SCIENCE PROGRESS

High-throughput DNA sequencing technologies for water and wastewater analysis

Science Progress
2019, Vol. 102(4) 351–376
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0036850419881855
journals.sagepub.com/home/sci
SAGE

Alexander WY Chan, James Naphtali and Herb E Schellhorn
Department of Biology, McMaster University, Hamilton, ON, Canada

Abstract
Conventional microbiological water monitoring uses culture-dependent techniques to screen indicator microbial species such as *Escherichia coli* and fecal coliforms. With high-throughput, second-generation sequencing technologies becoming less expensive, water quality monitoring programs can now leverage the massively parallel nature of second-generation sequencing technologies for batch sample processing to simultaneously obtain compositional and functional information of culturable and as yet uncultured microbial organisms. This review provides an introduction to the technical capabilities and considerations necessary for the use of second-generation sequencing technologies, specifically 16S rDNA amplicon and whole-metagenome sequencing, to investigate the composition and functional potential of microbiomes found in water and wastewater systems.

Keywords
Second-generation sequencing, metagenomics, 16S rDNA amplicon sequencing, whole-metagenome sequencing, wastewater, water quality monitoring, bioinformatics, microbial communities, microbiome

Introduction

Routine water and wastewater testing is critical for many public health and industrial monitoring programs. While traditional tests for microbiological contamination, such as fecal coliform counts, provide indirect abundance measurements of culturable microbes, the advent of inexpensive new DNA sequencing technologies offers the potential to dramatically increase the scope of monitoring through direct

Corresponding author:
Herb E Schellhorn, Department of Biology, McMaster University, Life Science Building, 1280 Main Street West, Hamilton, ON L8S 4K1, Canada.
Email: schell@mcmaster.ca

and comprehensive identification. For environmental surveillance, DNA sequencing can be used to detect microbes (bacteria and archaea) that are not currently routinely tested for, including difficult-to-culture organisms such as *Clostridium* spp.[1] and *Methanobrevibacter* spp.[2] This information can be used for microbial source tracking to identify point sources of contamination[3] and evaluate the effectiveness of remedial and control measures. Many industrial waste treatment processes including those applied to wastewater depend on complex microbial fermentation.[4] High-throughput DNA sequencing can be used to monitor the efficacy of such processes, provide ancillary information to optimize wastewater treatment technology (particularly those that have a proprietary component) and provide biological information regarding process failure.[5]

The challenges in employing DNA sequencing technology for characterizing microbial communities are manifold and include the development of robust sampling methodologies and the application of informative statistical analyses. Furthermore, between sampling and statistics lies the daunting task of selecting the correct sequencing technologies to accomplish water and wastewater monitoring goals. The plethora of options available for sequence data processing can be overwhelming for those unfamiliar with bioinformatics. In this review, we discuss accepted practices in implementing DNA sequencing technologies and describe the bioinformatic tools available for sequence data processing. We also outline the benefits of DNA sequencing technologies over culture-based microbiological methods in water and wastewater quality monitoring and provide recommendations for designing monitoring programs.

Culture-based microbiological methods in water and wastewater analysis

In North America, public health authorities use plate culturing methods to assess water microbial quality.[6,7] In these methods, standardized volumes of water are initially passed through 0.45 μm filters. Filters are then placed on selective agar media that facilitate exclusive growth of the bacteria of interest. The degree of microbial contamination is quantified in colony-forming units (CFUs) per unit volume by counting the number of colonies on a plate.[8] Routine testing comprises quantification of *Escherichia coli* and total coliforms. Unfortunately, waterborne illnesses are caused by a multitude of bacterial genera including the following: *Shigella, Leptospira, Legionella, Vibrio, Salmonella, Campylobacter*, and *Arcobacter*.[9] Genera-specific plate culturing methods exist for many pathogens. However, these methods require specialized facilities and expertise.[7,8] Culture-based tests can only detect the presence a few microbial groups at a time with limited taxonomic resolution. Standard water quality monitoring methods that simultaneously test for all microbial pathogens would greatly benefit water quality monitoring in the interest of public health.

Methods to simultaneously characterize all microbes in a system could similarly benefit wastewater treatment system design and monitoring. In wastewater treatment, organic waste is metabolized into gases such as methane and hydrogen in a

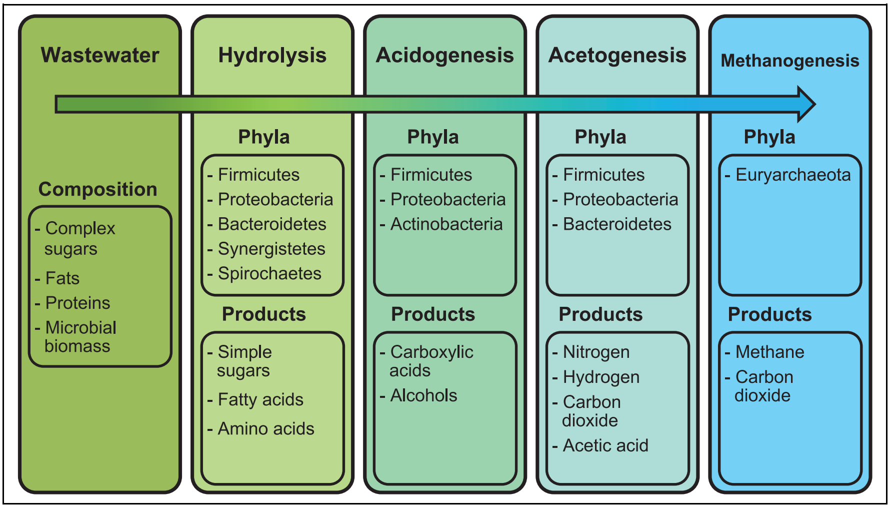

Figure 1. Microbial phyla responsible for anaerobic wastewater treatment processes—hydrolysis, acidogenesis, acetogenesis, and methanogenesis.
Decomposition of organic substances occurs in a stepwise fashion in which the metabolites of one group of microbes become the substrates of the next.

process called anaerobic digestion (AD; Figure 1).[4] AD is sustained by a diverse, core population of microbes which syntrophically metabolize complex molecules. In syntrophic metabolism, metabolites of certain taxa become the substrate for others (Figure 1).[10,11] Therefore, the species present in wastewater treatment systems are directly responsible for desired treatment outcomes.[12,13] Characterization of microbial community composition and metabolic capacity can facilitate the improvement of wastewater treatment systems through bioaugmentation or system modifications that promote the growth of effective waste-degrading microbes.[14,15] Culture-based assays have been used to test the metabolic activity of microbes in the effluent of onsite wastewater treatment system (OWTS).[16] However, like plate-culture water quality assays, these fail to fully characterize wastewater treatment systems as many relevant microbes are difficult to culture. This has given rise to a quest for culture-independent methods through second-generation sequencing (SGS).

SGS

SGS technologies are characterized by the sequencing of millions of short (<1000 bp) DNA fragments which are attributed to their sample of origin by appended index sequences.[17] Sequence information from each DNA fragment, or reads, are assigned taxonomy through alignment with sequences in microbial genome databases.[18] Therefore, SGS offers massively parallel methods for simultaneous and comprehensive identification of microbes in complex communities across several samples using DNA sequencing.[19] By removing the need for culturing, SGS allows for the identification of unculturable taxa that may play key roles in pathogenicity or wastewater treatment.[4,20] Furthermore, the relative abundance of microbes in a sample can be quantified using the number of reads assigned to each taxonomic group.[21,22] In addition to taxonomic assignment, reads can be aligned to gene databases to elucidate the functional gene pathways or utilized in de novo methods to construct genomes of novel microbial species.[23,24]

Illumina sequencing platforms have been widely adopted as the sequence platforms of choice[25] for SGS due to lower per-base costs and error rates and greater data output in comparison with other platforms.[26] Illumina utilizes sequencing by synthesis (SBS) technology in which DNA fragments are bound to a solid-phase flow cell, amplified and sequenced using fluorescently labeled nucleotides.[27] The maximum read length of any Illumina SGS platform is currently 300 bp.[28] Molecules can be sequenced from one end (single-end reads) or from both ends toward the middle (paired-end reads).[28] Paired-end reads can potentially be merged to create a longer contiguous sequence if there is overlap between the reads.[28] Many studies have also employed pyrosequencing,[29–31] a discontinued sequencing technology pioneered by 454 Life Sciences. Although this technology is no longer being advanced, studies employing pyrosequencing remain a valuable source of information as pyrosequencing results have been shown to be comparable to those obtained with SBS.[32]

The next frontier in DNA sequencing is long-read sequencing, also known as third-generation sequencing (TGS). TGS platforms can sequence hundreds of kilobases of a single DNA molecule.[33] Long reads have many applications including high-resolution taxonomic assignment, characterization of genome regions with repetitive sequences, and identification of epigenetic markers.[34] Unfortunately, TGS currently has relatively greater sequencing costs,[35] higher error rates, and lower sample throughput compared to SGS.[36] These shortcomings make TGS a poor alternative to SGS for metagenomic sequencing[37] of complex microbial communities at this time.

