing complex microbial communities and its application for MST is currently being investigated. The comprehensiveness and sensitivity of the PhyloChip allows for better characterization of low abundance organisms leading to improved description of microbial diversity (La Duc et al. 2009, Sagaram et al. 2009). The reproducibility of the PhyloChip data on microbial community compositions provides the opportunity to obtain results with high levels of statistical confidence (Brodie et al. 2006, DeSantis et al. 2007). A case study with PhyloChip-analyzed bacterial communities from an urban creek with known fecal pollution will be discussed in a later section.

11.3 Multivariate data analysis, interpretation, and presentation

11.3.1 Why multivariate techniques?

Multivariate analysis involves simultaneous analysis of multiple, often correlated, variables. Multivariate analysis of community profiles has been developed and is routinely used by ecologists who study animals or plants, yet the application of such tools has been limited in microbial ecology, both in terms of frequency and choice of multivariate methods (Ramette 2007). However, multivariate tools are necessary to analyze the multivariate datasets generated from community analysis-based methods for MST.

Dataset generated by microbial community analysis methods usually contain rows representing samples or sites and columns representing operational taxonomic units (OTUs). OTUs can be fatty acids from PLFA, TRFs from TRFLP, gel bands identifiers from either DGGE or TGGE, or sequences or species from either clone libraries or direct sequencing or PhyloChip analyses. Although each OTU could be treated as a single variable and analyzed by univariate statistical methods separately, separate univariate analyses not only are logistically difficult because there are hundreds to tens of thousands columns in a community profile, but also are scientifically undesirable because microbial communities evolve and adapt together, therefore these variables are not independent. Source-sink relationships that are not revealed when a single OTU is evaluated can be distinguished when a consortium of OTUs is analyzed simultaneously in an integrated fashion, hence the rationale of using community-based analysis for microbial source tracking (see section 11.1).

Many applications of MST are closely tied to TMDL assessment, which is currently based on FIB concentrations (Santo Domingo et al. 2007), and it is often desirable to evaluate correlations between MST and FIB concentration data. Correlations between multivariate community profiles with FIB concentrations can only be done through multivariate statistics such as direct gradient analysis (see section 11.3.2). Furthermore, when discoveries of OTUs that are indicative of source are desired, multivariate ordination techniques are more much efficient compared to manually counting OTUs that are shared among samples from the same source.

11.3.2 Selection of multivariate techniques and results interpretation

Common multivariate techniques for community profile analysis include cluster analysis, principle components analysis (PCA), correspondence analysis (CA), and nonmetric multidimensional scaling (NMDS). These techniques belong to a group called indirect gradient analysis, which aims to reveal community similarities among sites or samples through grouping or ordering the sites or samples into either dendrograms or on a two or three-dimensional plot. Direct gradient analysis such as canonical correspondence analysis (CCA), on the other hand, aims to correlate the overall multivariate community profile with environmental variables or FIB concentrations. More details on each individual technique can be found in this review and the references therein (Ramette 2007). However, methods differ in when and how they should be used, and proper selection of the methods is a very most important first step in data analysis. Selection of the multivariate methods must be based on data type (binary, compositional or abundance data), analysis objective, and strengths and limitations of the various multivariate methods (Ramette 2007). Standard statistical software such as R (R Core Development Team 2008) and SAS can be programmed to run multivariate analysis. Specialized multivariate software packages are also available: CANOCO (Microcomputer Power, Ithaca, NY), PC-ORD (MjM Software Design), and Primer (Primer-E Ltd, Plymouth Marine Laboratory, United Kingdom).

Microbial ecologists are likely most familiar with cluster analysis, which is historically the basis for constructing phylogenetic trees that reveal similarities between sequences (e.g., OTUs in clone libraries). When discovering similarities between sites or samples is the goal, cluster analysis essentially groups sites or samples according to a similarity coefficient based on OTU data, and its interpretation is mostly intuitive: samples or sites grouped in the same cluster are similar to each other (Ramette 2007). However, because cluster analysis forces the formation of

clusters, this method is most appropriate when groupings (i.e. discontinuous changes) of sites or samples are expected, such as communities from different known or suspected sources (e.g. animals, sewage etc.) (Legendre and Legendre 1998). Cluster analysis is not appropriate when changes in communities are either continuous or gradual, such as samples from upstream to downstream sites, and discrete groupings are not expected.

PCA is a very frequently used multivariate technique partially due to its elegant mathematical algorithms (ter Braak 1995); however, it may also be the least appropriate method for analyzing microbial community profiles. A basic assumption for PCA is that OTUs respond to environmental conditions (i.e. environmental gradients) in a linear fashion, which is rarely true because most species have an optimal environmental condition or ecological niche and their response curves to environmental condition are more similar to a unimodal model (ter Braak 1986). However, when the environmental gradient is very short, the unimodal response curve may appear linear, and PCA may be used appropriately. For example, most bacteria prefer a neutral pH and therefore exhibit a unimodal response to pH; yet, the response may be considered linear if the environmental pH conditions present a very short gradient ranging only from 5.5 to 6.0. Still, the absence of many OTUs in some sites or samples is a clear sign that the linear approximation is not valid and PCA is not an appropriate method. Improper usage of PCA may cause an artifact called the “horseshoe effect” where sites or samples are positioned on the 2-D PCA plot resembling the shape of a horse shoe; these positions do not represent either similarity or dissimilarity between sites or samples (Palmer 2006).

CA assumes unimodal species response curves which are more appropriate for analyzing ecological data such as microbial community profiles (ter Braak 1986). CA is also considered a

two-faced method in that it can accommodate a dataset even when the underlining gradient is short, and thus linear, as long as the composition data (i.e. relative abundance in percentages) instead of absolute abundance data are analyzed (ter Braak and Smilauer 2002). CA results are generally displayed in a 2-D plot called a “joint plot” where both sites and OTUs can be displayed as points on the plot, or a sample scatter plot where only sites are displayed.

Community similarity between sites is indicated by a close proximity of the site positions on the plot. The strong association of OTUs to certain sites (or sources) is implied by close proximity of the OTUs to the sites (or sources), and this can be used to reveal indicative OTUs for developing source-specific, i.e. qPCR, assays. Similar to the “horseshoe effect” in PCA, CA sometimes suffers an artifact called the “arch effect” when positioning of sites along the secondary axis could be arbitrary such that it resembles an arch. Removing the arch effect is achieved by a process called detrending, hence the term detrended correspondence analysis (DCA). Note that the axes on CA plots are meaningful, as they represent latent variables or gradients such as the distance to a point source (e.g. a storm drain discharge). This is useful for discovering trends and potential microbial contamination sources.

NMDS positions sites, or samples, into a 2-D (or 3-D) plot in a way that the ranks of dissimilarity between these sites are best preserved, much like positioning cities on a 2-D map where the relative, or ranks of, distances between those cities along the Earth’s spherical surface are preserved (Clarke and Warwick 2001). Therefore, sites in closer proximity to each other on a NMDS plot have more similar community profiles than those that are further apart on the plot. However, the distances between sites on the NMDS plot do not reflect the original dissimilarity in community profiles between those sites because only the ranks of the dissimilarity are preserved. NMDS has gained popularity because it does not either assume linear or unimodal

species response curves or produce “horseshoe” or “arch” artifacts, and its interpretation is intuitive. Drawbacks of NMDS include its great sensitivity to dissimilarity measures, i.e. distance metrics, which must be specified *a priori* by the user. NMDS also cannot simultaneously display sites and OTUs on one plot, thus associations between sites and OTUs may not be revealed (Palmer 2006). Therefore, if identifying site- or source-specific OTUs is the objective of the analysis, NMDS would not be the method of choice. NMDS is most useful for quickly assessing relative (dis)similarity among sites, or samples.

While indirect gradient analysis is generally considered exploratory, direct gradient analysis offers a means of specific hypothesis testing (ter Braak and Prentice 1988). A popular direct gradient analysis technique is canonical correspondence analysis (CCA), which is an extension of CA, therefore CCA shares many advantages and disadvantages of CA. CCA has been used to test whether community changes are influenced by environmental variations such as inundation (Cao et al. 2006). In the case of MST, for instance, CCA could be useful for testing whether community profiles correlate with FIB concentrations or if community profiles correlate with days after a sewage spill. An extension of CCA is partial CCA (pCCA), where influences from a co-variable can be excluded before evaluating effects on community profiles from another environmental variable. For example, pCCA was successfully used to identify the correlation between denitrifying community changes with heavy metal contamination after adjusting for influences of dissolved carbon (Cao et al. 2008). pCCA is potentially useful for MST when the effects of geographic location need to be accounted for before correlating changes in microbial community with the magnitude of a microbial pollution source, for instance, the volume of a sewage spill.

11.4 Two case studies

11.4.1 A case study using TRFLP

This case study examined bacterial communities using TRFLP during dry weather flow in the Arroyo Burro watershed (Santa Barbara, CA) where elevated FIB concentrations and human-specific *Bacteroides* markers were previously reported (Sercu et al. 2009). A laboratory spike-in experiment for calibrating the TRFLP technique, and a field study for investigating pollution sources in the Arroyo Burro watershed, were conducted. For the spike-in experiment, fecal samples were collected from suspected fecal sources such as dog, cat, and human (e.g. septage). Dog feces were acquired and composited from three healthy individuals of different breeds, each from a separate household. Cat feces were acquired from three healthy individuals of mixed breeds from two households. Septic solids, representing the composite material from several residential tanks, were obtained from a local pumping company (MarBorg, Santa Barbara). Relatively unimpacted creek water from a reference site in the watershed was collected in order to create spiked samples using the above fecal sources at various doses.

