CHAPTER 5

Metagenomic insights into the diversity and functions of microbial assemblages in lakes

Lateef B. Salam
Department of Biological Sciences, Microbiology unit, Summit University, Offa, Kwara State, Nigeria

Introduction

Water the most basic and important natural resource on the planet earth is used for a variety of purposes by humans. Though it is a renewable resource, population explosion, industrialization, and increased land use have imposed significant stress on the aquatic environments resulting in serious threat to the water quality (Elmhagen et al., 2015; Hinrichsen, 1998; Vorosmarty et al., 2010). Availability of freshwater is not only important for the terrestrial life forms but is critically important for the survival of all other forms of life. In spite of the fact that the greatest part of the earth's surface is covered with water, only 3% of the water is fresh while the remaining portion is salty and is not directly used by most terrestrial life forms (Sigee, 2005). Freshwater environments including lakes, ponds, rivers, streams, and groundwater are highly diverse in the resources and conditions available for microbial growth. Activities of autotrophic and heterotrophic microorganisms in freshwater environments create a balance between photosynthesis and respiration and control the natural cycles of oxygen, carbon, and other nutrients. Lakes are systems of enhanced biological activity that are central to many biogeochemical processes and represent a critical natural resource for human societies (Battin et al., 2009; Downing et al., 2006; Tranvik et al., 2009). The microbial community structure of lakes is influenced by natural stresses occasioned by variations in seasonal cycles in the epilimnion and the hypolimnion and anthropogenic stresses imposed by enrichment of lakes with different nutrients (eutrophication) and other pollutants disposed indiscriminately via different activities (Sigee, 2005).

The field of microbiology was heralded with the discovery of microscopes, which culminated in the design of powerful microscopes and

Freshwater Microbiology
ISBN 978-0-12-817495-1
https://doi.org/10.1016/B978-0-12-817495-1.00005-0

staining techniques that allowed culture-independent direct observation of microorganisms in diverse environments such as soil, water, air, and diseased tissues. However, the pioneering works of Robert Koch and other workers on isolation of pure culture and development of postulates that linked diseases with causative agents changed the perception of microbiology as a science predicated on culture-independent direct observation of natural environments to a science with strong leanings to culture-based investigation of microbial species.

The use of culture-based methods to elucidate the microbial community composition in aquatic environments has gained attraction for several decades because its analysis in these complex natural environments under an ever increasing influence of unknown variables is somewhat difficult. Obtaining the microbial species in pure culture and analyzing them in isolation engender reproducibility of results, control of external variables, and simple design of laboratory experiments (Garza & Dutilh, 2015). Notwithstanding the relative success of the culture-based methods, there was considerable evidence that the population of organisms appearing on isolation plates or the number of phage plaques that grew on a bacterial lawn were several orders of magnitude lower than total cell numbers, determined by microscopy and other methods. These numerical discrepancies termed as the "great plate count anomaly" and "great plaque count anomaly" showed that <1% of microorganisms observed microscopically in natural environments could be isolated on artificial media while the others termed "viable but non culturable" remain unamenable to culture-based techniques. Other shortcomings of the use of culture-based methods include lack of an essential biological interdependency with other species such as auxotrophs or obligate mutualists, inability to replicate the environmental growth conditions in the artificial culture media, lack of essential macronutrients and growth factors in isolation media, inability to culture large number of microbial species, and cultivation bias toward fast-growing organisms (Bollmann, Lewis, & Epstein, 2007; Graber & Breznak, 2005; Kalmbach, Manz, & Szewzyk, 1997; Nichols et al., 2008; Salam, Ilori, Amund, LiiMien, & Nojiri, 2018).

It is quite evident as enunciated above that studying single organisms in isolation for understanding the complex microbial communities that exist in nearly every environment does not allow for the full characterization of microbial interactions. In microbial communities, microorganisms may compete for nutrients, share functional genes through horizontal gene transfer, produce toxins and antibiotics that can kill or inhibit the growth of

other microbes, and produce various metabolites and signaling molecules for sharing and communication (Aguiar-Pulido et al., 2016). The intricate relationship and interaction among microorganisms in nature and the importance of these interactions for effective functioning of the ecosystem demand a culture-independent approach that provides access to a whole set of genes and genomes from a sample, which can be analyzed to elucidate the structure and function of the whole complex system. *Metagenomics* is defined as the culture-independent analysis of the collective genome of biological assemblages from an environmental sample to provide information on the community structure and function of a specific environment. It allows quantitative assessment of the microbial communities, reveals the presence of underrepresented nonculturable microorganisms, and eliminates cultivation bias as the DNA template used is extracted directly from the environmental samples. Metagenomics also has the potential to summarize the combined genetic blueprint of all organisms in each community (Riesenfeld, Goodman, & Handelsman, 2004) revealing the abundance of all represented genes and providing a synoptic description of the functional potential of the studied communities (Debroas et al., 2009; Fierer et al., 2007; Oh et al., 2011; Rusch et al., 2007).

Metagenomics: historical perspectives

As the discovery of microscopy allowed culture-independent observations of microorganisms in their natural environments, the advancement in the field of molecular biology has provided new tools and techniques to detect and quantify the uncultured microorganisms in these environments. It has further elucidated the importance of these silent and unappreciated workers in effective functioning of these environments.

Pioneering efforts by Carl Woese in the 1970s to unlock the world of this "hidden majority" relied heavily upon the small subunit ribosomal RNA (SSU rRNA) of prokaryotes (16S rRNA) and eukaryotes (18S rRNA) to infer evolutionary relationships. SSU rRNA genes are excellent candidates for phylogenetic analysis (Hugenholtz, 2002; Rastogi & Sani, 2011, chap. 2) because of their:

1. universal distribution,
2. structural and functional conservation,
3. variable and highly conserved regions, and
4. adequate length (~1.5 kbp) to provide a deep view of evolutionary relationships.

The development of polymerase chain reaction (PCR) and advancement in DNA sequencing technologies have allowed the PCR amplification of these SSU rRNA genes from total DNA extracted directly from the environmental sample known as *amplicon sequencing*. This technique involves sampling of a community, extraction of DNA from all cells in the sampled community, targeting and amplification of a taxonomically informative universal genomic marker (16S, 18S, *rpoB*, *gyrB*, *recA*, *hsp60*) by PCR, and sequencing and bioinformatic characterization of the resultant amplicons to determine the diversity and relative abundance of the microbes present in the sample (Sharpton, 2014). Amplicon sequencing revealed a tremendous amount of microbial diversity on earth and has been used to characterize the microbial community structure of diverse environments such as ocean thermal vent, hot springs, human gut, roots of plant, and Antarctic volcano mineral soils (Bowen De Leon, Gerlach, Peyton, & Fields, 2013; Lozupone & Knight, 2007; Lundberg et al., 2012; McCliment et al., 2006; Pace, 1997; Rappe & Giovannoni, 2003; Soo, Wood, Gryzmski, McDonald, & Cary, 2009; Yatsunenko et al., 2012). Other techniques premised on the use of PCR and universal genomic markers such as rRNA gene cloning and sequencing, fluorescent in situ hybridization (FISH), denaturing gradient gel electrophoresis (DGGE), temperature gradient gel electrophoresis (TGGE), restriction-fragment length polymorphism (RFLP) among others have also made remarkable impact in our understanding of the uncultured world of microbes. The use of rRNA genes as phylogenetic markers has attracted wider patronage resulting in a huge representation of these markers in many databases like Ribosomal Database Project (RDP) (Wang, Garrity, Tiedje, & Cole, 2007), Greengenes (DeSantis et al., 2006), and Silva (Quast et al., 2013) permitting taxonomic classification of microorganisms present in a metagenomic sample.

Despite the exciting revelations and informative insights offered by amplicon sequencing, there are inherent shortcomings in this technique which include short read lengths, sequencing errors, and incorrectly assembled amplicons, resulting in artificial sequences difficult to identify (Wylie et al., 2012). Others include varying estimates of diversity arising from targeting and amplifying different variable regions in the marker gene (Liu, DeSantis, Andersen, & Knight, 2008; Schloss, 2010), and inability of amplicon sequencing to resolve biological functions associated with the microbial community. It further includes overestimation of community diversity due to promiscuous transfer of 16S locus between distantly related taxa via horizontal gene transfer, difficulties in assessing operational taxonomic units,

and inadequate resolution of 16S rRNA genes at species and strain delineations (Konstantinidis, Ramette, & Tiedje, 2006; Poretsky, Rodriguez-R, Luo, Tsementzi, & Konstantinidis, 2014; Quince, Lanzen, & Curtis, 2009; Quince, Lanzen, Davenport, & Turnbaugh, 2011; Youssef et al., 2009). Additionally, the use of amplicon sequencing obliterates viral identification in environmental samples. Unlike prokaryotes and eukaryotes having 16 and 18S rRNA phylogenetic marker genes, viruses lack universally conserved phylogenetic markers, which limit their direct detection in the environment (Mohiuddin & Schellhorn, 2015; Rohwer & Edwards, 2002). Finally, in spite of the quantum of information revealed by amplicon sequencing of environmental samples as exemplified by the classical work of Newton, Jones, Eiler, McMahon, and Bertilsson (2011) and several other workers on natural history of freshwater lake bacteria, information relating to metabolic and ecological function of the microbial community of environmental samples is conspicuously absent.

To answer the "*what are they doing*?" question in microbial ecology, early workers relied on gene cloning, where specific genes of interest were cloned from the total DNA of the studied environment. The heterologously expressed product (cloned gene) is then associated with any given metabolic function (e.g., cellulases, oxidoreductases, nitrogenases, etc.). This approach birthed the development of gene expression techniques (using the systems of other microorganisms with desirable features) to test gene functions and roles in microbial communities. It also allows the discovery of new genes, functions, and metabolic products with biotechnological applications (Escober-Zepeda, Vera-Ponce de Leon, & Sanches-Flores, 2015). However, in spite of the discovery of novel molecules and other advantages of this method, the technique is laden with some challenges including cloning biases, sampling biases, incorrect promoter sites in genomes, and dispersion of genes involved in secondary metabolite production among others.

In contrast to amplicon sequencing and its drawbacks, shotgun metagenomics identifies information beyond the taxonomic composition and bypasses primer biases that are introduced through amplification. Shotgun metagenomics has the capacity to sequence most of the available genomes within an environmental sample and thus address the questions of *who* is present in a community (taxonomy/structure), *what* they are doing (function), and *how* these microorganisms interact to maintain a balanced ecological niche (Oulas et al., 2015). Originally, shotgun metagenomic studies require the extraction of environmental DNA, shearing of the extracted DNA followed by cloning into an appropriate vector and then traditional

Sanger sequencing. Investigation using shotgun metagenomics include the structural and functional characterization of microbial communities of water samples from different sites in the Sargasso Sea (Venter et al., 2004), whale falls, acid mine drainages, acidophilic biofilms, and soils (Tringe et al., 2005; Tyson et al., 2004).

Advances in high-throughput next generation sequencing (NGS) has revolutionized the techniques of shotgun metagenomics, as construction of clone library is no longer required and massive parallelization of NGS techniques ensure greater yield of sequence data and gives unprecedented insight into the genetic potentials of microbial communities as well as underrepresented populations (Handelsman, 2004; Newby, Marlowe, & Maier, 2009; Oulas et al., 2015). It also allows elucidation of metabolic properties of the microbial communities and enables the identification of novel molecules with significant functionalities and applications (Bashir, Singh, & Konwar, 2014; Streit & Schmitz, 2004).

Early works on metagenomics of aquatic environments focused primarily on marine environments such as oceans, seas, and coastal lagoons exploring biodiversity, genomic analysis of unknown taxa, expression of novel and useful genes, and detection of pathogenic bacteria (Ferreira et al., 2014; Ghai et al., 2012; Mohamed et al., 2013; Rivera et al., 2003; Sogin et al., 2006; Venter et al., 2004). Although the use of metagenomic approaches to characterize the microbial community structure and function of freshwater environments commenced much later, findings from metagenomic characterization of microorganisms from Lake Gatun (Rusch et al., 2007), Lac du Bourget (Debroas et al., 2009), and Lake Lanier (Oh et al., 2011) have provided some first snapshots of the functional diversity of freshwater bacterioplankton in the lake ecosystems. Furthermore, the metagenomic approaches have been employed to detect pathogenic bacteria in drinking water, decipher bacterial communities, assess water quality, and identify biomarkers for pollution detection and source attribution in freshwater ecosystems (Bai, Liu, Liang, & Qu, 2013; Chao et al., 2013; Damashek, Smith, Mosier, & Francis, 2014; Fisher, Newton, Dila, & McLellan, 2015; Lee, Lee, Sung, & Ko, 2011; Ocepek et al., 2011; Staley et al., 2014; Wu et al., 2010). Metagenomic analysis of viral communities in both marine and freshwater environments suffered serious setbacks due to lack of universal marker genes for viruses. However, the advent of shotgun metagenomic technique where viral fractions were isolated from freshwater for their analysis addressed this limitation, thus providing an unprecedented insight into viral diversity and overcoming many limitations of the culture-based and/or PCR-based

methods for virus identification. Based on this development, it offers an unrestricted access to the type and distribution of viruses in any host environment (Breitbart et al., 2002; Mohiuddin & Schellhorn, 2015).