Most published SGS microbiome studies utilize variants of Illumina's MiSeq and HiSeq platforms. The maximum output of the most recent MiSeq and HiSeq models are 2.5×10^7 and 5×10^9 reads, respectively.[28] Illumina also recently released the NovaSeq platform which has a maximum output of 2×10^{10} reads.[28] Illumina platform purchasing and service costs increase with data generation capacity.[28] For example, excluding library preparation and quality control costs, a paired-end, 150 bp read-length (2 × 150 bp) sequencing run on one flow cell lane can cost approximately US$1500 (2019 prices) on a MiSeq platform.[38] In contrast, for the benefit of more reads, the same run costs approximately US$3000 on the HiSeq[38,39] and US$9000 on the NovaSeq.[39,40] Platform choice is informed by study design and number of required reads per sample, also referred to as sequencing depth. To maintain sequencing depth, the total number of reads required increases with the number of samples to be analyzed. Sequencing depth requirements are dependent on the project goals and SGS technique. Currently, the two primary SGS techniques for the determination of microbial community composition and function are 16S ribosomal DNA (16S rDNA) sequencing and whole-metagenome sequencing (WMS), respectively.

Although government and industry have recently begun to explore the use of SGS for biomonitoring of aquatic environments,[6,41] high-throughput sequencing has yet to be widely adopted for monitoring of water and wastewater treatment systems. The primary barriers to wide-scale adoption of SGS are cost and expertise.[42] Culture-based methods, though limited in scope, require considerably less resources, training, and time. However, decreasing sequencing costs,[43] curation of bioinformatic protocols,[44,45] and development of user-friendly sequence analysis tools[46,47] continue to improve the feasibility of SGS for routine monitoring. The following sections provide recommendations and guidelines for the application of 16S rDNA and WMS analysis in water and wastewater analysis.

16S rDNA sequencing applications in wastewater treatment and water quality monitoring

16S rDNA amplicon sequencing is the de facto molecular method for microbial identification in complex environmental samples.[48] In a recent study, 16S rDNA sequencing was used to characterize and contrast microbial communities of anaerobic digesters in biogas plants (BPs) and sewage treatment plants (STPs).[49] Microbial diversity was greater in STPs than in BPs, while BP core community members were more metabolically linked than those of STPs. Differences in microbial interactions and community members between the two plant types were attributed to the greater variability in STP influent composition.[49] The simultaneous digestion of sewage and agricultural waste has been suggested as a process to increase biogas production and cost efficiency.[49] However, the results of this study indicate that the communities that degrade each substrate type are distinct and that co-digestion may not be optimal.

El-Chakhtoura et al.[50] employed SGS to assess the stability of microbial community structure from water treatment plant to a distribution endpoint. Plant and endpoint communities were significantly different which indicated that microbial populations underwent substantial changes within the water distribution network.[50] Specifically, the abundance of rare taxa (e.g. Nitrospirae, Acidobacteria, and Gemmatimonadetes) was greater at the endpoint in than at the water treatment plant.[50] Although the observed microbial community changes did not constitute a public health risk,[50] these 16S rDNA sequencing results support the need for water quality assessments throughout distribution networks.

16S rDNA sequencing principles

16S rDNA encodes for the ubiquitous and highly conserved 16S RNA subunit of bacterial and archaeal ribosomes.[51] Bacterial and archaeal phylogeny is based on levels of similarity between full-length (˜1540 bp) 16S rDNA sequences.[52] Within

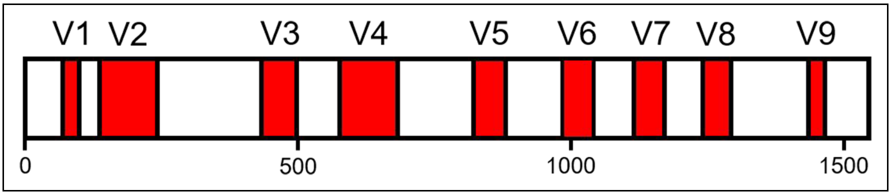

Figure 2. Distribution of V1–V9 hypervariable regions (HVRs) along a linear representation of the *Escherichia coli* 16S rDNA sense strand.
HVRs are shown in red, with widths to scale; white regions represent conserved regions and numbered tick marks indicate nucleotide position.

16S rDNA, there are nine (V1–V9) hypervariable regions (HVRs, Figure 2). Gene fragments with relatively diverse nucleotide compositions are used to distinguish microbial taxa.[53] HVRs are flanked by conserved sequences that allow them to be targeted and amplified through polymerase chain reaction (PCR) using universal primers that capture a broad range of taxa.[54] PCR amplification isolates 16S rDNA from complex mixtures of DNA by increasing their concentration. In addition, PCR can be used to attach adapter sequences that facilitate binding to sequencing machines and index sequences that identify the sample of origin for each amplicon. By convention, 16S rDNA primers are named according to their corresponding nucleotide positions (NP) in *E. coli* 16S rDNA[54] and their replication direction with respect to the 5′ to 3′ direction of the sense strand denoted by “f” and “r” for forward and reverse, respectively. For example, 341f/785r is a primer pair that spans the V3–V4 regions[55] (Figure 2).

Taxonomic profiles consisting of the bacterial and archaeal taxa within a sample and their relative abundances are constructed from the analysis of HVRs. Microbial groups are identified based on HVR composition, and the number of reads attributed to each group is used to calculate the relative abundance of those organisms in a sample. HVR-based taxonomic assignment is reliable down to the genus level.[56] However, taxonomic assignment accuracy substantially decreases at the species level.[56] Species assignment is only recommended for HVR sequences that match exactly with reference database sequences.[57] The presence of species of interest can also be confirmed with species-specific PCR as a follow-up to 16S rDNA sequencing. Conserved gene markers such as conserved signature protein and indel sequences provide effective PCR targets for species-level microbial detection.[58]

SGS platforms are currently unable to sequence the entire length of 16S rDNA in a single read. As a result, researchers must amplify and analyze segments of the 16S rRNA gene[48] that may include up to three HVRs. Furthermore, microbial community profiles can significantly differ depending on the HVR(s) sequenced[59] because the taxa present may be difficult to resolve due to a lack of nucleotide differences in the chosen HVRs. Therefore, the choice of HVR(s) can have significant impacts on sequencing results and may lead to erroneous conclusions regarding the systems being studied.

Table 1. 16S rDNA hypervariable regions sequenced in microbiome projects.

Hypervariable regions	Sample types	Primer pairs
V1–V2	Activated sludge[64]	27f-338r, 8f/357r
V1–V3	Freshwater[65]	8f/556r
V3–V4	Freshwater[55,66]	341f/805r, 341f/785r
V4	Freshwater[67]	515f/806r, 515f/808r
V4–V5	Anaerobic digester[10,68]	515f/909r, 563F/926R
V5–V6	Freshwater[3,69], Anaerobic digester[33]	807f/1050r
V6–V8	Anaerobic digester[70]	926f/1392r

HVR and primer choice

Unfortunately, there is a lack of consensus as to which HVRs are the most reliable for taxonomic profiling.[60] Ideally, studies should employ 16S rDNA regions that capture the greatest degree of diversity and obtain accurate estimates of relative abundance of individual taxonomic group members. In silico comparisons of taxonomic assignment with 16S rDNA regions and full-length 16S rDNA sequences found that the V1–V3 (NP 27–519) and V1–V4 (NP 63–685) regions produced the most similar bacterial species assignments and richness estimates.[52] Tremblay et al.[61] found that, in vitro, a primer pair (515f/806r) targeting the V4 HVR yielded results that were in better agreement with WMS results than V6–V8 (926f/1392r) and V7–V8 (1114f/1392r) pairs.

In addition to the HVR regions targeted, primers can differ based on sequence composition. 16S rDNA sequencing results obtained with different primers targeting the same HVR can significantly differ due to primer bias; greater primer binding affinity for sequences that have fewer base mismatches. Allowing for one primer mismatch greatly increases primer taxonomic coverage in silico.[62] However, experimental primer testing has shown that even a single primer mismatch can significantly bias results.[63] Substituting primer mismatches with degenerate bases significantly reduces primer bias.[63]

Despite the lack of consensus regarding HVR and primer choice for microbial taxonomic profiling, researchers can make informed decisions based on in silico and empirical testing of primer pairs. Consistency should be prioritized in ongoing microbiome studies to increase the likelihood that observed community differences are a result of system environment and biological activity opposed to changes in the HVRs targeted.[60,61] Table 1 shows the examples of HVRs used in literature studies of water and wastewater systems.