For the field study, nine sites in the lower AB watershed were selected. The nine sites (site 9 to 0 from upstream to downstream) included a storm drain (site 9) discharging into the AB creek (site 8 to 3) that flowed through the AB lagoon (site 2) then into the ocean at Henry's Beach, CA (site 1). Water samples were collected from these sites on three consecutive days (August 2005) as described previously (Sercu et al. 2009). No rain occurred at least 48 hours prior to or during sampling, and the creek flow rate was $0.013 \text{ m}^3 \text{ s}^{-1}$. Four sewage influent samples were also

collected from the El Estero Wastewater Treatment Plant (Santa Barbara, CA) during the period October 2004 to October 2005. Microbial communities were analyzed by TRFLP using universal primers targeting the domain *Bacteria*. Relative abundance TRFLP data were aligned and analyzed using detrended correspondence analysis (DCA) as before (Cao et al. 2006).

Community (dis)similarity among the spiked samples and the fecal sources reflected differences among the fecal sources and the spiking dose (Fig 11.6). Communities from water samples spiked with low doses of septic solids (0.01% and 0.1%) were more similar to the reference water sample than to the septic solids, while communities from those spiked with higher doses (1% and 5%) were more similar to the septic solids. Water samples that contained spikes from cat and dog feces grouped with cat and dog feces but not with the septic solids and the reference water. The results indicate that TRFLP can identify pollution sources but relative contributions from different sources affect its sensitivity for MST.

For the field study, bacterial communities showed temporal stability during the three sampling days and clear spatial source tracking related to hydrological connectivity (Fig 11.7).

Communities differed as the water flowed downstream from the storm drain (site 9), to site 8, shortly downstream of the drain discharge, to the further downstream creek and lagoon sites (site 7-2), and then to the ocean (site 1). The major DCA gradient (DCA axis 1) explained 20% of the total variation in TRFLP profiles and coincided with the creek flow direction. While sewage samples grouped together by themselves and separated from the water samples along the secondary DCA axis, storm drain samples grouped with sewage samples when profiles from dog and cat feces were also included in the DCA (data not shown). These results are consistent with frequent detection of human-specific *Bacteroides* markers at the drain and the creek site

following drain discharge (Sercu et al. 2009), and further implicated the storm drain as a source of human fecal pollution. The field study results also demonstrate the ability of community analysis to track source of pollution on a spatial scale, in addition to its ability to accomplish source identification.

11.4.2 A case study using PhyloChip

This case study examined bacterial communities during dry weather flow in the lower Mission Creek and Laguna watersheds (Santa Barbara, CA) where elevated FIB concentrations and human-specific *Bacteroides* markers were previously reported (Sercu et al. 2009). Communities from creek (including storm drains), lagoon and ocean sites, along with 3 fecal samples of human origin, were analyzed by the G2 PhyloChip (Wu et al. 2010). Mission Creek and Laguna Channel flow through an urbanized area of downtown Santa Barbara and discharge at a popular bathing beach. As described previously (Sercu et al. 2009), water column samples from 3 consecutive days were collected during the dry season (June 2005) from 9 locations within the Mission Creek and Laguna watersheds in Santa Barbara, California. No rain occurred at least 48 hours prior to or during the sampling. The creek flow rate in the Mission Creek was $0.016 \text{ m}^3\text{s}^{-1}$. Both watersheds discharged into the same lagoon and then flowed from the lagoon into the ocean.

NMDS with the Jaccard distance measure was used to visualize the dissimilarity between the bacterial communities in the samples (Fig 11.8). Ocean bacterial communities were different from creek, lagoon and fecal communities. For the most part, lagoon and creek communities

were different from each other, except for 2 samples. Fecal samples grouped separately from creek, lagoon and ocean samples illustrating the presence of distinct bacteria taxa. These bacteria, which were unique to fecal samples, could potentially be used as signature taxa for indicating fecal communities.

PhyloChip data resulted in a comprehensive description of community composition from each of the samples. Comparative analysis of relative richness, which is the number of OTUs detected normalized to total OTUs on the PhyloChip for each class, showed that 4 bacterial classes exhibited the greatest variation across the sample types. Relative richness of *Bacteroidetes*, *Bacilli* and *Clostridia* was higher in the fecal samples than the other sample types, while *Alphaproteobacteria* relative richness was lower in the fecal samples (Fig 11.9). The sum of the relative richness of *Bacteroidetes*, *Bacilli* and *Clostridia* was divided by the relative richness of *Alphaproteobacteria* to obtain a ratio (BBC:A). Thus, the BBC:A ratio incorporated microorganisms prevalent in fecal samples as well as those that were found in “pristine” environments. The BBC:A for fecal, creek, lagoon and ocean samples were 4.06, 0.92, 0.69, and 0.65, respectively. The BBC:A for fecal samples was ~6-fold higher than that of ocean samples. High concentrations of FIB and human-specific *Bacteroides* markers were also detected on all 3 days at the site with the highest BBC:A (Wu et al. 2010). With further development and validation, this ratio could potentially be a useful tool for identifying fecal pollution. Future research is moving towards identifying signature communities for MST.

11.5 Relationship of community analysis to multiple indicator approaches

As no single indicator can correctly identify microbial contamination sources 100% of the time, multiple indicators are often measured from the same sample and used in a tiered fashion. When measured in this manner, multiple indicators operate as a selected subset of the microbial community. Therefore, methods used to analyze and interpret community analysis results can be applied to analyzing multiple indicator datasets as well. For example, multivariate statistics are rarely used in multiple indicator MST studies but provide a means to simultaneously view all indicator results, and could greatly aid in discovering patterns and trends. A DCA joint plot where sample IDs and the multiple indicators are simultaneously displayed illustrates the concept, using data from the case study in section 11.4.1 (Fig 11.10). The close proximity of culturable *E. coli* and *Enterococcus* and of human-specific *Bacteroides* markers to sites 9 and 8 (drain and nearby creek site, respectively), and the proximity of salinity to site 1 (ocean) indicate relatively higher levels of these indicators at the corresponding sites, while total coliform, dissolved oxygen, and pH do not indicate a strong pattern or trend (Fig 11.10).

Similarly, methods commonly used for multiple indicators such as ratio and predictive modeling (Blanch et al. 2006) can also be applied to community analysis. In addition to using the overall community data, one can choose to reduce the data to a few OTU groups and investigate the ratios between groups as a source identification tool (e.g. as per the case study in section 11.4.2), or to focus on several specific OTUs that may be indicative of specific sources (Bernhard and Field 2000a). Furthermore, common OTUs (or cosmopolitan OTUs, if any) can be removed from the overall community data, and the community profiles can be focused into a more select dataset where the most predictive OTUs, which can be considered multiple indicators, are measured simultaneously in just one assay.

11.6 Summary and future directions of community analysis for MST

11.6.1 Advantages

Ideally, MST methodology should include assays targeting specific pathogens that have been identified via epidemiology studies as public health risks in the context of recreational water use (Field and Samadpour 2007). However, to date, epidemiology studies include a very limited number of pathogen measurements. There are two reasons for such limitation. First, prior knowledge about relevant pathogens in a particular water body is often lacking. Second, the cost to perform a complete survey of all possible pathogenic indicators via many single measurements is prohibitive. Nevertheless, sole reliance on specific pathogens could be inadequate for MST, since detection of pathogens would depend both on the presence of fecal material and on the health status of surrounding human or animal populations. Community analyses provide a cost-effective alternative that offers many advantages, most notably: 1) comprehensiveness and relevancy, and 2) data density.

Comprehensiveness and relevancy refer to the inclusion of pathogens, fecal indicators and other organisms when DNA or other biomarkers are fully extracted and analyzed from a water sample (Fig 11.3). If a single marker is labile and its environmental fate ill-defined, simultaneous reliance on many singly- or interactively-relevant markers from a community can enable waste detection even in the absence of the yet-to-be-discovered single, robust marker. Furthermore, when identification of fecal source(s) is based on the entire microbial community, by default it is also based on tracing pathogens within that community (Fig 11.3). Although resolution for

particular pathogens differs with variations of the community analysis techniques (Section 11.2.1), the public health relevance of data acquired from community analysis is higher for MST when compared to data from a few individual markers whose transport characteristics and fates are unlikely to mimic fates exhibited by a majority of fecal pathogens.

Data density refers to the inherently multivariate nature of community analysis data, which are extremely versatile in how and what information may be extracted. In an MST study, the multivariate community analysis data also represent multiple lines of evidence, which, as in a trial by jury, increase the certainty of a water quality diagnosis. While performing many different types of individual assays can also provide the needed multiple lines of evidence, e.g. by various chemical and biochemical host-specific markers, each requires separate procedures and expertise. An additional advantage provided by community analysis includes the capability of using community similarity between sites to conduct spatial source tracking (see the case study in Section 11.4.1). This advantage exists because community similarity analysis naturally combines data from source types and loading (Sercu et al. 2009), both of which are often needed for locating the source of contamination in MST studies. Lastly, by using community analysis approaches more broadly in MST, more insights across numerous studies and geographical locations can be obtained to define additional individual markers of waste or to define the suite of individual markers within the overall community that best resolve sources of concern.

11.6.2 Critical issues

Despite its advantages, there are technological, logistical, and implementation-related issues regarding usage of community analysis for MST.

Technologically, the sensitivity and resolution of certain community analysis methods may prevent source identification if the sources contributing to the receiving waters are very diluted or are of complex, mixed origins. Also, more biomass (or DNA) is often needed for community analysis than a single (q)PCR assay. Temporal variation in source concentrations, a concern for MST study in general, can add to the issue. It is therefore important to examine the characteristics and complexity of a watershed (or an MST study system) before 1) selecting a community analysis technique, because the different techniques vary in their sensitivity and resolution (Section 11.2) as well as cost and feasibility and 2) formulating a sampling design that may capture temporal and spatial variation in source contributions. Multiple community analysis techniques may also be used in one MST study to obtain more sensitivity and resolution in a cost-effective manner. For example, less expensive, high throughput TRFLP may be applied to screen an entire watershed for “hot zones,” where more expensive cloning and sequencing or PhyloChip analysis can be applied to obtain higher resolution source identification. While efficiency of concentration (such as water sample filtration) and DNA extraction methods are important to the reliability of all molecular techniques in general, such concerns and PCR bias may be less for community composition studies performed comparatively among samples or sites. However, such technical issues may hinder quantitative interpretation of community analysis results.

