Chapter 13

# Nucleic Acid-Based Methods of Analysis

Emily B. Hollister, John P. Brooks and Terry J. Gentry



When describing their discovery of the double-stranded structure of DNA, Watson and Crick (1953) famously wrote, "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material." This understated revelation not only enabled elucidation of the DNA replication process, but also ultimately provided the foundation for the plethora of molecular biology-based methods that have been developed over the past several decades. With these new approaches, microbiologists are now able to use a small sample of microbial nucleic acid to identify microorganisms, track genes, and evaluate genetic expression in the environment. Furthermore, the extraction of nucleic acids directly from environmental samples (soil, water, air), followed by nucleic acid-based analysis, has helped to overcome some of the biases of culture-based assays that require isolation and growth of organisms in the laboratory before they can be studied (see Case Study 4.2). These technologies include a variety of different methods such as: polymerase chain reaction (PCR)-based assays, microarrays, DNA sequencing and metagenomics. Given the breadth and rapid development of nucleic acid-based technologies over the past few decades, it is impossible to cover all of these methods in one textbook let alone one book chapter. Therefore, the goal of this chapter is to present some of the most commonly used nucleic acid-based methodologies, with a focus on their theoretical foundations and applications for detecting and characterizing environmental microorganisms.

## 13.1 STRUCTURE AND COMPLEMENTARITY OF NUCLEIC ACIDS

Molecular analyses are dependent on the type and sequence of the nucleic acids. Nucleic acids are in one of

I.L. Pepper, C.P. Gerba, T.J. Gentry: Environmental Microbiology, Third edition. DOI: http://dx.doi.org/10.1016/B978-0-12-394626-3.00013-2

two forms, either deoxyribonucleic acid (DNA) or ribonucleic acid (RNA). Nucleic acids consist of a phosphate sugar backbone, where the 5′ carbon atom of one sugar is covalently linked to the 3′ carbon atom of the adjacent sugar. The difference between the two forms is a single base substitution and the presence or absence of a hydroxyl group in the 3′ carbon of the sugar. DNA is made up of four deoxynucleotide bases: guanine (G); cytosine (C); adenine (A); and thymine (T), while uracil (U) substitutes for thymine in RNA. For convenience, the nucleotide bases are generally referred to by their single letter abbreviations. The bases are divided based on their chemical composition and structure into either purine (G and A) or pyrimidine bases (C, T, U). In DNA, the sugar is deoxyribose, while in RNA the sugar is ribose. As will be discussed shortly, the presence of a hydroxyl group on the 3′ carbon of deoxyribose plays a critical role in replication of DNA.

Structurally, DNA consists of two strands of deoxyribonucleic acids combined together to form a double helix. One strand of DNA is oriented 5′ to 3′, while the complementary strand is oriented 3′ to 5′. These two strands are linked by hydrogen bonds between corresponding pairs of bases. Specifically, G binds only to C, and A binds only to T. For each G-C pairing, there are three hydrogen bonds, whereas for the A-T pairing there are only two hydrogen bonds (Figure 13.1). Because these bases are specific in their ability to bind together to form a base pair (bp), they are said to be complementary to each other. Single-stranded pieces of DNA (ssDNA) form a double-stranded DNA helix (dsDNA) if it comes in contact with another ssDNA molecule that has a complementary sequence. For example, a DNA strand containing the nine bases 5′-A-T-T-C-G-G-A-A-T-3′ will anneal to its complementary strand 3′-T-A-A-G-C-C-T-T-A-5′ with the resulting dsDNA being nine base pairs in length. Ideally, the two strands of a DNA molecule will be 100% complementary; however, there are cases where mismatches or incorrect base matching occurs which is stabilized by the hydrogen bonds of the surrounding correctly paired bases.

Recall that most bacteria have a single circular DNA chromosome that is in excess of one million base pairs (see Section 2.2.3). This chromosome encodes all of the genetic information that makes each bacterial population unique. For this information to be of use it must be replicated and then translated into RNA and eventually transcribed into structural proteins or enzymes that carry out all the activities of the cell (see Information Box 13.1). In DNA replication, one strand of DNA serves as a template for the synthesis of its complementary strand. The enzyme responsible for replication, DNA polymerase, adds a nucleotide complementary to the template strand to the free hydroxyl at the 3′ carbon of the last sugar on the growing DNA chain, and thus DNA is synthesized only in the 5′ to 3′ direction. The double-stranded, complementary nature of DNA not only forms the basis of its replication within an organism, but is also the foundation

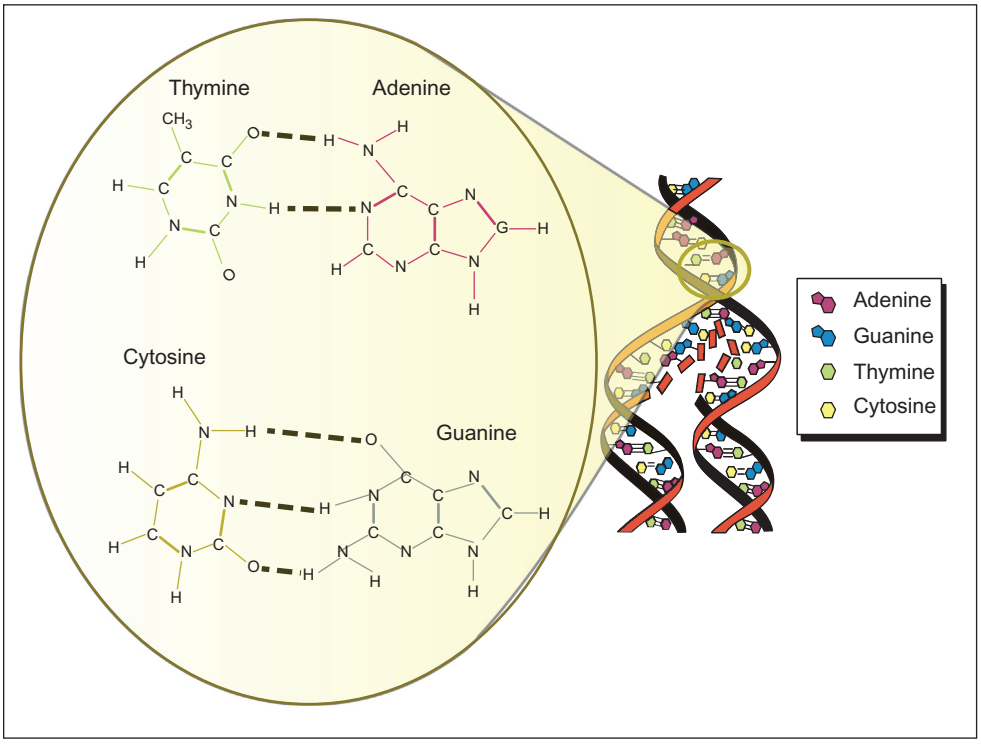


FIGURE 13.1 Hydrogen bonding between purine and pyrimidine bases in double-stranded DNA.

**Information Box 13.1 Transcription and Translation**

$$\text{DNA} \xrightarrow{\text{transcription}} \text{RNA} \xrightarrow{\text{translation}} \text{protein}$$

How does the cell use the information stored in its chromosome? It must transfer the information into useful activity. This is done in two steps, transcription and translation. In transcription, a DNA sequence is transcribed by the enzyme RNA polymerase to produce an RNA copy that is complementary to the DNA. During translation, the three different types of RNA (mRNA, tRNA, and rRNA) are used to synthesize a protein that is based on an mRNA sequence. In this process, a specific amino acid is inserted into a growing polypeptide chain based on the sequence of the mRNA being translated. Specific tRNA molecules deliver each amino acid to the ribosome, which is constructed of proteins and rRNA and is the site where translation or protein production occurs.

of several nucleic acid-based molecular analyses that will be discussed later, including gene probes and the polymerase chain reaction (PCR).

In RNA, G still binds to C, but A binds to U. There are three forms of RNA:

- ribosomal RNA (rRNA), which plays a structural role in creating the ribosome, the site of protein translation
- messenger RNA (mRNA), which is transcribed from a coding region of DNA and is subsequently translated into protein
- transfer RNA (tRNA), which transfers specific amino acids to the ribosome during mRNA translation

Two types of RNA, rRNA and mRNA, are used extensively in molecular genetic analyses to understand phylogenetic relationships, and to evaluate metabolic activity, respectively.

## 13.2 OBTAINING MICROBIAL NUCLEIC ACIDS FROM THE ENVIRONMENT

The first step in nucleic acid-based methods is often the extraction of nucleic acids from environmental samples. This step is critical for the downstream success and validity of the subsequent analyses. In this process, DNA is concurrently extracted from all of the populations within a sample, generating a mixture of DNA referred to as community DNA. The most common approach to extraction of community DNA from soil is to lyse the bacterial cells *in situ* (direct lysis) (see Section 8.1.3.1). Community DNA is ideally representative of all populations within the sample community; however, in reality, the extraction efficiency of different types of microorganisms can vary widely causing biases in subsequent analyses. Similar extraction approaches can be used to obtain community RNA for analysis, although there are additional challenges associated with extraction of community RNA from soil, including active RNases and the inherently short half-life of mRNA. Consequently, it is more difficult to assess microbial gene expression in soil. Community RNA extractions generally parallel those of *in situ* lysis of DNA, with the extra step of adding an RNA stabilization agent and/or snap-freezing immediately after sampling. As with soil, microorganisms from either water or air samples can be analyzed via molecular analysis following extraction and purification of DNA/RNA from the collected microbial biomass (see Sections 8.2 and 8.3).

Regardless of the source of biomass or the lysis method, it is often necessary for the nucleic acid extract to undergo one or more purification steps to remove coextracted impurities (e.g., humic acids, metals, carbohydrates), which could interfere with or completely inhibit subsequent analyses. Purification methods include cesium chloride density centrifugation, commercial purification kits employing spin filters and traditional phenol–chloroform extraction followed by ethanol precipitation (see also Chapter 8). It is highly advised that the quantity and quality of extracted nucleic acids be determined (e.g., via UV spectrophotometry) prior to use, and the extract purified as necessary for the desired downstream application. However, the number of purifications performed should be minimized as much as possible since large portions of the extracted nucleic acids may be lost during the purification process. In addition, the presence of humic residues, metals and other inhibitory materials from environmental samples can often still interfere with downstream analyses even after careful purification. Finally, it should be noted that community DNA/RNA extracts may also contain nonmicrobial DNA/RNA originating from larger organisms such as insects and plants in the extracted sample. This may result in misleading concentrations of nucleic acids extracted from samples that contain substantial amounts of nonmicrobial biomass, such as a plant root and its associated microbial community.

Another limitation of nucleic acid extraction methods is that DNA/RNA recovery can vary from sample to sample depending on sample characteristics (e.g., for a soil the clay and salt content) and extraction efficiency. Not surprisingly, this variation can be further magnified when researchers use different methods to extract nucleic acids. This may not be a major issue for individual studies where all of the samples are extracted and processed in the same manner, but it is potentially a major confounding factor for comparison of results across experiments and between research groups, such as in large, collaboration-based projects. In order to address this issue, researchers are attempting to establish international standards for nucleic acid extraction from soils and other matrices (Petric *et al.*, 2011). Once nucleic acids are

obtained from either cultured isolates or environmental samples, they can be analyzed with a variety of methods including those discussed in the following sections.

## 13.3 HYBRIDIZATION-BASED ASSAYS

Gene probe methodology takes advantage of the fact that DNA can be denatured and reannealed (Information Box 13.2). Gene probes consisting of single-stranded DNA can be used to identify the presence of a particular nucleic acid sequence within an environmental sample. Typically, probes are short sequences of DNA known as oligonucleotides, which are complementary to the target sequence of interest. These probes are labeled in some way that facilitates their detection. Probes can be used for a variety of environmental applications including: (1) examination of soil microbial diversity; (2) identification of a particular genotype; and (3) testing for virulence genes of suspected pathogens. Construction of gene probes and specific application methodologies are outlined in detail in manuals such as Green and Sambrook (2012). Here we briefly present the general concepts involved in the construction of gene probes along with some of their practical applications. These concepts are also critical to understanding some of the more sophisticated techniques discussed later, including microarrays.

**Information Box 13.2 Hybridization and Denaturation**

When two complementary DNA strands combine together, the process is known as DNA–DNA hybridization, because the resulting dsDNA is a hybrid of the two separate strands. The reverse process, in which dsDNA melts into two single strands, is called denaturation. This can be done chemically or simply by heating the DNA to 94°C. On cooling, the two single strands automatically hybridize back into a double-stranded molecule, a process known as reannealing. Complementarity is an important concept because if one strand sequence is known, the sequence of the other strand is easily deduced. This concept is the basis of many of the nucleic acid–based methodologies discussed in this chapter.

### 13.3.1 Marker Selection

Gene probes can be designed to target various functional or phylogenetic genes (Table 13.1) depending upon the objective of the study. The targeted sequence may be unique to a particular microbial species, in which case the gene probe would allow screening of an environmental sample for the presence of that microorganism. Alternatively, the target gene may code for the production of an enzyme unique to a metabolic pathway. In this case, positive gene probe results indicate that the environmental sample contains the genetic potential for that particular activity. This kind of probe can be defined as a functional gene probe. A good example of such a probe is one designed to be complementary to genes coding for enzymes involved in nitrogen fixation. It is important to

**TABLE 13.1 Commonly Used Marker Genes**

| Target Type | Examples | Information Obtained |
|---|---|---|
| Ribosomal RNA (rRNA) or internal transcribed spacer (ITS) | 16S rRNA gene, the small ribosomal subunit (SSU) in Bacteria and Archaea | Taxonomy and phylogeny of Bacteria and Archaea |
| | 18S rRNA gene, the small ribosomal subunit in eukaryotes | Taxonomy and phylogeny of Fungi and other eukaryotes |
| | ITS, the internal transcribed spacer | Taxonomy of Fungi |
| Other universal or "housekeeping" genes | *rpoB*, an RNA polymerase subunit | Phylogenetic information, similar to that obtained from rRNA or ITS sequences |
| | *gyrB*, a DNA gyrase | |
| | *recA*, DNA recombination and repair | |
| | HSP70, a heat shock protein | |
| Functional genes | *amoA*, *nifH*, *ntcA* | Detection of microorganisms involved in various nitrogen transformations |
| | *dsrAB* | Detection of sulfate-reducing Bacteria and Archaea |
| | *phnA*, *phnI*, *phnJ* | Detection of microorganisms involved in phosphorus transformations |

**TABLE 13.2 Selected Nucleic Acid Databases**

| Database and Web Address | Contents and Comments |
|---|---|
| GenBank® (National Center for Biotechnology Information, NCBI) http://www.ncbi.nlm.nih.gov/ | Annotated database of publicly available DNA sequences, various software tools for analyzing genome data, funded by National Institutes of Health |
| ENA, European Nucleotide Archive http://www.ebi.ac.uk/ena/ | Europe's primary resource for nucleotide sequence and annotation |
| DDBJ, DNA Databank of Japan http://www.ddbj.nig.ac.jp/ | Database of nucleic acid sequences generated in Japan |
| INSDC, International Nucleotide Sequence Database Collaboration http://www.insdc.org/ | Collaboration between GenBank, EMBL and DDBJ, which allows for sharing of data between member archives |
| Ribosomal Database Project (RDP) http://rdp.cme.msu.edu/ | Quality-controlled ribosomal sequence data and analysis services, including sequence classification and alignment of bacterial and archaeal 16S rRNA gene and fungal 28S rRNA gene sequences |
| UNITE http://unite.ut.ee/ | Quality-controlled molecular database and tools for the annotation and identification of fungi using ITS sequences |
| Genomes OnLine Database (GOLD) http://www.genomesonline.org/ | Comprehensive listing of completed and ongoing genomes with links to sequence information |
| JGI, Joint Genome Institute http://www.jgi.doe.gov/ | Archive for large-scale sequencing projects generated by the U.S. Department of Energy, plus tools for comparative analysis of genomes and metagenomes |
| Kyoto Encyclopedia of Genes and Genomes (KEGG) http://www.genome.jp/kegg/ | Annotated genes and metabolic pathways within organisms |

keep in mind that since DNA probes are being used, only the genetic potential and not expression of the gene (requires mRNA detection) or the actual activity (requires enzyme activity assays) is being detected. Probes designed against specific ribosomal RNA (rRNA) sequences are known as phylogenetic probes. Phylogenetic probes can be specific for groups of bacteria, for example Proteobacteria or even classes of Proteobacteria (e.g., $\alpha$, $\beta$, $\gamma$) or can be designed to detect an entire domain (Bacteria, Archaea or Eukarya) in which case they are called universal probes.

## 13.3.2 Probe Construction and Detection

The basic strategy in the construction of a gene probe, or PCR primer, is to obtain the sequence of a target gene and then select a portion of this sequence for use as a probe. Nucleic acid sequences can be obtained from a variety of databases (Table 13.2), with the U.S. National Institute of Health's genetic sequence database GenBank® and the European EMBL database being two of the largest and most commonly consulted. Several other smaller databases with particular foci are also available. The number of sequences resident within databases has expanded significantly each year (Figure 13.2A). The vast number of sequences currently available facilitates design of target-specific primers and probes. In addition, thousands of whole microbial genomes have been sequenced, serving as the foundation for a new area of research based on comparisons of whole genomes or comparative genomics (Figure 13.2B).

Once a target sequence is obtained, it and other sequences for the same target in other organisms are imported into another program, where the sequences are aligned. This step is critical since it allows identification of conserved and unique regions, which can be subjected to algorithms to design probes that meet appropriate criteria such as length, melting temperature ($T_m$) and minimal secondary structure. The size of the probe can range from $\approx$20 base pairs to as many as several hundred base pairs. A large number of probes for various organisms are publicly available. One of the best sources for rRNA-targeted probes is the probeBase database which currently contains information for $>$2700 probes and $>$170 PCR primers (Loy *et al.*, 2007).

Once a probe is selected, it is synthesized and labeled in such a way that it can be detected after hybridization

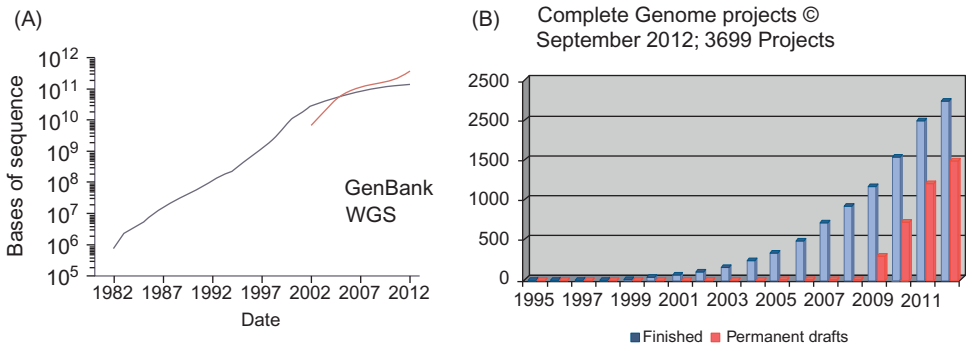


FIGURE 13.2 (A) Growth of the sequence data in GenBank and the Whole Genome Shotgun (WGS) databases. These databases are maintained by the National Center for Biotechnology Information, U.S. National Library of Medicine and are part of the International Nucleotide Sequence Database Collaboration. (B) Sequenced microbial genomes to 2012.

to the target sequence. A researcher has several options when constructing and labeling a probe. In the early days of molecular biology, the only labeling option was radioactivity, but several nonradioactive methods are now available. Radioactive labeling of a probe is typically done by labeling the sequence with a radioactive chemical, such as $^{32}$P. Nonradioactive alternatives include probes labeled with digoxigenin (DIG), biotin or fluorescein, which can be incorporated into the sequence by chemical synthesis. The different labels are detected by binding the respective antibody or streptavidin–alkaline phosphate conjugate, which, when reacted with the appropriate substrate, will give a signal (see also Chapter 12).