16S rDNA sequence data processing methods

16S rDNA amplicon sequence data are outputted in text format with quality scores which indicate the probability of error for each base call.[71] Reads are grouped into files according to their sample of origin based on index sequences in a process called demultiplexing. Before taxonomic assignment, sample sequence reads are trimmed to remove adapter sequences and low-quality bases to prevent spurious taxonomic assignment. Adapter-trimming programs include Cutadapt[72] and Trimmomatic.[73] Examples of quality-filtering programs are Sickle[74] and BBDuk.[75] Paired-end reads can be merged if they meet a threshold for sequence overlap using programs such as PANDAseq[76] and BBMerge.[75] In addition, reads should be processed to identify and remove chimeric sequences that can be mistakenly attributed to unique taxa.[77] Available 16S rDNA chimera removal programs include ChimeraSlayer and UCHIME.[78] Sequences are then taxonomically classified using tools such as RDP Classifier, UCLUST, VSEARCH, and BLAST[56] which align sample sequences with those found in 16S rDNA reference databases. The primary reference databases used in 16S rDNA microbiome studies are Greengenes, RDP, SILVA, and NCBI.[79]

A common method for grouping sequences for taxonomic classification is to cluster them into operational taxonomic units (OTUs) at a level of 97% similarity.[80] Read counts for each sequence in a group are summed and attributed to an OTU. Taxonomy is assigned to these OTUs by aligning sequences representative of each OTU with those found in reference databases.[71] Whelan and Surette[81] compared multiple clustering programs and found that some algorithms produced substantially erroneous results when characterizing human-sourced mock microbial communities. Most notably, for certain mock communities, DNACLUST overestimated diversity by nearly 4000% and UPARSE underestimated diversity by greater than 300%.[81] Although no single program outperformed all others for every mock community, AbundantOTU+ coupled with RDP Classifier and the Greengenes 2011 database provided the best overall performance.[81]

Errors associated with clustering algorithms can be circumvented with DADA2, an alternative amplicon processing program that uses Illumina error-modeling to resolve sequence reads.[44] Rather than OTUs, DADA2 produces assigned sequence variants (ASVs). Whereas OTUs may change with the sequences included in an analysis,[82] ASVs are characterized on a sequence-to-sequence basis using sequencing error profiles and are thus more stable.[83] Therefore, between-study comparisons of ASVs are more robust than those of OTUs because they are less dependent on study-specific data processing parameters. In a benchmarking study, DADA2 outperformed established OTU sequence clustering programs when characterizing mock, mouse fecal and human vaginal communities. DADA2 was able to identify sequence variants and had lower output residual error rates, fewer false positives, and more correct taxonomic assignments than UPARSE, MED, UCLUST, and Mothur.[44] ASV stability and greater taxonomic assignment accuracy make DADA2 a strong candidate for the sequence processing program of choice for future 16S rDNA microbiome studies.

Determining the optimal set of bioinformatics programs for all SGS applications is beyond the scope of this review. Prospective microbiome researchers are encouraged to compare the benefits and shortcomings of the programs available. A useful starting point for software selection is the suite of default programs utilized by the

16S rDNA data analysis platform, QIIME2 (quantitative insights into microbial ecology "2").[46] QIIME2 provides a framework in which to trim, quality-filter, resolve, and annotate 16S rDNA sequences and is the successor to the extensively utilized bioinformatics platform, QIIME.[84] In addition, protocols established by microbiome sequencing consortia, such as the Human Microbiome Project[85] and the Earth Microbiome Project,[86] can be valuable resources for designing a 16S rDNA sequence processing pipeline.

16S rDNA sequencing depth

Sequencing depth can differ by several orders of magnitude between samples because of poor mixing or improper DNA input standardization prior to sequencing. The number of taxa, particularly rare species, detected increases with sequencing depth.[86] This complicates community analyses because variation in read counts can lead to significantly different estimates of diversity between samples[87] that may be the result of differences in sequencing depth rather than biology.[88]

For 16S rDNA studies, a common practice to account for sequencing depth differences is to rarefy sample counts[88] by randomly removing sequence reads without replacement until all samples reach a chosen sequencing depth.[89] Sequencing depths should be chosen to balance the number of samples kept with the level of diversity captured. Examples of sequencing depth utilized in water and wastewater studies is shown in Table 2. In some cases, it is advisable to discard samples with lower read counts because these samples may contain higher proportions of contaminating sequences from DNA extraction kits, PCR reagents, and the lab environment.[90] However, lower read counts may be unavoidable in drinking water samples with low biomass. Measures to mitigate the contribution of contamination to sample reads include processing blank samples to identify contaminating sequences for in silico removal and concentration of sample biomass prior to DNA extraction.[90]

An alternative method to rarefying read counts for the normalization of sequencing depth is to apply a variance stabilizing transformation (VST) to count data.[87] Generally, rarefied sample communities have been reported to cluster more accurately according to sample origin, while VST has been reported to have more statistical power when differentiating sample groups.[88,87] Proponents of VST argue that rarefying data discards statistically relevant data, whereas practitioners of rarefying argue that VST fails to address library size effects and leads to higher false discovery rates.[88,87]

WMS applications in wastewater and water monitoring

An alternative to SGS analysis with 16S rDNA amplicon sequencing for taxonomic profiling is WMS. In contrast with single-gene sequencing techniques such as 16S rDNA, WMS is the analysis of all DNA sequences from a population of microorganisms in a given environment. WMS sequences are obtained from randomly fragmented genomes within a sample and are therefore not limited to taxonomic

Table 2. Second-generation sequencing study designs for water and wastewater monitoring projects.

Project design and community profiling	Sequencing platform and technique	No. of samples	Type and no. of replicates	Paired-end read length (bp)	Depth (reads per sample)
Longitudinal community analysis of bog lakes sampled weekly over 1–5 years.[91]	HiSeq, 16S rDNA	1387	Biological, 547 Technical, 2	2 × 150	5 × 104
Microbial source tracking of fecal bacteria inputs into the Lake Superior watershed.[3]	HiSeq/MiSeq, 16S rDNA	319	Technical, 3	2 × 150	1.9×10^5
Longitudinal community analysis of anaerobic digesters over 14 days mixed with differing salt concentrations.[69]	MiSeq, 16S rDNA	15	Biological, 3	2 × 250	4.5×10^4
Core microbiome community analysis of activated sludge systems for longitudinal characterization.[92]	HiSeq, 16S rDNA	39	Biological, 13	2 × 150	4.0×10^4
Metagenomic characterization of nitrogen-contaminated groundwater.[93]	HiSeq, WMS	5	0	2 × 100	5×10^7
WMS of shore-water and sand-water samples.[94]	HiSeq, WMS	32	Biological, 16	2×100	9.4×10^6
Taxonomic and functional analysis of anaerobic digesters processing sludge and manure.[95]	HiSeq, WMS	14	Biological, 6	2 × 100	2.8×10^7
Temporal study of cellulose-degrading lab-scale anaerobic digesters over 1 year. Assembled complete genomes.[23]	HiSeq, WMS	6	Biological, 3	2 × 150	1×10^8

WMS: whole-metagenome sequencing.

marker genes. This technique does not require primers and thus avoids issues with primer bias. WMS profiles have been shown to more accurately reflect expected community compositions.[96] Furthermore, diversity estimates obtained with WMS are often comparable or even greater than those obtained with 16S rDNA sequencing due to greater taxonomic resolution based on multiple gene markers.[97–99]

The greater breadth of sequencing information obtained by WMS also allows for more applications than 16S rDNA. For example, whereas 16S rDNA analysis cannot be used to identify species not present in sequence databases, DNA sequences obtained by WMS can be used to assemble previously uncharacterized genomes.[100] In addition to taxonomic profiling, WMS can identify metabolic gene pathways in a community that may be of functional significance for a given system.[4,101] Taxonomic and functional profiles of metagenomes can be compared, and chemical metadata enables insight into multiple elements of the microbiome. These include the variability of functional potentials across samples, the effect environmental parameters have on metagenome composition and functional potential, and the presence of microbial communities and genes (i.e. biomarkers) characteristic to the study environment.[4,102] Profiling metabolic functions in environmental microbiomes is important as microbial communities drive biogeochemical processes.[103] However, these additional applications require greater sequencing depths than those typically analyzed by required for 16S rDNA amplicon sequencing and thus incur greater costs than 16S rDNA amplicon sequencing.[67,104]

WMS has been used to study community growth patterns of cyanobacterial harmful algal blooms (cHABs).[105] Global proliferation of cyanobacteria in water environments has been linked to eutrophication by phosphorous and nitrogen species from residential and agricultural sources.[106] The use of WMS to profile metabolic genes in microbial communities reveals how cyanobacterial bloom communities utilize these nutrients and adapt to changing environmental conditions.[105,107] By sequencing cyanobacterial blooms over time, predictive growth models describing changes in community composition and function can be made to forecast water quality impairment due to increased nutrient loading.[108]

In anaerobic digesters, fluctuations in physiochemical parameters such as pH, temperature, total solids, and organic and inorganic chemical substances can inhibit the activity of some microbes, leading to a shift in community proportions and the accumulation of metabolites at inhibitory levels.[109–111] WMS analysis of anaerobic sludge digester by Li et al.[112] demonstrated a decrease in abundance of the hydrogenotrophic methanogen *Methanosaeta*, which produces methane from the reduction of carbon dioxide using hydrogen, and stable growth of the acetoclastic methanogen *Methanosarcina*, which produces methane from acetate, at elevated levels of ammonium.[112] This community shift was also reflected at the functional gene level, where gene abundances encoding for enzymes responsible for the acetoclastic pathway increased during ammonium stress.[112] As a result of ammonium stress, methane production decreased due to the inhibition of hydrogenotrophic methanogens and their genes.[112] Therefore, the ability for WMS to obtain species-level taxonomic abundances as well as gene abundances coding for enzymes and cellular mechanisms governing syntrophic interactions in the AD process is useful for evaluating the biological basis of digester performance.