Though metagenomics is a powerful tool, it has its own challenges. First, the reference databases used to classify and label microorganisms are limited, thus, rendering several sequence reads unresolved. Second, the metagenomics cannot reveal the dynamic properties, such as the spatiotemporal activity of the microbial communities and the impact of the environment on these activities. Third, the only functional information provided by metagenomics is the potential of the microbial communities to display functional properties associated with the presence of genes with no information about their expression levels (Aguiar-Pulido et al., 2016).

Sequencing technologies

DNA sequencing is defined as the sequencing of nucleotide within a DNA molecule using laboratory methods (Sanger, Nicklen, & Coulson, 1977). The pioneering efforts of Maxam and Gilbert (1977) and Sanger et al. (1997) led to the development of two contrasting methods for DNA sequencing. Maxam and Gilbert (1977) utilized chemical breakage, radioisotope labeling, and electrophoresis to sequence DNA, while Sanger and coworkers utilized the dideoxynucleotide analogs as specific chain-terminating inhibitors of DNA polymerase to sequence DNA (Sanger et al., 1977). Due to its high efficiency and non-involvement of toxic chemicals, Sanger's method outlived the Maxam and Gilbert method as the most preferred among researchers and scientists for nearly 40 years. Sanger sequencing is a familiar workflow that is relatively fast and cost-effective for lower number of targets.

Advances in molecular biology and democratization of sequencing technologies result in the development of high throughput next generation sequencing (NGS) platforms that trump Sanger sequencing on four grounds: speed, cost, sample size, and accuracy. For Sanger sequencing, a large amount of DNA template is needed for each read as several strands of template DNA are needed for each base being sequenced. The strand that terminates on each base is needed to construct a full sequence. Multiple staggered copies are taken for contig (a set of overlapping DNA segments that together represent a consensus region of DNA) construction and sequence validation. NGS is quicker than Sanger sequencing in two ways. First, the chemical reaction may be combined with the signal detection in some versions of

NGS, whereas in Sanger sequencing these are two separate processes. Second and more significantly, only one read (maximum ~1 kb) can be taken at a time in Sanger sequencing, whereas NGS is massively parallel and allows 300 Gb of DNA to be read on a single run on a single chip.

Repeats are intrinsic to NGS, as each read is amplified before sequencing, and because it relies on many short overlapping reads, so each section of DNA or RNA is sequenced multiple times. Also, because it is so much quicker and cheaper, it is possible to do more repeats than with Sanger sequencing. More repeats result in greater coverage, which leads to a more accurate and reliable sequence, even if individual reads are less accurate for NGS. The NGS platforms also obviate the use of labor-intensive cloning process (synonymous with Sanger sequencing) and its associated bias against genes, toxic for the cloning host (Sorek et al., 2007).

The first commercialized next-generation sequencing platform was 454 pyrosequencing platform introduced in 2004 (Mardis, 2008). In this system, template DNA are generated by either fragmentation where universal adaptors are ligated to the fragmented end or PCR using adaptor-built primers (Margulies et al., 2005; Moorthie, Mattocks, & Wright, 2011). The prepared fragments are hybridized onto special beads and a mix containing the fragment-hybridized beads, PCR reagents, and oil agitated to form tiny oil reaction chambers. The DNA present on the surface of each bead is then amplified via thermal cycling (Ambardar, Gupta, Trakroo, Lal, & Vakhlu, 2016; Pillai, Gapalan, & Lam, 2017). Sequencing reaction in this platform is carried out on Pico Titer Plate that comprises millions of microscopic wells. The beads and sequencing reagents are loaded onto these wells except the nucleotides, each of which are added chronologically in the order A, C, G, and T (Harrington, Lin, Olson, & Eshleman, 2013; Moorthie et al., 2011). As a nucleotide is added to the growing DNA strand by DNA polymerase, an inorganic phosphate ion is released resulting in the release of a flash of light detected by a camera across the whole plate, hence the name pyrosequencing (Ballester, Luthra, Kanagal-Shamanna, & Singh, 2016). The process is repeated several times and the resultant image sequence is computationally converted into sequence reads (Liu, Li, et al., 2012; Liu, Zhang, et al., 2012). The read length of Roche 454 platform increased from 100 to 150 bp,200,000 reads, 20 Mb output/run to 700 bp and output of 700 Mb with the introduction of 454 GS FLX Titanium system in 2008 (Liu, Li, et al., 2012; Liu, Zhang, et al., 2012). The selling point of this platform compared to Sanger sequencing are massive parallelization, longer read length (compared to other NGS), and speed as it takes 10 h sequencing

start to completion. However, the major drawbacks of this platform are the high error rate and high cost of reagents (Ballester et al., 2016; Liu, Li, et al., 2012; Liu, Zhang, et al., 2012).

Illumina sequencing platform, introduced in 2006, adopted sequencing by synthesis system, which uses specially designed fluorescent labeled terminator nucleotides to allow the chain termination process to be reversed (Moorthie et al., 2011). Library is prepared by DNA fragmentation followed by addition of custom adapters to the fragmented DNA. Template fragments bind to the solid surface of a flow cell as the library flows across it (Su et al., 2011). Bridge amplification follows, which creates ~one million copies of each template in tight physical clusters on the flow cell surface (Su et al., 2011). Sequencing reaction commences with the addition of a universal sequencing primer that hybridizes to the adaptor sequences added in the first stage (Liu, Li, et al., 2012; Liu, Zhang, et al., 2012). Modified dNTPs used, contain a terminator that prevent further polymerization and sequencing reaction proceeds simultaneously on a very large number of different template molecules spread out on a solid surface (Moorthie et al., 2011). A fluorescent label from the terminator generates a series of color images, which can be detected by a camera and then computationally converted into sequence reads (Quail et al., 2012). The overall error rates across all Illumina models are less than 1% and the most frequent type of error is substitution (Dohm, Lottaz, Borodina, & Himmelbauer, 2008). Illumina produces series of purpose-driven sequencers (MiSeq, NextSeq 500, HiSeq) optimized for a variety of throughputs and run times. The MiSeq intended for targeted sequencing and sequencing of small metagenomes is a fast, benchtop sequencer with run time as low as 4 h. The NextSeq 500, introduced in year 2014, is a fast benchtop sequencer that uses a novel two-channel sequencing strategy and is capable of producing 120 Gb in less than 30 h. The two-channel sequencing strategy of NextSeq 500 reduces data processing time and increases output unlike MiSeq and HiSeq 2500 that use four-channel sequencing strategy resulting in longer data processing time (HiSeq 2500, 1Tb in 6 days) (Reuter, Spacek, & Snyder, 2015).

Illumina launched HiSeq X 10, HiSeq 3000/HiSeq 4000 systems in year 2015 and they provide a remarkable level of throughput due to billions of nanowells at fixed locations compared to the normal flow cell (Ambardar et al., 2016). These new Illumina series sequencing systems boast of enhanced optics and computing capacity, incorporate a new patterned flow cell technology containing billions of nanowells that standardized cluster spacing and size, thus allowing higher cluster densities (Reuter et al., 2015).

Another NGS technology, SOLiD (Sequencing by Oligonucleotide Ligation and Detection), developed by Life Technologies and introduced in year 2007 uses ligation of fluorescently labeled hybridization probes to determine the sequence of a template DNA strand, two bases at a time as a program for sequencing (Mardis, 2013). In SOLiD, four colored dyes are used each representing four possible different two-base combinations resulting in 16 possible combinations (Quail et al., 2012). The first base in the sequence is always from the universal primer which is added initially, the rest of the sequence can be inferred from the raw color data, thus making SOLiD the only NGS platform with overall accuracy of >99.85% (Mardis, 2013; Moorthie et al., 2011). However, the platform is not widely used due to limited read lengths and expensive computational infrastructure (Ambardar et al., 2016).

Two widely used NGS platforms with compact sequencer are Ion Personal Genomic Machine (Ion PGM) later acquired by Life Technologies and MiSeq Illumina sequencer. Ion Torrent PGM is a compact benchtop sequencer that uses semiconductor sequencing technology, where incorporation of a nucleotide into the DNA molecules by DNA polymerase triggers release of a proton, which results in pH change to help the PGM to recognize whether a nucleotide is added or not (Pareek, Smoczynski, & Tretyn, 2011). It is the first commercial sequencing machine that does not need fluorescence and camera thus resulting in lower cost, higher speed and throughput, and smaller instrument size (Pareek et al., 2011). However a major drawback of this platform is lower coverage in genomes with very high A-T content (Ballester et al., 2016). The MiSeq Illumina sequencer stands out for its higher sequencing accuracy and as the only compact next generation sequencer that integrates amplification, sequencing, and data analysis in a single instrument (Pareek et al., 2011; Quail et al., 2012).

The need to analyze larger structural variants and the overall chromosome structure, which are outside the scope of NGS platforms (Schadt, Turner, & Kasarskis, 2010) necessitates the development of single-molecule sequencing technologies collectively termed *third generation sequencing* (TGS) (Pareek et al., 2011; Ståhl et al., 2016). The three commercially available third-generation DNA sequencing technologies are Pacific Biosciences (Pac Bio) Single Molecule Real Time (SMRT) sequencing, the Illumina Tru-seq Synthetic Long-Read technology, and the Oxford Nanopore Technologies sequencing platform (Pareek et al., 2011). Pac Bio uses single-molecule real-time (SMRT) sequencing technology, a parallelized single-molecule DNA sequencing by synthesis; provides much longer read lengths

(~10,000 bp), which are advantageous in annotation and assembly for shotgun metagenomics, and perform paired-end sequencing through a process called *strobing*. Pac Bio platform is however limited by high error rates and low coverage (Oulas et al., 2015). Oxford nanopore technologies premised on strand sequencing is a method of genome analysis where a single molecule of DNA or RNA passed through a protein nanopore is sequenced without the need for PCR amplification or chemical labeling of the sample. This eliminates erroneous sequencing caused by shotgun metagenomics and obviates the need for the error-prone assembly step during data analysis (Oulas et al., 2015). Thus, the major advantages of TGS or NGS include reduced cost, reduced time, longer read length, direct sequencing of single molecule of DNA or RNA, and elimination of PCR amplification step or chemical labeling of the sample (Oulas et al., 2015; Pareek et al., 2011; Pillai, Gopalan, & Lam, 2017; Schadt et al., 2010).