### 13.3.3 Dot, Southern and Northern Blots

Dot blotting or dot hybridization is a technique used to evaluate the presence of a specific nucleic acid sequence in a microbial culture. Nucleic acid is extracted from the culture, spotted on a nitrocellulose filter and subsequently probed. Dot blots can be used to indicate the presence or absence of a sequence, or can be used to quantify the target. For quantitation, the relative amount or intensity of hybridization will give an estimate of the quantity of sequences in a sample when compared with similarly spotted standards. Gene probes can also be used to detect a specific gene sequence within bacterial colonies on a Petri plate containing a mixed population of bacteria through use of a process termed colony hybridization or lifts. To perform a colony hybridization, a nylon membrane is lightly pressed onto the Petri plate so that some bacterial cells from each colony adhere to the membrane. Subsequently, in a series of steps, the cells are lysed directly on the membrane, and the DNA is fixed to the filter and then denatured into two single strands. The gene probe is similarly denatured and then added to the membrane. The DNA is allowed to reanneal, and ideally the single strand of the gene probe will anneal with the complementary target DNA sequence from the bacterial cells. After washing the membrane free of unhybridized probe, the filter undergoes probe detection. After this procedure, only the colonies that contain the specific DNA sequence give a signal. Since the original Petri plate contains the corresponding intact colonies, the viable colony of interest can now be identified, isolated and retained for further study.

In some cases, it may be necessary first to separate the target DNA or RNA into different size fractions before probing. For example, it may be important to know whether a gene is carried on a plasmid or the chromosome. To determine this, all of the plasmids within the microbe being studied can be extracted and separated by gel electrophoresis (Information Box 13.3; Figure 13.3). The plasmid DNA is then transferred onto a nylon membrane by blotting, and the membrane subsequently probed. Only the DNA molecules that contain the target sequence hybridize with the probe, thus allowing detection of those plasmids that contain the target sequence. When the target is DNA, as in the above case, this process is known as Southern blotting or hybridization. Similarly, Northern blotting, which detects RNA, can be used in gene expression studies to detect induction of a specific gene.

### 13.3.4 Microarrays

Scientists have modified and miniaturized the blotting process to construct "gene chips" or microarrays that enable simultaneous detection of thousands to even millions of different gene targets. Microarrays are basically a collection of oligonucleotides (or gene probes) that have been "arrayed" onto a glass slide or a chip. In contrast to the hybridizations discussed in Section 13.3.3, the probes are bound to the solid matrix, and the target nucleic acids are then hybridized to the bound probes. Depending upon the array design and target nucleic acids, microarrays can

### Information Box 13.3 Agarose Gel Electrophoresis

Agarose gel electrophoresis is a fundamental tool in nucleic acid analysis. It is a simple and effective technique for viewing and sizing DNA molecules such as plasmids or DNA fragments, including PCR products. The DNA samples are loaded into wells in an agarose gel medium. Voltage is applied to the gel, causing the DNA to migrate toward the anode because of the negatively charged phosphates along the DNA backbone. The gel is stained with a dye such as ethidium bromide, allowing visualization of the DNA when viewed under ultraviolet (UV) light. The smaller DNA fragments migrate faster through the gel matrix, and the larger fragments migrate more slowly. The molecular weight of the DNA (base pairs [bp]) determines the rate of migration through the gel and is estimated from comparison with standards (i.e., molecular weight markers) of known sizes that are run in parallel on the gel (Figure 13.3).

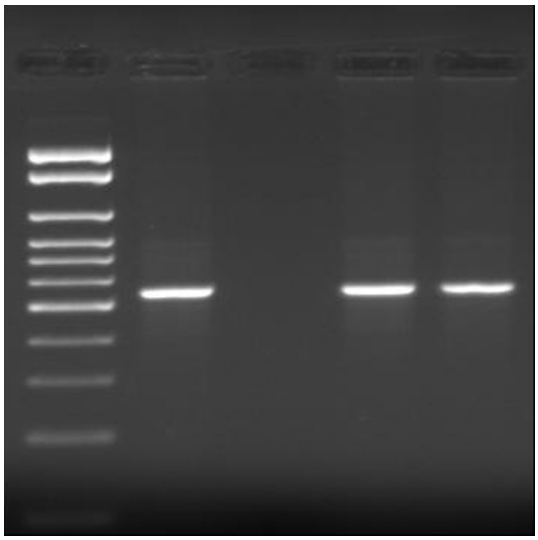

FIGURE 13.3 Agarose gel electrophoresis of PCR-amplified product DNA. Lane M includes a size ladder and lanes "+" and "−" are the positive and negative controls. Lanes 1 and 2 are samples that are positive for the target DNA as shown by presence of the ≈550 bp PCR product. Photo courtesy E.C. Martin.

be used to study the presence (DNA), abundance (DNA) and/or activity of microorganisms (mRNA). Although large arrays were initially cost-prohibitive for many laboratories, they are becoming widely used due to technological advances that have decreased the costs associated with probe synthesis, microarray fabrication and template labeling. Furthermore, the rapidly increasing availability of sequence information, including whole genome sequences, makes microarray approaches an enticing option not only for investigating the transcriptome (all genes expressed by an organism; Chapter 21) and subsets of the transcriptome targeted to specific functions, but also for the screening of environmental samples for expression of particular genes of interest by any member (s) of the microbial community.

Two types of microarray fabrication processes are commonly used: printed (or spotted) arrays and synthesized (*in silico*) arrays. There are two common ways to produce the probes for printed arrays. For printed cDNA arrays, genes of interest ranging in size from ≈0.6 kb to 2.4 kb are amplified using the polymerase chain reaction (PCR) to produce cDNA probes (see Section 13.4.5 for explanation of cDNA). For printed oligonucleotide arrays, the sequences of the genes of interest (in some cases, all genes in the genome) are used to design unique 35−70 nucleotide probes. These oligonucleotide probes are then synthesized commercially. Once the probes have been produced, they are robotically deposited onto the microarray slide and subsequently chemically linked to the slide. Some manufacturers synthesize their oligonucleotides directly on the microarray chip in a process called *in silico* synthesis. This involves the use of photolithography and solid-phase DNA synthesis, and has enabled the construction of high-density arrays containing over one million probes.

Microarrays can also vary based upon their targeted organisms and information provided (Table 13.3). The two types of microarrays most commonly used for environmental microbiology are phylogenetic oligonucelotide arrays (POAs) and functional gene arrays (FGAs). The POAs are primarily based upon rRNA (e.g., 16S rRNA) genes, and are used to detect the presence of organisms. This allows a researcher to follow population dynamics and community profile changes across a wide variety of species on the same array. The largest POA currently available is the PhyloChip developed by Gary Andersen's group at Lawrence Berkley National Laboratory. The PhyloChip G2 contains probes representing 9773 clusters of sequences (operational taxonomic units; OTUs) within Bacteria and Archaea, and also probes at higher levels of taxonomic classification (e.g., subfamilies). A newer, commercial version of the PhyloChip (G3) is currently available via Second Genome (San Bruno, CA). The PhyloChip has been used to characterize microbial communities in a variety of samples including urban air samples (Brodie *et al.*, 2007), air in commercial aircraft (Korves *et al.*, 2013), the Gulf of Mexico following the *Deepwater Horizon* oil spill (Hazen *et al.*, 2010), mine tailings (Wakelin *et al.*, 2012) and uranium-contaminated groundwater (Brodie *et al.*, 2006).

In contrast to POAs, FGAs are designed based upon genes directly tied to an environmentally relevant process (e.g., sulfate reduction). This enables researchers not only to use FGAs to detect the presence and potential activity of organisms which may contribute to processes of interest, but also as a measure of microbial activity if mRNA is assayed for the target sequences. The most

**TABLE 13.3 Microarrays for Characterizing Environmental Microorganisms**

| Array | Targeted Microorganisms | Information Provided | Probe Template |
|---|---|---|---|
| Phylogenetic oligonucleotide array | Cultured & uncultured | Phylogenetic | rRNA genes |
| Functional gene array | Cultured & uncultured | Functional | Functional genes |
| Community genome array | Cultured | Phylogenetic | Whole genome |
| Whole-genome open-reading-frame array | Cultured | Phylogenetic & functional | Open reading frames in whole genome |
| Metagenomic array | Cultured & uncultured | Functional | Environmental DNA |

Adapted from *Gentry* et al. *(2006)*.

**TABLE 13.4 Microbial Genes Represented on the GeoChip 3.0 Functional Gene Microarray**

| Gene Category | Number of Probes | Covered Sequences |
|---|---|---|
| Carbon cycling | 5196 | 10,573 |
| Nitrogen cycling | 3763 | 7839 |
| Phosphorus cycling | 599 | 1220 |
| Sulfur cycling | 1504 | 2042 |
| Organic contaminant degradation | 8614 | 17,441 |
| Metal resistance | 4870 | 10,962 |
| Antibiotic resistance | 1594 | 3944 |
| Energy processes | 508 | 671 |
| Phylogenetic marker (*gyrB*) | 1164 | 2298 |
| Total | 27,812 | 56,990 |

Adapted from *He* et al. *(2010)*.

comprehensive FGA currently available is the GeoChip developed by Jizhong Zhou's group at the University of Oklahoma. GeoChip 3.0 contains 27,812 probes covering 56,990 sequences—primarily representing genes involved in C, N and S biogeochemical cycling, organic chemical degradation and metal resistance (Table 13.4). The GeoChip has been used to characterize microbial communities in a variety of samples including grassland soil (He *et al.*, 2010), reclaimed surface-mine soil (Ng, 2012), the Gulf of Mexico following the *Deepwater Horizon* oil spill (Lu *et al.*, 2012) and remediation of uranium-contaminated groundwater (Van Nostrand *et al.*, 2011).

Other types of arrays include: (1) community genome arrays that use whole genomes of isolated organisms for probe material and can be used to determine the similarity of tested isolates to these strains; (2) whole-genome open-reading-frame arrays that contain probes for all potential genes in a genome and can be used for transcriptomics and comparative genomics; and (3) metagenomic arrays that use cloned environmental DNA as templates for probes and can be used as a high-throughput method to screen these libraries (Section 13.6.2).

Once constructed, microarrays can be employed to characterize microbial communities and determine how microorganisms respond to and interact with their environment. Figure 13.4 shows the steps in a microarray experiment in which the goal is to characterize the microbial community via detection of target sequences (e.g., 16S rRNA or functional genes). The first step is to extract DNA from the sample. If necessary, the DNA can be further purified and/or amplified (Section 13.4). Subsequently, the DNA is labeled, usually with a fluorescent molecule such as Cy3 or Cy5. The labeled DNA is then denatured and hybridized to the probes on the microarray, after which the array is scanned to indicate the extent of hybridization to the various probes. A cutoff value, such as a signal-to-noise ratio of 3.0, is commonly used to differentiate background noise from true target detection. These results are then compared to those for hybridization of other samples to ascertain community diversity, presence of microorganisms or genes of interest, and/or population dynamics.

Additionally, microarrays can be used to determine microbial activity. In this case, mRNA is extracted from two samples: the test or experimental sample which has been exposed to a treatment (e.g., an environmental pollutant) and a reference or control sample that has not been exposed to the treatment. Messenger RNA (mRNA) is extracted from both samples, converted to cDNA via reverse transcription (Section 13.4.5), labeled with two different molecules (e.g., Cy3 and Cy5) and hybridized to the

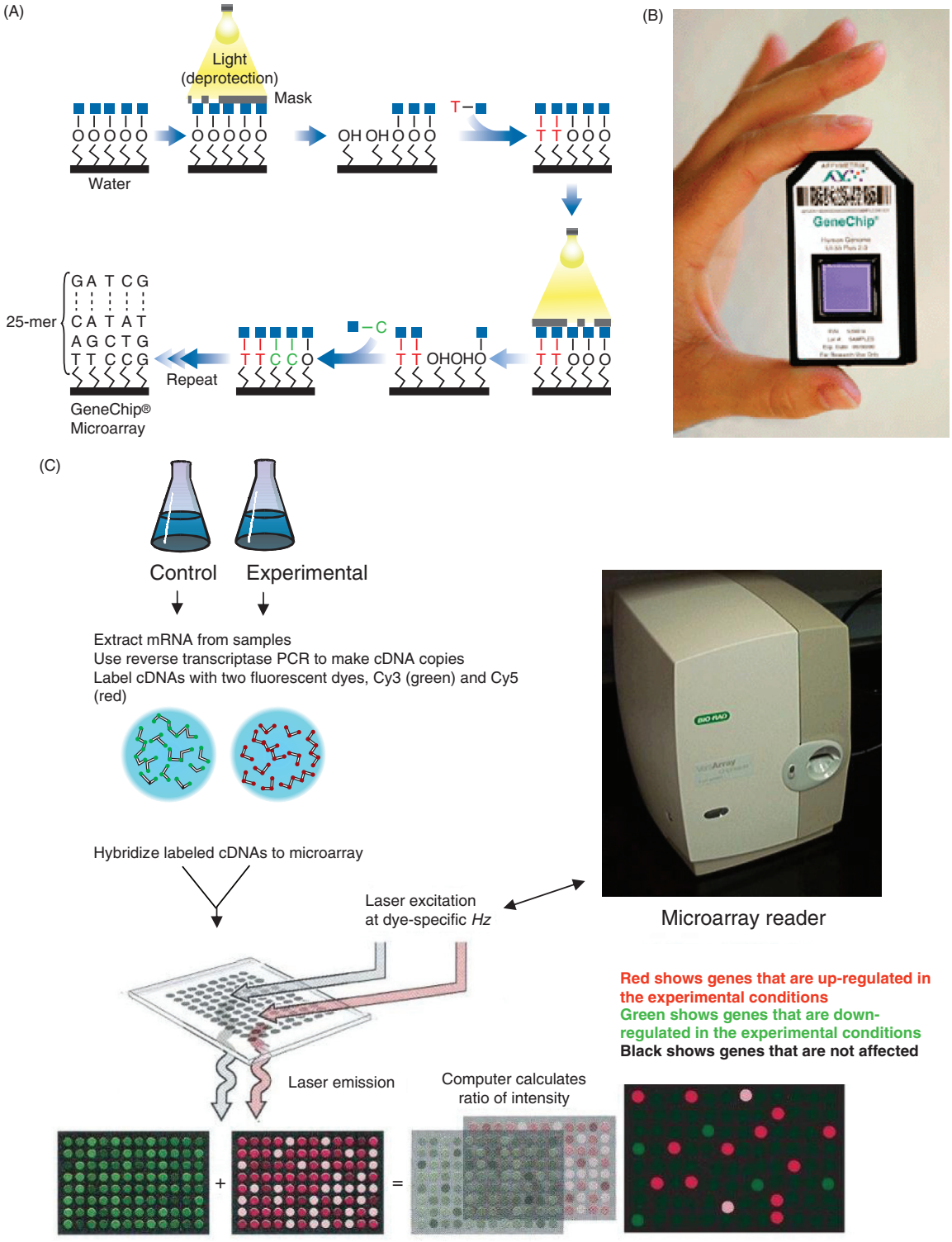


FIGURE 13.4 Steps in a microarray experiment. DNA is isolated from a sample and, if necessary, further purified and/or amplified. The resulting DNA is then labeled, denatured and hybridized to probes on a microarray. The array is then scanned to detect and quantify the binding of different probes on the array to the target DNA. The availability of different labels allows simultaneous testing of two or more different samples. This dual-labeling approach is commonly used for gene-expression studies where mRNA is extracted from a control and treated sample. The mRNA extracts are converted to cDNA and then labeled with different fluorescent labels, e.g., Cy3 and Cy5. The up- or down-regulation of genes in the treatment sample is then determined by comparison to hybridization in the control sample.

same array. The array is then scanned at the appropriate wavelength for each label, and the up- or down-regulation of each gene in the treatment sample is determined by comparison to hybridization in the control sample.

Although microarrays are powerful analytical tools, they face several challenges when being used to characterize environmental microorganisms (Van Nostrand *et al*., 2013). First, only targets with high enough similarity or homology to the probes will bind. This often results in a reduced sensitivity of detection. For example, when samples such as soil community DNA are analyzed, a reduced detection limit may result, presumably due to target and probe sequences being diverse and consequently not highly homologous to the designed probes. Due to these issues, microarrays are currently only capable of detecting dominant members (≥5%) of microbial communities. This can also be an issue if only small quantities of DNA or mRNA are available for microarray analysis. Target nucleic acids can be amplified prior to microarray analysis, but this can introduce additional artifacts and further complicate analysis (Wang *et al*., 2011; Section 13.4).

Although sensitivity can be a challenge, specificity is perhaps the greatest issue with use of microarrays for environmental samples. In contrast to using microarrays with pure cultures where the entire genome sequences of the tested organisms are often known and probes can thus be designed to reduce or eliminate the potential for cross-hybridization, little is known regarding the DNA sequences present in most environmental samples. Therefore, even though microarray probes are usually designed carefully and screened against all relevant sequences in public databases such as GenBank, these sequences may only represent a tiny fraction of the sequences present in a given sample (e.g., while there may be millions of sequences in GenBank for soil microorganisms collected from around the world, they may only represent a small portion of the microorganisms in a forest soil from Huntsville, TX, U.S.A.). Due to this, it is impossible to say that a positive microarray hybridization reflects true detection of that target and not cross-hybridization with an unknown sequence in the sample, although other techniques such as qPCR and DNA sequencing can be used to verify microarray detection of specific targets, and reduce this limitation. Additionally, although some researchers have shown that microarrays can be quantitative, amplification biases and potential cross-hybridization can make the quantitative detection of target sequences difficult when working with environmental samples.

### 13.3.5 Fluorescence *In Situ* Hybridization (FISH)

Probes with fluorescent labels can be used to investigate cells *in situ*, i.e., in a culture or in an environmental sample in a technique called fluorescence *in situ* hybridization (FISH). In this case, reagents are added that facilitate penetration of the probe through the cell membrane, where the probe ultimately hybridizes to its target sequence. Cells containing the target sequence are visualized under a fluorescence or confocal microscope. This technique has the advantage of allowing visualization of spatial relationships between populations within a community to be elucidated (see Section 9.4.3 and Figure 9.14). Since FISH analyses are conducted *in situ*, an inherent difficulty of this technique is differentiating signal from background noise such as probe nonspecifically sorbed to soil particles or the presence of naturally fluorescing compounds including metals and some soil minerals.

## 13.4 AMPLIFICATION-BASED ASSAYS

### 13.4.1 The Polymerase Chain Reaction (PCR)

The primary methodology used for most amplification-based assays is the polymerase chain reaction (PCR). Developed by Kerry Mullis and associates in 1985 (Saiki *et al*., 1985), PCR has now been available for three decades and, along with its various iterations, has become the "gold standard" assay in microbiology. PCR has become so synonymous with microbiology that the majority of readers of this textbook will not have known a time prior to its use; it is the cell phone of microbiology.