Since WMS is not subjected to amplification bias, low-abundance organisms such as pathogens and viruses can be accurately detected, quantified, and sourced to determine their presence and persistence in environments. [94,113–115] Culture-based and molecular-based (PCR) assays for pathogen detection require selective media and specific primers to detect pathogens. However, such methods are limited in sensitivity and breadth of detection due to primer mismatches, inability to culture rare and novel pathogens, and simply because such analyses screen pre-selected targets.[116,117] The high-throughput, sequence-independent nature of WMS enables the characterization of all offers comprehensive insight into the abundance, diversity, and composition of known and unknown pathogens in a given environment.

WMS has been used to source and quantify antibiotic resistance genes (ARGs) and virulence factors in waterbodies[118,119] and wastewater processing systems.[113,115] Effluent waste runoff from residential areas, pharmaceutical and agricultural industries, and wastewater treatment plants (WWTP) is discharged into lakes, resulting in the accumulation of ARG and pathogens in waterbodies.[120,119] By sequencing the total DNA within a sample, WMS provides abundance estimates of ARG types, virulence genes coding for bacterial motility, cell adherence and secretion, and mobile genetic elements (MGE) such as transposons, plasmids, and bacteriophages.[113,121,122] By analyzing the abundance of these genes, the potential of ARG and virulence factor propagation by horizontal gene transfer across environments can be determined.[113,121] In addition, novel pathogenic species and strains can be profiled by WMS for pre-emptive treatment of outbreaks in water and wastewater environments.[123–125]

WMS principles

In WMS, the genomic DNA is extracted, randomly sheared into fragments, and subsequently sequenced. DNA sequencing libraries are prepared by fragmentation, size-selection, labeling, and enrichment of DNA. Library preparation protocols used with Illumina sequencers include the TruSeq DNA series of library kits and the transposon-based Nextera XT Library Kit protocols.[28] DNA is first fragmented by nebulization, sonication, or enzymatic digestion.[126,25] After fragmentation, ends of DNA fragments are repaired and adapters are ligated to facilitate sample DNA binding to Illumina flow cells.[25] The fragments are then size-selected by either gel-electrophoresis or using solid-phase reversible immobilization (SPRI) beads.[127] After size-selection, DNA sequences can be enriched in concentration and purity by PCR using proprietary Illumina primers.[25] Selecting a fragmentation method depends on the type of library preparation kit used. Fragmentation by nebulization and sonication is compatible with Illumina's TruSeq Nano and PCR-free sample preparation kits with a recommended input DNA mass of approximately 1–2 μg of DNA.[25] Alternatively, the Illumina Nextera XT library preparation kit is more cost-effective and utilizes tagmentation, a process in which DNA is fragmented and simultaneously tagged using transposase enzymes while requiring only 50 ng of input DNA.[25] However, since transposases target sequence motifs, transposome-based sequence fragmentation and tagging cause biased read coverage against sequences with higher G and C nucleotide content compared to TruSeq PCR and PCR-free library preparation methods due to the random and therefore unbiased nature of DNA fragmentation by sonication or nebulization.[25,128] After optional PCR enrichment, the adapter-ligated sequences are then applied onto flow cells for sequence cluster generation and fluorescent-based nucleotide detection.

Taxonomic profiling using WMS

Before taxonomic and functional analysis, WMS reads are quality-trimmed and -filtered for adapters and low-quality base calls based on the phred scoring system. This process is similar to the filtering and trimming applied to 16S rDNA sequencing datasets with tools such as Trimmomatic[73] and BBDuk.[75] Taxonomic assignment based on whole-genomes requires different software tools than 16S rDNA sequencing. NCBI's BLAST+ (Basic Local Alignment Search Tool) and the accelerated BLAST tool, DIAMOND,[129] align sequenced reads to translated-nucleotide protein sequences stored in NCBI's non-redundant (nr) protein database. The aligned reads can then be binned and counted into taxonomic ranks using programs such as MEGAN6[130] and DUDes.[131] Marker-gene-based tools such as MetaPhlAn2,[132] PhyloSift,[133] and mOTU[134] in conjunction with alignment tools such as Bowtie2[135] and HMMER[136] assign sequenced reads to a database containing clade-specific marker genes from >7500 bacterial and archaeal species to estimate bacterial and archaeal relative abundances. Alternatively, programs such as Kraken[137] and CLARK[138] assign entire sequenced reads to genomes, often employing unique *k*-mer (sequences of user-defined length *k*) abundances for taxonomic classification. Read-based assignment tools generate slightly greater numbers of false positives in taxonomic assignment at greater read depths than marker-based assignment tools because reads that are initially falsely identified are further miscalled as more reads are sequenced.[139] The taxonomic false-positive rates produced using marker-based and read-based taxonomic assignment tools for WMS can be mitigated by classifying aligned reads using a lowest common ancestor (LCA) algorithm as provided by MEGAN.[130]

Whole-metagenome sequence assembly

Fragmented DNA sequences produced by WMS can also be aligned and assembled into contiguous sequences to form complete genomes in metagenomic samples. Assembly-based WMS analysis aligns raw reads at areas of overlap together de novo into contiguous sequences called contigs to assemble draft genomes of microbes in metagenomic samples using tools such as metaSPAdes and Megahit.[140,141] After assembly, contigs can either be binned into complete genomes for novel species discovery[100,142] or be directly mapped based on sequence homology using alignment algorithms such as BLAST and profile hidden Markov models (HMMs)[136] against nucleotide sequence databases such as NCBI's non-redundant ("nr") protein database or MetaPhlAn's clade-specific marker gene sequences[143] to quantify taxonomic and functional gene abundances. Metagenomic novel genome construction requires at least two sets of metagenomes from the same system that are differentiated by a treatment such as time of sampling or DNA extraction method.[100] The scaffolds are binned based on concerted changes in the frequency of *k*-mer sequences between treatments.[100] This binning method is called compositional-based binning, which enables the discovery of novel genomes and microbial species not by referencing known genes in existing databases, but instead by de novo assembly and binning.[100,23] Longer sequences are required for greater taxonomic and functional resolution during alignment to reference genomes.[144] However, like the misalignment of short DNA fragments, erroneous chimeric contigs can be constructed from fragments that have overlapping regions, such as sequence repeats, that do not originate from the same gene or species.[145]

The accuracy and quality of contig assemblies and the genomes produced varies substantially between available programs.[146] The quality of these genomes is evaluated by determining the presence of putative single-copy genes essential to the survival of microbes by tools such as CheckM.[142,147] The presence of the genes without duplication indicates high-quality metagenomic assembled genomes (MAGs).[142,146] Megahit and metaSPAdes are among the best performing metagenome assembly programs based on accuracy in mock community analyses, percentage of reads mapped to assemblies, computing power requirements, contig length, captured diversity, and assembly error rates.[146,148–150] Both programs capture greater than 75% of the expected diversity in mock microbial communities.[150,146,149] Megahit is recommended for studies where characterizing low-abundance diversity and microbial strain resolution are priorities.[149,150] metaSPAdes assembles longer contigs than Megahit and performs better when reconstructing expected open reading frames (ORF) of low-abundance species in mock communities with uneven species distributions.[148,149] The number of ORFs detected is indicative of the number of genes in a sample.[148] High sequencing depth for each sample is necessary to achieve complete construction and coverage of the diverse genomes present in metagenomes (Table 2).[96,104,151] High sequencing depth achieves a greater diversity of species profiles by covering low-abundance genomes and genes and mitigates the effect of sequencing errors and false-positive discovery rates generated during sequence analysis.[18,104] High sequencing depth is achieved by reducing the number of samples sequenced in flow cell lanes (Table 2).[152]