Metagenomics of freshwater lakes—basic steps (Fig. 5.1)

Sample collection and preparation

It is important to concentrate freshwater samples prior to DNA extraction because besides freshwater lakes being enriched with inorganic/organic nutrients, which stimulate growth and proliferation of microbial species, most of them are oligotrophic. So, there is a need to concentrate many liters of water samples to allow recovery of reasonable amount of total DNA for metagenomic analysis (Wilcox et al., 2016) and to achieve this, precipitation, followed by centrifugation or filtration can be used. Water sample can be filtered using various flow filter systems with different pore sizes depending on the target group of microorganisms (prokaryotes, eukaryotes, viruses). Several workers have reported that higher concentrations of DNA are obtained by filtration methods (Deiner, Walser, Maechler, & Alermatt, 2015; Felczykowska, Krajewska, Zielińska, & Łoś, 2015; Li, Yang, Lv, et al., 2015; Li, Yang, Ma, et al., 2015). However, to filter the required volume of water sample and prevent clogging of the filter, it is advisable to filter multiple small volumes of water through separate filters and later combine the DNA collected from each filter via ethanol precipitation (Lear et al., 2018; Santas, Persaud, Wolfe, & Bauman, 2013). Onsite filtration is recommended for DNA extraction from <1 L freshwater samples. Devices such as Sterivex filters (Wright, Lee, Zaikova, Walsh, & Hallam, 2009) or customized portable filtration devices as described by Yamanaka et al. (2016) could be used

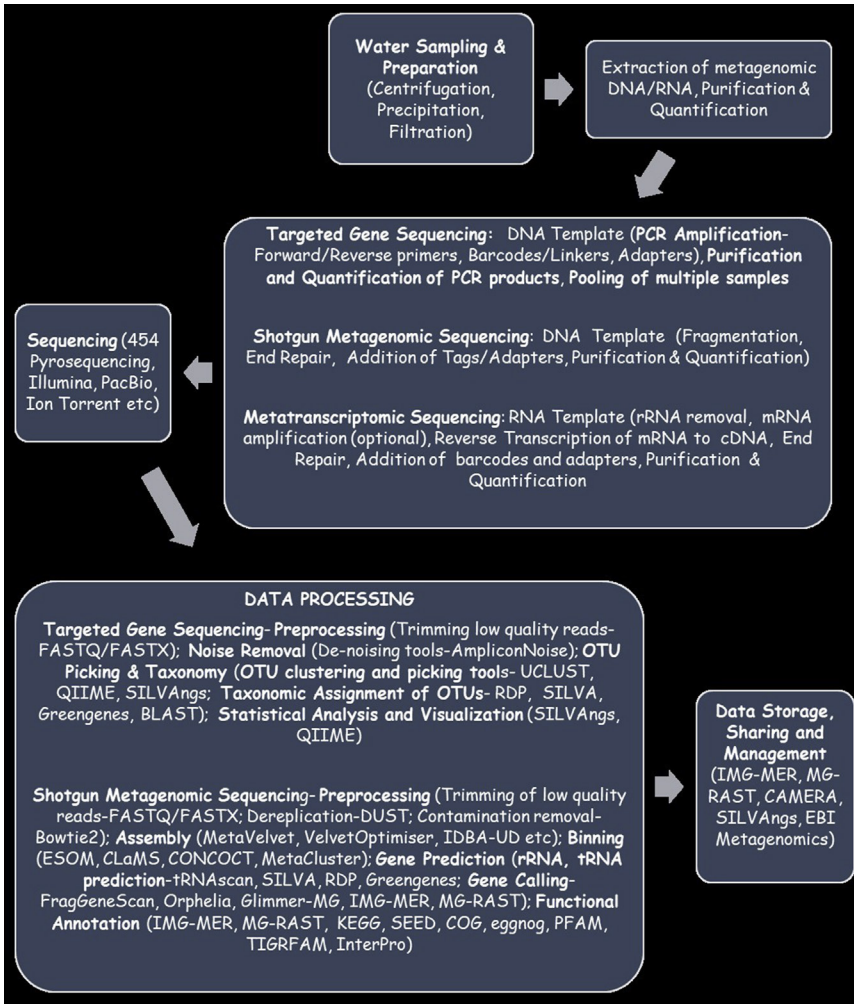

Figure 5.1 ***Flowchart of basic steps in metagenomics.*** The flowchart outlines the basic steps in metagenomic analysis of freshwater lakes starting with water sampling and preparation, DNA/RNA extraction, purification and quantification, preparation of the genome for sequencing depending on the analysis (Targeted gene, shotgun and/or metatranscriptomics), sequencing, data processing depending on the sequencing data, and data storage and management. Popular tools and algorithms used in metagenomic analysis are shown for every step.

for this purpose. However, for large water samples, it is important to filter and preserve the water sample at −20°C once it is collected because the quality and quantity of DNA present in water decreases rapidly (Gilpin et al., 2013; Maruyama, Nakamura, Yamanaka, Kondoh, & Minamoto, 2014; Thomsen et al., 2012).

Sample preparation for extraction of viral metagenome is quite different because of small size of viral genome. The most common method involves the use of tandem tangential flow filtration (TFF) system, which is preceded by a series of filtration using impact filters of different pore sizes to preclear the mixture. Two TFF columns are generally used, in which the first TFF column has a pore size of 0.22 or 0.45 μm that allows viruses to go through but retains most bacteria. The second TFF column unit has a 30, 70, or 300 kDa MW cutoff that retains and concentrates viral size particles. The viral size particle may be further pelleted by ultracentrifugation and thereafter preserved at −20°C/−80°C before extraction (Aguirre de Cárcer,López-Bueno, Pearce, & Alcamí, 2015; Djikeng, Kuzmickas, Anderson, & Spiro, 2009; Uyaguari-Diaz et al., 2016).

DNA extraction

Extraction of quality environmental DNA for downstream processes provides unrivalled insights into microbial community structure and function in complex environmental samples. It is important to obtain a high quality DNA (with high purity and low degree of fragmentation). Though there are several DNA extraction protocols, most of the protocols share similar aims (Lever et al., 2015) including:

1. comprehensive cell lysis and extraction of DNA into aqueous solution,
2. removal of organic and inorganic contaminants and enzymatic inhibitors from resultant aqueous DNA extracts, and
3. minimization of DNA losses throughout the purification process.

Cell lysis is achieved mechanically (by bead-beating, freeze-thawing, sonication), enzymatically (by incubating water samples with enzymes that hydrolyze cell wall and cell membrane components such as lysozyme, proteinase K), and/or chemically (by using detergents such as sodium dodecyl sulfate (SDS) which denature the lipid bilayers, or chaotropic agents that denature transmembrane proteins) (Hurt et al., 2001; Tsai & Olsen, 1991). Subsequent purification steps involve washing with organic solutions and/or detergents (e.g., phenol, chloroform, cetyl trimethylammonium bromide, CTAB); precipitation with isopropanol, ethanol, or polyethylene glycol; and/or filtration via silica-columns, ion-exchange resins, magnetic beads, or gels (Felczykowska et al., 2015; Li, Yang, Lv, et al., 2015; Li, Yang, Ma, et al., 2015; Ogram, Sayler, & Barkay, 1987; Pitcher, Saunders, & Owen, 1989; Smalla, Cresswell, Mendonca-Hagler, Wolters, & van Elsas, 1993; Stahl, Flesher, Mansfield, & Montgomery, 1988; Tanaka, Sakai, Kobayashi, Hatakeyama, & Matsunaga, 2009; Wang, Ran, Man, & Yang, 2011; Zhang, Xu, & Shi, 2012; Zhao et al., 2008; Zhou, Brons, & Tiedje, 1996).

While laboratory-prepared DNA extraction chemical solutions and extraction protocols held sway in the 1980 and 1990s, there has been a significant shift toward increased use of commercial DNA extraction kits. This is due to the inherent challenges associated with laboratory-prepared DNA extraction reagents such as possible contamination of extraction chemicals and reagents, inability to obtain high-quality DNA in higher yield for downstream applications and strenuous preparation of extraction solutions, among others. Commercial DNA extraction kits obviate the need for reagent preparation and contamination as the reagent is already prepared. It also has streamlined extraction procedures, produces adequate DNA yield across wide range of samples, and facilitates comparison of metagenomic datasets due to method standardization (Coolen, van de Giessen, Zhu, & Wuchter, 2011; Inagaki et al., 2006; Lear et al., 2018).

Amplification and sequencing

Amplification and sequencing depend largely on the type of metagenomic study conducted, that is, whether it is targeted gene sequencing with phylogenetic (e.g., 16S rRNA) (Caporaso et al., 2012; Sogin et al., 2006) or functional (e.g., *amoA*, *nifH*) (Gubry-Rangin et al., 2011; Pester et al., 2012) gene targets or on shotgun metagenome sequencing. After high-quality DNA is obtained, targeted gene is amplified with conserved primers. Each set of primers is generally barcoded with short oligonucleotide tags (6 to 12 mer), as well as sequencing adapters, so that multiple samples can be pooled and sequenced simultaneously (Caporaso et al., 2012; Sogin et al., 2006). Then, after removal of nontarget DNA fragments by gel electrophoresis, target DNA is quantified, sequenced, and analyzed using bioinformatic approaches, such as operational taxonomic unit (OTU) assignment, sequence assembly, phylogeny, and annotation. Though targeted gene sequencing provides greater depth of coverage for specific genes within a microbial community, it cannot be used to analyze the whole genetic and functional diversity of communities (Weinstock, 2012). So, this limitation necessitates the development of shotgun metagenomics. In this method, community DNA is randomly sheared using various methods (nebulization or sonication), sheared fragments are repaired prior to ligation to platform-specific adapters, which serve as the priming sites for template amplification, and subsequent sequencing produces vast amounts of short reads, which can be assembled and annotated for phylogenetic and functional characterization (Loman et al., 2012; Nagarajan & Pop, 2013). Shotgun metagenomics provides community-level information in complex environments, identifying novel genes, phylotypes, regulators, and pathways.

Amplicon-based targeted sequencing requires the amplification of a single taxonomic gene marker from the community DNA using specific primers. In amplicon sequencing, the choice of primers strongly determines the accuracy of 16S rRNA gene amplification and sequencing. To this end, primer pairs covering different regions and with varying specificities have been examined by several workers (Klindworth et al., 2013; Kozich, Westcott, Baxter, Highlander, & Schloss, 2013; Schloss, Gevers, & Westcott, 2011; Soergel, Dey, Knight, & Brenner, 2012; Tremblay et al., 2015) and the most acceptable primers have been designated as the ones that guarantee the highest coverage of the members of Bacteria and Archaea domains of prokaryotes. In addition, the primers should amplify regions that result in fewer chimeras, lower error rates, and higher specificity for prokaryotes (Haas et al., 2011; Kozich et al., 2013; Lear et al., 2018), as for example, Klindworth et al. (2013) carried out *in silico* evaluation of the overall coverage and phylum spectrum of 175 primers and 512 primer pairs for Bacteria and Archaea using SILVA 16S/18S rDNA nonredundant reference dataset and observed that out of 175 primers and 512 primer pairs checked, only 10 were recommended as broad range primers. Further, the majority of the commonly used single primers exhibited significant differences in overall coverage and phylum spectrum. Also, Caporaso et al. (2012) in the Earth Microbiome Project (EMP) recommended a primer pair for the universal amplification of all prokaryotic organisms consisting of forward primer 515F and reverse primer 806RB that target a 301 bp in the V3–V4 region of the 16S gene. *In silico* analysis of these primers revealed that they cover 94.8% of all Bacteria and 95.2% of Archaea sequences without losing specificity for prokaryotic organisms.

Talking about the challenges associated with function-based targeted sequencing, it has the following main challenges:

1. There is a widespread lack of sequence conservation across functionally homologous genes, which makes PCR primer design difficult for those genes and hence relevant functional genes in the environment are not detected.
2. In instances where fairly conserved primers are designed for some functional genes of interest (e.g., *amoA*, *nifH*, *nirS*, *nirK*), the success of amplification is relative as it is habitat/ecosystem dependent probably due to variations in the quality of extracted DNA, community complexity, sequence divergence, and target gene abundance. This makes comparative studies of the functional genes to be impossible or highly compromised (Faust & Raes, 2012).

3. It is often difficult to prepare high-quality libraries of amplified PCR products for various functional genes from multiple samples because non-specific amplification requires the tedious and time-consuming step of additional gel purification of PCR products prior to sequencing, which could substantially slow down the sequencing process (Zhou et al., 2015).

While targeted and shotgun sequencing of DNA offer glimpses of the gene content and genetic diversity of microbial communities, it cannot distinguish between expressed and nonexpressed genes in each environment. The development of metatranscriptomics premised on total RNA extracted from microbial communities allows random sequencing of expressed microbial community RNA. To obtain a high level of mRNA transcripts, rRNA that typically dominates the extracted total RNA is removed and the remaining mRNAs are reverse transcribed to complementary DNA (cDNA), ligated to adapters, and sequenced (Frias-Lopez et al., 2008; Moran et al., 2013; Sorek & Cossart, 2010). Metatranscriptomics studies have provided deep insights into microbial community functions and activities from diverse habitats and have aided in the discovery of novel genes, metabolic pathways, and regulatory elements hitherto undetected in DNA-based surveys (Zhou et al., 2015).