PCR harnesses the biological DNA replication process to produce multiple copies (up to millions in a single reaction) of a DNA sequence *in vitro*. The amplified DNA is then detected using gels or other fluorescence-based techniques and can be used in further applications such as cloning and DNA sequencing. Before PCR, environmental microbiologists relied on culture and physiological assays to unravel the microbiological mysteries of soil, water and other environmental matrices. The two greatest advantages of PCR are that it enables researchers to amplify a DNA target millions of times over and also eliminates the need to isolate and culture microorganisms before they can be detected or studied. Both of these features have greatly enhanced the detection and characterization of environmental microorganisms and revolutionized our understanding of the microbiological realm, far beyond what culture-based assays alone allowed. Furthermore, a wide variety of modified PCR assays have been developed that enable detection of RNA targets, measurement of microbial activity, differentiation of live and dead cells and quantitative detection of microorganisms. In the following sections, we will discuss the basic process of PCR along with several modified PCR assays and their applications to environmental microbiology.

### 13.4.2 The Steps of PCR

The basic PCR reaction consists of three major steps: (1) DNA melting; (2) primer annealing; and (3) DNA polymerization or elongation. These steps comprise a single PCR amplification cycle and are often repeated 30–40 times in order to amplify a target DNA sequence to produce millions of copies or amplicons (Information Box 13.4; Figure 13.5). The first step relies on adding sufficient heat to melt or separate the antiparallel DNA strands into two distinct molecules. Essentially, this is a process that has been lifted from cell biology; whereas in a cell, helicase is used to separate the strands, heat is used to separate the strands in the amplification tube. This process is usually carried out at 94–95°C for between 10 and 45 seconds. The second step of the amplification cycle involves single-stranded, ≈20 nucleotides (nt)-long DNA primers, which ideally would be complementary to the ends of the DNA target. One primer binds to the 3′ end of one of the separated DNA strands, while the other primer binds to the 3′ end of the other, antiparallel strand. The primers bind to the 3′ ends because DNA polymerase requires a free 3′ OH on the primer to polymerize DNA down the template DNA molecule in the opposite direction. This all occurs at a single temperature typically ranging from 45 to 60°C, depending upon the composition of the primers (i.e., the melting point of the primers). Ideally, the temperature is 2–4°C below the melting temperature of the primers. In theory, the higher the temperature at which binding occurs, the more specific the binding will be, though this can vary depending upon the primer and assay. It is important to state that higher temperature binding may also limit assay sensitivity. This step is highly specific and may be the most important step throughout the entire process. Primers are either highly specific or slightly degenerate (Information Box 13.5). When the primers are specific, the primer acts as a mirror opposite, and each nucleotide matches its mirrored partner on the DNA template. Other times, such as with 16S rRNA PCR, the primers are degenerate, meaning they are able to bind varying configurations of the primer binding site. In this case, the primers may be degenerate in up to ≈10% of their nt, which enables binding to variations of the primer site. This leads us to the third and final step of the amplification cycle, DNA polymerization. There are many DNA polymerases available, each specializing in specific duties (Information Box 13.6). Here we will focus on the typical DNA polymerase and amplification. Efficient DNA polymerization was not possible until *Taq* polymerase was discovered. *Taq*

**Information Box 13.4 The Polymerase Chain Reaction (PCR)**

Although advances in reagents and thermocyclers now allow polymerase chain reaction (PCR)-based amplifications to be performed much more rapidly, the basic idea behind PCR has not changed since its advent in the mid-1980s. The general PCR approach is as follows:

**Stage 1: DNA melting-hold step.** This is the process by which the DNA duplex strands are separated into single-stranded DNA molecules awaiting amplification. The energy added to the process separates the strands by breaking the hydrogen bonds between the individual nucleotides of opposite strands. The step is typically performed at 94–95°C for 5–10 min. Additionally, this step may be used to activate, or "turn on," some DNA polymerases that are designed to require this activation in order to prevent amplification by-products from being formed before the desired cycling conditions are initiated.

**Stage 2: Cycling.** This is the part of a PCR assay where DNA amplification occurs. This stage is broken into a three-step process (denature, anneal and extension), often repeated between 25 and 45 cycles, depending on the needs of the researcher (Figure 13.5). The denaturing step follows a similar process to the stage 1 hold step, the only difference is that the denaturing cycling step is typically conducted at 95°C for only 15–30 s. As before, this step is used to separate or melt the DNA strands, to prepare for the following step, the annealing step. In the annealing step, primers (small ≈20 bp DNA molecules) bind to the separated DNA strands. The temperature at which this is performed is primer dependent, and can vary from 45 to 65°C. Lower annealing temperatures generally result in less specific binding, and thus increase the potential for nontarget amplification. The third and final step in the cycle is the extension step, in which the DNA polymerase synthesizes the new DNA strand beginning from the 3′ end of the bound primer. During this step, DNA polymerase adds new nucleotides to the extending DNA molecule following the guidelines of the original DNA template. Typical conditions for this step are 72°C for approximately 60 s, depending on the size of the expected product and the DNA polymerase.

**Stage 3: DNA final extension–hold steps.** The final stage consists of a final DNA extension, as well as a final hold step at 4°C. The final extension step consists of DNA polymerase extending unfinished amplification products. This is then followed by a final hold step in which the thermocycler is set to hold a 4°C temperature to prevent DNA degradation until the user can proceed to further analysis of the PCR products.

In the end, the amplification proceeds in an exponential fashion, in which $X_n = 2^n * X_0$: where $X_n$ = the number of amplicons after $n$ cycles, beginning with $X_0$ DNA template copies.

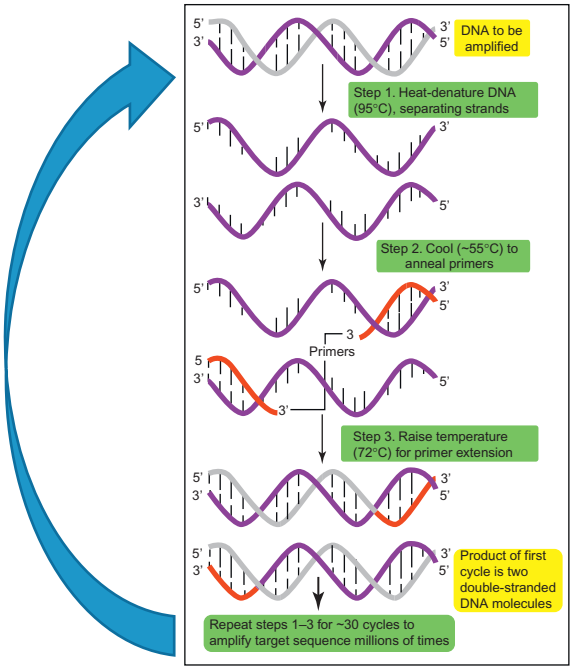


**FIGURE 13.5** The PCR cycle. Step 1, denaturation of the DNA template; Step 2, annealing of primers to the single-stranded template; Step 3, extension of the primer to make a complementary copy of the DNA template. These three steps comprise one cycle of PCR.

### Information Box 13.5 What Does The 'N' in My PCR Primer Sequence Mean? – Degenerate Primers

If DNA is formed from the nucleotides A, T, G and C, what do the other letters commonly seen in PCR primers and probes (e.g., M, W and N) represent? These symbols are part of the International Union of Pure and Applied Chemistry (IUPAC) standard nomenclature for nucleotide base sequence names and represent degenerate or nonstandard nucleotides that can bind multiple nucleotides. For example, an "M" in a PCR primer would consist of a mixture of A and C at that position, and thus could bind to either T or G in the complementary DNA strand. An "N" is the most extreme example and reflects of mixture of A, T, G and C at that position in the PCR primer, which would thus enable binding to any nucleotide at that position in the complementary DNA strand. These types of nucleotides are commonly used to design degenerate PCR primers or probes that thus enable detection of a wider spectrum of microorganisms by a single primer set or probe.

### Information Box 13.6 Not all DNA Polymerases are Created Equal

There are a many different DNA polymerases available from a variety of sources. These enzymes can vary greatly with respect to their sensitivity to reaction contaminants, their optimal reaction temperature, the type and length of template they will amplify, the speed of amplification, and the accuracy of the resulting amplicons. When choosing a DNA polymerase, it is crucial to consider the specific enzyme properties needed for each specific application. For example, many DNA polymerases, including *Taq*, add a single "A" to the 3′ ends of PCR products. This is invaluable for cloning PCR products using the TA cloning process (Section 13.6.1). However, some DNA polymerases produce blunt-ended products. If a researcher wants to clone PCR amplicons produced with these enzymes using a TA cloning system, the amplicons must first be "A"-labeled via a single amplification step with an enzyme such as *Taq* DNA polymerase.

polymerase is a heat-stable DNA polymerase from the thermophilic bacterium *Thermus aquaticus*. Prior to its discovery, the DNA polymerases used in PCR were labile enzymes which required constant addition throughout the PCR process. In addition to being heat stable, *Taq* polymerase can achieve processing rates of anywhere from 35 to >100 nt per second. This enables assays to occur in as little as 45 min (e.g., real-time PCR), though longer products require longer polymerization. This final step in the PCR cycle normally occurs at 72°C. The three steps in a PCR cycle are then repeated between 25 and 40 times, depending on the end-point of the assay. It is important to note that there are other steps involved in the PCR process. For instance, PCR often begins with a single initial melting step of 95°C for 10 min; this process serves two purposes, to activate a particular type of *Taq* DNA polymerase and to separate the initial DNA templates. In addition, following the final cycle of PCR, the user will often include a final elongation step, which serves to elongate the ends of any PCR product not finished during the previous 40-cycle PCR. Theoretically, a single copy of DNA could be amplified to produce $>10^{12}$ amplicons following 40 PCR cycles (Information Box 13.4); however, this level of amplification is typically not achieved under real-world conditions. For more information about generic PCR processes and steps, the reader is encouraged to view resources such as the DNA Learning Center website (http://www.dnalc.org/resources/animations/pcr.html) hosted by Cold Spring Harbor Laboratory.

Upon completion of the PCR assay, the results obtained are typically plus/minus (presence/absence), meaning that the gene or target of interest has either been detected in the template DNA or it has not; PCR amplicons are typically observed and compared using a gel (Figure 13.3). Here the products are loaded onto an agarose gel to which an electric current is applied allowing the negatively charged DNA molecules to migrate towards the positive end. The amplicons can be compared to positive and negative controls as well as a molecular weight marker. This marker provides a

means to compare multiple product sizes, normally from 100 to 1500 bp; the amplicons, of different sizes, migrate along with the electric current and settle out at precise locations, which enables the user to determine sizes in comparison to the marker bands.

In the early days of PCR, thermal cycling occurred in water baths with a user cycling tubes from one water bath to the next water bath; needless to say, this was a time-consuming and labor-intensive effort. The commercialization of PCR brought thermocyclers into the standard laboratory; thermocyclers are now capable of housing hundreds of individual reaction tubes in a single self-contained instrument, capable of rapidly and precisely altering and holding temperatures for exact time points. Newer thermocyclers allow for the use of smaller reagent amounts; often reactions can be completed in a $<25$ μl-reaction vessel. Many of these apparatuses also enable users to operate multiple temperatures on a single block, also known as gradient cycling, which allows for primer troubleshooting at various temperatures. Gradient cyclers are especially useful in the early stages of PCR assay development, and enable the user to test multiple temperatures corresponding to a single primer set to determine which temperature/primer combination produces the best amplification. Gradient cyclers enable the user to run completely different PCR assays, with completely different PCR primers all functioning at various temperatures. For instance, one row on the block could correspond to 16S rRNA PCR, while another row corresponds to *nosZ* PCR, and, finally, another to *uidA* PCR.

Another specialized thermocycler, one which has quickly become the "gold standard" in most environmental microbiological laboratories, is the real-time PCR thermocycler (aka quantitative PCR). This thermocycler is capable of quickly amplifying and reporting product results. The advantage of the real-time thermocycler is that no gel is needed post-run while simultaneously obtaining quantitative results. The use of this thermocycler and its accompanying PCR assays will be discussed in the section below.

### 13.4.3 PCR Assays

The basic PCR assay has been modified via changes to primer targets, cycling conditions and template material to greatly expand the utility of PCR-based assays. Quantitative PCR (qPCR), multiplex PCR, reverse transcriptase PCR (RT-PCR), integrated cell culture PCR (ICC-PCR) and enterobacterial repetitive intergenic consensus sequence-PCR (ERIC-PCR) are just a few of the variations of the traditional PCR approach available to the environmental microbiologist.

Perhaps the most important recent addition to the PCR repertoire is qPCR (Information Box 13.7). Quantitative PCR has enabled researchers to harness the sensitivity of PCR, but with the quantitative characteristics of culture-based approaches. The basic premise behind qPCR consists of using fluorescent reporter dyes to detect the amount of newly amplified DNA. Reporter dyes are detected following each reaction cycle, and when the fluorescent signal is plotted against a standard curve, the reaction can be quantified. Until the advent of qPCR, the PCR assay was, for the most part, limited to presence/absence or qualitative (semi-quantitative) assays. In this respect, qPCR has enabled environmental microbiologists to measure and quantify microorganisms that may otherwise be difficult or even impossible to identify by culture methodology.

Quantitative PCR approaches primarily use one of two chemistries for quantification, either TaqMan® or SYBR Green®; the former accommodates a number of fluorescent dyes on an additional labeled internal DNA probe, while the latter is more cost-effective but can only utilize SYBR Green as the fluorophore (reporter signal) (Figure 13.6). An assay with SYBR Green requires an additional step in the form of a melt-curve step. This step applies heat, from 65 to 90°C, to separate the newly amplified DNA strands; a reduction in the fluorescent signal indicates the melting temperature at which DNA denaturation occurs. This final step indicates to the researcher the specificity of the amplified product, provided that it matches with the positive controls. The TaqMan assay relies on the specificity provided by the internal fluorescent probe; this probe binds a small ($\approx 15$ bp) region of DNA, which is internal to the two primer-binding sites. The inclusion of this internal probe in TaqMan-based assays greatly increases specificity as compared to SYBR Green-based approaches. Both chemistries yield quantitative results, provided a standard curve with known DNA quantities is used (Figure 13.7). More information about qPCR can be found in references including Heid *et al.* (1996) and Dowd and Pepper (2007).

Multiplex PCR is another effective manipulation of the traditional PCR assay. This method enables the detection/measurement of multiple targets in a single reaction assay (Figure 13.8), something that few culture-based approaches can achieve. Multiplex PCR primers can be added to a PCR assay with multiple targets; in some cases gene targets belong to completely different taxonomic groups. In this respect, the user could add PCR primers corresponding to *Campylobacter* spp., 18S rRNA (e.g., eukaryotic DNA) and tetracycline resistance in a single PCR reaction, provided that the primers do not exhibit any cross-reactivity. Typically, though, the reactions address multiple genes of a single pathogen, or varying taxonomic hierarchal genes such as a reaction targeting the family Enterobacteriaceae, the genus *Escherichia* and toxin-specific primers corresponding to *E. coli* O157:H7. The multiplex PCR assay was first adapted to gel approaches where one would read a gel with multiple bands in a single lane, each corresponding to a specific gene of interest. This approach has since been modified to fit with the

**Information Box 13.7 Quantitative PCR (qPCR)**

The most commonly used qPCR chemistries are SYBR Green® and TaqMan® (Figure 13.6). SYBR Green involves the use of a fluorescent dye which, when bound to double-stranded DNA can be excited and fluoresce. In this case the fluorescent signal is non-specific, meaning that any dsDNA will fluoresce, including primer-dimers and other non-specific PCR products. SYBR Green is more cost-effective than other qPCR chemistries and in some ways easier to carry out, since only two primers, a polymerase and dye, are needed for the reaction (in addition to basic PCR reagents). TaqMan assays involve the use of *Taq* polymerase and exploits its 5′ nuclease activity. TaqMan assays use an upstream and downstream primer, as in any PCR, with the addition of an internal probe which binds between the two primer-binding sites. This probe is labeled with a fluorescent dye as well as a nonfluorescent quencher; when both are in close proximity, the fluorescent dye is quenched and cannot be seen. TaqMan uses its 5′ nuclease activity, through routine DNA amplification, to cleave this probe, thus liberating the fluorescent dye, and enabling visualization and subsequent quantification. TaqMan assays can be modified to accommodate multiplex PCR, since multiple dyes are available, and each primer/probe combination could accommodate an individual gene of interest. SYBR Green, on the other hand, is not as conducive to multiplex assays due to the difficulty in differentiating multiple products.

The quantitative capability of qPCR is enabled by the generation of a standard curve from different concentrations of target sequences. The standard curve is used for both SYBR Green and TaqMan-based chemistries; the use of the curve enables comparison between samples and known quantities of DNA standards. In Figure 13.7A, DNA standards decrease in quantity from left to right on the figure. In Figure. 13.7B, these DNA standards are regressed against the resulting strength of their fluorescent signals (SYBR Green or TaqMan probes). The curve fit of 0.995 ($r^2$) indicates a strong correlation between signal strength and DNA quantity; therefore, this curve and subsequent regression equation can be used to determine the amount of genomic units present in an unknown sample.

Another component of qPCR assays is the melt-curve (Figure 13.7C). This is used to determine the specificity of SYBR Green-based qPCR assays. Since SYBR Green yields a signal when bound to dsDNA, the presence of any dsDNA will yield a signal. The use of a melt-curve enables the user to compare the melting temperatures of unknown samples to known positive controls (e.g., the standard curve). During the melt-curve, the temperature is raised (65–90°C) ultimately resulting in denaturation of the double-stranded PCR amplicons, and thus a decrease in the SYBR Green fluorescent signal. Comparison of the PCR amplicon melt-curve to a standard helps to verify the specificity of the amplification. The use of TaqMan-based chemistry reduces the need for melt-curve analyses, since internal DNA probes are designed to bind between the upstream and downstream primer-binding sites thus providing an added layer of specificity.

advances of qPCR, where instead of reading multiple bands on a gel, the researcher would read multiple fluorescent signals, each relating to a specific gene target. This approach typically uses TaqMan chemistry, but has been modified to work with SYBR Green by using multiple melting points at the melt-curve step (Figure 13.7C).

The PCR assay can also be modified to provide a form of genotyping using either modified primers or a post-PCR step involving restriction enzymes. The former approach involves the exploitation of repetitive genetic elements within an organism's chromosome, while the latter can be applied to nearly any PCR amplicon. Environmental microbiologists have applied these techniques to a number of treatment-impacted environments to report broadscale changes in a microbial population, or to determine genotypic differences between isolated strains. The application of ERIC-PCR and related PCR fingerprinting approaches enables theoretical discrimination of strains belonging to single species (Section 13.5.1). Restriction digestion of PCR products can also yield fingerprints (e.g., T-RFLP or terminal-restriction fragment length polymorphisms; Section 13.5.2.1) of microbial communities; this method has been commonly used to determine differences in soil bacterial communities, such as in soils under varying agronomic management systems. Typically, DNA from a microbial community would be extracted, PCR-amplified with 16S rRNA primers, digested with one or two restriction enzymes and then visualized using either a high-resolution gel or capillary-based system. Fluorescently labeled primers facilitate visualization of the digested bands when using the capillary-based approach. As with ERIC-PCR and other PCR-based fingerprinting methods, these approaches carry with them their own caveats. Particularly, inter-lab variability and even inter-operator variability can limit the potential of these assays. Additionally, the presence of similar bands in two different samples which represent different or even multiple organisms. Furthermore, with the decreasing costs of DNA sequencing, many labs have moved away from community fingerprinting methods such as T-RFLP and instead use high-throughput sequencing-based analyses that offer greater resolution and better taxonomic characterization of microbial communities (Chapter 21).