Microbiome functional profiling

WMS also provides insight into the physiological processes and abilities of microbial communities. To obtain functional profiles of metagenomic sequences, genes are first identified as coding DNA sequences, noncoding RNA genes, or other sequence motifs such as clustered regularly interspaced short palindromic repeats (CRISPRs) in a process called gene prediction.[153] Coding DNA sequences can be identified and distinguished from noncoding RNA genes using tools such as MetaGeneMark,[154] Prodigal,[155] and Prokka[156] by detecting transcription initiation sequences or ORF.[155] Metagenomic sequences are functionally annotated based on the protein and protein families categorized into protein sequence clusters of orthologous groups (COGs) they encode using software such as MG-RAST,[157] MEGAN,[130] IMG/M,[153] HUMAnN for human microbiomes,[158] and the package for the R statistical language ShotgunFunctionalizeR.[159] These tools map reads that are homologous to protein sequences previously curated and recorded into databases such as KEGG,[160] UniProt,[161] MetaCyc,[162] and SEED.[163] The number of reads that map to functional sequences can be used to quantify the abundances of genes coding for cellular and metabolic mechanisms relative to the entire metagenome, providing insight into the potential biological functions in a metagenomic sample.[101]

Determinations of microbial activity

A major limitation of 16S rDNA sequencing and WMS is that they cannot assess microbial activity because they cannot differentiate between DNA from live cells and DNA from lysed dead cells.[69,164] Methods for assessing the overall microbial activity that can accompany 16S rDNA sequencing and WMS include measurement of ATP concentrations[165] and differential staining assays that target live cells.[166] Meta-transcriptomic sequencing (MTS), also referred to as RNA-seq, is an SGS method that can assess microbial activity through determination of microbial expression levels. In MTS, the complementary DNA (cDNA) of RNA transcripts is sequenced, quantified, and annotated according to function and taxonomy.[69,167] RNA extraction and sequence processing methods for MTS have been reviewed elsewhere.[168,169]

Combining 16S rDNA and WMS approaches

WMS is a versatile tool for identifying the microbes present in water and wastewater systems, characterizing their metabolic capacity and determining their potential pathogenicity. The utility of WMS outstrips that of 16S rDNA, and researchers should consider investing in WMS if they have the resources and require a thorough understanding of microbial community structure and function.[96] However, application of 16S rDNA and WMS does not need to be mutually exclusive. 16S rDNA surveys can be used to identify unique samples for more in-depth investigation at greater sequencing depths with WMS.[152] Samples of interest are identified based on microbial community characteristics such as ecological diversity, dissimilarity between communities, or abundance of specific taxa. The combination of 16S rDNA and WMS offers a potential for greater return of investment on sequencing. The need to identify samples of interest can arise from the inherent heterogeneity of water and wastewater systems.

Heterogeneity of freshwater and wastewater systems

Microbial communities differ significantly across spatial and temporal scales in freshwater watersheds[170] and wastewater treatment systems.[69] Even within a mixed wastewater treatment system, biomass communities can differ from granule to granule.[68] Spatial heterogeneity requires researchers to consider which sampling points are of highest priority and the significance of the microbial community at each potential sampling point. For example, in drinking water monitoring programs, sampling plans may need to prioritize points closer to the consumer as microbial communities have been found to significantly vary in composition throughout distribution systems.[50] Longitudinal studies are required to obtain results that are an accurate representation of fluctuating microbial populations within a system.[69] Longitudinal studies should be designed to capture periods of environmental change or system perturbation which may have significant impacts on microbial community and system function. Furthermore, system heterogeneity necessitates biological replication to determine variability within system types and technical replication to assess the efficacy of sample homogenization measures (Table 2).[171] Finally, efforts to address system heterogeneity must be balanced with the sequencing depth and sequencing costs.

Conclusion

This review introduces SGS and provides guidelines for monitoring water and wastewater environments using SGS. Despite the complexity of implementing DNA sequencing strategies for water and wastewater quality monitoring, 16S rDNA and WMS offer comprehensive methods for the characterization of microbial communities. Using SGS, water quality professionals can explore the potential of new wastewater treatment technologies, inform drinking water quality surveys, and track the spread of pathogenic genes throughout aquatic environments. In the future, climate change and growing populations are likely to increase the frequency of water shortages and waterborne disease outbreaks around the world.[172,173] It is the responsibility of water quality professionals to utilize all the tools at their disposal for improving wastewater treatment and water resource stewardship.

Author contributions

A.C. and J.N. take co-authorship of this article. H.E.S. is the corresponding author and conceived of the topic and aided in the proof-reading and organization of the article. All authors have read and approved the final manuscript.

Acknowledgements

We also acknowledge support from the Southern Ontario Water Consortium (SOWC) Advancing Water Technologies program. We thank Craig Jowett and Chris Jowett of Waterloo Biofilter Systems (Rockwood, Ontario) and members of the Schellhorn laboratory for helpful discussions.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported with funding from the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery and Engage programs to HES.

ORCID iD

Herb E Schellhorn https://orcid.org/0000-0002-3800-3245

References

1. Cai L and Zhang T. Detecting human bacterial pathogens in wastewater treatment plants by a high-throughput shotgun sequencing technique. *Environ Sci Technol* 2013; 47(10): 5433.
2. Kirkegaard RH, McIlroy SJ, Kristensen JM, et al. The impact of immigration on microbial community composition in full-scale anaerobic digesters. *Sci Rep* 2017; 7(1): 9343.
3. Brown CM, Staley C, Wang P, et al. A High-throughput DNA-sequencing approach for determining sources of fecal bacteria in a lake superior estuary. *Environ Sci Technol* 2017; 51(15): 8263–8271.
4. Campanaro S, Treu L, Kougias PG, et al. Metagenomic analysis and functional characterization of the biogas microbiome using high throughput shotgun sequencing and a novel binning strategy. *Biotechnol Biofuels* 2016; 9: 26.
5. Pore SD, Shetty D, Arora P, et al. Metagenome changes in the biogas producing community during anaerobic digestion of rice straw. *Bioresour Technol* 2016; 213: 50–53.
6. United States Environmental Protection Agency. *Approved CWA microbiological methods for wastewater and sewage sludge* (2006–2014 ed.). Washington, DC: USEP Agency, 2017.
7. Health Canada. *Water quality—reports and publications* (2004–2019 ed.). Ottawa, ON, Canada: Government of Canada, 2019.
8. American Public Health Association, American Water Works Association, Water Environment Federation. *Standard methods for the examination of water*. Washington, DC: APHA, AWWA, WEF, 2012.
9. Pandey PK, Kass PH, Soupir ML, et al. Contamination of water resources by pathogenic bacteria. *AMB Express* 2014; 4: 51.
10. Rui J, Li J, Zhang S, et al. The core populations and co-occurrence patterns of prokaryotic communities in household biogas digesters. *Biotechnol Biofuels* 2015; 8: 158.
11. Campanaro S, Treu L, Kougias PG, et al. Metagenomic binning reveals the functional roles of core abundant microorganisms in twelve full-scale biogas plants. *Water Res* 2018; 140: 123–134.
12. Sun L, Liu T, Muller B, et al. The microbial community structure in industrial biogas plants influences the degradation rate of straw and cellulose in batch tests. *Biotechnol Biofuels* 2016; 9: 128.