Fate (including die-off, persistence, and growth) and transport of microorganisms contributed from various fecal sources in the environment would also affect application of community

analysis for MST. The microbial communities contained in fecal sources such as sewage undergo alterations when migrating through soil or groundwater. Wastewater changes biochemically when passed through reactive porous media (An et al. 2004) like soil, such that nutrient (Hua et al. 2004, Stogbauer, Berner and Stuben 2004) and microbial (Hua et al. 2003) concentrations may change in predictable ways (Brinkmann et al. 2004). However, how microbial communities in fecal sources change during their mixing, and migration in the environment is yet unknown. This is consistent with the state of knowledge for other, host-specific, individual markers (Field and Samadpour 2007) and warrants further research.

Logistically, as most methods need specialized equipment and expertise, community analyses are often more expensive than a (q)PCR assay of a single indicator, and are performed in (sometimes very specialized) research laboratories or commercial service laboratories. Routine use of such methods may be limited by the availability of expertise and cost, although high demand of applying such analysis (from MST or other fields such as biomedical research) will drive the availability up and the cost down.

A common issue related to implementation in the MST field is the lack of standard protocol for performing MST studies. While lack of standardized sample processing and laboratory protocol is common to most MST methods, for community analysis based methods, there are even more needs for standard protocols of data processing, analysis, and interpretation, particularly because of data complexity. Research aimed at developing standardized MST protocols for sampling design, sample processing (including DNA extraction), and data analysis are needed. Lastly, and of largely practical concern, is the potential difficulty for water quality managers in communicating microbial community data to the public. There may be multiple dimensions to

this issue, including the fact that the non-scientific community is generally unaware that microbial communities exist in nature, and thus could become unnecessarily alarmed by data that reveals the richness of microbial taxa in water even in the absence of fecal contamination. As with the public consumption of voluminous “personal genomic” data (from genetic testing for susceptibility to disease) for which full interpretation is lacking (Wright and Kroese 2010), there is a possibility for public misunderstanding and data misuse. The evolution of microbial community analysis in microbiological water quality from a research tool into a monitoring tool will require consideration of this and other issues described in this section.

11.6.2 Future directions

Future directions for community analysis may include 1) incorporation of these methods into epidemiology studies, 2) assisting in research on indicator persistence and survival, and 3) development of new indicators, customized community analysis, and automation. While advanced molecular technologies did not exist for early epidemiology studies, modern epidemiology studies often archive genetic samples for future analysis. These archived samples can be analyzed by community analysis such that the comprehensive water quality data can be correlated to human health data that are already collected. This would provide a means to discover pathogens or nonpathogenic indicators that are highly predictive of health risk.

Comprehensive community analysis is also used to study the succession of microbial communities along pollution gradient of sewage discharge (Zhang et al. 2009) or after sewage spills (Dubinsky et al. 2009) which can provide insight into indicator survival and persistence

after a sewage source is introduced into receiving waters. Similar studies can be conducted on other fecal sources in other environments to provide similar information that are greatly needed for revealing age and contributions of various fecal sources in MST (Blanch et al. 2006).

Although limitations such as read length and complexities of data analysis still exist, the next generation sequencing technology and bioinformatics are advancing very rapidly to provide truly high-efficiency and low-cost sequencing in the near future (Shendure and Ji 2008). Sequencing the whole microbial community associated with each source covering diverse geographic locations and individuals would soon be feasible, which will lead to identification of more and better source-specific indicators for development of qPCR-based methods and specialized MST microarrays. Because of the versatility of the community analysis techniques, “customized” community analysis can be designed to provide higher sensitivity and resolution for MST. For example, high throughput techniques such TRFLP can be paired with more specific and variable marker genes such as functional genes to offer higher sensitivity by tuning down the background that is generated when universal 16S rRNA primers are used. While the PhyloChip is one form of custom microarray, other microarrays can also be designed (Shiu and Borevitz 2008) to specifically target pathogens and pathogenic genes. Source identification microarrays (i.e., “MST on a chip”) can be designed to include thousands of source-specific assays on the microarray such that each sector on the microarray represents a particular pollution source (human, sewage, gulls etc.). The “MST on a chip”, combined with automated data analysis software, may give a probability estimate of contributions from each source and enable fast diagnosis in a watershed. Ultimately, a high throughput, close to real-time (within 6-12 hrs) pipeline (array or qPCR) for processing water samples and obtaining results may be developed for real-time source tracking.

Acknowledgements

The authors acknowledge the City of Santa and the Switzer Foundation for support, and the NSF-funded Santa Barbara Long Term Ecological Research project (NSF OCE 9982105 and OCE 0620276) for assistance including stream flow data in Santa Barbara, and the work of Laurie C. Van De Werfhorst and Bram Sercu in sampling, and sample and data processing for the AB and MC case studies herein. Other attributions for the Arroyo Burro and Mission Creek fecal source and sample acquisition plus analysis are as per Sercu et al. (2009). Part of this work was performed at Lawrence Berkeley National Laboratory under Department of Energy contract number DE-AC02-05CH11231.

References

- An PL, Hua JM, Franz M et al. (2004) Changes of chemical and biological parameters in soil caused by trickling sewage. *Acta Hydrochim Hydrobiol* **32**: 296-303.
- Andersen GL, He Z, DeSantis TZ et al. (2010) The use of microarrays in microbial ecology. In *Environmental Molecular Microbiology*. Liu W-T, and Jansson JK (eds). Caister Academic Press, Norfolk, UK, pp. 87-109.
- Bernhard AE, and Field KG (2000a) Identification of nonpoint sources of fecal pollution in coastal waters by using host-specific 16S ribosomal DNA genetic markers from fecal anaerobes. *Appl Environ Microbiol* **66**: 1587-1594.
- Bernhard AE, and Field KG (2000b) A PCR assay to discriminate human and ruminant feces on the basis of host differences in *Bacteroides-Prevotella* genes encoding 16S rRNA. *Appl Environ Microbiol* **66**: 4571-4574.
- Blanch AR, Belanche-Munoz L, Bonjoch X et al. (2006) Integrated analysis of established and novel microbial and chemical methods for microbial source tracking. *Appl Environ Microbiol* **72**: 5915-5926.
- Bodrossy L, and Sessitsch A (2004) Oligonucleotide microarrays in microbial diagnostics. *Current Opinion in Microbiology* **7**: 245-254.
- Boehm AB, Fuhrman JA, Mrse RD et al. (2003) Tiered approach for identification of a human fecal pollution source at a recreational beach: Case study at Avalon Cay, Catalina Island, California. *Environ Sci Technol* **37**: 673-680.
- Braker G, Ayala-del-Rio HL, Devol AH et al. (2001) Community structure of denitrifiers, *bacteria*, and *archaea* along redox gradients in pacific northwest marine sediments by terminal restriction fragment length polymorphism analysis of amplified nitrite reductase (*nirS*) and 16S rRNA genes. *Appl Environ Microbiol* **67**: 1893-1901.

Breitbart M, Hewson I, Felts B et al. (2003) Metagenomic analyses of an uncultured viral community from human feces. *J Bacteriol* **185**: 6220-6223.

Brinkmann T, Abbt-Braun G, Karle E et al. (2004) Transformation of wastewater-derived dissolved organic matter below leaky sewers - Fate of amino acids and carbohydrates. *Acta Hydrochim Hydrobiol* **32**: 316-327.

Brodie EL, DeSantis TZ, Parker JPM et al. (2007) Urban aerosols harbor diverse and dynamic bacterial populations. *Proceedings of the National Academy of Sciences of the United States of America* **104**: 299-304.

Brodie EL, DeSantis TZ, Joyner DC et al. (2006) Application of a high-density oligonucleotide microarray approach to study bacterial population dynamics during uranium reduction and reoxidation. *Appl Environ Microbiol* **72**: 6288-6298.

Cabelli VJ, Dufour AP, McCabe LJ et al. (1982) Swimming-associated gastroenteritis and water quality. *Am J Epidemiol* **115**.

Cao Y, Green PG, and Holden PA (2008) Microbial community composition and denitrifying enzyme activities in salt marsh sediments. *Appl Environ Microbiol* **74**: 7585-7595.

Cao Y, Cherr GN, Córdova-Kreylos AL et al. (2006) Relationships between sediment microbial communities and pollutants in two California salt marshes. *Microb Ecol* **52**: 619-633.

Chivian D, Brodie EL, Alm EJ et al. (2008) Environmental genomics reveals a single-species ecosystem deep within earth. *Science* **322**: 275-278.

Clarke KR, and Warwick RM (2001) *Change in marine communities: An approach to statistical analysis and interpretation, 2nd edition*. PRIMER-E, Plymouth.

Clement BG, Kehl LE, DeBord KL et al. (1998) Terminal restriction fragment patterns (TRFPs), a rapid, PCR-based method for the comparison of complex bacterial communities. *J Microbiol Methods* **31**: 135-142.

Córdova-Kreylos AL, Cao Y, Green PG et al. (2006) Diversity, composition and geographical distribution of microbial communities in California salt marsh sediments. *Appl Environ Microbiol* **72**: 3357-3366.

Cruz-Martinez K, Suttle KB, Brodie EL et al. (2009) Despite strong seasonal responses, soil microbial consortia are more resilient to long-term changes in rainfall than overlying grassland. *ISME J* **3**: 738-744.

DeAngelis KM, Brodie EL, DeSantis TZ et al. (2009) Selective progressive response of soil microbial community to wild oat roots. *ISME J* **3**: 168-178.