### 13.4.4 PCR Target Sequences

PCR has become a tool of choice for environmental microbiology, and its assays and uses vary substantially

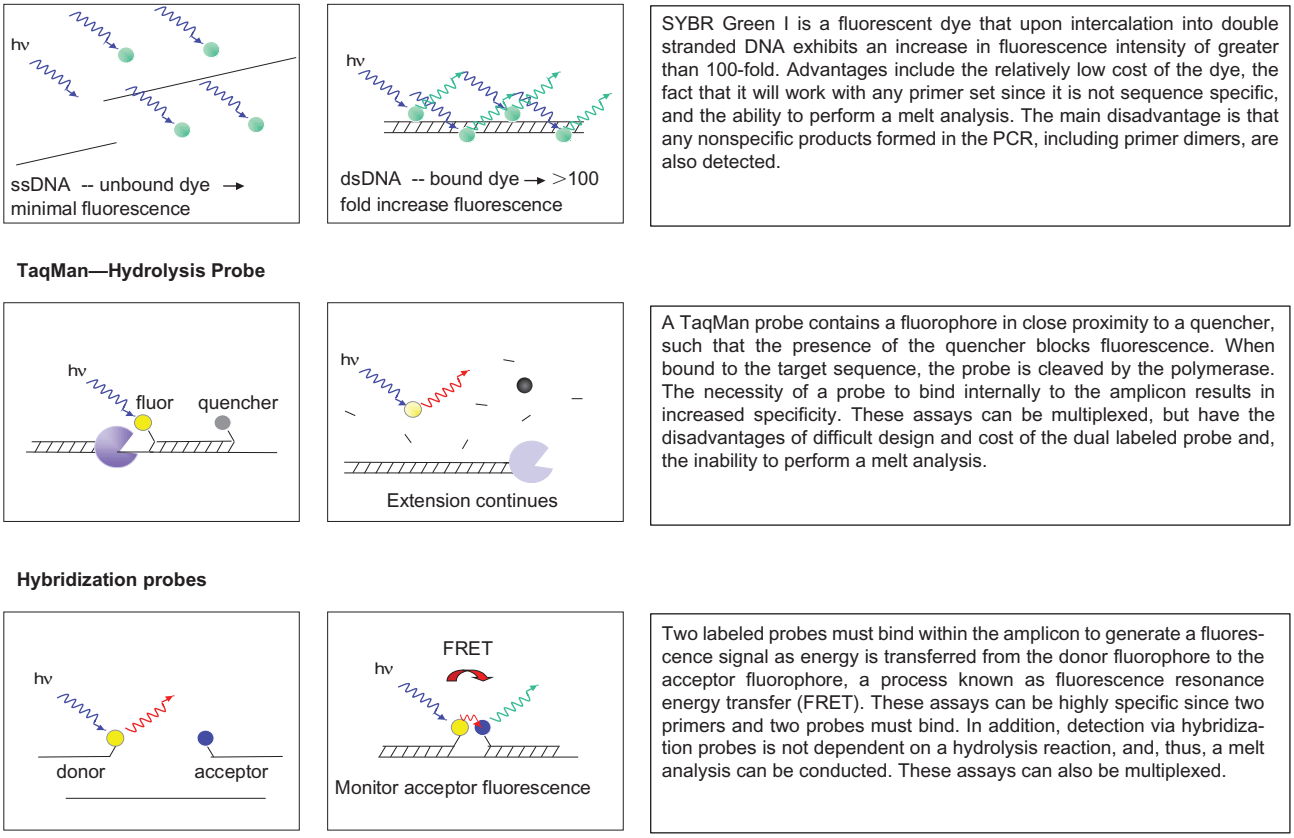


FIGURE 13.6 Schematic of SYBR Green I, TaqMan and hybridization probe fluorogenic detection approaches for quantitative (real-time) PCR.

given the end product. Products can either be qualitatively measured, such as presence or absence measurements, or quantitatively assessed. Primer design is one factor that drives this process. Target sequences commonly used in environmental microbiology include: 16S and 18S rRNA genes, genes involved in biogeochemical cycling such as nitrogen fixation, antibiotic resistance genes and genes related to pathogenicity (Table 13.1). Additionally, each of these primer targets can be modified for quantitative PCR (qPCR) assays. The 16S rRNA gene is one of the most exploited gene targets in environmental microbial PCR. The 16S rRNA gene is highly conserved due to its importance in cell functionality; thus, any changes to this target are related to phylogenetic similarity (Chen *et al.*, 1989). Some have referred to the gene as a molecular chronometer (Woese, 1987). This gene has proven to be versatile, particularly for bacterial taxonomy, qPCR and library analysis. At its most basic level, the 16S rRNA gene assay can be used in a presence/absence format to identify the taxonomy of a given bacterial isolate. A simple "colony PCR" followed by sequence analysis can be used to classify an isolate. Quantitative PCR involving 16S rRNA primers, on the other hand, provides quantitative data which can be used to enumerate a specific bacterial population in a given environment. There are some caveats associated with the assay, but for the most part, 16S rRNA qPCR can be used to estimate the total number of bacteria in a sample. The qPCR 16S rRNA assay has been applied to a number of clinical and environmental samples, often replacing traditional plating techniques in many environmental microbiology labs. This method can be used in environments where traditional culture may not yield accurate results, such as environments with fastidious or noncultivable pathogens (Matsuki *et al.*, 2004); qPCR can provide numbers which are comparable to cell counts, typically 10−100-fold greater than culture techniques (Matsuki *et al.*, 2004). The 16S rRNA gene can also be used to determine phylogeny via traditional clone libraries or high-throughput DNA sequencing (Section 13.7; Chapter 21).

Additionally, PCR has many advantages for detecting environmental pathogens and antibiotic resistance (Table 13.5). Various PCR assays have been successfully applied to measure pathogens in environmental matrices such as air, soil, water, food and fomites. The PCR assay can be applied to either isolated microorganisms or complex environmental samples, depending on the purpose of the assay. In the case of isolated microorganisms, a pathogen PCR assay could be used to confirm the pathogen, confirm pathogenicity traits or confirm other functional

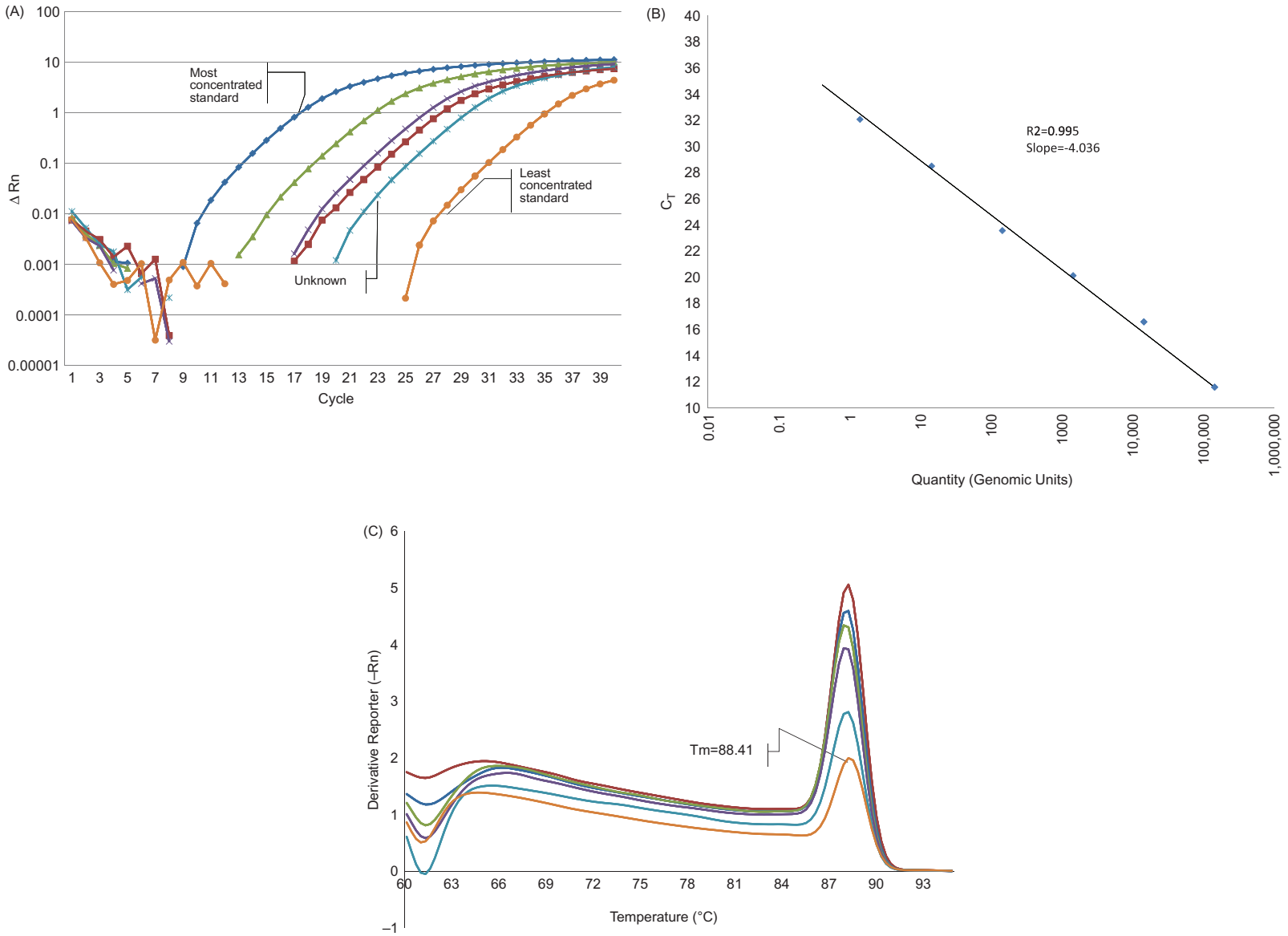


FIGURE 13.7 (A) A typical qPCR amplification plot generated using a 10-fold dilution series of genomic DNA. (B) A standard curve generated from the amplification plot. (C) A melt-curve analysis. Courtesy R.K. Smith.

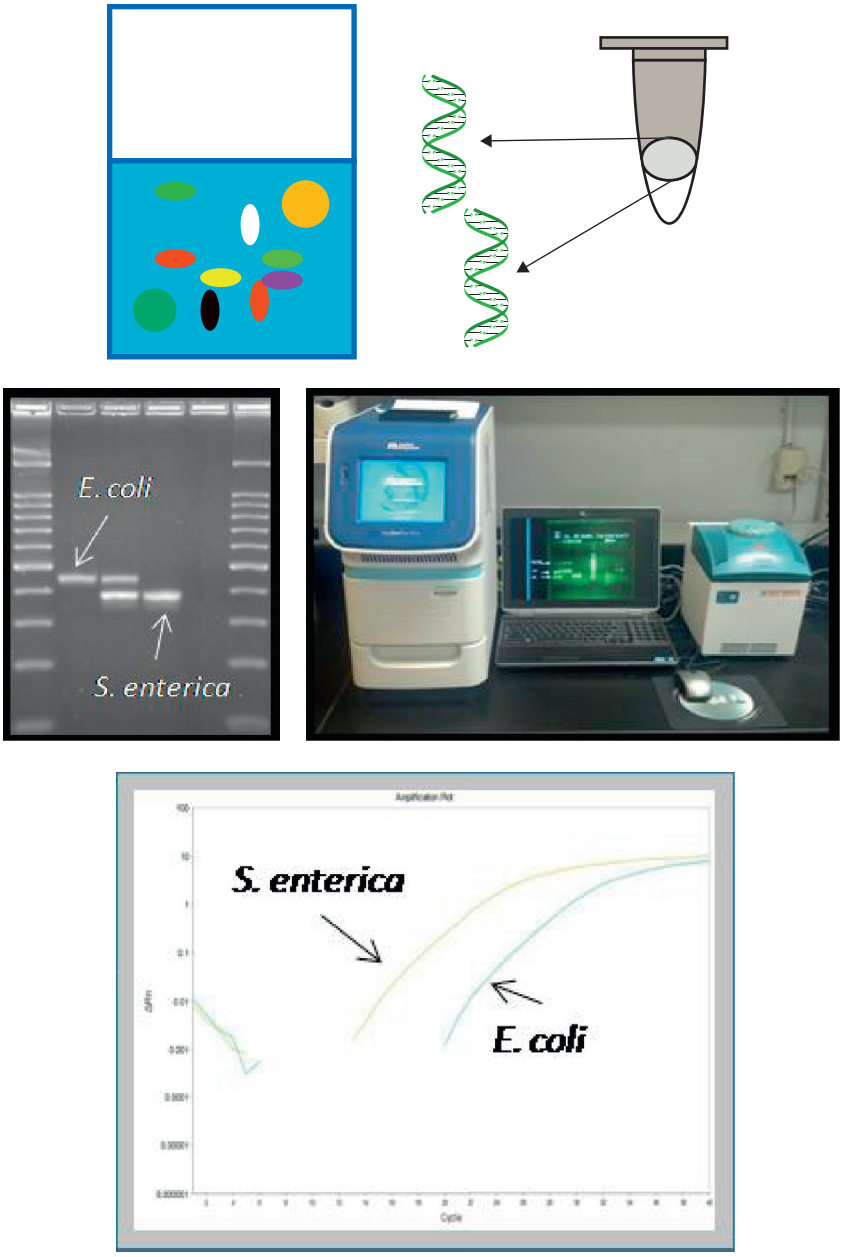


FIGURE 13.8 Schematic representation of multiplex PCR in which multiple sets of primers allow simultaneous amplification of more than one sequence (organism).

genetic traits. In an environmental sample, the PCR assay could be used to confirm the presence of a pathogen as a means to eliminate negative samples before more costly culture techniques are used; for instance this could be applied to reduce the costs associated with virus culture. Common pathogen PCR targets in the environmental lab might include *Escherichia coli* (*uidA*), *Campylobacter jejuni* (*hipO*), *Salmonella* spp. (*invA*), *Norovirus* (capsid), tetracycline resistance (*tetR*) and methicillin-resistant *Staphylococcus aureus* (*mecA*). In these cases, particularly for *C. jejuni* and *Norovirus*, the measurement of difficult-to-culture pathogens can be facilitated by PCR. However, for organisms whose genomic material consists of only RNA, such as *Norovirus*, this requires additional steps as discussed in the section below.

## 13.4.5 RNA Amplification Using PCR

Since DNA polymerase is the basis for PCR, all PCR approaches use DNA as their starting template; however, RNA instead of DNA is the genomic material for many viruses, including several pathogens of concern to environmental microbiologists. In addition, since mRNA is the precursor molecule to proteins, mRNA can be used as a reporter molecule for microbial protein production and

**TABLE 13.5 Comparison of Various PCR-Based Methods with Traditional, Culture-Based Methods for Pathogen Detection**

| Advantage/Disadvantage | PCR-Based Method | | | | Culture |
|---|---|---|---|---|---|
| | qPCR | RT-PCR | ICC-PCR | Enrichment-PCR | |
| Reduced time of detection | Yes | Yes | Yes | Yes | No |
| Increased sensitivity | No | No | Yes | Yes | Yes |
| Affected by PCR inhibitory substances | Yes | Yes | No | Yes[a] | No |
| Reduced costs | Yes | Yes | Yes | Yes | Yes[b] |
| Viable pathogens detected | Yes | Yes | Yes | Yes | Yes |
| Detects **only** viable/culturable pathogens | No | No[c] | Yes | Yes | Yes |
| Detects viable but not culturable pathogens | Yes | Yes | No[d] | No[d] | No |
| Quantitative results | Yes | No | Yes[e] | Yes[e] | Yes |

[a]*Some enrichment media contains PCR inhibitory substances.*
[b]*Costs for various pathogens can increase significantly.*
[c]*If mRNA is the starting template, then RT-PCR will detect only viable pathogens.*
[d]*Assumes that a viable-but-not-culturable pathogen would not grow effectively in any of the known culture systems.*
[e]*A most-probable-number approach must be set up during the culture phase.*

an indicator of microbial activity. There are a number of RNA-focused kits which can extract RNA from a host of environments focusing solely on viral RNA or more broad-spectrum extraction of generic microbial RNA (Chapter 8). In considering viral RNA, or any RNA product requiring PCR for detection, the researcher must turn towards a specialized set of enzymes known as reverse transcriptases. In a process known as reverse transcriptase PCR (RT-PCR), these enzymes reverse transcribe RNA to DNA, hence the name, which can then be used in a number of downstream assays, particularly PCR (Figure 13.9). The newly converted cDNA (complementary DNA) serves as the template for PCR. A number of viral detection assays utilize this pretreatment, including assays for enteroviruses and *Norovirus*. In the case of the latter, if not for reverse transcriptase, microbiologists would still be struggling to assay *Norovirus* since no culture-based detection method is available. The RT-PCR process can be modified to accommodate qPCR; for *Norovirus*, RT-qPCR is the only means available for quantifying *Norovirus* levels in the environment (Hill *et al*., 2010).

Approaches have also been developed to randomly amplify whole-community (environmental) mRNA. This process is relatively straightforward for eukaryotic mRNA due to the presence of a poly(A) tail on the mRNA, which can serve as a primer site. However, this does not work for bacterial mRNA since it does not have a poly(A) tail. Gao *et al*. (2007) developed a method to amplify whole-community mRNA (including bacteria) by using a fusion primer (random primer with an attached T7 RNA polymerase promoter sequence) to generate cDNA from environmental mRNA. After the cDNA was generated, it was amplified via linear RNA amplification with T7 RNA polymerase prior to conversion to cDNA for microarray analysis. This process resulted in >1000-fold amplification of the RNA template and produced representative and reproducible results, although some amplification biases were detected.