13. Yao Y, Lu Z, Zhu F, et al. Successful bioaugmentation of an activated sludge reactor with Rhodococcus sp. YYL for efficient tetrahydrofuran degradation. *J Hazard Mater* 2013; 261: 550–558.
14. Tale VP, Maki JS and Zitomer DH. Bioaugmentation of overloaded anaerobic digesters restores function and archaeal community. *Water Res* 2015; 70: 138–147.
15. Connelly S, Shin SG, Dillon RJ, et al. Bioreactor scalability: laboratory-scale bioreactor design influences performance, ecology, and community physiology in expanded granular sludge bed bioreactors. *Front Microbiol* 2017; 8: 664.
16. Jalowiecki L, Chojniak JM, Dorgeloh E, et al. Microbial community profiles in wastewaters from onsite wastewater treatment systems technology. *PLoS ONE* 2016; 11(1): e0147725.
17. Kozich JJ, Westcott SL, Baxter NT, et al. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl Environ Microbiol* 2013; 79(17): 5112–5120.
18. McIntyre ABR, Ounit R, Afshinnekoo E, et al. Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. *Genome Biol* 2017; 18: 182.
19. Clooney AG, Fouhy F, Sleator RD, et al. Comparing apples and oranges? Next generation sequencing and its impact on microbiome analysis. *PLoS ONE* 2016; 11(2): e0148028.
20. Albanese D and Donati C. Strain profiling and epidemiology of bacterial species from metagenomic sequencing. *Nat Commun* 2017; 8(1): 2260.
21. Su C, Lei L, Duan Y, et al. Culture-independent methods for studying environmental microorganisms: methods, application, and perspective. *Appl Microbiol Biotechnol* 2012; 93(3): 993–1003.
22. Cocolin L, Alessandria V, Dolci P, et al. Culture independent methods to assess the diversity and dynamics of microbiota during food fermentation. *Int J Food Microbiol* 2013; 167(1): 29–43.
23. Vanwonterghem I, Evans PN, Parks DH, et al. Methylotrophic methanogenesis discovered in the archaeal phylum Verstraetearchaeota. *Nat Microbiol* 2016; 1: 16170.
24. Franzosa EA, McIver LJ, Rahnavard G, et al. Species-level functional profiling of metagenomes and metatranscriptomes. *Nat Methods* 2018; 15(11): 962–968.
25. Schirmer M, D'Amore R, Ijaz UZ, et al. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics* 2016; 17: 125.
26. Liu L, Li Y, Li S, et al. Comparison of next-generation sequencing systems. *J Biomed Biotechnol* 2012; 2012: 251364.
27. Slatko BE, Gardner AF and Ausubel FM. Overview of next-generation sequencing technologies. *Curr Protoc Mol Biol* 2018; 122(1): e59.
28. Inc I, Illumina sequencing editor, https://www.illumina.com/systems/sequencing-platforms.html (2019, accessed 29 May 2019).
29. Ye L and Zhang T. Bacterial communities in different sections of a municipal wastewater treatment plant revealed by 16S rDNA 454 pyrosequencing. *Appl Microbiol Biotechnol* 2013; 97(6): 2681–2690.
30. Zhan AB, Hulak M, Sylvester F, et al. High sensitivity of 454 pyrosequencing for detection of rare species in aquatic communities. *Methods Ecol Evol* 2013; 4: 558–565.
31. Shchegolkova NM, Krasnov GS, Belova AA, et al. Microbial community structure of activated sludge in treatment plants with different wastewater compositions. *Front Microbiol* 2016; 7: 90.

32. Luo C, Tsementzi D, Kyrpides N, et al. Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. *PLoS ONE* 2012; 7(2): e30087.
33. Jain M, Koren S, Miga KH, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol* 2018; 36(4): 338–345.
34. Van Dijk EL, Jaszczyszyn Y, Naquin D, et al. The third revolution in sequencing technology. *Trends Genet* 2018; 34(9): 666–681.
35. Pootakham W, Mhuantong W, Yoocha T, et al. High resolution profiling of coral-associated bacterial communities using full-length 16S rRNA sequence data from PacBio SMRT sequencing system. *Sci Rep* 2017; 7(1): 2774.
36. Besser J, Carleton HA, Gerner-Smidt P, et al. Next-generation sequencing technologies and their application to the study and control of bacterial infections. *Clin Microbiol Infect* 2018; 24(4): 335–341.
37. Driscoll CB, Otten TG, Brown NM, et al. Towards long-read metagenomics: complete assembly of three novel genomes from bacteria dependent on a diazotrophic cyanobacterium in a freshwater lake co-culture. *Stand Genomic Sci* 2017; 12: 9.
38. Institute of Biotechnology, Cornell University. Next generation sequencing, http://www.biotech.cornell.edu/node/555 (2019, accessed 29 May 2019).
39. Medicine Stanford Center for Genomics and Personalized Medicine. Illumina services, http://med.stanford.edu/gssc/rates.html (2019, accessed 29 May 2019).
40. Harvard University, Illumina sequencing, https://bauercore.fas.harvard.edu/illumina-sequencing (2019, accessed 29 May 2019).
41. Littlefair JE, Carreau J, Webb M, et al. Environmental DNA(eDNA) as a next-generation biomonitoring tool. *Environment Canada*, 20 February 2017, https://www.wsp.com/en-CA/insights/environmental-dna-edna-as-a-next-generation-biomonitoring-tool
42. Bogaerts B, Winand R, Fu Q, et al. Validation of a bioinformatics workflow for routine analysis of whole-genome sequencing data and related challenges for pathogen typing in a European National Reference Center: *Neisseria meningitidis* as a proof-of-concept. *Front Microbiol* 2019; 10: 362.
43. Wetterstrand KA, DNA sequencing costs: data from the NHGRI Genome Sequencing Program (GSP), https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data?fbclid=IwAR2lXeAl7i02DS6YO0TU53ONiNNmr23KW7sI7_3NYDi3RP HpUBKEJkNpmQg (2018, accessed 29 May 2019).
44. Callahan BJ, McMurdie PJ, Rosen MJ, et al. DADA2: high-resolution sample inference from Illumina amplicon data. *Nat Methods* 2016; 13(7): 581–583.
45. Bağcı C, Beier S, Górska A, et al. Introduction to the analysis of environmental sequences: metagenomics with MEGAN. In: Anisimova M (ed.) *Evolutionary genomics: statistical and computational methods*. New York: Springer, 2019, pp. 591–604.
46. Bolyen ERJ, Dillon MR, Bokulich NA, et al. QIIME 2: reproducible, interactive, scalable, and extensible microbiome data science. *PeerJ*: 2018; 53.
47. Kultima JR, Coelho LP, Forslund K, et al. MOCAT2: a metagenomic assembly, annotation and profiling framework. *Bioinformatics* 2016; 32(16): 2520–2523.
48. Yang B, Wang Y and Qian PY. Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. *BMC Bioinformatics* 2016; 17: 135.
49. Buettner C and Noll M. Differences in microbial key players in anaerobic degradation between biogas and sewage treatment plants. *Int Biodeter Biodegr* 2018; 133: 124–132.

50. El-Chakhtoura J, Prest E, Saikaly P, et al. Dynamics of bacterial communities before and after distribution in a full-scale drinking water network. *Water Res* 2015; 74: 180–190.
51. Li GW, Oh E and Weissman JS. The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* 2012; 484(7395): 538–541.
52. Kim M, Morrison M and Yu Z. Evaluation of different partial 16S rRNA gene sequence regions for phylogenetic analysis of microbiomes. *J Microbiol Methods* 2011; 84(1): 81–87.
53. Barb JJ, Oler AJ, Kim HS, et al. Development of an analysis pipeline characterizing multiple hypervariable regions of 16S rRNA using mock samples. *PLoS ONE* 2016; 11(2): e0148047.
54. Chakravorty S, Helb D, Burday M, et al. A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *J Microbiol Methods* 2007; 69(2): 330–339.
55. Klindworth A, Pruesse E, Schweer T, et al. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res* 2013; 41(1): e1.
56. Bokulich NA, Kaehler BD, Rideout JR, et al. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome* 2018; 6: 90.
57. Edgar RC. Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics* 2018; 34(14): 2371–2375.
58. Patel S and Gupta RS. Robust demarcation of fourteen different species groups within the genus Streptococcus based on genome-based phylogenies and molecular signatures. *Infect Genet Evol* 2018; 66: 130–151.
59. Walters W, Hyde ER, Berg-Lyons D, et al. Improved bacterial 16S rRNA gene (V4 and V4-5) and fungal internal transcribed spacer marker gene primers for microbial community surveys. *mSystems* 2016; 1(1): e00009.
60. Fouhy F, Clooney AG, Stanton C, et al. 16S rRNA gene sequencing of mock microbial populations- impact of DNA extraction method, primer choice and sequencing platform. *BMC Microbiol* 2016; 16(1): 123.
61. Tremblay J, Singh K, Fern A, et al. Primer and platform effects on 16S rRNA tag sequencing. *Front Microbiol* 2015; 6: 771.
62. Parada AE, Needham DM and Fuhrman JA. Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environ Microbiol* 2016; 18(5): 1403–1414.
63. Apprill A, McNally S, Parsons R, et al. Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton. *Aquat Microb Ecol* 2015; 75: 129–137.
64. Guo F, Ju F, Cai L, et al. Taxonomic precision of different hypervariable regions of 16S rRNA gene and annotation methods for functional bacterial groups in biological wastewater treatment. *PLoS ONE* 2013; 8(10): e76185.
65. Navarro-Noya YE, Suarez-Arriaga MC, Rojas-Valdes A, et al. Pyrosequencing analysis of the bacterial community in drinking water wells. *Microb Ecol* 2013; 66(1): 19–29.
66. Parulekar NN, Kolekar P, Jenkins A, et al. Characterization of bacterial community associated with phytoplankton bloom in a eutrophic lake in South Norway using 16S rRNA gene amplicon sequence analysis. *PLoS ONE* 2017; 12(3): e0173408.