Bioinformatics

Assembly

A common first step prior to assembly of metagenomic sequence reads is to run a variety of computational tools for quality control, which identify and remove low-quality reads and contaminants. This includes FastQC, which shows sequence statistics such as quality values, length distribution, and GC content; FastQ Screen, Cutadapt, Trimmomatic, BBDuk, among others (Breitwieser, Lu, & Salzberg, 2017). These tools trim and filter low-quality reads, screen the reads for contamination, remove adapter sequences, primers, poly A tails and normalize the coverage by down-sampling reads. For 16S metagenomic dataset, denoising step, which is platform-specific is important especially in 454 pyrosequencing data, due to intrinsic errors generated from pyrosequencing. Denoising is performed very efficiently by AmpliconNoise (Quince et al., 2011), a tool that performs four basic denoising steps:

1. filtering of noisy reads,
2. removing of pyrosequencing noise,
3. removing of PCR noise (PCR errors), and
4. chimera identification and removal.

Sequence assembly refers to aligning and merging fragments from a longer DNA sequence to reconstruct the original sequence. In metagenomic studies, the need to recover the genome of uncultured organisms or to obtain full-length coding sequence (CDS) for subsequent characterization necessitates the assembly of short sequence read fragments to obtain longer genomic contigs (Thomas, Gilbert, & Meyer, 2012). The assemblage of short-sequence reads into contigs adopts a two-pronged strategy:

1. **Reference-based assembly:** It refers to the use of one or more reference genomes as a "map" to create contigs, which can represent genomes or parts of genomes belonging to a particular species or genus. The software packages, which are not computationally intensive such as Newbler (Roche), MIRA4, AMOS, and MetaAMOS are commonly used in metagenomics for performing referenced-based assemblies (Chevreux et al., 2004; Treangen et al., 2013). To be effective, sequences from closely related organism have already been deposited in online data repositories and databases, allowing them to be used as references for the assembly process.
2. **De novo assembly**: It refers to the generation of assembled contigs using no prior reference to known genome(s) (Paszkiewicz & Studholme, 2010). It requires large computational resources and relies heavily on sophisticated graph theory algorithms called de-Bruijn graphs. Varieties of de-Bruijn assemblers such as EULER, Velvet, SOAP, and Abyss were among the first algorithms to perform de novo assembly and are still widely used today (Li, Li, Kristiansen, & Wang, 2008; Pevzner, Tang, & Waterman, 2001; Simpson et al., 2009; Zerbino & Birney, 2008). However, there are serious shortcomings associated with these assemblers, as they are designed with the assumption of assembling a single genome and are often inefficient when used for metagenome assemblies. These shortcomings necessitate the development of next generation of assembly tools such as MetaVelvet, MetaVelvet-SL, and Meta-IDBA (Afiahayati, Sato, & Sakakibara, 2014; Namiki, Hachiya, Tanaka, & Sakakibara, 2012; Peng, Leung, Yiu, & Chin, 2012; Sato & Sakakibara, 2015). MetaVelvet and Meta-IDBA employ a combined binning and assembly approach to create more accurate assemblies from datasets containing a mixture of multiple genomes.

Phylogenetic binning

Binning that can be either composition-based or similarity-based is a process of clustering sequence reads or contigs into groups and assigning each

subset of each group to a specific biological taxon, that is, species, subspecies, or genus. Theoretically, each bin represents a single genome and is assembled separately, removing in the process some of the problems of erroneous assemblies connecting contigs from diverse taxa (Roumpeka, Wallace, Escalettes, Fotheringham, & Watson, 2017). Composition-based binning is hypothesized on the observation that individual genomes have a unique k-mer sequence distribution or a certain GC content. By making use of this conserved species-specific nucleotide composition, these methods are capable of grouping sequences into their respective genomes (Oulas et al., 2015; Thomas et al., 2012). In similarity-based binning, the similarity of a specific gene encoded by an unknown genomic fragment with known genes in a reference database is used to classify and bin the sequence (Thomas et al., 2012). Several softwares including PhyloPythia (McHardy, Martin, Tsingos, Hugenholtz, & Rigoutsos, 2007), S-GSOM (Chen et al., 2008), PCAHIER (Zheng & Wu, 2010), PhyloPythiaS (Patil, Roune, & McHardy, 2012), TETRA (Teeling, Waldmann, Lombardot, Bauer, & Glockner, 2004), ESOM (Dick et al., 2009), TACOA (Diaz, Krause, Goesmann, Niehaus, & Nattkemper, 2009), and ClaMS (Pati, Heath, Kyrpides, & Ivanova, 2011) that are composition-based; SOrt-ITEMS (Monzoorul Haque, Ghosh, Komanduri, & Mande, 2009), MetaPhyler (Liu, Gibbons, Ghodsi, Treangen, & Pop, 2011) and CARMA (Krause et al., 2008) that are similarity-based; and PhymmBL (Brady & Salzberg, 2009) and MetaCluster (Leung et al., 2011) that are hybrid (composition- and similarity-based) are the available binning algorithms. Some metagenomic analysis pipelines such as MG-RAST (Glass, Wilkening, Wilke, Antonopoulos, & Meyer, 2010), IMG/MER 4 (Markowitz et al., 2014), and MEGAN (Huson & Weber, 2013) also contain tools that utilize similarity-based binning algorithms.

Another phylogenetic binning program of note is CONCOCT (Clustering cONtigs with Coverage and ComposiTion), which uses GMM (Gaussian Mixture Model), sequence composition, and the coverage across multiple samples to cluster metagenomic data (Alneberg et al., 2014). Following sequence reads assembly, the longer contigs are fragmented and the reads are mapped back onto contigs to determine the coverage across all samples. To form a combined profile for each contig, the coverage and sequence composition vectors are joined, and a GMM is used to describe the entire dataset. The efficiency of CONCOCT was tested using mock and real metagenomic data, and the results indicate high precision of CONCOCT on the mock data and clustering complicated microbial communities (Alneberg et al., 2014; Roumpeka et al., 2017).

One of the most commonly used tools for analysis of 16S sequence dataset is QIIME (Quantitative Insights into Microbial Ecology), a bioinformatic pipeline for analysis of microbial communities sampled through marker gene (16S, 18S) amplicon sequencing (Caporaso et al., 2010). The pipeline performs quality control of sequence reads, clusters the reads at a requested phylogenetic level into OTUs, and taxonomically assigns the OTUs using a variety of algorithms such as Basic Local Alignment Search Tool (BLAST), the Ribosomal Database Project (RDP) classifier, RTAX, Mothur classifier, and Uclust, to search for the closest match to an OTU from which a taxonomic lineage is inferred (Oulas et al., 2015). Commonly utilized reference databases for marker genes include Greengenes (16S), Ribosomal Database Project (16S), Silva (16S, 18S), and Unite (ITS) (Kõljalg et al., 2013). The QIIME pipeline also performs statistical analysis of the metagenomic dataset such as rarefaction, alpha-diversity analysis (Phylogenetic Diversity (PD), Chao, etc.), and beta-diversity analysis (e.g., UniFrac, PCoA). The *Biom* (Biological Observation Matrix) file generated in QIIME, which represent OTU tables can be imported into MEGAN or other matrix-type data compliant statistical software for visualization. However, QIIME is mostly implemented using the programming language PYTHON; hence, proficiency in programming language is required (Oulas et al., 2015).

Other widely used pipelines for analysis of microbial communities include Mothur, created from an amalgam of DOTUR, SONS, and Treeclimber (Schloss & Handelsman, 2005, 2006a, 2006b) along with other algorithms currently incorporated; and SILVAngs (Quast et al., 2013), a web-based fully automated analysis pipeline for rRNA marker gene amplicon sequencing datasets. The pipeline workflow includes: alignment of reads, quality assessment and filtering of low-quality reads, dereplication, clustering, and OTU picking using previously defined thresholds, and taxonomic assignment of OTUs using the SILVA rDNA database.

Metagenome gene prediction and functional annotation

Annotation of the assembled data and identification of genomic features such as genes and regulatory elements is the next step in the metagenomics analysis pipeline. Prior to annotation, a series of preprocessing steps are usually carried out on the sequence read to remove low-quality reads (FASTX-Toolkit, Su, Xu, & Ning, 2012; DynamicTrim, Cox. Peterson, & Biggs, 2010), dereplicate, and to remove sequencing artifacts and host DNA contaminants (Bowtie-2, Langmead & Salzberg, 2012). *Gene calling*, which is

the identification of genes within the reads or assembled contigs, is the stage that follows in the annotation pipeline and here genes are labeled as coding DNA sequences (CDSs) and noncoding RNA genes. Dedicated software programs such as Metagene (Noguchi, Park, & Takagi, 2006), Orphelia (Hoff, Lingner, Meinicke, & Tech, 2009), MetaGeneMark (Zhu, Lomsadze, & Borodovsky, 2010), Prodigal (Hyatt et al., 2010), and FragGeneScan (Rho, Tang, & Ye, 2010) have been developed to identify CDSs using ab initio gene prediction algorithms. Gene prediction tools utilize codon information (i.e., start codon—AUG) to identify potential open reading frames and hence label sequences as coding or noncoding. Other gene prediction tools that are gaining attention include Glimmer-MG (Kelley, Liu, Delcher, Pop, & Salzberg, 2012), which not only predicts genes but also identifies insertions/deletions more accurately than FragGeneScan and can also predict substitution errors affecting stop codons (Roumpeka et al., 2017). Noncoding RNAs such as ribosomal RNA (rRNA) genes are predicted using locally developed rRNA models for IMG/MER, and MG-RAST using BLAT algorithm that predicts rRNA genes by comparing three known databases: SILVA (Quast et al., 2013), Greengenes (DeSantis et al., 2006), and Ribosomal Database Project-RDP (Cole et al., 2007). For transfer RNAs (tRNAs) they are predicted using programs like tRNAscan (Lowe & Eddy, 1997).

Functional annotation of the predicted genes is achieved by homology-based searches of query sequences against annotation databases containing known functional and/or taxonomic information. Due to large datasets involved in metagenomic analysis, manual annotation is not feasible and automated annotation such as BLASTX, which runs on high-performance computer cluster, is time-consuming, especially for large datasets, and computationally expensive (Oulas et al., 2015). Many reference data repositories are available that provide functional annotation to metagenomic datasets. This includes Kyoto Encyclopedia of Genes and Genomes (KEGG), (GhostKOALA) (Kanehisa et al., 2008; Kanehisa, Sato, & Morishima, 2016), SEED subsystems (Overbeek et al., 2005), EuKaryotic Orthologous Groups (KOG)/Clusters of Orthologous Groups (COG) (Tatusov, Galperin, Natale, & Koonin, 2000), EggNOG (Powell et al., 2014), PFAM (Bateman et al., 2000), and TIGRFAM (Haft, Selengut, & White, 2003). Most metagenomic annotation pipelines such as IMG/MER, MG-RAST, Prokka (Seaman, 2014), among others, use their single framework to merge and visualize the interpretations of multiple annotation database or composite protein domain database (InterPro,

InterProScan) searches to obtain a comprehensive biological functions annotation.

Web-based metagenomic annotation resources such as IMG/MER and MG-RAST are fully automated pipelines embedded with diverse bioinformatic tools for read quality control, gene prediction, and functional annotation (Oulas et al., 2015). They are widely used particularly by researchers who are less adept in handling programming languages used to write most metagenomic analysis tools. While both pipelines are data management repositories and are widely used for comparative metagenomics, important differences between the two exist, and this clearly reflects in the abundance profiles generated by the two pipelines. The similarities between MG-RAST and IMG-MER end with gene prediction, as both pipelines predict all the genes in the metagenome (Oulas et al., 2015). However, MG-RAST, after gene prediction, identifies the best homologs of those genes in the isolate genomes (in its database) using BLAT tools, which only consider sequence similarities above or equal to 70%, thereby missing strong hits to other genes below 70% similarity. After identifying the best hits to genes from isolated genomes, subsequent analyses in MG-RAST is done using the genes of the isolated genomes, not the genes of the metagenome being analyzed. Though the method is fast, as the only computationally intensive step is to identify best hits of the metagenome against the isolated genomes, it creates lot of limitations as novel genes in the metagenome that do not have homologs in the isolated genomes are completely missed. Contrastingly, IMG-MER runs all its analysis on the predicted genes from the metagenomes rather than on the homologous genes from the isolated genomes which allow identification of PFAM hits, provide comprehensive functional information compared to COG used in MG-RAST. Further, as IMG-MER keeps the original metagenome genes and contigs, it provides synteny information, which can be used in identifying novel biosynthetic gene clusters (BGCs) in the metagenome (Oulas et al., 2015). Other web-based metagenomic analysis pipelines such as EBI MetaGenomics by EBI (Hunter et al., 2014) which recently became MGnify (www.abi.ac.uk 2018) and EDGE platform (Li et al., 2017) contain several software tools for bioinformatic analyses such as quality control of datasets, assembly, gene prediction, functional annotation, taxonomic classification, and phylogenetic analysis of metagenomic reads.