## 13.4.6 Detection of Live vs. Dead Microorganisms

The DNA-based PCR reactions that we have discussed thus far are capable of many things, the quick identification of a genetic trait (e.g., hours vs. days/weeks for pathogen culture confirmation), quantifying organisms, detection of unculturable microorganisms, etc., but they cannot be used to determine whether the nucleic acid originated from a viable or dead microorganism. Traditionally, this has been the primary argument against the use of PCR as a standard detection method. At least, this is true when considering PCR alone; however, some enterprising researchers have adapted PCR to live, culture-based approaches to solve this issue (Case Study 13.1). ICC-PCR (Blackmer *et al*., 2000) and broth enrichment-based PCR for bacteria (Thomas *et al*., 1991) have been used to simultaneously detect genetic traits in a sample while ensuring these traits came from live, theoretically infectious, pathogens. The methods ensure that

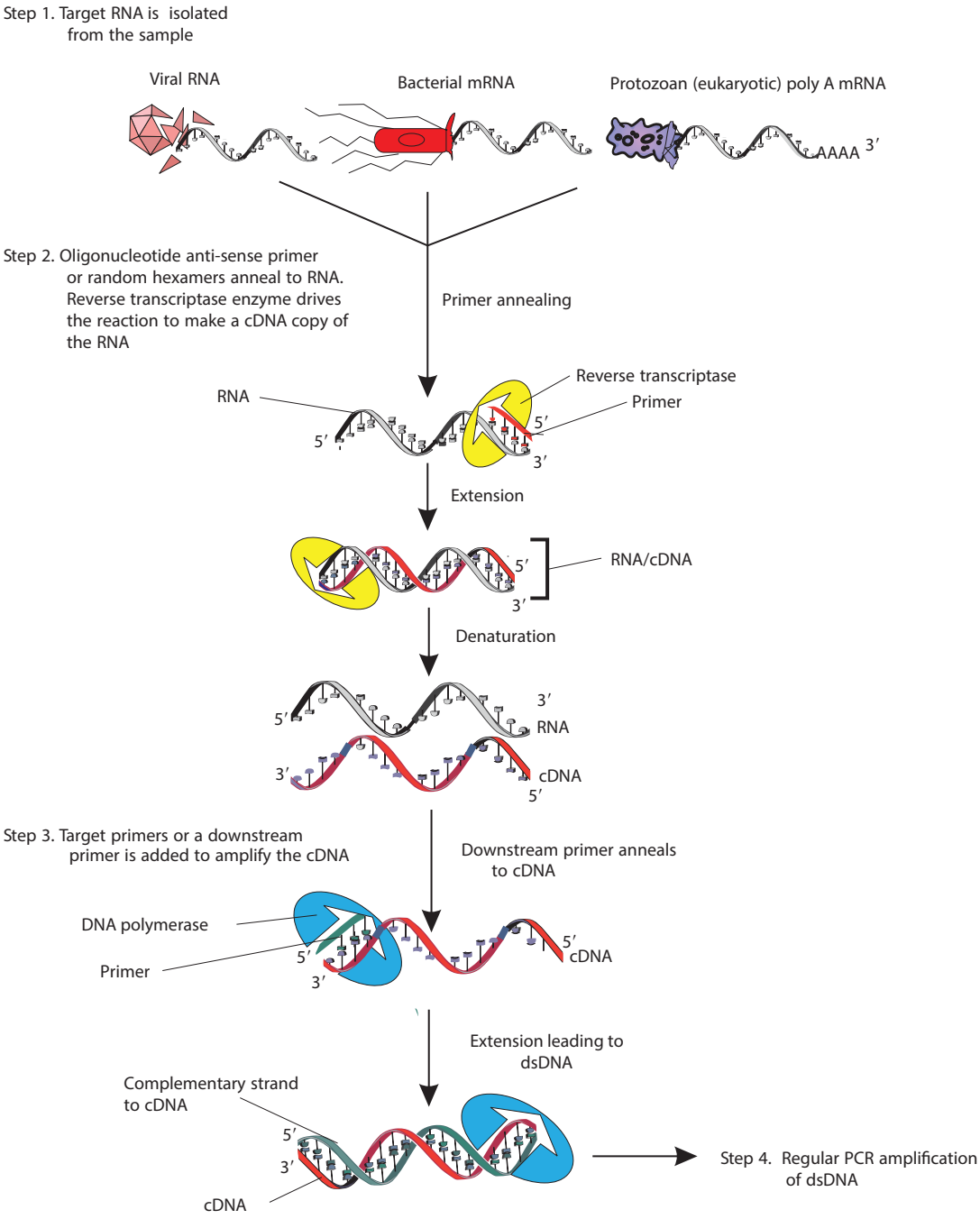


FIGURE 13.9 Reverse transcriptase-PCR (RT-PCR) amplification of RNA. RNA is reverse transcribed to synthesize cDNA by random hexamers or a specific antisense primer. PCR can then be performed on the cDNA template.

at least one of the major disadvantages associated with PCR can be overcome (Table 13.5).

Another approach that has been developed to differentiate between PCR detection of live and dead microorganisms is based upon the use of propidium monoazide (PMA; Nocker *et al.*, 2006). PMA only penetrates dead cells. Once inside the cell, it intercalates into the DNA. The PMA can then be photo-induced to covalently cross-link with the DNA, which renders it insoluble and prevents its subsequent extraction. Therefore, the DNA extracted from a treated culture would only originate from live cells, at least theoretically. Like all PCR-based assays, PMA-PCR can have some challenges when being used on environmental samples, but it shows great potential for allowing quantitative, PCR-based detection of viable organisms (van Frankenhuyzen *et al.*, 2011).

**Case Study 13.1 To Live or Not to Live: Detection of Viable Pathogens with PCR**

Since PCR arrived on the microbiological scene nearly 30 years ago, environmental microbiologists have envisioned a time in which difficult-to-detect pathogens could be easily quantified. Although PCR enabled sensitive pathogen detection from a number of matrices, one caveat has always remained: "... so is the DNA that you detected, from live or dead cells?" The very nature of PCR is based on the *in vitro* amplification of template DNA extracted from cells – either living or dead; thus PCR lacks the discriminating power to only consider DNA from "viable" organisms. Consider the fact that PCR is slowly replacing culture assays and one can see why this issue needs to be addressed. That is not to say that culture is no longer needed; culture techniques are still far more sensitive than PCR, given the sample aliquots allowable (>10 g) in a culture assay versus a 25 μl PCR assay with 2 μl of template. There are a few techniques available which enable the researcher to remove "dead" DNA from the equation:

1. The use of intercalating molecules, such as propidium monoazide (Nocker *et al.*, 2006), has enabled researchers to selectively detect DNA from live bacterial cells as these chemicals will only penetrate a compromised cell wall/membrane.
2. The use of mRNA as a starting template in a reverse transcriptase PCR has been shown to be effective as well. In this case, we exploit the cell's natural production of mRNA as an indicator of a live cell; the theory is, if the cell is living it will continue to produce mRNA as it continues with protein synthesis.
3. The combination of culture and PCR methods can yield powerful results; enrichment-based and integrated cell-culture (ICC) PCR have been utilized for the measurement of viable environmental bacteria, viruses, and parasites, and in some ways allows for an extremely sensitive assay (Blackmer *et al.*, 2000).

The theory here relies on the ability for one to culture or enrich these pathogens prior to the PCR assay, thus increasing sensitivity and ensuring that only viable pathogens are detected. However, these enrichment methods are not without their drawbacks; particularly, the fact that quantifiable results can only be obtained through the use of a most-probable-number approach, making the method quite labor intensive. Additionally, the time it takes to enrich may add an extra day to bench work, thus delaying the availability of test results.

### 13.4.7 Other Nucleic Acid Amplification Methods

In addition to the PCR-based amplification methods discussed in the previous sections, there are also other methods for amplifying DNA. One of the most common uses of the bacteriophage ϕ29 DNA polymerase is a process often referred to as multiple displacement amplification (MDA), since a single strand of DNA is simultaneously being replicated at different points. Random hexamer oligonucleotides are commonly used as the primers for MDA. In contrast to PCR, the reaction can occur at 30°C and does not require temperature cycling. MDA has several advantages over PCR including having a lower error rate and the ability to generate longer amplicons. Like any amplification process, it may introduce biases (Wang *et al.*, 2011), but it has great potential for whole-genome amplification, especially for single-cell sequencing (Lasken, 2012; Chapter 21).

### 13.4.8 Challenges for PCR-Based Amplification Methods

The presence of metals, chelating agents and humic acids can limit the effectiveness of a PCR assay, particularly when these traits inhibit the reaction to a point where no amplicons can be obtained. This can be especially troublesome when reporting negative samples; "Did the sample fail to amplify because no genetic target existed, or because of the presence of an inhibitor?" Additionally, the PCR assay can enhance detection limits, making the assay more sensitive than culture; however, PCR is limited by sample aliquot volume. A PCR is typically conducted in volumes of <25 μl, thus only a small fraction of the original sample can be accounted for in each tube; therefore, while a theoretical single molecule of DNA can be detected, the fact that such a small amount of sample is accounted for reduces the overall sensitivity of the assay. Assay sensitivity can be increased using broth enrichment approaches, in which the target bacterium can be biologically amplified and detected via PCR assay; this approach is similar to the ICC-PCR virus assay. The assay can be quantified using a most probable number (MPN) approach (McLaughlin *et al.*, 2009). Additionally, sample aliquots can be concentrated prior to PCR, utilizing flocculation, molecular weight filters, membrane filters or centrifugation (Chapter 8). As with any scientific method, PCR is not without its limitations, but it continues to play an important role in the environmental microbiology laboratory, and will continue to do so for the foreseeable future.

## 13.5 DNA FINGERPRINTING

A variety of DNA fingerprinting approaches have been developed over the past few decades. Although these

approaches produce a similar output—a fingerprint or barcode for a microorganism or microbial community—they can vary widely in their approach. The basis for generating the fingerprints has primarily been: (1) digestion of genomic DNA with restriction enzymes; (2) PCR amplification of repeating elements or specific genes; or (3) some combination of PCR amplification and digestion. While the DNA fingerprinting of microbial isolates remains a cornerstone of microbial detection, DNA fingerprinting of microbial communities has been largely replaced by newer technologies such as high-throughput DNA sequencing. In the following sections we will highlight some of the major methods used for DNA fingerprinting and their application to characterization of environmental microorganisms.

## 13.5.1 Fingerprinting of Microbial Isolates

### *13.5.1.1 Restriction Fragment Length Polymorphism Analysis*

Restriction fragment length polymorphism (RFLP) analysis is frequently used to characterize or differentiate whole genomes or specific PCR products from bacterial isolates. In general, this process involves extracting and purifying genomic DNA from a pure culture of a bacterial isolate, and in some cases PCR amplifying the target sequence. The technique is often used in conjunction with cloning (Section 13.6.1) where primers are used to amplify the target gene from a mixed population of microorganisms. Clones are generated (each containing the target sequence from a single environmental microbe), and then the clones are subjected to RFLP analysis. In all cases, DNA is cut into smaller fragments by the use of restriction enzymes. The fragments of DNA are usually separated by gel electrophoresis. If the restriction enzymes cut at many sites within the genomic DNA, several or even hundreds of DNA fragments can result. The pattern of these fragments following electrophoresis and staining can be used to differentiate genomes or clones from one another. In addition, if gene probes are used to interrogate the bands (i.e., Southern hybridization), then the location of a particular gene or sequence within fragments can be determined, which may be beneficial for subsequent molecular analyses or for simplification of complex banding patterns. This approach is the basis for the automated RiboPrinter® System (DuPont™, Wilmington, DE, U.S.A.) for typing microbial isolates (Figure 13.10A).

### *13.5.1.2 PCR of Repetitive Genomic Sequences*

PCR fingerprinting techniques employ primers that anneal to sequences that are repeated at multiple locations throughout the genome of an organism. Consequently, amplification results in the generation of multiple products of varying size. Analysis of the resulting amplicons via gel electrophoresis allows differentiation of many environmental isolates. Several PCR fingerprinting techniques have been applied to environmental samples. One PCR fingerprinting method is arbitrarily primed PCR (AP-PCR), also referred to as random amplified polymorphic DNA (RAPD) (Welsh and McClelland, 1990). AP-PCR uses one random primer (10−20 bp), which is annealed at a low temperature, resulting in nonspecific amplification of a genome. Thus, this reaction requires no prior sequence information, and will generate a fingerprint based on the uniqueness of the genome of the isolate or species. Depending on the random primer used, fingerprint patterns can be simple (1−2 bands) or complex (>10 bands).

Other fingerprinting methods use two different primers to PCR amplify repetitive sequence elements in microbial genomes. These elements include REP (repetitive extragenic palindromic), ERIC (enterobacterial repetitive intergenic consensus) (Figure 13.10B; Versalovic *et al.*, 1991) and BOX sequences (Burr *et al.*, 1997). These PCR-based fingerprinting approaches, and their variations, are commonly used to characterize microorganisms. For example, a patented REP-PCR-based approach has been commercialized and is available as the

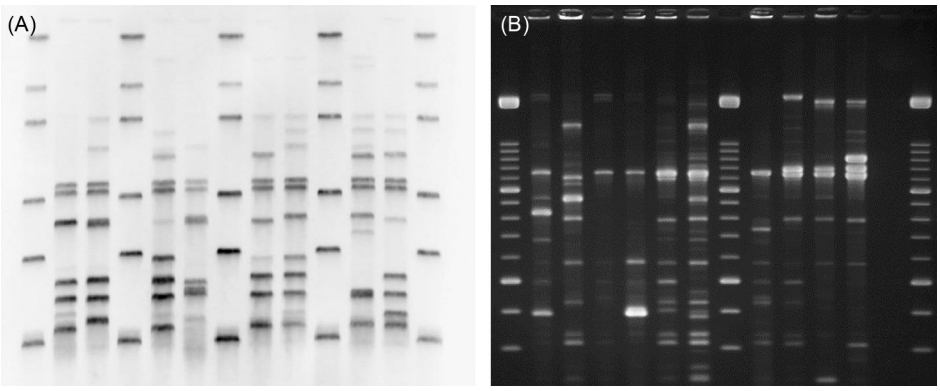


FIGURE 13.10 Examples of DNA fingerprinting approaches for characterizing microbial isolates: (A) RiboPrinter® and (B) enterobacterial repetitive intergenic consensus (ERIC) sequence-PCR fingerprinting of environmental *E. coli* isolates. RiboPrinter fingerprinting is an automated Southern hybridization-based method; whereas ERIC-PCR fingerprinting is a PCR-based method. Photos courtesy E.C. Martin.

DiversiLab™ System for genotyping of bacteria and fungi (bioMérieux, Marcy l'Etoile, France).

In comparison to RFLP fingerprinting, PCR fingerprinting is simple and rapid. However, it may be too sensitive for some applications, and even slight changes in experimental conditions can result in nonreproducible fingerprints. Such conditions include DNA quality and PCR temperature profiles. Reliability must be taken into consideration when making comparisons between runs, between laboratories and in some cases even between different DNA templates within the same laboratory.

### 13.5.1.3 Amplified Fragment Length Polymorphism Analysis

In addition to the more traditional fingerprinting techniques described above, some newer fingerprinting approaches rely on fluorescently labeled oligonucleotides. Amplified fragment length polymorphisms (AFLP), originally developed as a nonfluorescent technique, uses selective PCR amplification of genomic restriction fragments to generate fingerprints of individual isolates (Vos *et al*., 1995). A limited number of primer sets are used for amplification and no prior sequence knowledge of the target is required. In contrast to some of the PCR fingerprinting techniques described earlier, stringent PCR conditions are used in this technique, which results in the generation of robust and reliable DNA fingerprints. The number of fingerprint bands generated is dependent upon the template and primer set used, the latter of which is determined experimentally during an optimization process. There are essentially four steps involved in this technique: (1) restriction digestion of genomic DNA; (2) ligation of oligonucleotide adapters; (3) nonselective and selective amplification of sets of restriction fragments; and (4) gel or capillary electrophoresis analysis of amplified fragments.

AFLP polymorphisms are generally a single nucleotide polymorphism in the restriction sites or in the selective nucleotides adjacent to the restriction sites. The primers used in AFLP target a chimeric DNA that includes the adapter and restriction fragment sequences, with selective nucleotides added to the 3′ ends of some of the primers. In this case, chimeric DNA refers to DNA that is from two sources that have been spliced together. Addition of the selective nucleotides is done so that the primers only prime a subset of the restriction sites, limiting the number of generated fragments. A preselective PCR amplification can be done using nonlabeled primers (with zero or one selective nucleotides) to reduce the number of digestion fragments followed by a "selective" PCR amplification in which the primers have additional selective nucleotides added to their 3′ ends. One of the primers is fluorescently labeled; thus amplification results in the formation of fluorescently labeled amplicons. AFLP primers targeting restriction enzyme sites *Eco*RI (this is a 6 base pair base cutter that results in rare cuts) and *Mse*I (a 4 base pair base cutter that results in frequent cuts) are the most commonly used, but the approach can be used with other primer pairs designed to target other restriction fragment sites. Typically, the *Eco*RI primer is labeled, and since amplification is accomplished using the *Eco*RI and *Mse*I primers, only *Eco*RI−*Mse*I fragments are detected. A typical AFLP fingerprint contains between 50 and 100 amplified fragments. AFLP technology has been widely used in environmental microbiology and is useful for identifying and typing specific organisms such as plant pathogens (Portier *et al*., 2006).

### 13.5.1.4 Pulsed-Field Gel Electrophoresis

Pulsed-field gel electrophoresis (PFGE) was first described by Schwartz and Cantor in 1984. PFGE is similar to the RFLP analyses described above, in that it is another way to detect polymorphisms using restriction enzymes. In addition, PFGE is also used to get good estimations of genome size, by cutting the chromosomal DNA into fragments that can be sized accurately. PFGE is used to detect fragments of higher molecular weight than those in normal RFLP, ranging from 10 to 800 kb. These larger fragments are often generated by the use of rare cutting enzymes, the selection of which are organism dependent (e.g., *Xba*I is often used for typing of *E. coli* O157:H7, *Salmonella* and *Shigella* (Ribot *et al*., 2006) while *Sfi*I and *Not*I are used for typing *Vibrio cholerae* (Cooper *et al*., 2006)). Overall cell lysis and restriction digestion occur in plugs made from bacterial broth cultures and molten agarose. The plugs are embedded into wells in the gel prior to electrophoresis. This procedure reduces shearing of the DNA, which is important since large fragments are needed for analysis. The restriction digest is separated by gel electrophoresis of alternately pulsed, perpendicularly oriented electrical fields and stained, resulting in a large-fragment fingerprint. Due to its ability to differentiate very closely related microorganisms, PFGE is used to type microorganisms tied to foodborne outbreaks like hemorrhagic *E. coli*, *Salmonella* and *Shigella* (Information Box 13.8).

### 13.5.1.5 Plasmid Analysis

Many bacterial isolates contain one or more plasmids of variable size (see Section 2.2.3). In some cases, detection of these plasmids allows identification of specific bacterial isolates. The size and number of plasmids associated with a given bacterium are often unique. Note, however, that plasmids can be transferred to other species or in some cases lost from a bacterium. Plasmids are self-replicating, circular, extrachromosomal DNA molecules that encode genes nonessential for cell survival (e.g., metal and/or antibiotic resistance genes) under nonselective

**Information Box 13.8 PulseNet – A System for Tracking Microorganisms Causing Foodborne Disease Outbreaks**

The U.S. Centers for Disease Control and Prevention (CDC) sponsors the PulseNet program (http://www.cdc.gov/pulsenet/) for identifying and tracking microorganisms involved in foodborne disease outbreaks. The goals of the program are to enhance early detection of foodborne disease outbreaks, and provide information to assist epidemiologists in identifying the source of the outbreaks. The CDC maintains a database of fingerprints for a variety of microorganisms, such as *E. coli* O157:H7, which were developed using pulsed-field gel electrophoresis (PFGE). PulseNet participants isolate and DNA fingerprint suspected illness-causing microorganisms using standardized PFGE procedures, and the fingerprints are ultimately uploaded to the national database at CDC. Database managers then compare the uploaded fingerprints to other microorganisms in the database, and look for clusters of patterns representing potential outbreaks caused by the same microorganism. Results are reported back to the original lab and local, state and federal agencies as appropriate. As nucleic acid-based methods continue to rapidly advance, PulseNet plans to use new methods such as multilocus sequence typing (MLST) to identify microorganisms in the future.

conditions. Some plasmids are also capable of integrating into the host genome. Curing is the process whereby a plasmid is removed from a bacterium, and it can often be achieved by the application of a stress such as heat shock, or simply if the selective pressure such as a metal or antibiotic is not present in the growth medium. For example, pJP4 is an 80 kb plasmid that encodes some of the enzymes used to degrade 2,4-dichlorophenoxyacetic acid as well as sequences responsible for mercury resistance. This plasmid has been shown to undergo horizontal gene transfer from *Ralstonia eutropha* (now classified as *Cupriavidus necator*) to other soil organisms (see Case Study 13.1) and is not stable within the *Ralstonia* isolate if grown in laboratory culture without selective pressure (Di Giovanni *et al*., 1996).

Despite the potential for plasmid transfer or loss, plasmids have been used to detect specific bacteria or genes. Specific gene sequences associated with a plasmid or homologous gene sequences among plasmids can be detected using gene probes. Bacteria that are associated with specific plasmids can be identified by plasmid profiles or plasmid fingerprints. Plasmid profiles are prepared by electrophoresis and staining of whole plasmids, whereas plasmid fingerprints are produced by restriction digestion of plasmid DNA, followed by electrophoresis and staining of restriction fragments. For gene probe analysis, bacterial isolates can be lysed and probed via colony lifts using labeled plasmid DNA sequences. Alternatively, plasmid extracts free of chromosomal DNA can be prepared, electrophoresed and probed, in essence becoming a Southern hybridization.