DeSantis TZ, Brodie EL, Moberg JP et al. (2007) High-density universal 16S rRNA microarray analysis reveals broader diversity than typical clone library when sampling the environment. *Microb Ecol* **53**: 371-383.

DeSantis TZ, Hugenholtz P, Larsen N et al. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* **72**: 5069-5072.

Dick LK, Stelzer EA, Bertke EE et al. (2010) Relative decay of *Bacteroidales* microbial source tracking markers and cultivated *Escherichia coli* in freshwater microcosms. *Appl Environ Microbiol* **76**: 3255–3262.

Dick LK, Bernhard AE, Brodeur TJ et al. (2005) Host distribution of uncultivated fecal *Bacteroidales* bacteria reveal genetic markers for fecal source identification. *Appl Environ Microbiol* **71**: 3184-3191.

Dorigo U, Volatier L, and Humbert JF (2005) Molecular approaches to the assessment of biodiversity in aquatic microbial communities. *Water Res* **39**: 2207-2218.

Dubinsky E, Wu C, Hulls J et al. (2009) A complete microbial community approach to tracking fecal pollution in coastal waters. In *State of the Estuary Conference*. Oakland, CA.

Dunbar J, Ticknor LO, and Kuske CR (2001) Phylogenetic specificity and reproducibility and new method for analysis of terminal restriction fragment profiles of 16S rRNA genes from bacterial communities. *Appl Environ Microbiol* **67**: 190-197.

Eaton AD, Clesceri LS, Greenberg AE et al. (1998) *Standard Methods for The Examination of Water and Wastewater*. American Public Health Association, Washington, DC.

Eckburg PB, Bik EM, Bernstein CN et al. (2010) Diversity of the human intestinal microbial flora. *Science* **308**: 1635-1638.

Ercolini D (2004) PCR-DGGE fingerprinting: novel strategies for detection of microbes in food. *J Microbiol Methods* **56**: 297-314.

Esseili MA, Kassen II, and Sigler V (2008) Optimization of DGGE community fingerprinting for characterizing *Escherichia coli* communities associated with fecal pollution. *Water Res* **42**: 4467-4476.

Farnleitner AH, Kreuzinger N, Kavka GG et al. (2000) Simultaneous detection and differentiation of *Escherichia coli* populations from environmental freshwaters by means of sequence variations in a fragment of the beta-D-glucuronidase gene. *Appl Environ Microbiol* **66**: 1340-1346.

Field KG, and Samadpour M (2007) Fecal source tracking, the indicator paradigm, and managing water quality. *Water Res* **41**: 3517-3538.

Field KG, Bernhard AE, and Brodeur TJ (2003a) Molecular approaches to microbiological monitoring: Fecal source detection. *Environmental Monitoring and Assessment* **81**: 313-326.

Field KG, Chern EC, Dick LK et al. (2003b) A comparative study of culture-independent, library-independent genotypic methods of fecal source tracking. *J Water Health* **1**: 181-194.

Flanagan JL, Brodie EL, Weng L et al. (2007) Loss of bacterial diversity during antibiotic treatment of intubated patients colonized with *Pseudomonas aeruginosa*. *J Clin Microbiol* **45**: 1954-1962.

Fogarty LR, and Voytek MA (2005) Comparison of Bacteroides-Prevotella 16S rRNA genetic markers for fecal samples from different animal species. *Appl Environ Microbiol* **71**: 5999-6007.

Frostegard A, Tunlid A, and Baath E (1993) Phospholipid fatty acid composition, biomass, and activity of microbial communities from two soil types experimentally exposed to different heavy metals. *Appl Environ Microbiol* **59**: 3605-3617.

- Hazen TC (1988) Fecal coliforms as indicators in tropical waters--a review. *Toxicity Assessment* **3**: 461-477.
- Hua JM, An PL, Winter J et al. (2003) Elimination of COD, microorganisms and pharmaceuticals from sewage by trickling through sandy soil below leaking sewers. *Water Res* **37**: 4395-4404.
- Hua JM, An PL, Winter J et al. (2004) Elimination of organic compounds of sewage or sewage fractions in sandy soil below leaking sewers. *Acta Hydrochim Hydrobiol* **32**: 287-295.
- Hwang CC, Wu WM, Gentry TJ et al. (2009) Bacterial community succession during in situ uranium bioremediation: spatial similarities along controlled flow paths. *ISME J* **3**: 47-64.
- Ibekwe AM, Bold RM, Lyon SR et al. (2008) Microbial community composition in middle Santa Ana River watershed impacted by non-point source pollutants. In *the 108th general meeting for American Society of Microbiology*. Boston, MA.
- Ishii S, Hansen DL, Hicks RE et al. (2007) Beach sand and sediments are temporal sinks and sources of *Escherichia coli* in Lake Superior. *Environ Sci Technol* **41**: 2203-2209.
- Izbicki JA, Swarzenski PW, Reich CD et al. (2009) Sources of fecal indicator bacteria in urban streams and ocean beaches, Santa Barbara, California. *Annals of Environmental Sciences* **3**: 139-178.
- Jeter SN, McDermott CM, Bower PA et al. (2009) Bacteroidales Diversity in Ring-Billed Gulls (*Larus delawarensis*) Residing at Lake Michigan Beaches. *Appl Environ Microbiol* **75**: 1525-1533.
- Kaur A, Chaudhary a, Kaur A et al. (2005) Phospholipid fatty acid - A bioindicator of environment monitoring and assessment in soil ecosystem. *Curr Sci* **89**: 1103-1112.
- Kent AD, Smith DJ, Benson BJ et al. (2003) Web-based phylogenetic assignment tool for analysis of terminal restriction fragment length polymorphism profiles of microbial communities. *Appl Environ Microbiol* **69**: 6768-6776.
- La Duc MT, Osman S, Vaishampayan P et al. (2009) Comprehensive census of bacteria in clean rooms by using DNA microarray and cloning methods. *Appl Environ Microbiol* **75**: 6559-6567.

Lamendella R, Santo Domingo JW, Oerther DB et al. (2007) Assessment of fecal pollution sources in a small northern-plains watershed using PCR and phylogenetic analyses of Bacteroidetes 16S rRNA gene. *FEMS Microbiol Ecol* **59**: 651-660.

LaMontagne MG, Schimel JP, and Holden PA (2003) Comparison of subsurface and surface soil bacterial communities in California grassland as assessed by terminal restriction fragment length polymorphisms of PCR-amplified 16S rRNA genes. *Microb Ecol* **46**: 216-227.

LaMontagne MG, Griffith JF, and Holden PA (2003) Comparative analysis of animal fecal bacterial communities using terminal restriction fragment length polymorphisms of bacterial 16S rDNA PCR-amplified from fecal community DNA. In *American Society for Microbiology General Meeting*. Washington, DC.

LaMontagne MG, Leifer I, Bergmann S et al. (2004) Bacterial diversity in marine hydrocarbon seep sediments. *Environ Microbiol* **6**: 799-808.

Leadbetter ER (1997) Prokaryotic diversity: form, ecophysiology, and habitat. In *Manual of Environmental Microbiology*. Hurst CJ, Knudsen GR, McInerney MJ et al. (eds). American Society for Microbiology, Washington, D. C., pp. 14-24.

Legendre P, and Legendre L (1998) *Numerical Ecology*. Elsevier Science BV, Amsterdam.

Ley RE, Hamady M, Lozupone C et al. (2008) Evolution of mammals and their gut microbes. *Science* **32**: 1647-1651.

Li F, Hullar MAJ, and Lampe JW (2007) Optimization of terminal restriction fragment polymorphism (TRFLP) analysis of human gut microbiota. *J Microbiol Methods* **68**: 303-311.

Liu W-t, and Jansson JK (eds) (2010) *Environmental Molecular Microbiology*. Caister Academic Press, Norfolk, U.K.

Liu W-T, Marsh TL, Cheng H et al. (1997) Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. *Appl Environ Microbiol* **63**: 4516-4522.

Loffhagen N, Härtig C, and Babel W (2004) *Pseudomonas putida* NCTC 10936 balances membrane fluidity in response to physical and chemical stress by changing the saturation degree and the trans/cis ratio of fatty acids. *Biosci Biotechnol Biochem* **68**: 317-323.

Lu JR, Santo Domingo JW, Lamendella R et al. (2008) Phylogenetic diversity and molecular detection of bacteria in gull feces. *Appl Environ Microbiol* **74**: 3969-3976.

Macalady JL, Mack EE, Nelson DC et al. (2000) Sediment microbial community structure and mercury methylation in mercury-polluted Clear Lake, California. *Appl Environ Microbiol* **66**: 1479-1488.

Maidak BL, Cole JR, Lilburn TG et al. (2001) The RDP-II (Ribosomal Database Project). *Nucleic Acids Res* **29**: 173-174.

Muyzer G, and Smalla K (1998) Application of Denaturing Gradient Gel Electrophoresis (Dgge) and Temperature Gradient Gel Electrophoresis (Tgge) in Microbial Ecology. *Antonie van Leeuwenhoek International Journal of General & Molecular Microbiology* **73**: 127-141.

National Research Council (1993) *Managing Wastewater in Coastal Urban Areas*. National Academy Press, Washington, D. C.

Nocker A, Burr M, and Camper AK (2007) Genotypic microbial community profiling: A critical technical review. *Microb Ecol* **54**: 276-289.

Osborn AM, and Smith CJ (2005) *Molecular Microbial Ecology*. Taylor & Francis, New York, NY, USA.

Palmer MW (2006). Ordination methods for ecologists: The ordination webpage.

<http://ordination.okstate.edu/>

Pennanen T, Frostegard A, Fritze H et al. (1996) Phospholipid fatty acid composition and heavy metal tolerance of soil microbial communities along two heavy metal-polluted gradients in coniferous forests. *Appl Environ Microbiol* **62**: 420-428.