Metatranscriptomics focuses on what genes are expressed by the entire microbial community and provides a snapshot of the gene expression from the total mRNA in a given sample at a given time, under specific

conditions. Metatranscriptomics analysis pipeline uses two approaches: (1) reads mapping to a reference genome (2) de novo assembly of the reads into transcript contigs and supercontigs. In reads mapping to a reference genome, the metatranscriptomic reads are mapped to specialized databases using alignment tools such as BWA, Bowtie2, and BLAST and the result annotated using functional annotation tools such as KEGG, COG, GO, and Swiss-Prot (Duran-Pinedo et al., 2014; Jorth et al., 2014; Leimena et al., 2013; Xiong et al., 2012; Yost, Duran-Pinedo, Teles, Krishnan, & Frias-Lopez, 2015). However, in de novo assembly, the metatranscriptomic reads are assembled using de novo sequence assemblers such as Trinity, MetaVelvet, Oases, AbySS, SOAPden-ovo into longer fragment called contig and the relative expression of genes are inferred using functional annotation tools (Birol et al., 2009; Grabherr et al., 2011; Li et al., 2010; Namiki et al., 2012; Robertson et al., 2010; Schulz, Zerbirno, Vingram, & Birney, 2012). The major limitation to large-scale application of metatranscriptomics includes domination of rRNA in the harvested RNA that dramatically reduces the coverage of mRNA, which is the main target of metatranscriptomics studies. Other limitations include instability of mRNA, which could compromise the integrity of the sample before sequencing, difficulty in differentiating between host and microbial RNA, and limited coverage of transcriptome reference databases (Aguiar-Pulido et al., 2016; Peano et al., 2013; Perez-Losada, Castro-Nallar, Bendall, Freishtat, & Crandall, 2015).

Metagenomic data sharing, storage, and management

Web-based metagenomic analysis pipelines such as IMG/MER, CAMERA, MG-RAST, and EBI metagenomics provide an integrated environment for analysis, management, storage, and sharing of metagenome projects. This means that a generally accepted annotation scheme is developed to allow for efficient data exchange, integration, sharing, and visualization between different platforms and reduce the need for reprocessing of metagenomic datasets, which is computationally expensive (Oulas et al., 2015).

The Genomic Standards Consortium (GSC) is making giant strides in developing a widely accepted language that shares ontologies and nomenclatures thereby providing a common standard for exchange of intermediate and processed results derived from the analysis of metagenomic projects. To achieve this, MIMS (Minimum Information about a Metagenome Sequence) and MIMARKS (Minimum Information about a MARKer Sequence) have been invented (Yilmaz et al., 2011), which provide a scheme of standard languages for metadata annotation.

Freshwater lakes metagenomic studies

Freshwater lakes and toxic cyanobacterial blooms

Understanding the ecology and control of cyanobacterial harmful algal blooms (cHABs) is important for preservation of aquatic ecosystems, sustenance of surface water quality, and improved health for humans and livestock. Several members of cHABs produce cyanotoxins (microcystins, cylindrospermopsin, anatoxin) and off-flavor/odorous compounds (geosmin or 2-methylisoborneol), which negatively impact the aquatic ecosystems and render the water unsafe for human and livestock use (Ferrão-Filho, Da, & Kozlowsky-Suzuki, 2011; Li et al., 2012). Several workers have used NGS technologies to decipher the community structure and functions of cHABs and provided useful information that could aid in control of this menace. Using 454 GS FLX Titanium system, Steffen et al. (2012) conducted metagenomic studies during bloom events in three lakes [Lake Taihu (China), Lake Erie (N. America), and Grand Lake St. Marys (GLSM; OH, USA)] covering two continents. The studies revealed a high proportion of cyanobacterial sequences (25%–88%) including common bloom members of Chroococcales (e.g., *Cyanothece*, *Synechococcus*, *Crocosphera;* c.40%–45%), Nostocales (c.10%), and Oscillatoriales (e.g., *Lyngbya*, *Trichodesmium*-like, *Arthrospira*, etc; c.17%–38%). Later studies of other freshwater systems using metatranscriptomics suggest that the most active cyanobacterial taxa were also from the same cyanobacterial orders, dominated by *Microcystis* in a eutrophic Singaporean reservoir (Penn, Wang, Fernando, & Thompson, 2014) and in Lake Erie with codominant taxa including Synechococcales and Gloeobacterales (Steffen et al., 2015). Noncyanobacterial members of cHAB communities revealed by recent studies were dominated by members of Proteobacteria (>90% of noncyanobacterial sequences) with a smaller portion of Bacteroidetes (Li et al., 2011; Penn et al., 2014; Steffen et al., 2012). Additionally, using MPS of the plastid 23S rRNA gene specific for cyanobacterial and eukaryotic algal species (e.g., Streptophyta, Euglenids, Chlorophyta, Bacillariophyta) were also observed within blooms (Steven, McCann, & Ward, 2012). Furthermore, viruses/phages, protozoa, and fungi have also been detected in cHAB and are increasingly recognized as key players in bloom persistence and decay (Ger, Hansson, & Lürling, 2014; Gerphagnon, Latour, Colombet, & Sime-Ngando, 2013; Xia, Li, Deng, & Hu, 2013).

There is a strong belief that nutrient enrichment via eutrophication in bloom-associated aquatic community heightens genome evolution rates

and possible niche expansion of cyanobacterial species. Steffen et al. (2012) while comparing the genome of *Microcystis aeruginosa* strain NIES843 to metagenomes from Lake Erie, Lake Taihu, and GLSM identified metagenomic islands within *M. aeruginosa* strain NIES 843 genome (defined as regions of ≥10 kb with low coverage in the metagenomes) that were not observed in these three sites and which contained transposase and mobile genetic elements. Further studies revealed that transposase expression in strain NIES 843 is upregulated by nutrient-enrichment, especially in presence of organic nitrogen, urea (Steffen et al., 2014). In addition, metagenomic studies conducted by several workers have reported the detection of cyanotoxin degraders in toxic bloom–associated environments. Steffen et al. (2012) reported the presence of the *mlrC* gene, the gene involved in microbial degradation of microcystin in both lakes (Taihu and Erie). Mou, Lu, Jacob, Sun, and Heath (2013) analyzed the biodegradation of microcystin by natural communities from Lake Erie through metagenomic analysis and reported the enrichment of Methylophilales and Burkholderiales in microcystin-amended microcosm. However, *mlr*, the only known gene responsible for microcystin biodegradation was not detected, suggesting the organisms utilized a new, undiscovered pathway for microcystin biodegradation (Mou et al., 2013). The detection of these cyanotoxin degradation pathways and the hunt for other novel biodegradation pathways is important to water quality managers and environmental experts, as it would aid in designing effective strategies to control the expression of these toxins, improve surface water quality, and safeguard the health and wellbeing of humans and livestock.

Freshwater lakes and viral metagenomics

Perhaps, viruses are the largest reservoirs of underexplored microbial components of the entire biosphere, as they along with phages outnumber microbial cells ten to one in most aquatic environments (Chibani-Chennoufi, Bruttin, Dillmann, & Brussow, 2004; Mohiuddin & Schellhorn, 2015).Viruses control microbial abundance and community structure, influence bacterial virulence and pathogenesis, and shape microbial genetic diversity and evolution via virus-mediated gene transfer (transduction) and host range (Aguirre de Cárcer, López-Bueno, Pearce, & Alcamí, 2015; Mohiuddin & Schellhorn, 2015). Additionally, they are a major player in food web interactions, affect global biogeochemical cycles, and some such as plant viruses and bacteriophages have been used in water quality monitoring as markers for human fecal contamination, and microbial source

tracking (Mohiuddin & Schellhorn, 2015; Uyaguari-Diaz et al., 2016). Recently, freshwater viral metagenomics that is largely unexplored is gaining traction as the need for water quality monitoring, improved water treatment strategies, and deep understanding of viral ecology and physiology is becoming imperative for the general wellbeing.

Skvortsov et al. (2016) characterized the viral community of Lough Neagh, the largest, nutrient-rich, eutrophic freshwater lake in Ireland using Illumina shotgun metagenomics and found that only 15% of the reads (334,507 reads) had homologs in RefSeq database, <0.5% of the reads had ssDNA, majority (97%) had dsDNA viruses of which Caudovirales (tailed bacteriophage) accounted for 79.9% of reads while unclassified dsDNA phages and dsDNA viruses constitute 15.8% and 1.0% of the reads, respectively. Functional characterization of the unassembled metagenomic reads using SEED subsystems showed that 68.3% of all classified reads belong to the subsystem "phages, prophages, transposable elements and plasmids" of which 66.4% belong to phages and prophages and 1.4% belong to gene transfer agents.

Observations of Djikeng et al. (2009) on viral communities in Lake Needwood (USA) clearly show that human activities impact significantly on the diversity and composition of microbial composition in freshwater ecosystems. They identified several viruses of terrestrial origin in the lake with potential agricultural and public health implications. Aside from detection of viruses originating from animals, birds, farmed plants and fish, they also detected Banna virus, a mosquito-borne zoonotic virus predominantly in tropical climates of South East Asia in the water samples collected from the lake. In addition, they also detected several aquatic viruses such as fish viruses, pathogenic to farmed fish (Atlantic salmon nervous necrosis virus, ASNNV; Atlantic halibut nodavirus, AHNV; red spotted grouper nervous necrosis virus, RSGNNV; and striped jack nervous necrosis virus, SJNNV) and viruses, pathogenic to farmed shrimp (Taura syndrome virus, TSV; White spot syndrome virus, WSSV).

Roux et al. (2012) compared the viromes of two lakes located in two different continents and observed that freshwater clades were similar between lakes in different continents but only differ in terms of relative abundance of viral species. Phylogenetic analysis of the major groups of viral communities in the two lakes were closely related despite significant ecological differences between the two lakes showing that viruses in freshwater habitats cluster together regardless the vast geographical distances between sample locations. This observation is further corroborated by two

independent studies conducted by Kim, Aw, Teal, and Rose (2015) who reported the detection of SJNNV, WSSV, and TSV in the Great Lakes and Djikeng et al. (2009) who detected similar viruses in Lake Needwood and the two lakes are several kilometres apart in the USA and they do not share a common water flow.

Several workers have posited that there is a positive correlation between seasonal changes and viral composition in different freshwater ecosystems. Using a metagenomic approach, Tseng et al. (2013) conducted a 2-year survey in an enclosed freshwater reservoir subjected to episodes of typhoon in Feitsui, Taiwan, and observed that addition of terrestrial viruses, which increases in summer and decreases in winter contributed to the increase in diversity of viral communities. They attributed higher host abundance and activity as the reason for high viral diversity in summer unlike in winter when the host abundance and activity was low. Similar trend was observed by Djikeng et al. (2009) in Lake Needwood where higher viral diversity was detected in samples collected in June than in November and observed that some viruses were season-specific, for example, influenza A virus were only detected in winter while the mosquito-borne Banna virus were largely detected in summer. These studies clearly show that seasonal variations and extreme climatic changes significantly alter the relative abundance and diversity of viral communities in freshwater ecosystems.

Freshwater lakes and prokaryotic metagenomics

Prokaryotes are considered as the main drivers of transformation and the cycling of most biologically active elements in the aquatic ecosystems via degradation and mineralization of organic compounds to their inorganic constituents. As the advent of metagenomics allows deeper insight into the community structure and function of prokaryotes in freshwater lakes, several metagenomic studies on freshwater lakes and the roles of prokaryotes have been documented.

Newton et al. (2011) leaning heavily on 16S rRNA gene-based phylogenetic datasets and FISH, reviewed the global bacterial diversity in epilimnetic layers of freshwater lakes and posited that Proteobacteria, Actinobacteria, Cytophaga-Flavobacterium-Bacteroidetes (CFB), Cyanobacteria and Verrucomicrobia, were the most recovered phyla. They also averred that the class Betaproteobacteria are by far the most studied and often the most abundant bacteria inhabiting the upper waters of lakes and sometimes constituting up to 60%–70% of the total number of cells. Using NGS platforms, Roche 454 FLX Titanium and the Illumina Genome Analyzer (GA) II,

Oh et al. (2011) also reported the dominance of the phyla Proteobacteria (37%), Actinobacteria (32%), and Verrucomicrobia (14%) and the predominance of the class Beta-Proteobacteria in the 10 Gb of community whole-genome shotgun (WGS) DNA sequence data obtained from the epilimnion layer of the temperate lake, Lake Lanier.