Regardless of whether plasmid profiles or plasmid fingerprints are generated, this analysis provides a relatively quick method for identification of specific DNA sequences associated with a bacterium. Often these DNA sequences can be correlated with specific phenotypic characteristics such as resistance to antibiotics or metals, or the ability to degrade specific organic contaminants. In addition, plasmids have been used to identify specific bacterial isolates. However, in this case, the greatest disadvantage of plasmid analysis is the potential loss of the plasmid from the bacterium of interest. This can occur through plasmid curing, which could result in false-negative results. On the other hand, transfer of the plasmid to other isolates through horizontal gene transfer could result in false-positive results.

## 13.5.2 Fingerprinting of Microbial Communities

### *13.5.2.1 Terminal-Restriction Fragment Length Polymorphism Analysis*

A method which has frequently been used to generate molecular fingerprints of microbial communities is terminal-restriction fragment length polymorphism (T-RFLP) analysis. For a T-RFLP analysis, DNA is extracted from a microbial community and then amplified with a primer pair for a specific gene of interest (most commonly the 16S rRNA gene), where one or both of the primers is fluorescently labeled. The amplicon is subsequently digested with one or more restriction enzymes. Fragments are then separated on an automated DNA analyzer, and only fragments containing the fluorescently labeled primer are detected. Primer binding sites are located at the ends (or termini) of the amplicon, and fragments are differentiated based on sequence differences in regions extending from that binding site of the labeled primer; thus, the name T-RFLP. Microbial diversity in a community can be estimated based on the number and peak heights of terminal restriction fragments (T-RF), which are easily visualized on electropherograms. A variety of software programs are available for T-RFLP data analysis, facilitating reproducible results. T-RFLP has been used largely to assess diversity and shifts in microbial communities.

### *13.5.2.2 Denaturing/Temperature Gradient Gel Electrophoresis*

Microbial diversity can also be assessed by subjecting PCR-amplified DNA fragments to denaturing gradient

gel electrophoresis (DGGE) or temperature gradient gel electrophoresis (TGGE). Originally developed for the medical field to detect point mutations in genetic linkage studies, it was introduced to microbial ecology by Muyzer *et al.* (1993). In essence, DGGE or TGGE can be used for analysis of PCR-amplified gene sequences obtained from community DNA extractions or mixtures of different bacterial isolates. The DNA amplicons generated are of nearly identical lengths, but with variable sequence composition. These analyses are most frequently applied to 16S rRNA gene amplicons, since regions of sequence variability are bounded by conserved primer sites. However, the techniques can be applied to any PCR where it is expected that generated amplicons will have sufficient sequence variability to allow separation.

Separation is based on changes in the electrophoretic mobility of DNA amplicons as they migrate through a gel containing a linearly increasing gradient of DNA denaturants (urea/formaldehyde or temperature). Changes in amplicon mobility result from partial melting of the double-stranded DNA in discrete regions or domains less resistant to denaturants or temperature. In the case of DGGE, this occurs because the temperature of the gel is held constant, so that the melting of domains varies according to the concentration of the denaturant and therefore position in the gel. When the DNA enters a region of the gel containing sufficient denaturant, a transition from helical to partially melted molecules occurs, with a resultant branching of the molecule that sharply decreases the mobility of the DNA amplicons. Sequence variation within particular domains alters their melting behavior; thus, different PCR amplicons essentially stop migrating at different positions in the denaturing gradient. The use of a GC-rich sequence or GC clamp, which consists of 40–45 bases of a GC-rich sequence, acts as a high-temperature-melting domain and prevents complete melting of the amplicons. The GC clamp is normally attached to the 5′ end of the forward primer. Using these techniques, amplicons with only a few base pair variants can be separated efficiently in a linearly increasing denaturing gradient of urea and formaldehyde at 60°C. In contrast, a linearly increasing temperature gradient in the presence of a high constant concentration of urea and formaldehyde is used for separation of PCR amplicons in TGGE. Theoretically, each band in a DGGE/TGGE profile represents one population within the community. However, in practice, some populations are represented by multiple bands and some bands represent multiple populations. Once separated, specific bands can be excised from the gel and subjected to further analysis such as sequencing to identify dominant members in a community.

## 13.6 RECOMBINANT DNA TECHNIQUES

### 13.6.1 Cloning

Recombinant DNA technology or DNA cloning has been widely used to examine the genetics of individual bacteria, archaeans and fungi, as well as mixed communities of microorganisms. Cloning, the process of creating identical copies of a gene, has enabled scientists to find new or closely related genes, as well as characterize and identify unculturable or unknown isolates. Cloning may also be used to examine the activity of specific genes, or in the case of functional metagenomics it may be used to screen for activities within large fragments of DNA isolated directly from the environment.

The process of cloning creates a population of organisms that contains recombinant DNA molecules (Figure 13.11). That is, these organisms maintain and express genes contained within their own genomes while at the same time carrying and utilizing genetic material from other organisms. During the cloning process, a single fragment of "source" DNA is ligated to a cloning vector, and a single vector is introduced into each host. As a result, each transformed host only contains one fragment, or sequence, of source DNA. This is one of the most powerful aspects of cloning—the ability to isolate single sequences of source DNA within a host cell. This host cell can then be propagated, generating many copies of the DNA sequence of interest, which can then be studied further.

A cloning vector is a self-replicating DNA molecule, such as a plasmid or phage, which transfers a DNA fragment between host cells (Figure 13.11). A useful cloning vector has four key properties: (1) it can replicate within the desired host; (2) it has a multiple cloning site (MCS) or basis for insertion of foreign DNA; (3) it contains genes that enable selection of vector-containing host cells; and (4) it provides a quick screening mechanism to allow clones containing a source DNA insert to be identified. The origin of replication impacts both host range and copy number of the vector. Most cloning vectors have a ColE1 (pUC)-derived origin of replication that enables production of high copy numbers of the vector in an *E. coli* host. The multiple cloning site (MCS) contains a concentration of restriction enzyme sites for inserting DNA and/or removing DNA inserts, and the insertion site is flanked by sites complementary to common primer sequences (e.g., M13 and T7) that allow for easy amplification of inserted DNA. For selection, vectors typically encode one or more antibiotic resistance genes. As a result, successfully transformed host cells are selected for based upon their ability to grow on a medium containing the appropriate antibiotic(s). There are a variety of ways

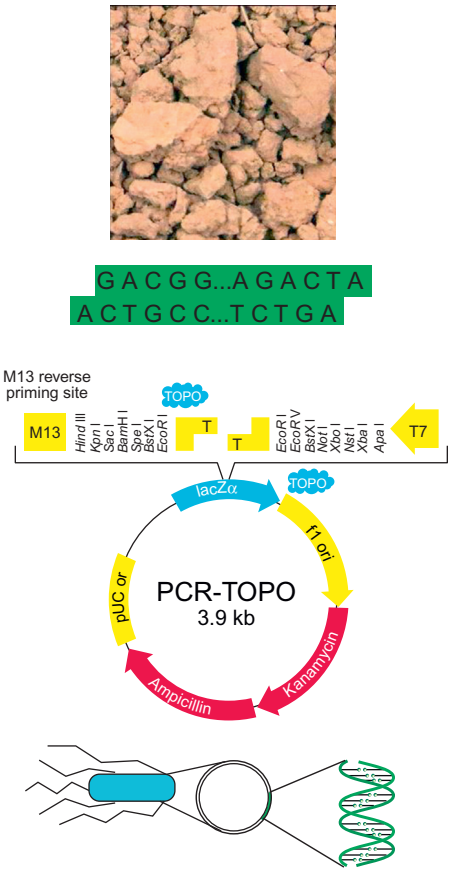


FIGURE 13.11 Cloning of a target gene via PCR amplification. Once the target gene is PCR amplified, the resulting amplicon is cloned into a plasmid vector—the TOPO® TA Cloning System (Life Technologies) in this example. The basis for cloning into this vector is the single "A" added to the 3′ ends of PCR products by many DNA polymerases including *Taq* polymerase. Each terminal "A" overhang can then complement a corresponding "T" in the vector backbone. In the TOPO TA Cloning System, the "ligation" of the insert into the vector backbone is done using a topoisomerase enzyme (TOPO). The illustrated vector contains several common components including: (1) origins of replication (PUC and f1) allowing the vector to multiply in a given host(s); (2) a multiple cloning site (MCS) with a concentration of restriction enzyme sites for inserting DNA and/or removing DNA inserts; (3) antibiotic resistance genes for selection of transformed cells containing the vector; and (4) the *lacZ*α gene containing the site for DNA insertion in order to differentiate transformed cells containing DNA inserts from those that do not via the blue–white screening process (Information Box 13.9). The MCS region also contains sites complementary to common primer sequences (e.g., M13 and T7) that allow for easy amplification of inserted DNA. Once the target gene is inserted, the cloning vector can be transformed into a competent bacterial host, which will grow and divide. Subsequently, all progeny bacteria will carry the foreign DNA as long as selection pressure for maintaining the plasmid (i.e., an antibiotic) is included in the growth medium. Plasmid vector illustration adapted from Life Technologies Corporation, Grand Island, NY, U.S.A.

to screen for the presence of a source DNA insert within the cloning vector, but one of the most common is the blue–white selection process that allows recombinant hosts to be identified on the basis of the interruption, or insertional inactivation, of a reporter gene (Information Box 13.9; Figure 13.12).

PCR has had a dramatic impact on the process of cloning DNA fragments. PCR can be used to rapidly produce large amounts of the source DNA fragment, and the use of appropriately designed primers provides a mechanism of introducing restriction enzyme sites at the ends of the fragments, simplifying the first two steps of the cloning process. A single PCR amplicon can then be directly inserted into a cloning vector and replicated in a host, such as *E. coli*. In addition, cloning vectors have been designed with a "T" on both 3′ ends of a linearized vector backbone that enables binding to the single "A" added to the 3′ ends of PCR products (Figure 13.11) by many DNA polymerases including *Taq* DNA polymerase (Information Box 13.6). Cloning of PCR amplicons has become a very common research tool, and commercial kits for PCR-based cloning are widely available.

## 13.6.2 Metagenomics

The term metagenomics was first coined by Handelsman *et al.* (1998) in reference to the collective analysis of the gene content of a community of soil microbes. Metagenomics involves the manipulation and analysis of DNA fragments obtained from a mixed community of microorganisms. Two main approaches exist for the characterization of metagenomes: sequence-based analysis and functional screening (Handelsman, 2004) (Figure 13.13). Metagenomics can be used to provide information regarding the structure, function and metabolic potential of a microbial community. Metagenomics studies are currently being performed in a wide variety of environments, ranging from soil and water to bioreactors, built environments and mammalian or insect hosts.

Early metagenomics studies typically used functional screening as a starting point to identify genes of interest for sequencing. This approach relied on cloning large DNA fragments into specialized vectors known as bacterial artificial chromosomes (BACs) or yeast artificial chromosomes (YACs). This process allows multiple genes encoded on a continuous piece of DNA to be cloned into a host together and screened for functional activity. One of the potential pitfalls of this approach, however, is that heterologous expression in the host is required. That is, the host must be able to express the cloned gene content, some of which may require coexpression of other genes. In addition, the codon usage of the host must be compatible with that of the inserted DNA. The major advantage of this approach is that, when heterologous gene expression is achieved, it does not require that the genes of interest be recognized by

**Information Box 13.9 Blue–White Screening Process for Identifying Clones Containing Recombinant DNA**

Many cloning vectors allow recombinant hosts to be identified on the basis of the interruption, or insertional inactivation, of a reporter gene. One widely used reporter is *lacZ*, a gene which encodes for β-galactosidase, which catalyzes the hydrolysis of lactose into glucose and galactose. Conveniently, it can also cleave the compound X-Gal (5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside) into galactose and 5-bromo-4-chloro-3-hydroxyindole, an insoluble compound with an indigo/blue color. These *lacZ*-based screening systems rely on a phenomenon known as α-complementation. In this process, a mutant β-galactosidase gene that produces a protein lacking several N-terminal residues is included on the host chromosome. Expression of this mutant gene alone produces an inactive β-galactosidase enzyme. However, this mutant enzyme can be made active by complementation of the missing portion of the chromosomal gene via inclusion of the first half of the *lacZ* gene, encoding the N-terminal (α) portion of β-galactosidase, on a cloning vector. When both portions of the gene are expressed, as in a host cell transformed with a cloning vector, an active β-galactosidase enzyme is formed. However, the insertion of source DNA into the cloning vector disrupts the vector-borne portion of the *lacZ* gene, prevents α-complementation, and blocks the formation of a functional β-galactosidase enzyme. As a result, clones containing a vector with inserted DNA can easily be identified when grown on media containing X-Gal, since they will form white colonies (Figure 13.12). In contrast, cells without inserted DNA produce active β-galactosidase, cleave X-Gal and thus produce blue colonies.

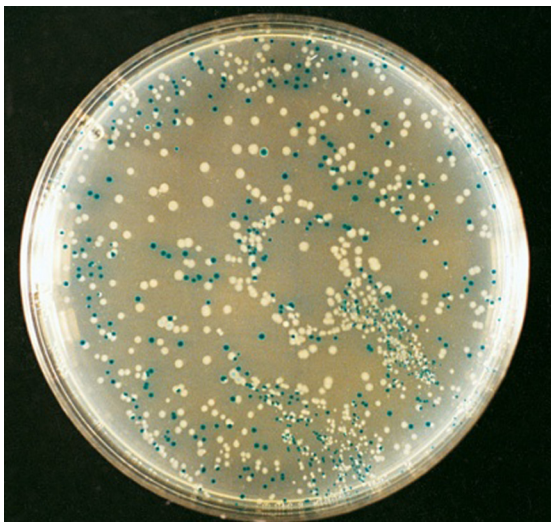

FIGURE 13.12 The blue–white screen for identification of *E. coli* clones containing recombinant DNA. White colonies contain inserted DNA while blue colonies do not.

sequence analysis, and thus allows potentially novel genetic elements to be identified.

The cloning of DNA fragments into BACs or YACs also provides a relatively easy means of targeting genes of interest for sequencing, since BACs and YACs contain known insertion site sequences and can serve as the targets for PCR-based amplification and sequencing primers. BACs were originally developed for cloning of eukaryotic genome fragments, can accommodate DNA inserts >300 kb and are maintained in low copy number in *E. coli*. YACs work similarly, and can incorporate inserts as large as 2 Mb in size; however, YAC transformation efficiencies are 100 times lower than BACs.

Most metagenomic studies today begin at the sequencing stage. Rather than generating DNA fragments and cloning them into BAC or YAC libraries, community DNA is fractionated into small pieces and sequenced directly via high-throughput sequencing (e.g., 454, Illumina and similar platforms). This approach is often described as shotgun sequencing. Using high-throughput sequencing (i.e., pyrosequencing) platforms, DNA fragments are fixed onto small beads and undergo sequencing by synthesis, a process in which thousands to millions of DNA fragments are amplified by PCR and their sequences are determined as each new base is incorporated into their respective PCR products. The advantages of this approach are that it can be used to sequence large amounts of DNA, over a relatively short time span, and at low cost relative to BAC sequencing. The disadvantages are that the length of the sequences that are generated is relatively short (approximately 150–400 bp, depending on the sequencing platform used) compared to what can be generated from BAC fragments (>300 kbp); the function of genes is inferred bioinformatically rather than tested empirically; and the analysis of such large numbers of short sequence fragments can be computationally challenging.

Despite the challenges associated with metagenomic library construction and analysis, metagenomic studies are providing unprecedented insight into the structure, function and genetic potential of microbial communities. Metagenomic sequencing provides access to both culturable and unculturable populations, as well as to underrepresented populations that otherwise go undetected. Because of this, metagenomics is allowing new insights to be gained with respect to how microbial communities function, and it is helping to unlock the vast genetic potential held within microbial communities.

## 13.6.3 Reporter Gene Systems

Nucleic acid sequences can be utilized not only to detect specific organisms but also to detect the expression of a particular gene. Reporter genes, often simply referred to

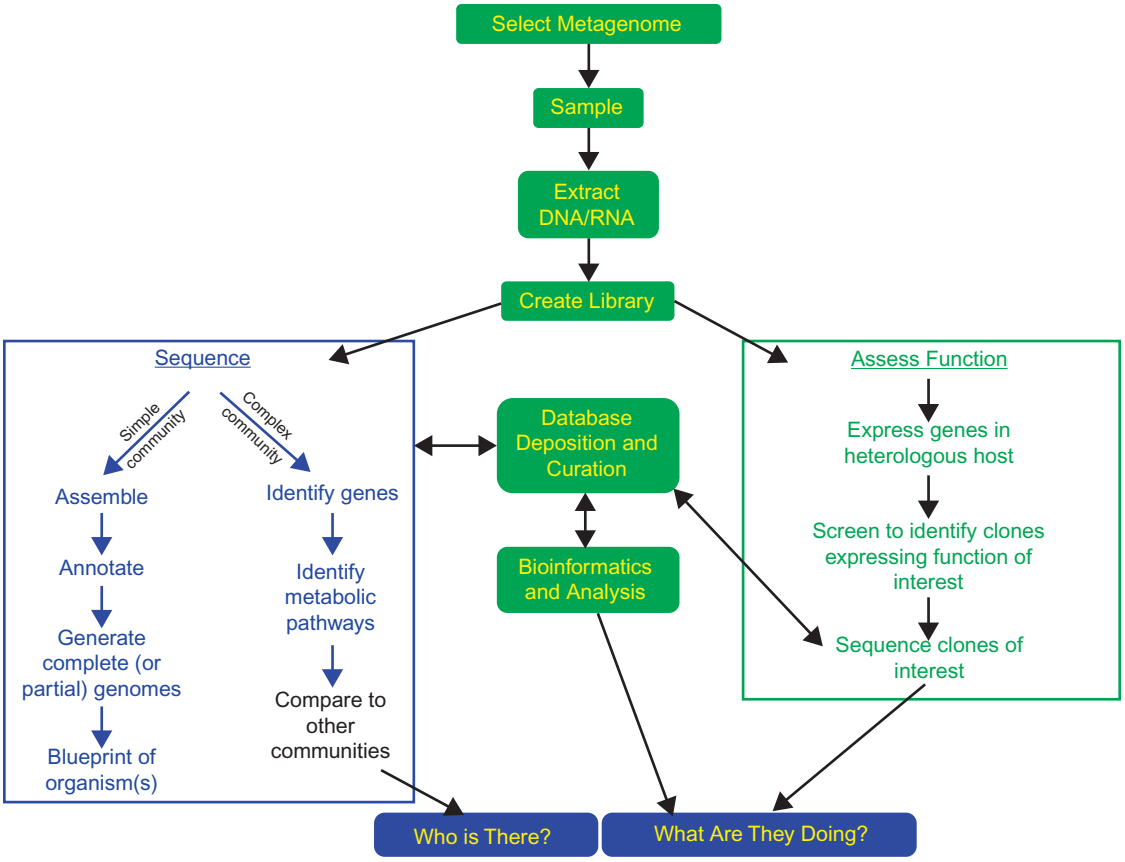


FIGURE 13.13 Metagenomics-based analyses can be used to answer the questions "Who is there?" and "What are they doing?" These projects involve the extraction of nucleic acids from an environmental sample followed by sequence-driven (light blue box) and/or function-driven (light green box) analyses. Techniques such as rRNA gene surveys are often used to determine the degree of diversity in a sample, and thus the appropriate approach for further analysis (e.g., the degree of sequencing or size of clone library necessary). An important point for any metagenomic project is the collection of environmental metadata such as location, vegetation, and soil data. This information is invaluable for environmental interpretation of metagenomic data and comparing results with other projects from around the world. For simpler communities comprised of only a few populations of microorganisms, such as a highly contaminated groundwater sample, it may be possible to assemble metagenomic sequence data to generate complete (or near-complete) genomes and metabolic blueprints of the dominant organisms. For more complex communities, sequence data is usually screened for the presence of genes of interest and compared to data from other communities, without assembly of sequencing reads. Adapted from National Academy of Sciences (2007).

as reporters, are genes that confer an easily detected signal when expressed by the host. The theory behind a reporter gene is that most genes encode products that are not easily detected, whereas the inserted reporter gene produces a product that is easy to detect. Reporters can be used for a variety of purposes including selection, gene expression analyses, promoter activity assessments, determining interactions between proteins such as in two-hybrid screening and determining the presence or bioavailability of pollutants. Antibiotic and metal resistance genes encoded on plasmid vectors can be considered reporters since they provide a selective mechanism for vector-bearing hosts. Reporter genes are also frequently constructed to allow gene expression of a target gene. In this case, a reporter is designed creating a DNA construct where the reporter gene is inserted between the promoter of the gene of interest and the gene of interest (Figure 13.14). Thus, the reporter is turned on (expressed) only when the target gene of interest is expressed. The reporter gene and the gene of interest are thus transcribed together and ultimately translated into a single polypeptide chain (fusion protein). A segment of DNA coding for a flexible polypeptide linker region is generally incorporated into the construct between the reporter and target gene, allowing for proper folding into active conformations of the two fused proteins. The activity of a particular promoter can also be determined using reporter genes. In this case, the reporter gene is placed under control of a target promoter and the activity of the reporter gene product is quantitatively measured.