R Core Development Team (2008) *R: A language and environment for statistical computing*.

Ramette A (2007) Multivariate analyses in microbial ecology. *FEMS Microbiol Ecol* **62**: 142-160.

Rastogi G, Osman S, Vaishampayan PA et al. (2010) Microbial diversity in uranium mining-impacted soils as revealed by high-density 16S microarray and clone library. *Microb Ecol* **59**: 94-108.

Sagaram US, DeAngelis KM, Trivedi P et al. (2009) Bacterial diversity analysis of Huanglongbing pathogen-infected citrus, using PhyloChip arrays and 16S rRNA gene clone library sequencing. *Appl Environ Microbiol* **75**: 1566-1574.

Santo Domingo JW, Bambic DG, Edge TA et al. (2007) Quo vadis source tracking? Towards a strategic framework for environmental monitoring of fecal pollution. *Water Res* **41**: 3539-3552.

Schütte UME, Abdo Z, Bent SJ et al. (2008) Advances in the use of terminal restriction fragment length polymorphism (T-RFLP) analysis of 16S rRNA genes to characterize microbial communities. *Appl Microbiol Biotechnol* **80**: 365-380.

Sercu B, Van De Werfhorst LC, Murray J et al. (2009) Storm drains are sources of human fecal pollution during dry weather in three urban southern California watersheds. *Environ Sci Technol* **43**: 293-298.

Seurinck S, Defoirdt T, Verstraete W et al. (2005) Detection and quantification of the human-specific HF183 *Bacteroides* 16S rRNA genetic marker with real-time PCR for assessment of human faecal pollution in freshwater. *Environ Microbiol* **7**: 249-259.

Shendure J, and Ji HL (2008) Next-generation DNA sequencing. *Nature Biotechnology* **26**: 1135-1145.

Shiu SH, and Borevitz JO (2008) The next generation of microarray research: Applications in evolutionary and ecological genomics. *Heredity* **100**: 141-149.

Shyu C, Soule T, Bent SJ et al. (2007) MiCA: A web-based tool for the analysis of microbial communities based on terminal-restriction fragment length polymorphism of 16S and 18S rRNA genes. *Microb Ecol* **53**: 562-570.

Sigler V, and Pasutti L (2006) Evaluation of denaturing gradient gel electrophoresis to differentiate *Escherichia coli* populations in secondary environments. *Environ Microbiol* **8**: 1703-1711.

Simpson JM, Santo Domingo JW, and Reasoner DJ (2004) Assessment of equine fecal contamination: the search for alternative bacterial source-tracking targets. *FEMS Microbiol Ecol* **47**: 65.

Stogbauer A, Berner Z, and Stuben D (2004) Redox control on the isotopic composition of dissolved sulfate in percolating sewage - An experimental study. *Acta Hydrochim Hydrobiol* **32**: 304-315.

Sunagawa S, DeSantis TZ, Piceno YM et al. (2009) Bacterial diversity and White Plague Disease-associated community changes in the Caribbean coral *Montastraea faveolata*. *ISME J* **3**: 512-521.

ter Braak CJF (1986) Canonical correspondence analysis: A new eigenvector technique for multivariate direct gradient analysis. *Ecology* **67**: 1167-1179.

ter Braak CJF (1995) Ordination. In *Data analysis in community and landscape ecology*. Jongman RHG, ter Braak CJF, and van Tongeren OFR (eds). Cambridge University Press, New York, pp. 91-173.

ter Braak CJF, and Prentice IC (1988) A theory of gradient analysis. . *Advances in Ecological Research* **18**.

ter Braak CJF, and Smilauer P (2002) *CANOCO reference manual and CanoDraw for windows user's guide: Software for canonical community ordination (version 4.5)*. Microcomputer Power, Ithaca, NY, USA.

Thies JE (2007) Soil microbial community analysis using terminal restriction fragment length polymorphisms. *Soil Sci Soc Am J* **71**: 579-591.

Tsiamis G, Katsaveli K, Ntougias S et al. (2008) Prokaryotic community profiles at different operational stages of a Greek solar saltern. *Res Microbiol* **159**: 609-627.

U.S. Environmental Protection Agency (2005) Microbial Source Tracking Guide Document. In. Cincinnatti, OH: Office of Research and Development, p. 150.

Vogel JR, Stoeckel DM, Lamendella R et al. (2007) Identifying fecal sources in a selected catchment reach using multiple source-tracking tools. *J Environ Qual* **36**: 718-729.

Wade TJ, Calderon RL, Sams E et al. (2006) Rapidly measured indicators of recreational water quality are predictive of swimming-associated gastrointestinal illness. *Environ Health Perspect* **114**: 24-28.

Weidhaas JL, Macbeth TW, Olsen RL et al. (2010) Identification of a *Brevibacterium* marker gene specific to poultry litter and development of a quantitative PCR assay. *J Appl Microbiol* **109**: 334 - 347.

White DC (1994) Is there anything else you need to understand about the microbiota that cannot be derived from analysis of nucleic acids? *Microb Ecol* **28**: 163-166.

Wright CF, and Kroese M (2010) Evaluation of genetic tests for susceptibility to common complex diseases: why, when and how? *Hum Genet* **127**: 125-134.

Wrighton KC, Agbo P, Warnecke F et al. (2008) A novel ecological role of the *Firmicutes* identified in thermophilic microbial fuel cells. *ISME J* **2**: 1146-1156.

Wu CH, Sercu B, Van De Werfhorst LC et al. (2010) Characterization of coastal urban watershed bacterial communities leads to alternative community-based indicators. *PLoS One* **Published online on June 23, 2010**. <http://dx.plos.org/10.1371/journal.pone.0011285>.

Yergeau E, Schoondermark-Stolk SA, Brodie EL et al. (2009) Environmental microarray analyses of Antarctic soil microbial communities. *ISME J* **3**: 340-351.

Zelles L (1999) Fatty acid patterns of phospholipids and lipopolysaccharides in the characterisation of microbial communities in soil: A review. *Biol Fertil Soils* **29**: 111-129.

Zhang R, Lau SCK, Ki J-S et al. (2009) Response of bacterioplankton community structures to hydrological conditions and anthropogenic pollution in contrasting subtropical environments. *FEMS Microbiol Ecol* **69**: 449-460.

Zoetendal EG, Collier CT, Koike S et al. (2004) Molecular ecological analysis of the gastrointestinal microbiota: A review. *J Nutr* **134**: 465-472.

Figure Captions

Fig 11.1 Relatedness of the fecal indicator bacteria (FIB, shown in bold), cultured as indicators of microbiological water quality.

Fig 11.2 The “chain of inference” that is the basis for traditional, cultivation-based microbiological water quality monitoring. In **bold**, the inference is: If FIB are present, then waste is also present, and so on. Epidemiology has related FIB directly to disease (right arrow) for some cases. To the left, culture-independent, PCR-based assays (dotted lines) include *Enterococcus*, host-specific markers (e.g. human), and host-specific pathogen genes (e.g. *E. coli*). Culture-independent, DNA-based whole microbial community analysis (far left), is comprehensive, encompassing most indicators along the chain and also bacterial pathogens, depending on the PCR primers employed.

Fig 11.3 Summary of microbial community analysis applied to MST. All bacteria (*Bacteria* and *Archaea*) along with *Eukarya* are recovered in environmental sampling; viruses, including coliphage, are recovered and analyzed separately (right). Bacteria, inclusive of culturable and non-culturable fecal indicators, can be analyzed holistically using total extractable DNA encoding 16S rRNA. The as-conceived associations of qPCR-based approaches assume that PCR primers are specific. By PhyloChip, bacterial pathogens, culturable or not, are automatically captured and can be specifically accounted for. Because pathogens are included, community analysis is conceptually highly-responsive to public health concerns. Not shown, but conceptually similar, are other community analysis approaches such as DGGE, TGGE, and the analysis of membrane fatty acids by PLFA.

Fig 11.4 Diagram to summarize microbial community analysis techniques which differ on the analysis target (**bold**) and target profiling technique (*italic in parentheses*). PLFA: Phospholipid fatty acid (**analysis**), TRFLP: Terminal restriction fragment length polymorphism, RFLP: Restriction fragment length polymorphism, ARDRA: Amplified ribosomal DNA restriction analysis, LH-PCR: Length heterogeneity polymerase chain reaction, ARISA: Automated ribosomal intergenic spacer analysis, DGGE: Denaturing gradient gel electrophoresis, TGGE: Temperature gradient gel electrophoresis, SCCP: Single strand conformation polymorphism.

Fig 11.5 PhyloChip sample processing schematic.

Fig 11.6 DCA plot of TRFLP profiles from un-spiked and spiked reference water samples (filled circles) from AB watershed and potential fecal sources (open stars). Fecal sources are denoted by S, C, and D for septic solids, cat feces, and dog feces, respectively. Water samples are denoted with a “w” followed (if spiked) by the source(s). Numeric numbers following the septic solids (“S”) source indicate dose of spiking (e.g., 001,01,1, and 5 indicate 0.01%, 0.1%, 1%, and 5%). “L” and “H” indicate a low and a high dose of spiking.

Fig 11.7 DCA plot of TRFLP profiles from water samples (numbers) from Arroyo Burro watershed and sewage samples (open stars). Water samples are denoted by increasing numbers from downstream to upstream (e.g. “1” for ocean, “2” for the Arroyo Buroo Lagoon, and “9” for the most upstream urban drain).

Fig 11.8 Nonmetric multidimensional scaling (NMDS) plot of Mission Creek Watershed samples, using the Jaccard distance measure. Stress = 8.14. Fecal samples grouped separately from ocean, lagoon and creek samples.