Several studies have been conducted to understand the effects of natural and anthropogenic perturbation on the microbial community dynamics and metabolic potentials of freshwater lakes. Such studies assist in discerning the sensitivity and resilience of the microbial community to potential perturbations. Poretsky et al. (2014) monitored the effects of strong summer storm and seasonal mixing on the microbial community dynamics of a temperate lake (Lake Lanier, GA, USA), using high throughput 16S rRNA amplicon sequencing and shotgun sequencing approaches. DNA was extracted from samples collected from four time points (three during storm event, August–September; one during mixing event, November) and sequenced via GS-FLX 454 Titanium shotgun sequencing. For amplicon sequencing, V1–V3 regions of the 16S gene in the DNA were PCR-amplified and sequenced using the same sequencing platform. The authors observed that though 16S amplicon sequencing captures broad shift in community diversity with time (Proteobacteria and Actinobacteria dominate at each time point; variation in the relative abundance of specific phyla and genera in each time point as well as individuals comprising the groups; lower relative abundance of Verrucomicrobia and higher relative abundance of Bacteroidetes in November samples compared to storm water samples), it has limited resolution and lower sensitivity compared to shotgun data, which has ×15 and ×10 more phyla and genera at each time point than 16S amplicon data. They also observed that while a strong summer storm has less of an effect on community composition, seasonal mixing revealed a distinct succession of microorganisms, which may be due to introduction of nutrient from the hypolimnion to the epilimnion during mixing, resulting in dramatic shift in microbial community composition (Poretsky et al., 2014; Shade et al., 2011, 2012; Shade et al., 2012). Recently, Morrison et al. (2017) monitored the microbial community dynamics during seasonal stratification events in Grand Lake, Oklahoma, USA, using amplicon-based Illumina high throughput sequencing. Water samples were taken at three different sites within the lake in March (prestratification samples), June (onset of phototrophic blooming, stratified water column) and September (late Summer, postalgal bloom, stratified water column) from the epilimnion, thermocline, and hypolimnion layers. The V4 hypervariable regions of the 16S gene were

PCR-amplified from the extracted DNA (from each sample) using the prokaryotic-specific primer pair 515F and 806R and sequenced using pair-end Illumina MiSeq platform. The authors observed a homogenous microbial community dominated by Actinobacteria and Bacteroidetes (Flavobacterium class) throughout the oxygenated water column of the prestratified March samples. June samples, taken at the onset of phototrophic blooming showed the dominance of distinct microbial communities in each of the layers. The oxic epilimnion layer was dominated by the phototrophic (majorly Prochlorococcus) and heterotrophic (Planctomycetes, Verrucomicrobia, and Betaproteobacteria) microbial communities. In the oxygen-deficient thermocline and hypolimnion layers, the sedimentation of surface biomass triggers the development of a highly diverse community with the enrichment of *Chloroflexi*, *Latescibacteria*, *Armatimonadetes,* and *Deltaproteobacteria* in the particle-associated fractions, and *Gemmatimonadetes* and *Omnitrophica* in the free-living fractions. While the study revealed the effects of anthropogenic perturbations (eutrophication-excessive nutrient run-off into lakes), that is, algal bloom on the dynamics of the lake's microbial community structure, resulting in multiple, spatiotemporally distinct niches during lake stratification, it also reveals the enrichment of multiple, uncultured, and poorly characterized microbial communities in lake's deeper oxygen-deficient thermocline and hypolimnion layers.

Combination of metagenomics and metaproteomics approach was used by Ng et al. (2010) and Lauro et al. (2011) to decipher microbial community functions in Ace Lake, Antarctica. While Ng et al. (2010) targeted a dominant green sulfur bacterium (Chlorobiaceae) believed to be prevalent at 12–14 m depth in the water column, Lauro et al. (2011) investigated the microbial communities throughout the water column (5–23 m). In both studies, microbial fractions were recovered from one to 10 L samples obtained after drilling the ice cover of the lake, whole protein extracts were analyzed using 1D-PAGE followed by LC-MS/MS, and mass spectra were searched directly against the assembled metagenome from the same sample. Metagenomic analysis of samples taken from 12.7 m depth of the lake showed the predominance of Chlorobiaceae, as 76% of the predicted open reading frame (ORF) was assigned to this green sulfur bacteria. The resulting composite genome was used directly to construct a database for searching metaproteomic mass spectra to infer functional activities of Chlorobiaceae at the time of sampling. This led to identification of 504 proteins corresponding to about 31% of the total predicted proteome of Chlorobiaceae. The authors observed that in the distribution of COG categories between

the metaproteome and metagenome, some functional categories such as proteins involved in translation and energy production and conversion were statistically overrepresented while those involved in defense mechanisms and inorganic ion transport and metabolism were underrepresented in the metaproteome. They also detected several genes and their corresponding proteins that are potential adaptive features such as specific polysaccharide structures attributed to cold adaptation, and genes encoding DNA restriction and modification system, which could serve as a potential protective feature against bacteriophages. Also noteworthy is the lack of evidence for assimilatory sulfate reduction in Chlorobiaceae, which suggests strict dependence of this organism on sulfate-reducing bacteria, located at 14 m depth of the lake. The authors through the combined use of metagenomics and metaproteomics were able to access the biology of this organism in its natural habitat by identifying the proteins necessary for the success and survival of Chlorobiaceae in Ace Lake under cold, oligotrophic, oxygen-limited, and extreme light conditions.

Lauro et al. (2011) investigated the structure and functions of the microbial communities in Ace Lake by sampling from six depths with the objective of capturing the interactions between microbial population that defined nutrient cycling. Metagenomic analysis assigned about 28% of ORF to COG categories, while metaproteomic study identified 1824 proteins. Phylogenetic diversity was found to increase with depth and was associated with an increased amount of hypothetical proteins, which is about 67% of the identified proteins at 23 m. Most of the hypothetical proteins seem to belong to novel functional pathways as they did not match any orthologs from known microorganisms. Aggregating the information available in the physicochemical, metagenomic, and metaproteomic data, the authors were able to describe the carbon, nitrogen, and sulfur cycles throughout the water column. For carbon cycling, cyanobacteria were found to carry out aerobic carbon fixation in the epilimnion layer of the lake; green sulfur bacteria and sulfate-reducing bacteria mediated anaerobic carbon fixation further down the water column, while fermentation was thought to take place at the hypolimnion layer of the lake. Remineralization of particulate organic carbon to dissolved organic carbon, which occurs at the surface of the lake was found to be mediated by heterotrophic microorganisms Actinobacteria and members of SAR11 clade. Remineralization of particulate organic matter to CO_2 and CH_4 in the lower stratum was thought to result from joint activities of fermentative, sulfate-reducing, and methanogenic microorganisms. The detection of

CO dehydrogenase genes also indicates that carbon monoxide oxidation could be an important energy generating pathway throughout the water column. Nitrogen assimilation seemed to be the fate process for nitrogen throughout the lake as indicated by the detection of glutamine and glutamate synthetases in the metaproteome, with remineralization localized at the lower stratum. No evidence of nitrification was observed throughout the water column as the metagenome did not contain any ammonia oxidation genes or signatures of known nitrifying bacteria or Archaea, leading the authors to suggest that the absence of nitrification might be a strategy to conserve bioavailable nitrogen as the lake is associated with low level of nitrate. For the sulfur cycle, green sulfur bacteria were found to consume the sulfide produced by the sulfate-reducing bacteria. The sulfate generated provides replenishment for sulfate-reducing bacteria and ensures the continuous turnover of sulfur compounds in the lake. Though genes for assimilatory sulfate reduction were detected in the metagenome, the authors found no evidence of their expression in the metaproteome. However, the identification of proteins from sulfide reductase complex from green sulfur bacteria was indicative of dissimilatory sulfide reduction being an active pathway in the lake. The authors also observed lower numbers of viral particles at the depth (12.7 m) where Chlorobiaceae was located, which is due to the influence of extreme light conditions on the organism's biology and posited that the persistence of the organism in the lake was because of the absence of the phage. The authors concluded that the emergence of phage predators could deplete the populations of green sulfur bacteria in the lake, which could have disastrous consequences for the whole lake community due to their central role in the recycling of carbon, nitrogen, and sulfur in the lake.

Wurzbacher et al. (2017) analyzed four replicate sediment cores taken from 30 m depth in oligo-mesotrophic Lake Stechlin in northern Germany by examining a full suite of biogeochemical parameters and microbial community composition of Archaea, Bacteria, and Eukarya using 454 GS high throughput sequencing that sequenced PCR products generated by amplifying the SSU V6–V8 region (which has high variability for all the three domains) using the primer pair 926F, 1392R. They observed that there was a near complete turnover within the uppermost 30 cm as indicated in the community β-diversity with a pronounced shift from Eukarya- and Bacteria-dominated upper layers (<5 cm) to Bacteria-dominated intermediate layers (5–14 cm) to Archaea dominated deep layers (>14 cm). Interestingly, aside from the typical methanogenic Archaea, which has been

reported in freshwater sediments, the authors also detected typical marine lineages MGI, MCG (Bathyarchaeota), MHVG, DHVEG-1, DHVEG-6 (Woesearchaeota), DSEG, MBG-A, MBG-B (Lokiarchaeota), which are believed to be exclusively resident in marine or deep-sea habitats. Dominance of these Archaeal groups in the sediment was attributed to cellular state of low activity, abundance of detrital proteins caused by cell lysis, and recycling of dormant cells, among others.

Using Metagenomic 2.0 approach (McMahon, 2015), Cuadrat, Ionescu, Davila, and Grossart (2018) used draft genomes generated by MetaBAT (Kang, Froula, Egan, & Wang, 2015) to recover genomic clusters of secondary metabolites from Lake Stechilin, North-Eastern Germany. The DNA was extracted from 26 metagenomic samples from the lake using phenol/chloroform extraction protocol, sequenced on an Illumina Hiseq 2500 using the V3 chemistry and binned using MetaBAT, an efficient binning tool that integrates empirical probabilistic distances of genome abundance, and tetranucleotide frequency for accurate metagenomic binning. Screening the genomes using Anti-SMASH and NAPDOS workflows for secondary metabolism genes, they identified 243 secondary metabolite clusters from 121 genomes with 18 nonribosomal peptide synthases (NRPS), 19 polyketide synthases (PKS), and 3 hybrid PKS/NRPS clusters. The approach also makes possible prediction of partial structure of several secondary metabolite clusters and allows taxonomic classifications and phylogenetic inferences. PKS and NRPS are two families of modular megasynthases that are important for biotechnological and pharmaceutical industry due to their broad spectrum of products ranging from antibiotics, antitumor drugs, food pigments, and harmful toxins such as Anatoxin-a, and Microcystins (Cuadrat et al., 2018; Dadheech et al., 2014).

Freshwater lakes and antibiotic resistance genes

Freshwater lakes inundated with discharges from industries, wastewater treatment plants, hospitals, animal husbandry, and aquaculture laden with antibiotics (Liu, Li, et al., 2012; Liu, Zhang, et al., 2012; Yang, Xu, Cao, Lin, & Wang, 2017; Yin, Yue, Peng, Liu, & Xiao, 2013) has led to massive pollution of the freshwater lakes resulting in impairments of ecosystems and human health, bioaccumulation in food webs, and dissemination and transfer of antibiotic resistance genes (ARGs) between environmental microorganisms and human pathogens (Bengtsson-Palme & Larsson, 2015; Du & Liu, 2012; Li, Yang, Lv, et al., 2015; Li, Yang, Ma, et al., 2015). Diverse environmental studies have enunciated the effects of the discharged antibiotics

on aquatic microorganisms, the nitrogen cycle, natural ecosystems, as well as the environmental fate and distribution of these antibiotics in aquatic environments (Bu, Wang, Huang, Deng, & Yu, 2013; Grenni, Ancona, & Barra Caracciolo, 2018; Liu & Wong, 2013; Roose-Amsaleg & Laverman, 2016; Valitalo, Kruglova, Mikola, & Vahala, 2017).