67. Tessler M, Neumann JS, Afshinnekoo E, et al. Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing. *Sci Rep* 2017; 7(1): 6589.
68. Kuroda K, Nobu MK, Mei R, et al. A single-granule-level approach reveals ecological heterogeneity in an upflow anaerobic sludge blanket reactor. *PLoS ONE* 2016; 11(12): e0167788.
69. De Vrieze J, Raport L, Roume H, et al. The full-scale anaerobic digestion microbiome is represented by specific marker populations. *Water Res* 2016; 104: 101–110.
70. Vanwonterghem I, Jensen PD, Dennis PG, et al. Deterministic processes guide long-term synchronised population dynamics in replicate anaerobic digesters. *ISME J* 2014; 8(10): 2015–2028.
71. Bokulich NA, Subramanian S, Faith JJ, et al. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat Methods* 2013; 10(1): 57–59.
72. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *Embnet J* 2011; 17: 10–12.
73. Bolger AM, Lohse M and Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014; 30(15): 2114–2120.
74. NAJ and Fass J. 2011 Sickle: a sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1*.33)*, https://github.com/najoshi/sickle
75. Joint Genome Institute. BBDuk guide, https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbduk-guide/ (2019, accessed 2019).
76. Masella AP, Bartram AK, Truszkowski JM, et al. PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinformatics* 2012; 13: 31.
77. Edgar RC, Haas BJ, Clemente JC, et al. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 2011; 27(16): 2194–2200.
78. Mysara M, Saeys Y, Leys N, et al. CATCh, an ensemble classifier for chimera detection in 16S rRNA sequencing studies. *Appl Environ Microbiol* 2015; 81(5): 1573–1584.
79. Balvociute M and Huson DH. SILVA, RDP, Greengenes, NCBI and OTT—how do these taxonomies compare? *BMC Genomics* 2017; 18(Suppl. 2): 114.
80. Nguyen NP, Warnow T, Pop M, et al. A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity. *NPJ Biofilms Microbiomes* 2016; 2: 16004.
81. Whelan FJ and Surette MG. A comprehensive evaluation of the sl1p pipeline for 16S rRNA gene sequencing analysis. *Microbiome* 2017; 5(1): 100.
82. Westcott SL and Schloss PD. De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. *PeerJ* 2015; 3: e1487.
83. Callahan BJ, McMurdie PJ and Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J* 2017; 11(12): 2639–2643.
84. Caporaso JG, Kuczynski J, Stombaugh J, et al. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 2010; 7(5): 335–336.
85. Chu DM, Antony KM, Ma J, et al. The early infant gut microbiome varies in association with a maternal high-fat diet. *Genome Med* 2016; 8(1): 77.
86. Thompson LR, Sanders JG, McDonald D, et al. A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* 2017; 551(7681): 457–463.
87. McMurdie PJ and Holmes S. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput Biol* 2014; 10: e1003531.

88. Weiss S, Xu ZZ, Peddada S, et al. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* 2017; 5(1): 27.
89. Caporaso JG, Lauber CL, Walters WA, et al. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc Natl Acad Sci USA* 2011; 108(Suppl. 1): 4516–4522.
90. Salter SJ, Cox MJ, Turek EM, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol* 2014; 12: 87.
91. Linz AM, Crary BC, Shade A, et al. Bacterial community composition and dynamics spanning five years in freshwater bog lakes. *mSphere* 2017; 2: e00169.
92. Saunders AM, Albertsen M, Vollertsen J, et al. The activated sludge ecosystem contains a core community of abundant organisms. *ISME J* 2016; 10(1): 11–20.
93. Ludington WB, Seher TD, Applegate O, et al. Assessing biosynthetic potential of agricultural groundwater through metagenomic sequencing: a diverse anammox community dominates nitrate-rich groundwater. *PLoS ONE* 2017; 12(4): e0174930.
94. Mohiuddin MM, Salama Y, Schellhorn HE, et al. Shotgun metagenomic sequencing reveals freshwater beach sands as reservoir of bacterial pathogens. *Water Res* 2017; 115: 360–369.
95. Luo G, Fotidis IA and Angelidaki I. Comparative analysis of taxonomic, functional, and metabolic patterns of microbiomes from 14 full-scale biogas reactors by metagenomic sequencing and radioisotopic analysis. *Biotechnol Biofuels* 2016; 9: 51.
96. Jovel J, Patterson J, Wang W, et al. Characterization of the gut microbiome using 16S or shotgun metagenomics. *Front Microbiol* 2016; 7: 459.
97. Guo J, Cole JR, Zhang Q, et al. Microbial community analysis with ribosomal gene fragments from shotgun metagenomes. *Appl Environ Microbiol* 2016; 82(1): 157–166.
98. Chan CS, Chan KG, Tay YL, et al. Diversity of thermophiles in a Malaysian hot spring determined using 16S rRNA and shotgun metagenome sequencing. *Front Microbiol* 2015; 6: 177.
99. Poretsky R, Rodriguez-R LM, Luo C, et al. Strengths and limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial community dynamics. *PLoS ONE* 2014; 9(4): e93827.
100. Albertsen M, Hugenholtz P, Skarshewski A, et al. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat Biotechnol* 2013; 31(6): 533–538.
101. Carr R and Borenstein E. Comparative analysis of functional metagenomic annotation and the mappability of short reads. *PLoS ONE* 2014; 9(8): e105776.
102. Bosse M, Heuwieser A, Heinzel A, et al. Biomarker panels for characterizing microbial community biofilm formation as composite molecular process. *PLoS ONE* 2018; 13(8): e0202032.
103. Sunagawa S, Coelho LP, Chaffron S, et al. Ocean plankton. Structure and function of the global ocean microbiome. *Science* 2015; 348(6237): 1261359.
104. Zaheer R, Noyes N, Ortega Polo R, et al. Impact of sequencing depth on the characterization of the microbiome and resistome. *Sci Rep* 2018; 8(1): 5890.
105. Li Q, Lin F, Yang C, et al. A large-scale comparative metagenomic study reveals the functional interactions in six bloom-forming *Microcystis*-Epibiont communities. *Front Microbiol* 2018; 9: 746.
106. Yan ZB, Han WX, Penuelas J, et al. Phosphorus accumulates faster than nitrogen globally in freshwater ecosystems under anthropogenic impacts. *Ecol Lett* 2016; 19(10): 1237–1246.

107. Bukaveckas PA, Franklin R, Tassone S, et al. Cyanobacteria and cyanotoxins at the river-estuarine transition. *Harmful Algae* 2018; 76: 11–21.
108. Otten TG, Graham JL, Harris TD, et al. Elucidation of taste- and odor-producing bacteria and toxigenic cyanobacteria in a midwestern drinking water supply reservoir by shotgun metagenomic analysis. *Appl Environ Microbiol* 2016; 82(17): 5410–5420.
109. Amha YM, Anwar MZ, Brower A, et al. Inhibition of anaerobic digestion processes: applications of molecular tools. *Bioresour Technol* 2018; 247: 999–1014.
110. Grohmann A, Fehrmann S, Vainshtein Y, et al. Microbiome dynamics and adaptation of expression signatures during methane production failure and process recovery. *Bioresour Technol* 2018; 247: 347–356.
111. Mosbaek F, Kjeldal H, Mulat DG, et al. Identification of syntrophic acetate-oxidizing bacteria in anaerobic digesters by combined protein-based stable isotope probing and metagenomics. *ISME J* 2016; 10(10): 2405–2418.
112. Li N, He J, Yan H, et al. Pathways in bacterial and archaeal communities dictated by ammonium stress in a high solid anaerobic digester with dewatered sludge. *Bioresour Technol* 2017; 241: 95–102.
113. Guo JH, Li J, Chen H, et al. Metagenomic analysis reveals wastewater treatment plants as hotspots of antibiotic resistance genes and mobile genetic elements. *Water Res* 2017; 123: 468–478.
114. Spirito CM, Daly SE, Werner JJ, et al. Redundancy in anaerobic digestion microbiomes during disturbances by the antibiotic monensin. *Appl Environ Microbiol* 2018; 84(9): e02692.
115. Bengtsson-Palme J, Hammaren R, Pal C, et al. Elucidating selection processes for antibiotic resistance in sewage treatment plants using metagenomics. *Sci Total Environ* 2016; 572: 697–712.
116. Wilson MR, Naccache SN, Samayoa E, et al. Actionable diagnosis of neuroleptospirosis by next-generation sequencing. *N Engl J Med* 2014; 370(25): 2408–2417.
117. Hamner S, Brown BL, Hasan NA, et al. Metagenomic profiling of microbial pathogens in the little bighorn river, montana. *Int J Environ Res Public Health* 2019; 16(7): 1097.
118. Fang TT, Wang H, Cui QJ, et al. Diversity of potential antibiotic-resistant bacterial pathogens and the effect of suspended particles on the spread of antibiotic resistance in urban recreational water. *Water Res* 2018; 145: 541–551.
119. Fresia P, Antelo V, Salazar C, et al. Urban metagenomics uncover antibiotic resistance reservoirs in coastal beach and sewage waters. *Microbiome* 2019; 7(1): 35.
120. Chu BTT, Petrovich ML, Chaudhary A, et al. Metagenomics reveals the impact of wastewater treatment plants on the dispersal of microorganisms and genes in aquatic sediments. *Appl Environ Microbiol* 2018; 84(5): e02168.
121. Gatica J, Jurkevitch E and Cytryn E. Comparative metagenomics and network analyses provide novel insights into the scope and distribution of beta-lactamase homologs in the environment. *Front Microbiol* 2019; 10: 146.
122. Ju F, Beck K, Yin X, et al. Wastewater treatment plant resistomes are shaped by bacterial composition, genetic exchange, and upregulated expression in the effluent microbiomes. *ISME J* 2019; 13(2): 346–360.
123. Bibby K and Peccia J. Identification of viral pathogen diversity in sewage sludge by metagenome analysis. *Environ Sci Technol* 2013; 47(4): 1945–1951.
124. Coutinho FH, Silveira CB, Gregoracci GB, et al. Marine viruses discovered via metagenomics shed light on viral strategies throughout the oceans. *Nat Commun* 2017; 8: 15955.