Fig 11.9 The relative richness of 4 bacterial classes for each of the sample types. The number in parentheses represent the total richness (number of OTUs) detected. Fecal samples have higher relative richness of *Clostridia*, *Bacteroidetes*, *Bacilli*, and lower *Alphaproteobacteria* than the other three sample types.

Fig 11.10 DCA joint plot to illustrate multivariate analysis on multiple indicator data, displaying both samples (numeric numbers) and indicators (italic characters). Samples are labeled as in Fig 11.9. Indicators are salinity, DO (dissolved oxygen), pH, TC (total *Coliform*), EC (*E. coli*), ENT (*Enterococcus*), and HBM (human *Bacteroides* marker). All indicator values are normalized to mean across all samples before DCA.

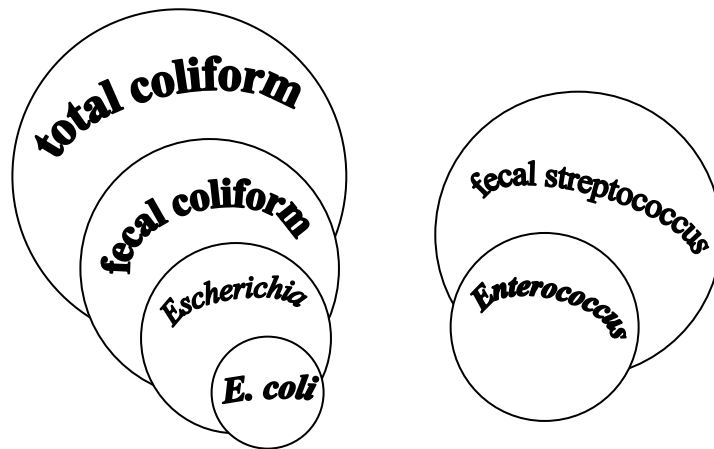


Fig 11.1 Relatedness of the fecal indicator bacteria (FIB, shown in bold), cultured as indicators of microbiological water quality.

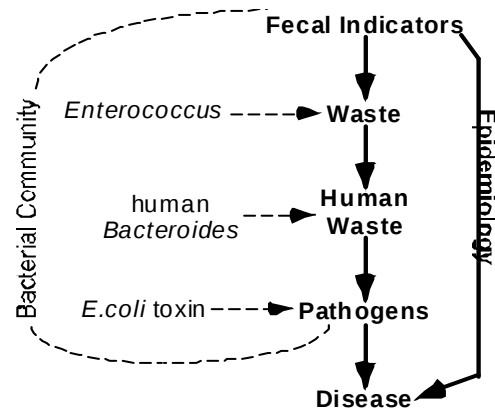


Fig 11.2 The “chain of inference” that is the basis for traditional, cultivation-based microbiological water quality monitoring. In **bold**, the inference is: If FIB are present, then waste is also present, and so on. Epidemiology has related FIB directly to disease (right arrow) for some cases. To the left, culture-independent, PCR-based assays (dotted lines) include *Enterococcus*, host-specific markers (e.g. human), and host-specific pathogen genes (e.g. *E. coli*). Culture-independent, DNA-based whole microbial community analysis (far left), is comprehensive, encompassing most indicators along the chain and also bacterial pathogens, depending on the PCR primers employed.

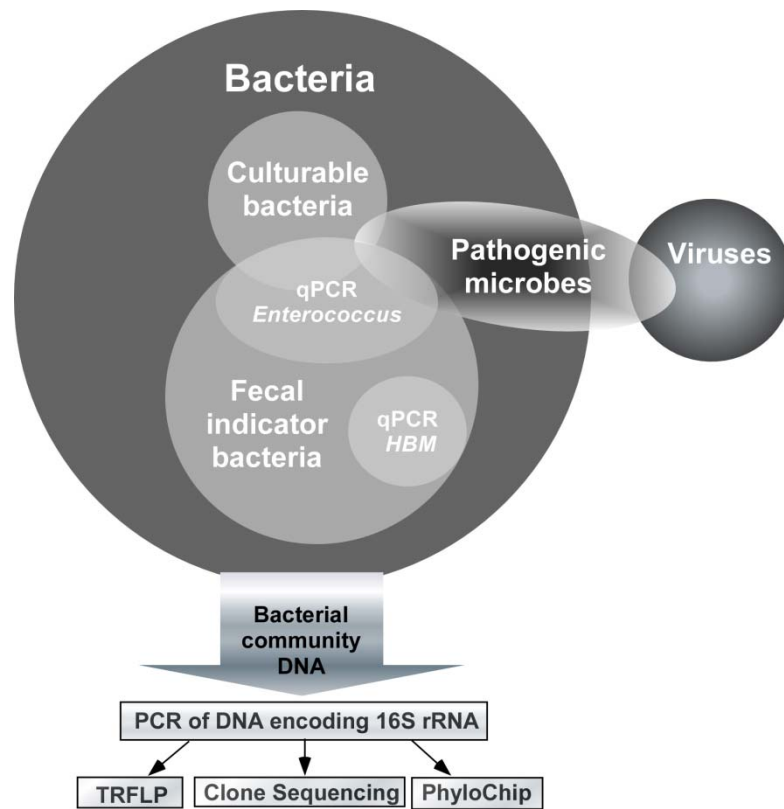


Fig 11.3 Summary of microbial community analysis applied to MST. All bacteria (*Bacteria* and *Archaea*) along with *Eukarya* are recovered in environmental sampling; viruses, including coliphage, are recovered and analyzed separately (right). Bacteria, inclusive of culturable and non-culturable fecal indicators, can be analyzed holistically using total extractable DNA encoding 16S rRNA. The as-conceived associations of qPCR-based approaches assume that PCR primers are specific. By PhyloChip, bacterial pathogens, culturable or not, are automatically captured and can be specifically accounted for. Because pathogens are included, community analysis is conceptually highly-responsive to public health concerns. Not shown, but conceptually similar, are other community analysis approaches such as DGGE, TGGE, and the analysis of membrane fatty acids by PLFA.

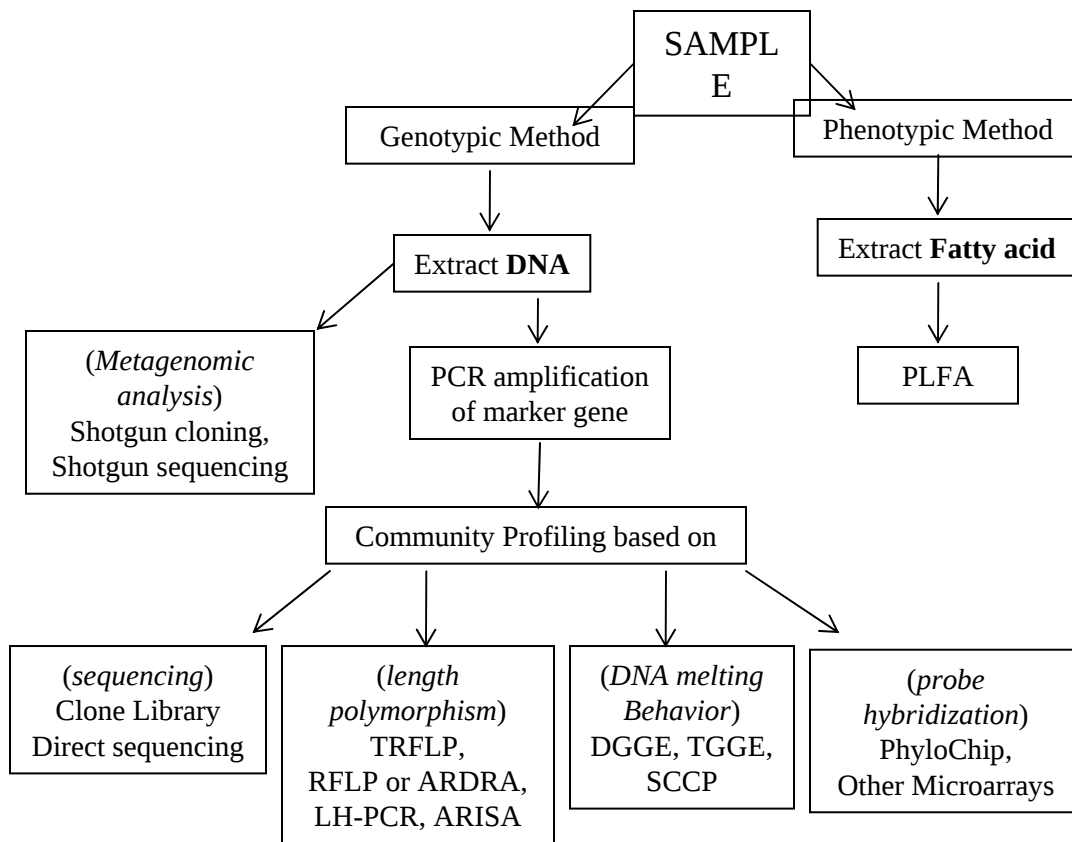


Fig 11.4 Diagram to summarize microbial community analysis techniques which differ on the analysis target (**bold**) and target profiling technique (*italic in parentheses*). PLFA: Phospholipid fatty acid (analysis), TRFLP: Terminal restriction fragment length polymorphism, RFLP: Restriction fragment length polymorphism, ARDRA: Amplified ribosomal DNA restriction analysis, LH-PCR: Length heterogeneity polymerase chain reaction, ARISA: Automated ribosomal intergenic spacer analysis, DGGE: Denaturing gradient gel electrophoresis, TGGE: Temperature gradient gel electrophoresis, SCCP: Single strand conformation polymorphism.