Sequence-based and functional metagenomics have been used to gain better understanding of the types of ARGs in environmental microorganisms and their colocalization with mobile genetic elements (MGEs). Sequence-based metagenomics involves direct extraction of total DNA from lake samples followed by random sequencing and the sequenced metagenomic reads are then interrogated against a reference database housing known ARG sequences to predict the sequences in the metagenome containing resistance genes (Tan et al., 2015). Functional metagenomics involved cloning randomly sheared DNA fragments into an expression vector, transforming them into a host and selecting transformants, which demonstrate resistance to the selected antibiotics (Allen, Moe, Rodbumrer, Gaarder, & Handelsman, 2009; Sommer, Dantas, & Church, 2009). This dual approach allows identification of highly divergent genes from known ARGs, showing direct evidence of resistance phenotypes associated with expressed genes, and obviate the need for reference genes for gene identification (Pehrsson, Forsberg, Gibson, Ahmadi, & Dantas, 2013). The increasing democratization, sophistication, accuracy, and speed of downstream processing pipelines of NGS datasets such as the Comprehensive Antibiotic Resistance Database (CARD, McArthur et al., 2013), the Beta-Lactamase Database (BLAD, Danishuddin, Hassan Baig, Kaushal, & Khan, 2013), the Antibiotic Resistance Database (Liu & Pop, 2009), and ResFinder have helped in promoting large-scale environmental studies to assess the threat posed by antibiotic resistance.

Recent studies employing NGS and bioinformatic tools to characterize the antibiotic resistance profiles of microbial communities in aquatic environments (water, wastewater, sludge) posited that the abundant ARGs detected in the environments were associated with antibiotics commonly administered in human or veterinary medicine (e.g., aminoglycoside, bacitracin, beta-lactam, chloramphenicol, sulfonamide, and tetracycline) (Li, Yang, Lv, et al., 2015; Li, Yang, Ma, et al., 2015; Tan et al., 2015). Several factors including chemical pollution such as antibiotics and heavy metals have been fingered to influence ARG distribution in lakes. The level of heavy metal contamination in lakes is believed to exert coselection on ARGs and is regarded as the main cause of ARG contamination

in sediments and bulk water of lakes (Baker-Austin, Wright, Stepanauskas, & McArthur, 2006; Devarajan et al., 2015; Seiler & Berendonk, 2012; Yang, Xu, et al., 2017). Other chemicals such as biocides, chemical preservatives, and polycyclic aromatic hydrocarbons have also been found to enrich ARGs in aquatic environments (Romero, Grand Burgos, Perez-Pulido, Golvez, & Lucas, 2017; Wang et al., 2017; Yang & Wang, 2018). Other factors such as physicochemical factors (organic matter of lake sediments, dissolved oxygen, total organic carbon, dissolved nitrogen) and human activities such as effluents from hospitals and wastewater treatment plant, runoff/discharge from agriculture, aquaculture, and animal husbandry have also been reported to influence ARG distribution in lakes (Devarajan et al., 2015; Di Cesare et al., 2015; Guo, Li, Chen, Bond, & Yuan, 2017; Wu, Yu, Yue, Liu, & Yang, 2016; Zhou et al., 2017; Zhu et al., 2013).

Several ARGs have been detected via metagenomic analysis of freshwater lakes. The sulfonamide resistance gene (*sul2*) and quinolone resistance gene (*qnrD*) have been reported to be preponderant in Kazipally Lake, India (Bengtsson-Palme, Bouloud, Fick, Kristiansson, & Larsson, 2014), while the *bacA* (bacitracin resistance) gene was the most prevalent in pristine lakes on the Tibetan Plateau (Chen et al., 2016). Globally, sulfonamide resistance gene (*sul*) and tetracycline resistance gene (*tet*) seem to be highly abundant in lakes with *sul1* reported to be abundant in urban lakes in China, 21 Swiss lakes, and Lake Geneva at 10^{-3} to 10^{-2} copies per 16S rRNA level (Czekalski, Berthold, Caucci, Egli, & Buergmann, 2012; Czekalski, Sigdel, Birtel, Matthews, & Burgmann, 2015; Yang, Liu, et al., 2017; Yang, Xu, et al., 2017). In lake sediments, abundance of *sul1* gene are in the range of 10^{-6} to 10^{-10} copies/gram of sediment with Lake Geneva having a high *sul1* abundance of 2^{-2} x 10^{9} copies/gram of sediment (Czekalski, Diez, & Buergmann, 2014). The *tet* genes are reported to be abundant in Lake Maggiore (10^{6} copies/mL) and Taihu Lake (10^{5} copies/mL) (Di Cesare et al., 2015; Zhang, Sturm, Knapp, & Graham, 2009). *tet* genes were reported to be abundant in Taihu Lake and Lake Geneva sediments attaining 10^{6} copies/gram (Czekalski et al., 2014; Zhang et al., 2009).

One of the major factors facilitating the propagation of ARGs in freshwater lakes is horizontal gene transfer (HGT) via mobile genetic elements (MGEs). HGT is higher in aquatic bacteria than in terrestrial bacteria and MGEs such as plasmids, transposons, integrons, and insertion sequences are important carriers of ARGs (Hu et al., 2016; Stokes & Gillings, 2011). Chu et al. (2017) reported the detection of high proportions of

plasmid-associated ARGs in Lake Michigan sediments, ranging from 32% to 100% of the total identified ARGs. Bangtsson-Palme et al. (2014) also reported the recovery of 26 known and 21 putative novel plasmids in the metagenome of an antibiotic-polluted Indian lake. In addition, Lekunberri, Balcázar, and Borrego (2018) observed a drastic increase in the abundance of integron integrases and ISCR elements following discharge of treated wastewater in Ter River, Spain. The class 1 integrons have been reported as the main vector for HGT of ARGs in several lakes and a direct relationship is believed to exist between the detection of class 1 integron and the preponderance of ARGs in lakes (Yin et al., 2013; Yang, Liu, et al., 2017; Yang, Xu, et al., 2017).

Countering the emerging threats posed by preponderance of antibiotics and antibiotic resistance genes in freshwater lakes demands urgent measures such as development of priority lists of antibiotics and their metabolites in lakes, identifying some of the major pollutants influencing ARG propagation in lakes, identifying the persistence of ARGs in freshwater lakes and their resistance to conventional drinking water treatment methods and instituting strict regulatory policies that protect freshwater lakes from antibiotic-laden effluents and discharges.

Conclusions

The importance of freshwater lakes as a renewable resource, a source of drinking water, recreational activities, and a sink for global biogeochemical cycling and regeneration mediated by diverse physiological groups of microorganisms is well documented. Natural and anthropogenic perturbations of freshwater lakes have impacted the richness, distribution, composition, and functions of lake microbial communities with consequences on the cycling of essential nutrients and health and safety of all life forms. Culture-dependent assessment of microbial community structure and functions in pristine and perturbed lakes barely scratches the surface, making it difficult to have a deeper insight into dominant and rare microorganisms at different strata of the lake, predict the effect of perturbations on community structure and functions, and detect novel biomolecules of immense industrial and biotechnological applications.

The advent of metagenomics and the democratization of sequencing technologies allow unparallel access into the world of the "hidden majority" that permeates different strata of freshwater lakes playing diverse roles and contributing immensely in the cycling and regeneration of nutrients

and maintenance of ecological balance. Early works on freshwater microbial ecology have focused on the epilimnion, believed to be the most active layer of the lake, since the number of microbial cells in lakes decreases with depth. Several workers have reported the predominance of the phyla Proteobacteria (particularly β-Proteobacteria), Actinobacteria, and Cyanobacteria; the microbial communities in algal and cyanobacterial blooms, the toxins generated from the blooms, and the debilitating effects of these toxins on humans and livestock. Later works have extended analysis of freshwater to include other layers of stratified lakes such as thermocline and the hypolimnion and their findings unraveled several rare and important members of the microbial communities and their distinctive functions in these freshwater lakes. Microbial lineages initially thought to be recovered exclusively from marine and deep-sea environments have been recovered via metagenomics from lake sediments, thus deepening our understanding on microbial ecology of stratified lakes. The importance of microbial interactions such as syntrophism, commensalism, predation, among others, in nutrient sequestration, degradation, and mineralization; sustenance of global carbon fluxes; and maintenance of ecological balance and water quality also dominated several studies.

The combined use of metagenomics, metaproteomics, and metatranscriptomics allows elucidation of functional genes used by the members of microbial communities in freshwater lakes to carry out essential metabolic activities and the taxonomic affiliations of the genes, thus illuminating the reason d'etre for the abundance of specific physiological group of microorganisms at a stratum of freshwater lakes. It also highlights the importance of various physicochemical factors in contributing to spatial distribution of microorganisms in lakes. Furthermore, recovery of several secondary metabolite clusters produced by microorganisms in lakes coding for novel biomolecules that have biotechnological, pharmaceutical, and industrial applications also indicates the importance of metagenomics in freshwater microbiology.

Lastly, while metagenomics has enriched our understanding of fresh water microbial ecology, there is a need to combine various -omic tools to further elucidate the structure and functions of microbial assemblages in freshwater lakes. Leveraging on these technologies will facilitate deeper understanding of freshwater lake environment and accelerate the detection of novel biomolecules and rare species with interesting features that may serve the critical needs of the society and save humanity from the resurging menace of toxic blooms and antibiotic resistance.

References

Afiahayati, Sato, K., & Sakakibara, Y. (2014). MetaVelvet-sl: An extension of the Velvet assembler to a de novo metagenomic assembler utilizing supervised learning. *DNA Research*, *22*(1), 69–77.

Aguiar-Pulido, V., Huang, W., Suarez-Ulloa, V., Cickovski, T., Mathee, K., & Narasimhan, G. (2016). Metagenomics, metatranscriptomics, and metabolomic approaches for microbiome analysis. *Evolutionary Bioinformatics Online*, *12*(Suppl. 1), 5–16.

Aguirre de Cárcer, D., López-Bueno, A., Pearce, D. A., & Alcamí, A. (2015). Biodiversity and distribution of polar freshwater DNA viruses. *Science Advances*, *1*, e1400127.

Allen, H. K., Moe, L. A., Rodbumrer, J., Gaarder, A., & Handelsman, J. (2009). Functional metagenomics reveals diverse beta-lactamases in a remote Alaskan soil. *The ISME Journal*, *3*, 243–251.

Alneberg, J., Bjarnason, B. S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., et al. (2014). Binning metagenomic contigs by coverage and composition. *Nature Methods*, *11*, 1144–1146.

Ambardar, S., Gupta, R., Trakroo, D., Lal, R., & Vakhlu, J. (2016). High throughput sequencing: An overview of sequencing chemistry. *Indian Journal of Microbiology*, *56*, 394–404.

Bai, Y., Liu, R., Liang, J., & Qu, J. (2013). Integrated metagenomic and physiochemical analyses to evaluate the potential role of microbes in the sand filter of a drinking water treatment system. *PLoS One*, *8*, e61011.

Baker-Austin, C., Wright, M. S., Stepanauskas, R., & McArthur, J. (2006). Co-selection of antibiotic and metal resistance. *Trends in Microbiology*, *14*, 176–182.

Ballester, L. Y., Luthra, R., Kanagal-Shamanna, R., & Singh, R. R. (2016). Advances in clinical next-generation sequencing: Target enrichment and sequencing technologies. *Expert Review of Molecular Diagnosis*, *16*, 357–372.

Bashir, Y., Singh, S. P., & Konwar, B. K. (2014). Metagenomics: An application-based perspective. *Chinese Journal ofBiology*, 1–7. https://doi.org/10.1155/2014/146030.

Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L., & Sonnhammer, E. L. (2000). The Pfam protein families' database. *Nucleic Acids Research*, *28*(1), 263–266.

Battin, T. J., Luyssaert, S., Kaplan, L. A., Aufdenkampe, A. K., Richter, A., & Tranvik, L. J. (2009). The boundless carbon cycle. *Nature Geoscience*, *2*, 598–600.

Bengtsson-Palme, J., Boulund, F., Fick, J., Kristiansson, E., & Larsson, D. G. J. (2014). Shotgun metagenomics reveals a wide array of antibiotic resistance genes and mobile elements in a polluted lake in India. *Frontiers in Microbiology*, *5*, e648.

Bengtsson-Palme, J., & Larsson, D. G. J. (2015). Antibiotic resistance genes in the environment: Prioritizing risks. *Nature Reviews Microbiology*, *13*, 396.

Birol, I., Jackman, S. D., Nielsen, C. B., Qian, J. Q., Varhol, R., Stazyk, G., et al. (2009). De novo transcriptome assembly with abyss. *Bioinformatics*, *25*(21), 2872–2877.