Several different reporter genes have been used including: *lacZ*, which encodes β-galactosidase (Information Box

13.9); *gusA* (*uidA*), which encodes β-glucuronidase; *xylE*, which encodes catechol 2,3-dioxygenase from the TOL plasmid of *Pseudomonas putida*; *inaZ*, which encodes an ice nucleation protein from *P. syringae*; *lux* genes, which catalyze light production; and genes for various fluorescent proteins such as the green fluorescent protein (GFP). The latter two are among the reporters most commonly used to characterize environmental microorganisms.

### 13.6.3.1 Luminescent Reporter Genes

Luciferases are light-generating enzymes produced by a wide variety of organisms (e.g., fireflies and bioluminescent bacteria such as *Aliivibrio fischeri* (Section 20.2.1)). Luciferase and its luciferin substrate are coded by *lux* (*luxCDABE*) genes and are active in the presence of oxygen and a source of reducing power such as flavin mononucleotide ($FMNH_2$). The *luxAB* genes encode the active form of luciferase, and the *lux*CDE genes encode the synthesis of the substrate. Use of the full *luxCDABE* gene cassette is therefore advantageous because it does not require the addition of any exogenous substrate. The basis of the *lux* system is that *lux* genes are inserted into the operon being studied, and when that operon is induced, luminescence is given off. The *lux* system is widely used because it is uncommon to find large numbers of indigenous microorganisms that emit light, and the bioluminescent signal generated is typically short-lived allowing for repeated sampling, which is required in time course studies. Most *lux* systems require $\approx 10^5$ cells per sample for the detection of light emission, and detection can be done directly in a nondestructive manner. Different *lux*-based systems have been used in numerous experiments. For example, Dorn *et al.* (2005) used a *lux* reporter and a fiber optic detection system to study the real-time *in situ* response of microbes in a soil column system to addition of naphthalene. In this system, optic fibers were emplaced into the soil column and used to collect luminescence over several days to study the response to naphthalene addition. This lux reporter was successfully able to provide real-time information about where naphthalene degradation occurred in the soil column. However, these reporters are typically very sensitive to changes in environmental conditions and it is important that the reporter be extensively characterized prior to use (Sørensen *et al.*, 2006).

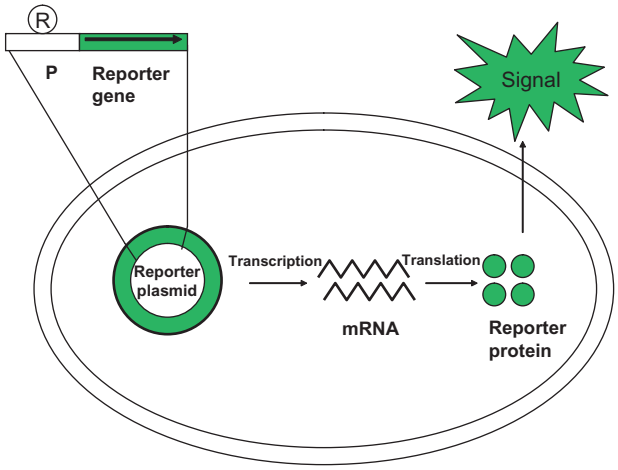


FIGURE 13.14 Regulation of a reporter gene by a regulatory protein. Binding of the regulatory protein R to the promoter P controls transcription, followed by translation of the mRNA to produce the protein. Both of these steps produce multiple copies of the reporter protein, leading to an increased protein concentration and a measurable signal.

### 13.6.3.2 Fluorescent Reporter Genes

There are now several fluorescent reporter systems including the green fluorescent protein (GFP) and the red fluorescent protein (DsRed or RFP). The genes for GFP and RFP were isolated from the jellyfish *Aequorea victoria* and from *Discosoma* sp. coral, respectively. These bioreporters rely on fluorescent light production instead of luminescence. The difference is that to obtain fluorescence, the molecule must first be excited by a specific wavelength of light. Following excitation, the fluorescent molecule emits, or fluoresces, at a different wavelength that can be detected and measured. Excitation of GFP with UV light (395 nm) results in a bright green fluorescence (509 nm), while for RFP, excitation occurs at 558 nm, and emission is at 583 nm. Through mutation, these genes have been modified to produce a broad spectrum of fluorescent proteins that have different excitation and emission patterns (Figure 13.15).

The fluorescent reporter systems have several advantages over *lac*Z and *xyl*E fusions. First, as with the *lux* system, because of the abundance of fluorescent proteins

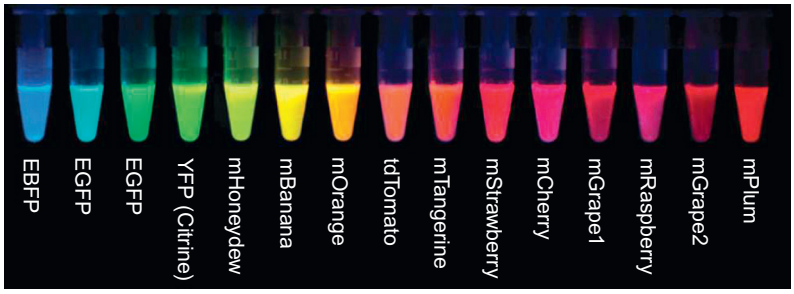


FIGURE 13.15 Examples of the wide variety of fluorescent proteins available for use as reporter genes. From Tsien (2010).

### Case Study 13.2 Use of Reporter Genes to Track Horizontal Gene Transfer in a Microcolony

Seoane *et al.* (2011) used three different fluorescent reporter genes to track transfer of plasmids between *Pseudomonas putida* KT2440 cells in a bacterial microcolony. The cells were inoculated on a nutrient agar slab, and images were taken with confocal laser scanning microscopy (CLSM) every 20 min. Figure 13.16 contains time-lapse images of the microcolony where donor *P. putida* KT2440 cells expressing a red fluorescent protein gene (DsRed) and producing LacIq (to repress GFP expression in donor cells) transferred a plasmid containing the green fluorescent protein (GFP) gene (pWW0 TOL::GFP) to recipient *P. putida* KT2440 cells that also expressed the yellow fluorescent protein (YFP) gene. Therefore, the transconjugants simultaneously expressed GFP and YFP. The GFP signal (transconjugant cells) and the corresponding overlay of all fluorescence signals (all cell types) are displayed on the top and on the bottom rows of Figure 13.16, respectively. Non-dividing, inoculated donors (bottom left of microcolony) have higher red intensities due to previous DsRed maturation. Thick arrows mark the individual cell transferring the plasmid while thin arrows indicate the resulting new transconjugant cell. After 160 min of donor–recipient contact, conjugative transfer was detected (A). It then took < 40 min for this transconjugant to retransfer twice (B, C) and < 20 min for the new transconjugants to retransfer again (D). After 480 min (approximately five division cycles), most of the potential recipient cells in the microcolony contained the plasmid (E).

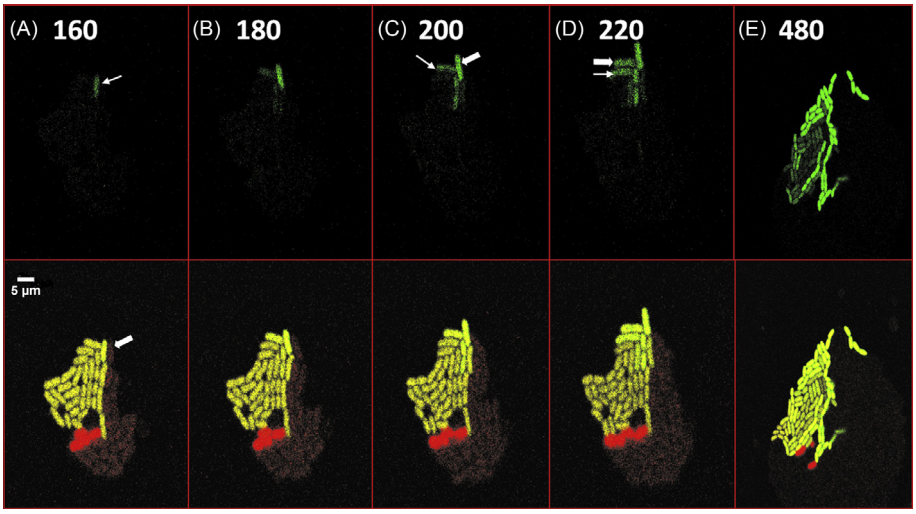


FIGURE 13.16 Use of reporter genes to track horizontal gene transfer in a bacterial microcolony. From Seoane *et al.* (2011).

in a single cell, transcription of a gene in individual cells rather than the average transcription of a population of cells can be measured. However, unlike the case of the *lux* system, cellular metabolism is not required for reporter fluorescence, thus allowing *in situ* gene expression by cells that are not actively growing, which is likely to be the case in many environments. Reporter fluorescence is also stable at 37°C, whereas the luciferase may be heat labile. Finally, both fluorescent proteins and *lux* are affected by oxygen levels. Although the fluorescent protein reporter systems function in the absence of $O_2$, the GFP protein must be oxidized in order to fluoresce. The *lux* system is affected when $O_2$ levels drop below 5 mg/L; however, luminescence occurs even as low as 0.5 mg $O_2$/L. Another important factor is that luminescence from a *lux* reporter stops as soon as gene expression stops, while the wild-type GFP protein will continue to fluoresce for as long as it remains intact. New variants of fluorescent proteins have been developed that have decreased stability to allow more accurate detection of real-time gene expression. Additionally, several color variants have been developed with altered excitation and emission spectra that allow for simultaneous tracking of multiple organisms (Case Study 13.2; Figure 13.16). These modifications have greatly increased the utility of fluorescent protein-based reporter systems for applications such as tracking dynamics of protein expression and protein localization within living cells. It should be kept in mind that it is difficult to track low levels of expression with fluorescent proteins and that the intrinsic fluorescence of some bacterial species can preclude their use as a reporter. Thus, the degree of potential background interference along with the desired characteristics of the reporter including ease of detection, sensitivity, ability to

quantify and degree of stability should be considered when selecting a reporter gene system. More detailed discussions regarding the advantages and disadvantages of various reporter genes are available in reviews including Burlage (2007) and Leveau *et al*. (2007).

## 13.7 SEQUENCE ANALYSIS

Advances in DNA sequencing technology have allowed sequence-based analyses to become routine and relatively cost-effective. DNA samples may be sent to commercial laboratories or core facilities for sequencing, but the use of benchtop sequencers in individual laboratories is becoming more common. Multiple databases store and catalog information from sequence-based projects and make it accessible to the entire scientific community (Table 13.2). The sequence data stored in these databases and archives carry accession numbers to facilitate identification, organization and communication about particular sequences or projects. The database entries also typically contain the DNA, RNA or protein coding regions for a given sequence, relevant metadata (e.g., where a sample was collected, when it was collected, the environmental characteristics of the collection site) and citations for associated published research. Software programs, such as BLAST (the Basic Local Alignment Search Tool; Altschul *et al*., 1997), allow researchers to identify sequences of interest with these databases, translate a DNA sequence into protein and look for homology or relatedness based on sequences. Some of these databases also provide links to gene maps, algorithms that identify restriction enzyme sites and relevant scientific literature.

One of the most common uses of sequence data in the field of environmental microbiology is in the identification of microbes through marker gene sequences. In the case of Bacteria and Archaea, the 16S ribosomal RNA (rRNA) gene sequence is used. For Fungi, the ribosomal internal transcribed spacer (ITS) tends to be used most frequently, but the 18S and 25−28S rRNA genes are also sometimes used. Multiple databases and sequence analysis software programs are available to aid in sequence searches and identification. The National Center for Biotechnology Information (NCBI) provides large, community-generated sequence archives. Other specialized databases, such as the Ribosomal Database Project and the UNITE fungal ITS database, obtain data from NCBI but implement additional quality filtering and annotation.

Researchers typically interact with these databases using BLAST or another similar software package. Programs like BLAST allow researchers to submit a query sequence and compare it to all other sequences in the database. The results of this comparison are then scored for similarity and identity. Searches like this are used when researchers want to find out which sequences in the databases are similar to their newly-generated, unknown sequences. In some cases, this information may be used for phylogenetic analyses (Information Box 13.10), allowing researchers to characterize the genetic relationships and ancestry of new sequences relative to those contained in reference databases. For example, a 16S rRNA gene sequence from an unknown soil isolate can be compared with other sequences in the database in order to characterize and identify the organism. By analyzing numerous sequences from a sample, this approach can also be used to determine community composition and diversity (Case Study 13.3; Figure 13.17). In other cases, sequences from multiple, related organisms may be aligned with one another to identify regions of homology (i.e., similarity) as a part of the design process for PCR primers and/or probes.

**Information Box 13.10 Phylogenetic Analysis**

In the 16S rRNA gene, there are regions that are conserved among all bacteria. These conserved regions flank a sequence region that is variable and that can be used to study the relatedness among different bacteria. These attributes have been exploited to study complex bacterial communities in environmental samples using a combination of PCR and sequencing analysis. Using universal 16S rRNA primers, PCR is performed on community DNA extracted from an environmental sample. The PCR results in amplicons of roughly the same size but varying sequence composition. Once the sequences are obtained, they are aligned and then analyzed using bioinformatics (Chapter 21) to determine the diversity and composition of the community and compare it to communities from other samples.

Molecular databases are built from sequences submitted by researchers all over the world. Sequence data can be found by searching for an accession number assigned to a specific sequence, a functional description such as "hydrolase," a known sequence pattern (e.g., "GGTACCTTGAG") or the name of a gene or organism of interest (e.g., *lacZ*, *E. coli*). A number of search algorithms are available for a wide scope of nucleic acid and protein sequence comparisons. FASTA (Pearson, 1990) and BLAST are well-established algorithms that facilitate effective comparisons of unidentified sequences with those contained in the current databanks. One must take care in choosing database sequences for comparison because quality control of sequence data varies from database to database and misannotation is a common occurrence.

While genomic and metagenomic studies often target the total gene content of an organism or a community, many times it is advantageous to target a single gene or collection of genes instead. The characterization of a target gene (or genes) is known as a marker gene study. As mentioned in Section 13.4.4, the 16S rRNA gene is commonly used as a marker gene for bacterial communities, and the

## Case Study 13.3 Use of DNA Sequencing to Determine Microbial Diversity in an Antarctic Freshwater Ecosystem

The 16S rRNA gene has been used to describe the composition and diversity of a wide variety of biological systems ranging from soil, water, and air to kitchen surfaces, hotel rooms, and insect or mammalian hosts. Prior to the advent of high-throughput sequencing technologies, such as the 454 and Illumina platforms, diversity surveys typically relied on cloning and Sanger sequencing (i.e., dye-terminator sequencing), PCR-fragment-based characterization (e.g., T-RFLP or DGGE), culture-based techniques, or some combination thereof. Although there are advantages and disadvantages to each approach, the general consensus is that high-throughput sequence-based approaches tend to give greater insight into community structure and diversity.

In a recent survey of Lake Tawani, a freshwater lake located in East Antarctica, Huang *et al.* (2013) used a combination of 454 sequencing, cloning and Sanger sequencing, and plate-based cultures to describe community composition and bacterial diversity and compared the outcomes obtained from each of the different approaches. The authors collected mixed water and sediment samples, using one subsample for community DNA extraction and another subsample for the enumeration of bacteria via plate counts. The community DNA was used to generate partial 16S rRNA gene amplicons for 454 sequencing, as well as near-full-length 16S rRNA gene amplicons for cloning and Sanger sequencing. The plate counts yielded $\sim 2 \times 10^3$ colony forming units per ml, and from this 270 bacterial colonies were selected for identification. The cloning experiment yielded 232 clones with the correct insert size, and the 454 sequencing run yielded 11,235 high-quality sequence reads. The identities of both the culture-derived colonies and 16S clones were confirmed through Sanger sequencing (Figure 13.17).

As in many soil and water-based ecosystems, bacteria belonging to the phylum Proteobacteria were encountered with the greatest frequency, regardless of the technique that was used. The authors found that the three different approaches successfully detected the most dominant taxa in Lake Tawani, but the culture-based, cloning, and 454-based techniques different in their coverage, estimates of relative abundance (i.e., ranking) and ability to detect infrequent, or rare, members of the community. Among the 270 culture-derived colonies, 16 operational taxonomic units (OTUs, a sequence-based approximation of bacterial species) were found. The clone library detected 64 OTUs, and the 454 sequence library detected 498 OTUs. Among these, only 9 OTUs were detected by all three approaches. The authors concluded that high-throughput sequencing provided greater sensitivity and greater coverage than the clone or culture-based techniques, but also noted that the combination of multiple technologies gave a more complete picture than any single approach alone.

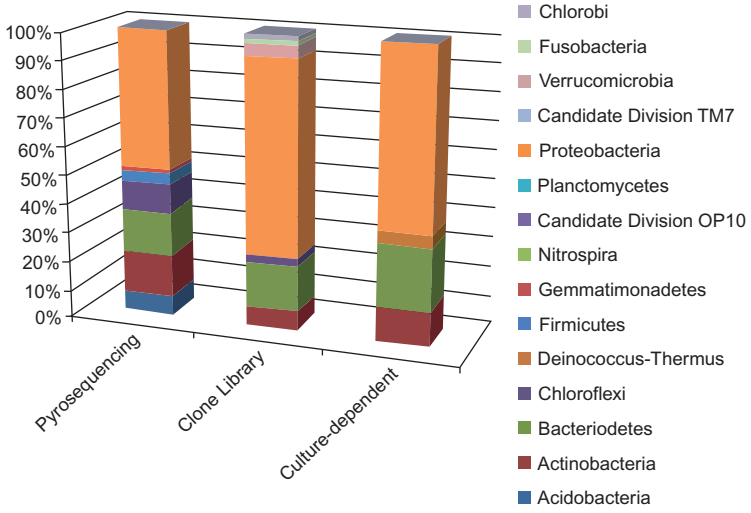


**FIGURE 13.17** Relative abundance of bacterial phyla in an East Antarctica freshwater ecosystem. Bacteria were identified through culture-dependent and culture-independent (pyrosequencing and clone library) methodologies. From Huang *et al.* (2013).

fungal ITS sequence is often used to characterize the composition of fungal communities. Other common targets include "housekeeping" genes, functional genes and collections of genes sequenced simultaneously in a process known as multi-locus sequencing (Table 13.1).