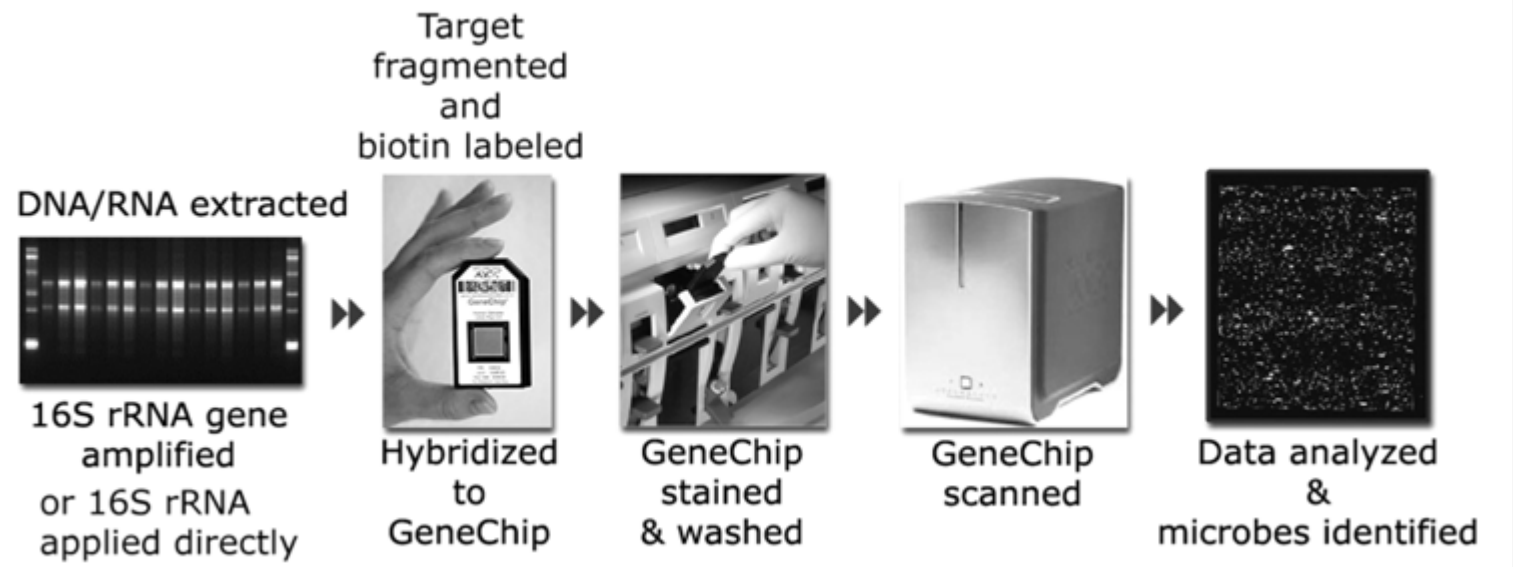


Fig 11.5 PhyloChip sample processing schematic.

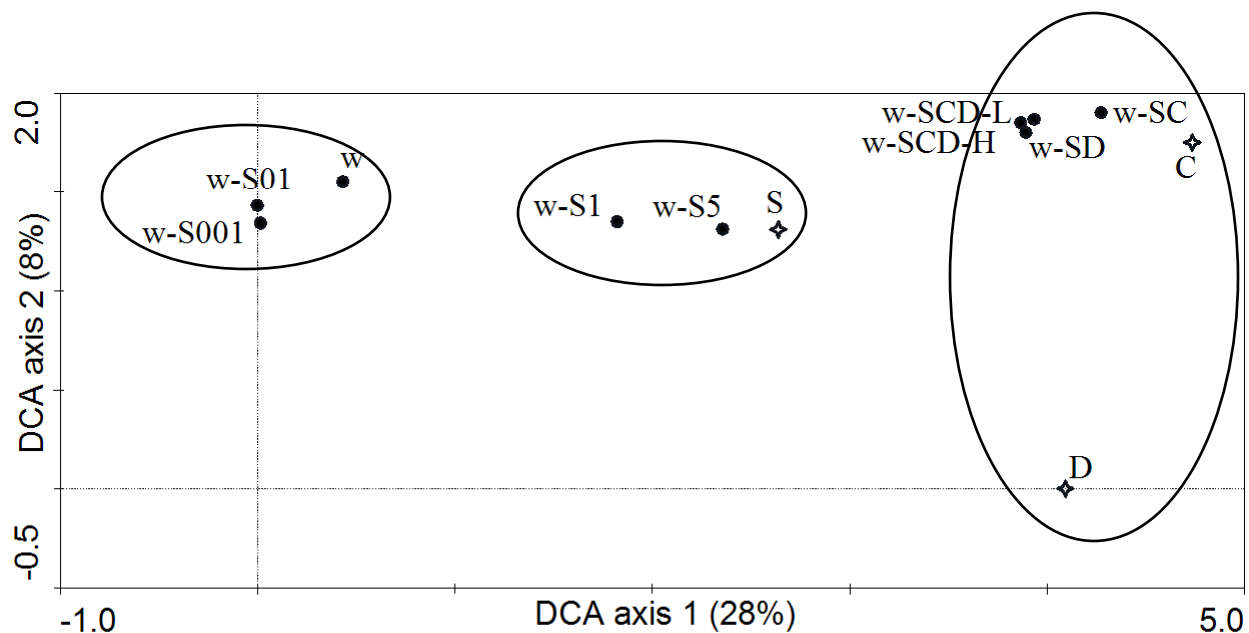


Fig 11.6 DCA plot of TRFLP profiles from un-spiked and spiked reference water samples (filled circles) from AB watershed and potential fecal sources (open stars). Fecal sources are denoted by S, C, and D for septic solids, cat feces, and dog feces, respectively. Water samples are denoted with a “w” followed (if spiked) by the source(s). Numeric numbers following the septic solids (“S”) source indicate dose of spiking (e.g., 001,01,1, and 5 indicate 0.01%, 0.1%, 1%, and 5%). “L” and “H” indicate a low and a high dose of spiking.

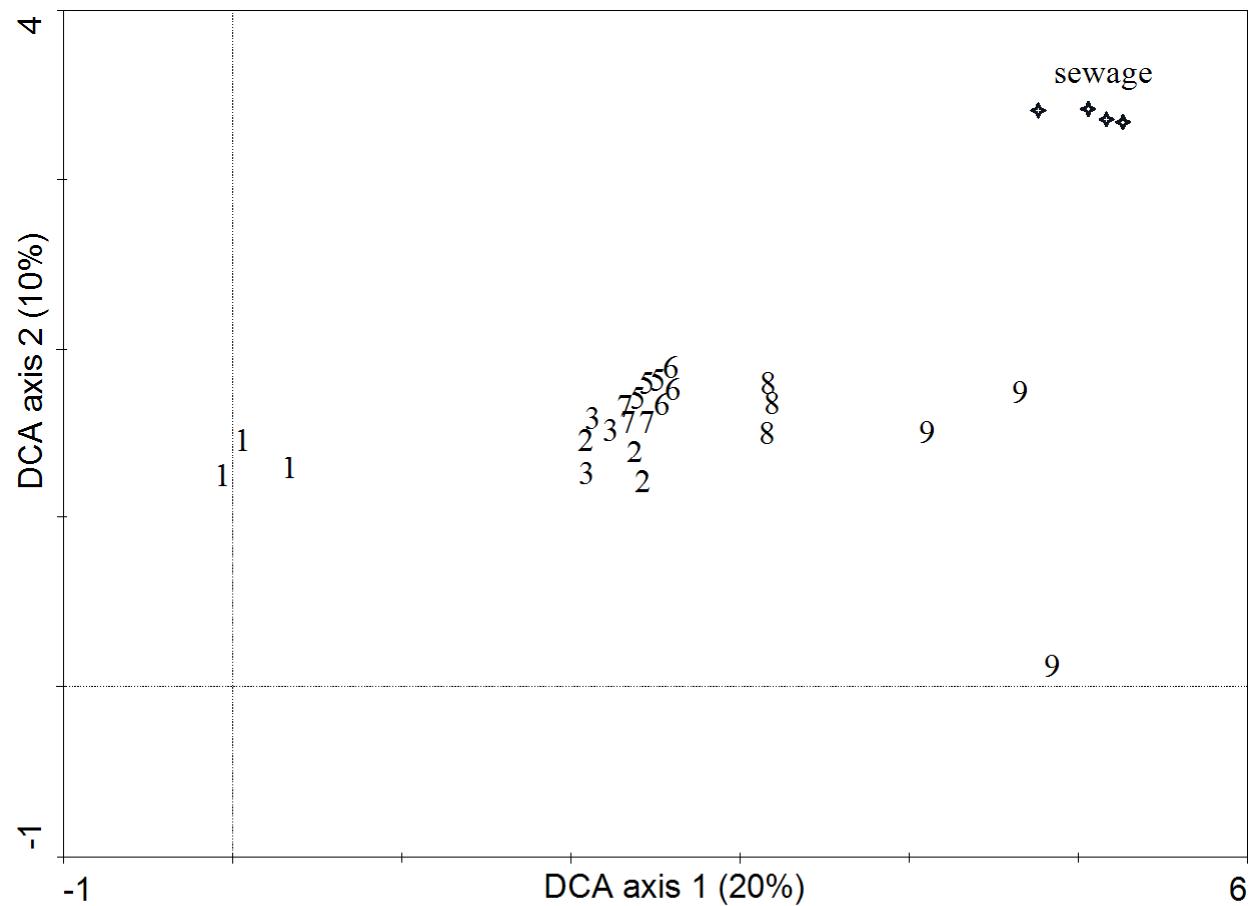


Fig 11.7 DCA plot of TRFLP profiles from water samples (numbers) from AB watershed and sewage samples (open stars). Water samples are denoted by increasing numbers from downstream to upstream (e.g. “1” for ocean, “2” for the AB Lagoon, and “9” for the most upstream urban drain).