125. Ju F, Li B, Ma L, et al. Antibiotic resistance genes and human bacterial pathogens: co-occurrence, removal, and enrichment in municipal sewage sludge digesters. *Water Res* 2016; 91: 1–10.
126. Poptsova MS, Il'icheva IA, Nechipurenko DY, et al. Non-random DNA fragmentation in next-generation sequencing. *Sci Rep* 2014; 4: 4532.
127. Mardis E and McCombie WR. Agarose gel size selection for DNA sequencing libraries. *Cold Spring Harb Protoc* 2017; 2017(8): 094698.
128. Lan JH, Yin Y, Reed EF, et al. Impact of three Illumina library construction methods on GC bias and HLA genotype calling. *Hum Immunol* 2015; 76(2–3): 166–175.
129. Buchfink B, Xie C and Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 2015; 12(1): 59–60.
130. Huson DH, Beier S, Flade I, et al. MEGAN community edition—interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS Comput Biol* 2016; 12(6): e1004957.
131. Piro VC, Lindner MS and Renard BY. DUDes: a top-down taxonomic profiler for metagenomics. *Bioinformatics* 2016; 32(15): 2272–2280.
132. Segata N, Waldron L, Ballarini A, et al. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Methods* 2012; 9(8): 811–814.
133. Darling AE, Jospin G, Lowe E, et al. PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ* 2014; 2: e243.
134. Sunagawa S, Mende DR, Zeller G, et al. Metagenomic species profiling using universal phylogenetic marker genes. *Nat Methods* 2013; 10(12): 1196–1199.
135. Langmead B and Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012; 9(4): 357–359.
136. Finn RD, Clements J and Eddy SR. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* 2011; 39: W29–W37.
137. Wood DE and Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* 2014; 15: R46.
138. Ounit R, Wanamaker S, Close TJ, et al. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics* 2015; 16: 236.
139. Shakya M, Quince C, Campbell JH, et al. Comparative metagenomic and rRNA microbial diversity characterization using archaeal and bacterial synthetic communities. *Environ Microbiol* 2013; 15(6): 1882–1899.
140. Nurk S, Meleshko D, Korobeynikov A, et al. metaSPAdes: a new versatile metagenomic assembler. *Genome Res* 2017; 27(5): 824–834.
141. Li L, He Q, Ma Y, et al. Dynamics of microbial community in a mesophilic anaerobic digester treating food waste: relationship between community structure and process stability. *Bioresour Technol* 2015; 189: 113–120.
142. Alneberg J, Bjarnason BS, De Bruijn I, et al. Binning metagenomic contigs by coverage and composition. *Nat Methods* 2014; 11: 1144–1146.
143. Truong DT, Franzosa EA, Tickle TL, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat Methods* 2015; 12(10): 902–903.
144. Magoc T and Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* 2011; 27(21): 2957–2963.
145. Treangen TJ and Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet* 2011; 13(1): 36–46.

146. Quince C, Walker AW, Simpson JT, et al. Shotgun metagenomics, from sampling to analysis. *Nat Biotechnol* 2017; 35: 833–844.
147. Parks DH, Imelfort M, Skennerton CT, et al. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 2015; 25(7): 1043–1055.
148. Olson ND, Treangen TJ, Hill CM, et al. Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes. *Brief Bioinform.* Epub ahead of print 7 August 2017. DOI: 10.1093/bib/bbx098.
149. Greenwald WW, Klitgord N, Seguritan V, et al. Utilization of defined microbial communities enables effective evaluation of meta-genomic assemblies. *BMC Genomics* 2017; 18(1): 296.
150. Vollmers J, Wiegand S and Kaster AK. Comparing and evaluating metagenome assembly tools from a microbiologist's perspective—not only size matters! *PLoS ONE* 2017; 12: e0169662.
151. Rodriguez-R LM and Konstantinidis KT. Estimating coverage in metagenomic data sets and why it matters. *ISME J* 2014; 8(11): 2349–2351.
152. Tickle TL, Segata N, Waldron L, et al. Two-stage microbial community experimental design. *ISME J* 2013; 7(12): 2330–2339.
153. Markowitz VM, Chen IM, Palaniappan K, et al. IMG 4 version of the integrated microbial genomes comparative analysis system. *Nucleic Acids Res* 2014; 42: D560–D157.
154. Zhu W, Lomsadze A and Borodovsky M. Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res* 2010; 38(12): e132.
155. Hyatt D, Chen GL, Locascio PF, et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 2010; 11: 119.
156. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014; 30(14): 2068–2069.
157. Keegan KP, Glass EM and Meyer F. MG-RAST, a metagenomics service for analysis of microbial community structure and function. *Methods Mol Biol* 2016; 1399: 207–233.
158. Abubucker S, Segata N, Goll J, et al. Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput Biol* 2012; 8(6): e1002358.
159. Kristiansson E, Hugenholtz P and Dalevi D. ShotgunFunctionalizeR: an R-package for functional comparison of metagenomes. *Bioinformatics* 2009; 25(20): 2737–2738.
160. Kanehisa M and Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 2000; 28: 27–30.
161. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 2017; 45: D158–D169.
162. Caspi R, Billington R, Ferrer L, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res* 2016; 44: D471–480.
163. Overbeek R, Olson R, Pusch GD, et al. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res* 2014; 42: D206–D114.
164. Lennon JT, Muscarella ME, Placella SA, et al. How, when, and where relic DNA affects microbial diversity. *MBio* 2018; 9(3): e00637.

165. Magic-Knezev A and Van Der Kooij D. Optimisation and significance of ATP analysis for measuring active biomass in granular activated carbon filters used in water treatment. *Water Res* 2004; 38(18): 3971–3979.
166. Emerson JB, Adams RI, Roman CMB, et al. Schrodinger's microbes: tools for distinguishing the living from the dead in microbial ecosystems. *Microbiome* 2017; 5(1): 86.
167. Hassa J, Maus I, Off S, et al. Metagenome, metatranscriptome, and metaproteome approaches unraveled compositions and functional relationships of microbial communities residing in biogas plants. *Appl Microbiol Biotechnol* 2018; 102(12): 5045–5063.
168. Bashiardes S, Zilberman-Schapira G and Elinav E. Use of metatranscriptomics in microbiome research. *Bioinform Biol Insights* 2016; 10: 19–25.
169. Niu SY, Yang JY, McDermaid A, et al. Bioinformatics tools for quantitative and functional metagenome and metatranscriptome data analysis in microbes. *Brief Bioinform* 2018; 19: 1415–1429.
170. Mohiuddin MM, Botts SR, Paschos A, et al. Temporal and spatial changes in bacterial diversity in mixed use watersheds of the Great Lakes region. *J Great Lakes Res* 2019; 45: 109–118.
171. Knight R, Jansson J, Field D, et al. Unlocking the potential of metagenomics through replicated experimental design. *Nat Biotechnol* 2012; 30(6): 513–520.
172. Gosling SN and Arnell NW. A global assessment of the impact of climate change on water scarcity. *Climatic Change* 2016; 134: 371–385.
173. Alexander KA and Blackburn JK. Overcoming barriers in evaluating outbreaks of diarrheal disease in resource poor settings: assessment of recurrent outbreaks in Chobe District, Botswana. *BMC Public Health* 2013; 13: 775.

Author biographies

Alexander WY Chan is an MSc candidate at McMaster University in Hamilton, ON, Canada. He obtained his BSc with Honors and Distinction from Queen's University and has technical experience in the petrochemical manufacturing industry. He is currently researching microbial communities in onsite wastewater treatment systems using 16S rDNA sequencing.

James Naphtali is an MSc candidate at McMaster University in Hamilton, ON, Canada. He received his BSc in Biology with Honors and Distinction from Redeemer University College. He is currently studying microbial communities and their function in onsite wastewater treatment systems using whole-metagenome sequencing.

Herb E Schellhorn is a Professor of Biology at McMaster University. He has interests in bacterial physiology/gene regulation and developing methods to monitor microorganisms in the industrial, municipal, and agricultural environments.