Marker genes are commonly used to identify individual sequences or isolates, but they can also be used to provide insight into the composition and diversity of entire microbial communities. Informative marker genes harbor several key traits, and these include:

- Ubiquity—the marker should be present in most, if not all, species.
- Genetic conservation—the sequence of the marker should be conserved enough that it can be targeted with PCR primers.
- Variability—in combination with genetic conservation, the marker should also contain variable regions of sequence that allow for the detection of differentiation between species, among lineages and within populations.

For some organisms, single marker genes do not provide enough information to distinguish species or strains from one another. For example, species within the genera *Escherichia* and *Shigella* share so much similarity among their 16S rRNA gene sequences that it is difficult to distinguish them from one another. Likewise, it is often necessary to distinguish pathogenic strains from their less harmful relatives in infection, disease outbreaks or source-tracking scenarios. In cases like this, multilocus sequence typing (MLST) can be used to provide the necessarily additional resolution. MLST involves the sequencing of multiple target genes (often seven or eight genes) from the same organism in order to provide a more complete taxonomic and phylogenetic profile (Pérez-Losada *et al.*, 2013). Once these sequences are generated, they can be compared to reference databases in order to obtain sequence- or strain-type identities (e.g., Aanensen and Spratt, 2005).

## 13.8 CHOOSING THE APPROPRIATE NUCLEIC ACID-BASED METHOD

With so many nucleic acid-based methods available, you may ask "How can I choose which one to use for my research?" Table 13.6 overviews environmental microbiology applications of several nucleic acid-based methods. Due to the wide variety of methods (and their modifications) available, this table should not be considered to be comprehensive with respect to either the methods listed or their applications. Other reference materials, such as the *Manual of Environmental Microbiology* (Hurst *et al.*, 2007), also provide detailed discussions regarding the applications and advantages and disadvantages of numerous nucleic acid-based methods. Ultimately, the choice of method will be driven by the available research budget and several factors including:

- Research objectives—Detection or quantification of microorganisms? Measurement of microbial activity? Characterization of microbial community composition? Elucidation of environmental function? Identification of genes conferring a specific process?
- Target organisms—Isolated or uncultured microorganisms? Pure culture or mixed community? Simple or complex microbial communities?
- Target material—DNA or RNA?

In many cases, the discovery power of nucleic acid-based methods can be magnified even further when used in combination with physiological methods such as stable isotope probing (Chapter 11). These combined approaches can enable not only the identification of microorganisms but also their connection to a specific environmental process.

The past several decades have seen dramatic advances in the use of nucleic acid-based methods to study environmental microorganisms. These methods have provided insights into the microbial world that were unimaginable just a few years ago. It is exciting to think about the continued methodological advances that will occur over the next few decades. If there is not a current method that will achieve your research objectives, we look forward to you modifying a current method or even developing an entirely novel approach for detecting and characterizing environmental microorganisms.

## QUESTIONS AND PROBLEMS

1. What are some of the advantages of using nucleic acid-based methods to detect and characterize environmental microorganisms?
2. What are some of the challenges of using nucleic acid-based methods to detect and characterize environmental microorganisms?
3. What are the major advantages and disadvantages of PCR when it is applied to environmental samples?
4. If you wanted to use PCR to assess microbial activity, what type of PCR would you use and why?
5. There are several features of qPCR that make its use advantageous over conventional PCR, list three and explain why each feature is beneficial.
6. Discuss the differences between the SYBR Green and TaqMan approaches for qPCR. Which qPCR approach is more specific? Why?
7. Compare and contrast ERIC and PFGE methods of DNA fingerprinting.
8. Compare and contrast the *lux* reporter system and the GFP reporter system. Based on these characteristics, which would be more useful for evaluating contaminant transport in a soil column study?
9. Describe what is meant by selection versus screening of a clone library. Explain how the *lacZ* reporter works and incorporate its use into your description.
10. Which nucleic acid-based method(s) would you use to accomplish each of the following objectives? Explain why you chose each method.
    a. Detect *Salmonella* spp. in a soil sample.
    b. Detect both *Salmonella* spp. and *E. coli* in the same sample.
    c. Quantify *Norovirus* in a water sample.
    d. Characterize the presence and diversity of dissimilatory sulfite reductase genes (*dsrAB*) in a groundwater sample.
    e. Determine the bacterial community composition of a fecal sample.
    f. Quantify expression of nitrite reductase genes (*nirK* and *nirS*) in a sediment sample.

**TABLE 13.6 Environmental Microbiology Applications of Various Nucleic Acid-based Methods**

| Technology | Application | | | | Notes |
|---|---|---|---|---|---|
| | Microbial Detection | Microbial Quantification | Microbial Activity | Community Characterization | |
| Polymerase chain reaction (PCR) | Yes | No | No | No | PCR is the most commonly used approach for microbial detection. It is also the basis for many of the other forms of detection and characterization. |
| Reverse transcriptase PCR (RT-PCR) | No | No | Yes | No | Reverse transcription of mRNA followed by PCR amplification of the resulting cDNA is a primary means of detecting microbial activity. |
| Quantitative PCR (qPCR) | Yes | Yes | Yes | No | Unlike PCR which is semi-quantitative at best, a variety of qPCR approaches can be used to quantify microorganisms (DNA) and their activity (if RT-PCR is used to convert mRNA to cDNA prior to qPCR). |
| DNA fingerprinting | Yes | No | No | Yes | DNA fingerprinting approaches such as pulsed-field gel electrophoresis are the basis of microbial typing used for detection of specific strains of microorganisms. These approaches are best used with pure cultures of organisms. Some DNA fingerprinting approaches such as T-RFLP have been used on microbial communities but these approaches have been largely replaced by DNA sequencing. |
| DNA/RNA sequencing | Yes | No | Yes | Yes | The current standard for microbial community characterization. If RNA template is sequenced (transcriptomics) this can indicate relative activity. Target DNA/RNA is often amplified via PCR methods prior to sequencing. |
| Fluorescence *in situ* hybridization (FISH) | Yes | Yes | Yes | No | FISH can be used to detect, quantify and determine activity of microorganisms *in situ*. However, because it is primarily a microscopy-based approach using labeled probes, only a small number of organisms can be assayed for at a given time. |
| Microarrays | Yes | No | Yes | Yes | Microarrays can be used to detect microorganisms (DNA) and their activity (usually after RT-PCR of mRNA). Due to the number of organisms that can be tested simultaneously, microarrays can also be used to characterize microbial communities. Although some microarray-based approaches can be quantitative, they are generally not the best approach for microbial quantification. |
| Reporter genes | No | No | Yes | No | The primary use of reporter genes is to determine activity of organisms engineered to contain the reporter. These genes can also be used to track the survival and movement of a host organism engineered to contain the reporter. |

**g.** Isolate a microbial gene from a soil sample that confers resistance to a newly discovered antibiotic.

**h.** Determine *in situ* expression of a bacterial gene in a biofilm.

**i.** Determine if a new pesticide impacts soil microbial community composition.

## REFERENCES AND RECOMMENDED READING

Aanensen, D. M., and Spratt, B. G. (2005) The multilocus sequence typing network: mlst.net. *Nucleic Acids Res.* **22**, W728–W733.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.

Blackmer, F., Reynolds, K. A., Gerba, C. P., and Pepper, I. L. (2000) Use of integrated cell culture-PCR to evaluate the effectiveness of poliovirus inactivation by chlorine. *Appl. Environ. Microbiol.* **66**, 2267–2268.

Brodie, E. L., DeSantis, T. Z., Joyner, D. C., Baek, S., Larsen, J. T., Andersen, G. L., *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.

Brodie, E. L., Desantis, T. Z., Parker, J. P., Zubietta, I. X., Piceno, Y. M., and Andersen, G. L. (2007) Urban aerosols harbor diverse and dynamic bacterial populations. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 299–304.

Burlage, R. S. (2007) Bioreporters, biosensors, and microprobes. In "Manual of Environmental Microbiology" (C. J. Hurst, R. L. Crawford, D. A. Lipson, A. L. Mills, and L. D. Stetzenbach, eds.), 3rd ed., American Society for Microbiology Press, Washington, DC, pp. 169–181.

Burr, M. D., Josephson, K. L., and Pepper, I. L. (1997) An evaluation of DNA based methodologies for subtyping *Salmonella*. *Crit. Rev. Environ. Sci. Technol.* **28**, 283–323.

Chen, K., Neimark, H., Rumore, P., and Steinman, C. R. (1989) Broad range DNA probes for detecting and amplifying eubacterial nucleic acids. *FEMS Microbiol. Lett.* **48**, 19–24.

Cooper, K. L. F., Luey, C. K. Y., Bird, M., Terajima, J., Nair, G. B., Kam, K. M., *et al*. (2006) Development and validation of a PulseNet standardized pulsed-field gel electrophoresis protocol for subtyping of *Vibrio cholerae*. *Foodborne Pathog. Dis.* **3**, 51–58.

Di Giovanni, G. D., Neilson, J. W., Pepper, I. L., and Sinclair, N. A. (1996) Gene transfer of *Alcaligenes eutrophus* JMP134 plasmid pJP4 to indigenous soil recipients. *Appl. Environ. Microbiol.* **62**, 2521–2526.

Dorn, J. G., Brusseau, M. L., and Maier, R. M. (2005) Real-time in situ monitoring of bioactive zone dynamics in heterogeneous systems. *Environ. Sci. Technol.* **39**, 8898–8905.

Dowd, S. E., and Pepper, I. L. (2007) PCR: agricultural and environmental applications for soil microbes. In "Manual of Environmental Microbiology" (C. J. Hurst, R. L. Crawford, D. A. Lipson, A. L. Mills, and L. D. Stetzenbach, eds.), 3rd ed., American Society for Microbiology Press, Washington, DC, pp. 676–686.

Gao, H., Yang, Z. K., Gentry, T. J., Wu, L., Schadt, C. W., and Zhou, J. (2007) Microarray-based analysis of microbial community RNAs by whole community RNA amplification. *Appl. Environ. Microbiol.* **73**, 563–571.

Gentry, T. J., Wickham, G. S., Schadt, C. W., He, Z., and Zhou, J. (2006) Microarray applications in microbial ecology research. *Microb. Ecol.* **52**, 159–175.

Green, M. R., and Sambrook, J. (2012) "Molecular Cloning: A Laboratory Manual," 4th ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Handelsman, J. (2004) Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.* **68**, 669–685.

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., and Goodman, R. M. (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem. Biol.* **5**, R248–R249.

Hazen, T. C., Dubinsky, E. A., DeSantis, T. Z., Andersen, G. L., Piceno, Y. M., Singh, N., *et al*. (2010) Deep-sea oil plume enriches indigenous oil-degrading bacteria. *Science* **330**, 204–208.

He, Z., Deng, Y., Van Nostrand, J. D., Tu, Q., Xu, M., Hemme, C. L., *et al*. (2010) GeoChip 3.0 as a high-throughput tool for analyzing microbial community composition, structure, and functional activity. *ISME J.* **4**, 1167–1179.

Heid, C. A., Stevens, J., Livak, K. J., and Williams, P. M. (1996) Real time quantitative PCR. *Genome Res.* **6**, 986–994.

Hill, V. R., Mull, B., Jothikumar, N., Ferdinand, K., and Vinjé, J. (2010) Detection of GI and GII noroviruses in ground water using ultrafiltration and TaqMan real-time RT-PCR. *Food Environ. Virol.* **2**, 218–224.

Hurst, C. J., Crawford, R. L., Lipson, D. A., Mills, A. L., and Stetzenbach, L. D. (eds.) (2007) Manual of Environmental Microbiology. 3rd ed., American Society for Microbiology Press, Washington, DC.

Huang, J. P., Swain, A. K., Thacker, R. W., Ravindra, R., Andersen, D. T., and Bej, A. K. (2013) Bacterial diversity of the rock-water interface in an east antarctic freshwater ecosystem, Lake Tawani(P). *Aquat. Biosyst.* **9**, 4.

Korves, T. M., Piceno, Y. M., Tom, L. M., DeSantis, T. Z., Jones, B. W., Andersen, G. L., *et al*. (2013) Bacterial communities in commercial aircraft high-efficiency particulate air (HEPA) filters assessed by PhyloChip analysis. *Indoor Air* **1**, 50–61.

Lasken, R. S. (2012) Genomic sequencing of uncultured microorganisms from single cells. *Nat. Rev. Microbiol.* **10**, 631–640.

Leveau, J. H. J., Loper, J. E., and Lindow, S. E. (2007) Reporter gene systems useful in evaluation in situ gene expression by soil- and plant-associated bacteria. In "Manual of Environmental Microbiology" (C. J. Hurst, R. L. Crawford, D. A. Lipson, A. L. Mills, and L. D. Stetzenbach, eds.), 3rd ed., American Society for Microbiology Press, Washington, DC, pp. 734–747.

Loy, A., Maixner, F., Wagner, M., and Horn, M. (2007) probeBase—an online resource for rRNA-targeted oligonucleotide probes: new features 2007. *Nucleic Acids Res.* **35**, D800–D804.

Lu, Z., Deng, Y., Van Nostrand, J. D., He, Z., Voordeckers, J., Zhou, A., *et al*. (2012) Microbial gene functions enriched in the deepwater horizon deep-sea oil plume. *ISME J.* **6**, 451–460.

Matsuki, T., Watanabe, K., Fujimoto, J., Takada, T., and Tanaka, R. (2004) Use of 16S rRNA gene-targeted group-specific primers for real-time PCR analysis of predominant bacteria in human feces. *Appl. Environ. Microbiol.* **70**, 7220–7228.

McLaughlin, M. R., Brooks, J. P., and Adeli, A. (2009) Characterization of selected nutrients and bacteria from anaerobic swine manure lagoons on sow, nursery, and finisher farms in the Mid-South USA. *J. Environ. Qual.* **38**, 2422–2430.

Muyzer, G., de Waal, E. C., and Uitterlinden, A. (1993) Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Appl. Environ. Microbiol.* **59**, 695–700.

National Academy of Sciences (2007) "The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet," The National Academies Press, Washington, DC.

Ng, J. (2012) Recovery of carbon and nitrogen cycling and microbial community functionality in a post-lignite mining rehabilitation chronosequence in east Texas. Ph.D. Dissertation, Texas A&M University, College Station, TX.

Nocker, A., Cheung, C.-Y., and Camper, A. K. (2006) Comparison of propidium monoazide with ethidium monoazide for differentiation of live vs. dead bacteria by selective removal of DNA from dead cells. *Appl. Environ. Microbiol.* **67**, 310–320.

Pearson, W. R. (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol.* **183**, 63–98.

Pérez-Losada, M., Cabezas, P., Castro-Nallar, E., and Crandall, K. A. (2013) Pathogen typing in the genomics era: MLST and the future of molecular epidemiology. *Infect. Gen. Evol.* **16**, 38−53.

Petric, I., Philippot, L., Abbate, C., Bispo, A., Chesnot, T., Hallin, S., *et al*. (2011) Inter-laboratory evaluation of the ISO standard 11063 "Soil quality—Method to directly extract DNA from soil samples." *J. Microbiol. Methods* **84**, 454−460.

Portier, P., Saux, M. F-L., Mougel, C., Lerondelle, C., Chapulliot, D., Thioulouse, J., *et al*. (2006) Identification of genomic species in *Agrobacterium* Biovar 1 by AFLP genomic markers. *Appl. Environ. Microbiol.* **72**, 7123−7131.

Ribot, E. M., Fair, M. A., Gautom, R., Cameron, D. N., Hunter, S. B., Swaminathan, B., *et al*. (2006) Standardization of pulsed-field gel electrophoresis protocols for the subtyping of *Escherichia coli* O157:H7, *Salmonella*, and *Shigella* for PulseNet. *Foodborne Pathog. Dis.* **3**, 59−67.

Saiki, R. K., Scharf, S., Faloona, F., Mullis, K. B., Horn, G. T., Erlich, H. A., *et al*. (1985) Enzymatic amplification of β-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* **230**, 1350−1354.

Schwartz, D. C., and Cantor, C. R. (1984) Separation of yeast chromosome-sized DNAs by pulsed-field gradient gel electrophoresis. *Cell* **37**, 67−75.

Seoane, J., Yankelevich, T., Dechesne, A., Merkey, B., Sternberg, C., and Smets, B. F. (2011) An individual-based approach to explain plasmid invasion in bacterial populations. *FEMS Microbiol. Ecol.* **75**, 17−27.

Sørensen, S., Burmølle, M., and Hansen, L. H. (2006) Making bio-sense of toxicity: new developments in whole-cell biosensors. *Curr. Opin. Biotechnol.* **17**, 11−16.

Thomas, E. J., King, R. K., Burchak, J., and Gannon, V. P. (1991) Sensitive and specific detection of *Listeria monocytogenes* in milk and ground beef with the polymerase chain reaction. *Appl. Environ. Microbiol.* **57**, 2576−2580.

Tsien, R. Y. (2010) Nobel lecture: constructing and exploiting the fluorescent protein paintbox. *Integr. Biol.* **2**, 77−93.

van Frankenhuyzen, J. K., Trevors, J. T., Lee, H., Flemming, C. A., and Habash, M. B. (2011) Molecular pathogen detection in biosolids with a focus on quantitative PCR using propidium monoazide for viable cell enumeration. *J. Microbiol. Methods* **87**, 263−272.

Van Nostrand, J. D., Wu, L., Wu, W., Huang, Z., Gentry, T. J., Deng, Y., *et al*. (2011) Dynamics of microbial community composition and function during in situ bioremediation of a uranium-contaminated aquifer. *Appl. Environ. Microbiol.* **77**, 3860−3869.

Van Nostrand, J. D., Gentry, T. J., and Zhou, J. (2013) Microarray-based amplification product detection and identification. In "Advanced Techniques in Diagnostic Microbiology" (Y.-W. Tang, and C. W. Stratton, eds.), 2nd ed., Springer, New York, NY, pp. 397−412.

Versalovic, J., Koeuth, T., and Lupski, R. (1991) Distribution of repetitive DNA sequences in eubacteria and application to fingerprinting of bacterial genomes. *Nucleic Acids Res.* **19**, 6823−6831.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., van de Lee, T., Hornes, M., *et al*. (1995) AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Res.* **23**, 4407−4414.

Wakelin, S. A., Anand, R. R., Reith, F., Gregg, A. L., Noble, R. R. P., Goldfarb, K. C., *et al*. (2012) Bacterial communities associated with a mineral weathering profile at a sulphidic mine tailings dump in arid Western Australia. *FEMS Microbiol. Ecol.* **79**, 298−311.

Wang, J., Van Nostrand, J. D., Wu, L., He, Z., Li, G., and Zhou, J. (2011) Microarray-based evaluation of whole-community genome DNA amplification methods. *Appl. Environ. Microbiol.* **77**, 4241−4245.

Watson, J. D., and Crick, F. H. C. (1953) Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature* **171**, 737−738.

Welsh, J., and McClelland, M. (1990) Fingerprinting genomes using PCR with arbitrary primers. *Nucleic Acids Res.* **19**, 861−866.

Woese, C. R. (1987) Bacterial evolution. *Microbiol. Rev.* **52**, 221−271.