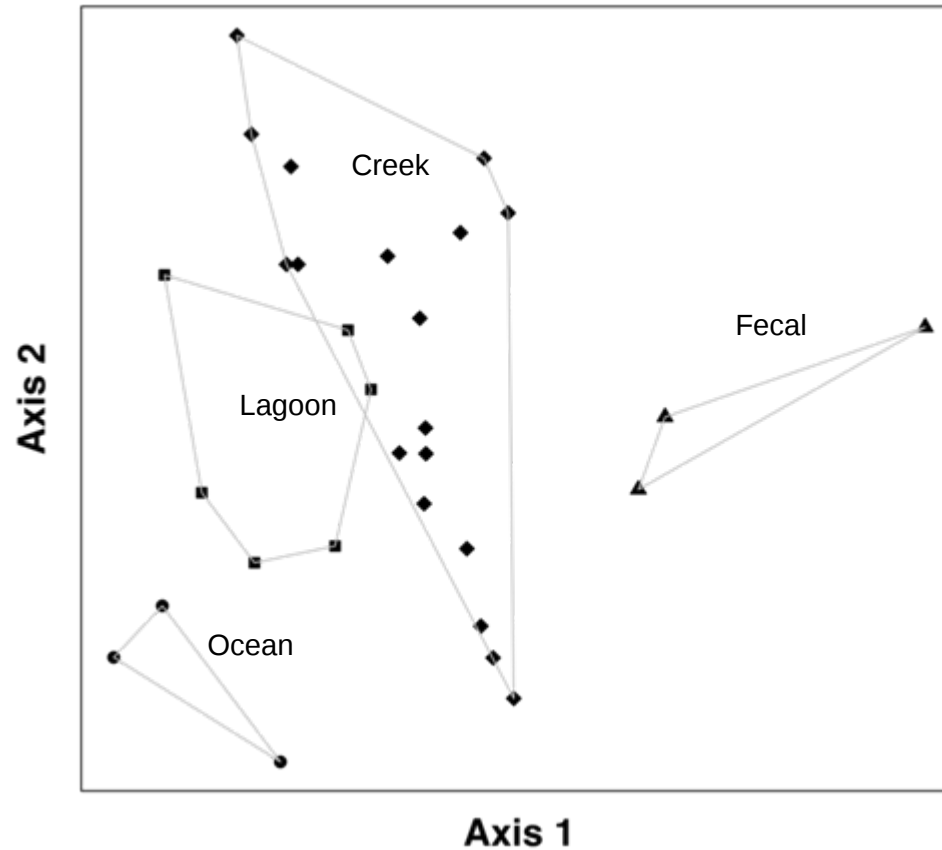


Fig 11.8 Nonmetric multidimensional scaling (NMS) plot of Mission Watershed samples, using the Jaccard distance measure. Stress = 8.14. Fecal samples grouped separately from ocean, lagoon and creek samples.

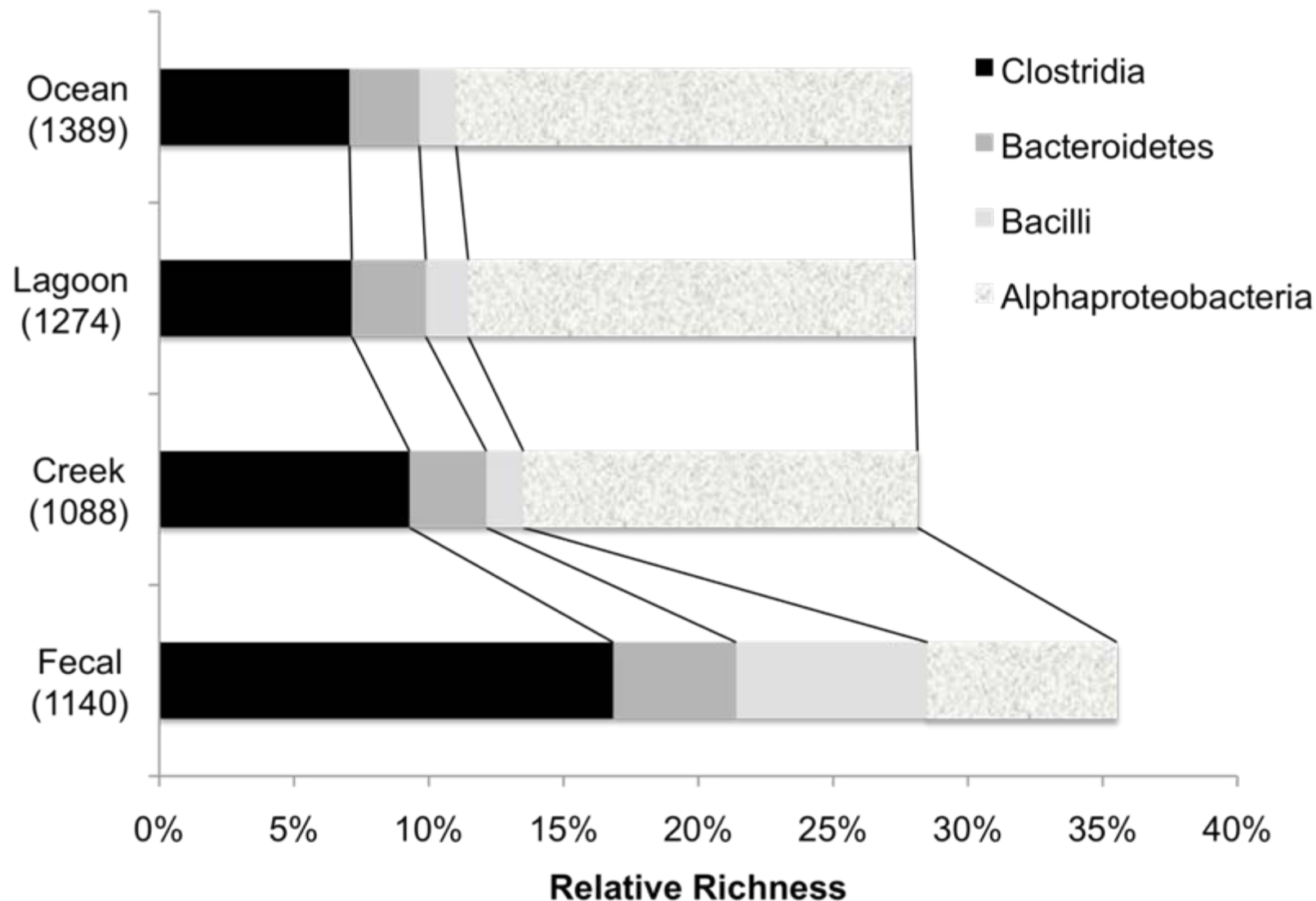


Fig 11.9 The relative richness of 4 bacterial classes for each of the sample types. The number in parentheses represent the total richness (number of OTUs) detected. Fecal samples have higher relative richness of *Clostridia*, *Bacteroidetes*, *Bacilli*, and lower *Alphaproteobacteria* than the other 3 sample types.

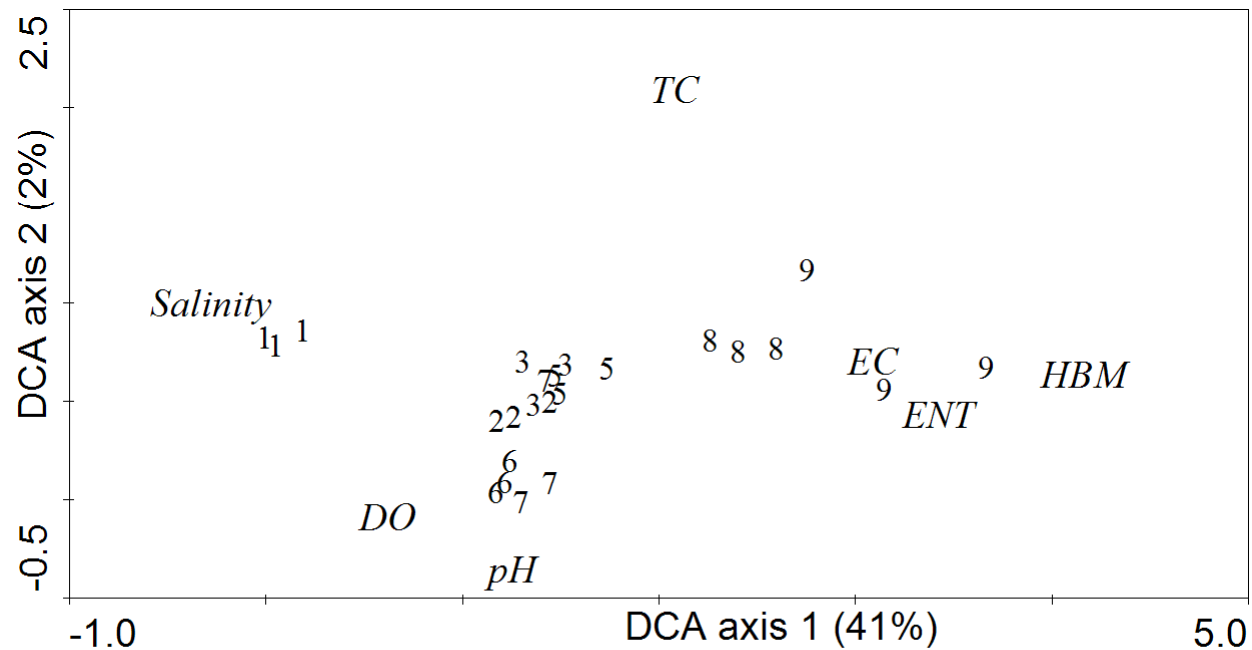


Fig 11.10 DCA joint plot to illustrate multivariate analysis on multiple indicator data, displaying both samples (numeric numbers) and indicators (italic characters). Samples are labeled as in Figure 9. Indicators are salinity, DO (dissolved oxygen), pH, TC (total *Coliform*), EC (*E. coli*), ENT (*Enterococcus*), and HBM (human *Bacteroides* marker). All indicator values are normalized to mean across all samples before DCA.