Bollmann, A., Lewis, K., & Epstein, S. S. (2007). Incubation of environmental samples in a diffusion chamber increases the diversity of recovered isolates. *Applied and Environmental Microbiology*, *73*, 6386–6390.

Bowen De Leon, K., Gerlach, R., Peyton, B. M., & Fields, M. W. (2013). Archaeal and bacterial communities in three alkaline hot springs in heart lake geyser basin, yellowstone national Park. *Frontiers in Microbiology*, *4*, 330.

Brady, A., & Salzberg, S. L. (2009). Phymm and PhymmBL: Metagenomic phylogenetic classification with interpolated markov models. *Nature Methods*, *6*(9), 673–676.

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J. M., Segall, A. M., Mead, D., et al. (2002). Genomic analysis of uncultured marine viral communities. *Proceedings of the National Academy of Sciences of the United States of America*, *99*, 14250–14255.

Breitwieser, F. P., Lu, J., & Salzberg, S. L. (2017). A review of methods and databases for metagenomic classification and assembly. *Briefings in Bioinformatics*, 1–15 2017.

Bu, Q., Wang, B., Huang, J., Deng, S., & Yu, G. (2013). Pharmaceuticals and personal care products in the aquatic environment in China: A review. *Journal of Hazardous Materials*, *262*, 189–211.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, *7*(5), 335–336.

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Huntley, J., Fierer, N., et al. (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *The ISME Journal*, *6*, 1621–1624.

Chao, Y., Ma, L., Yang, Y., Ju, F., Zhang, X., Wu, W., et al. (2013). Metagenomic analysis reveals significant changes of microbial compositions and protective functions during drinking water treatment. *Scientific Reports*, *3*.

Chen, I. M., Grechkin, Y., Dubchak, I., Anderson, I., Lykidis, A., Mavromatis, K., et al. (2008). IMG/M: A data management and analysis system for metagenomes. *Nucleic Acids Research*, *36*(Database), D534–D538.

Chen, B. W., Yuan, K., Chen, X., Yang, Y., Zhang, T., Wang, Y. W., et al. (2016). Metagenomic analysis revealing antibiotic resistance genes (ARGs) and their genetic compartments in the Tibetan environment. *Environmental Science and Technology*, *50*, 6670–6679.

Chevreux, B., Pfisterer, T., Drescher, B., Driesel, A. J., Muller, W. E., Wetter, T., et al. (2004). Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Research*, *14*(6), 1147–1159.

Chibani-Chennoufi, S., Bruttin, A., Dillmann, M. L., & Brussow, H. (2004). Phage-host interaction: An ecological perspective. *Journal of Bacteriology*, *186*, 3677–3686.

Chu, B. T. T., Petrovich, M. L., Chaudhary, A., Wright, D., Murphy, B., Wells, G., et al. (2017). Metagenomic analysis reveals the impact of wastewater treatment plants on the dispersal of microorganisms and genes in aquatic sediments. *Applied and Environmental Microbiology*. https://doi.org/10.1128/AEM.02168-17.

Cole, J. R., Chai, B., Farris, B. J., Wang, Q., Kulam-Syed-Mohideen, A. S., & McGarrell, D. M. (2007). The ribosomal database project (RDP-II): Introducing myRDP space and quality controlled public data. *Nucleic Acids Research*, *35*(Database issue), D169–D172.

Coolen, M. J. L., van de Giessen, J., Zhu, E. Y., & Wuchter, C. (2011). Bioavailability of soil organic matter and microbial community dynamics upon permafrost thaw. *Environmental Microbiology*, *13*, 2299–2314.

Cox, M. P., Peterson, D. A., & Biggs, P. J. (2010). SolexaQA: At-a-glance quality assessment of illumina second-generation sequencing data. *BMC Bioinformatics*, *11*, 485.

Cuadrat, R. R. C., Ionescu, D., Davila, A. M. R., & Grossart, H.-P. (2018). Recovering genomics clusters of secondary metabolites from lakes using genome-resolved metagenomics. *Frontiers in Microbiology*, *9*, 251.

Czekalski, N., Berthold, T., Caucci, S., Egli, A., & Buergmann, H. (2012). Increased levels of multiresistant bacteria and resistance genes after wastewater treatment and their dissemination into Lake Geneva, Switzerland. *Frontiers in Microbiology*, *3*, e106.

Czekalski, N., Diez, E. G., & Buergmann, H. (2014). Wastewater as a point source of antibiotic-resistance genes in the sediment of a freshwater lake. *The ISME Journal*, *8*, 1381–1390.

Czekalski, N., Sigdel, R., Birtel, J., Matthews, B., & Bürgmann, H. (2015). Does human activity impact the natural antibiotic resistance background? Abundance of antibiotic resistance genes in 21 Swiss lakes. *Environment International*, *81*, 45–55.

Dadheech, P. K., Selmeczy, G. B., Vasas, G., Padisak, J., Arp, W., Tapolczai, K., et al. (2014). Presence of potential toxin-producing cyanobacteria in an oligo-mesotrophic Lake in baltic lake district, Germany: An ecological, genetic and toxicological survey. *Toxins*, *6*, 2912–2931.

Damashek, J., Smith, J. M., Mosier, A. C., & Francis, C. A. (2014). Benthic ammonia oxidizers differ in community structure and biogeochemical potential across a riverine delta. *Frontiers in Microbiology*, *5*, 743.

Danishuddin, M., Hassan Baig, M., Kaushal, L., & Khan, A. U. (2013). BLAD: A comprehensive database of widely circulated beta-lactamases. *Bioinformatics*, *29*, 2515–2516.

Debroas, D., Humbert, J., Enault, F., Bronner, G., Faubladier, M., & Cornillot, E. (2009). Metagenomic approach studying the taxonomic and functional diversity of the bacterial community in a mesotrophic lake (Lac du Bourget — France). *Environmental Microbiology*, *11*, 2412–2424.

Deiner, K., Walser, J.-C., Maechler, E., & Altermatt, F. (2015). Choice of capture and extraction methods affect detection of freshwater biodiversity from environmental DNA. *Biological Conservation*, *183*, 53–63.

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., et al. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology*, *72*(7), 5069–5072.

Devarajan, N., Laffite, A., Graham, N. D., Meijer, M., Prabakar, K., Mubedi, J. I., et al. (2015). Accumulation of clinically relevant antibiotic-resistance genes, bacterial load, and metals in freshwater lake sediments in Central Europe. *Environmental Science and Technology*, *49*, 6528–6537.

Di Cesare, A., Eckert, E. M., Teruggi, A., Fontaneto, D., Bertoni, R., Callieri, C., et al. (2015). Constitutive presence of antibiotic resistance genes within the bacterial community of a large subalpine lake. *Molecular Ecology*, *24*, 3888–3900.

Diaz, N. N., Krause, L., Goesmann, A., Niehaus, K., & Nattkemper, T. W. (2009). TACOA: Taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. *BMC Bioinformatics*, *10*, 56.

Dick, G. J., Andersson, A. F., Baker, B. J., Simmons, S. L., Thomas, B. C., Yelton, A. P., et al. (2009). Community-wide analysis of microbial genome sequence signatures. *Genome Biology*, *10*(8), R85.

Djikeng, A., Kuzmickas, R., Anderson, N. G., & Spiro, D. J. (2009). Metagenomic analysis of RNA viruses in a fresh water lake. *PLoS One*, *4*(9), e7264.

Dohm, J. C., Lottaz, C., Borodina, T., & Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, *36*(16), e105.

Downing, J. A., Prairie, Y. T., Cole, J., Duarte, C. M., Tranvik, L. J., Striegl, R. G., et al. (2006). The global abundance and size distribution of lakes, ponds, and impoundments. *Limnology and Oceanography*, *51*, 2388–2397.

Du, L. F., & Liu, W. K. (2012). Occurrence, fate, and ecotoxicity of antibiotics in agro-ecosystems. A review. *Agronomy for Sustainable Development*, *32*, 309–327.

Duran-Pinedo, A. E., Chen, T., Teles, R., Starr, J. R., Wang, X., Krishnan, K., et al. (2014). Community-wide transcriptome of the oral microbiome in subjects with and without periodontitis. *The ISME Journal*, *8*(8), 1659–1672.

Elmhagen, B., Destouni, G., Angerbjörn, A., Borgstrom, S., Boyd, E., Cousins, S. A. O., et al. (2015). Interacting effects of change in climate, human population, land use, and water use on biodiversity and ecosystem services. *Ecology and Society*, *20*(1), 23.

Escober-Zepeda, A., Vera-Ponce de Leon, A., & Sanchez-Flores, A. (2015). The road to metagenomics: From microbiology to DNA sequencing technologies and bioinformatics. *Frontiers in Genetics*, *6*, 348.

Faust, K., & Raes, J. (2012). Microbial interactions: From networks to models. *Nature Reviews Microbiology*, *10*, 538–550.

Felczykowska, A., Krajewska, A., Zielińska, S., & Łoś, J. M. (2015). Sampling, metadata and DNA extraction – important steps in metagenomic studies. *Acta Biochimica Polonica*, *62*(1), 151–160.

Ferrão-Filho, A., Da, S., & Kozlowsky-Suzuki, B. (2011). Cyanotoxins: Bioaccumulation and effects on aquatic animals. *Marine Drugs*, *9*, 2729–2772.

Ferreira, A. J. S., Siam, R., Setubal, J. C., Moustafa, A., Sayed, A., Chambergo, F. S., et al. (2014). Core microbial functional activities in ocean environments revealed by global metagenomic profiling analyses. *PLoS One*, *9*, e97338.

Fierer, N., Breitbart, M., Nulton, J., Salamon, P., Lozupone, C., Jones, R., et al. (2007). Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of bacteria, archaea, fungi and viruses in soil. *Applied and Environmental Microbiology*, *73*, 7059–7066.

Fisher, J. C., Newton, R. J., Dila, D. K., & McLellan, S. L. (2015). Urban microbial ecology of a freshwater estuary of Lake Michigan. *Elementa: Science of the Anthropocene*, *3*, 000064.

Frias-Lopez, J., Shi, Y., Tyson, G. W., Coleman, M. L., Schuster, S. C., Chisholm, S. W., et al. (2008). Microbial community gene expression in ocean surface waters. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 3805–3810.

Garza, D. R., & Dutilh, B. E. (2015). From cultured to uncultured genome sequences: Metagenomics and modeling microbial ecosystems. *Cellular and Molecular Life Sciences*, *2*, 4287–4308.

Ger, K. A., Hansson, L.-A., & Lürling, M. (2014). Understanding cyanobacteria-zooplankton interactions in a more eutrophic world. *Freshwater Biology*, *59*, 1783–1798.

Gerphagnon, M., Latour, D., Colombet, J., & Sime-Ngando, T. (2013). Fungal parasitism: Lifecycle, dynamics and impact on cyanobacterial blooms. *PLoS One*, *8*, e60894.

Ghai, R., Hernandez, C. M., Picazo, A., Mizuno, C. M., Ininbergs, K., Diez, B., et al. (2012). Metagenomes of Mediterranean coastal lagoons. *Scientific Reports*, *2*.

Gilpin, B. J., Devane, M., Nourozi, F., Robson, B., Scholes, P., & Lin, S. (2013). Recommendations for the processing and storage of water samples before polymerase chain reaction (PCR) analysis. *New Zealand Journal of Marine and Freshwater Research*, *47*, 582–586.

Glass, E. M., Wilkening, J., Wilke, A., Antonopoulos, D., & Meyer, F. (2010). Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. *Cold Spring Harbour Protocols* (1), dbrot5368.

Graber, J. R., & Breznak, J. A. (2005). Folate cross-feeding supports symbiotic homoacetogenic spirochetes. *Applied and Environmental Microbiology*, *71*, 1883–1889.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–652.

Grenni, P., Ancona, V., & Barra Caracciolo, A. (2018). Ecological effects of antibiotics onnatural ecosystems: A review. *Microchemical Journal*, *136*, 25–39.

Gubry-Rangin, C., Hai, B., Quince, C., Engel, M., Thomson, B. C., James, P., et al. (2011). Niche specialization of terrestrial archaeal ammonia oxidizers. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 21206–21211.

Guo, J. H., Li, J., Chen, H., Bond, P. L., & Yuan, Z. G. (2017). Metagenomic analysis reveals wastewater treatment plants as hotspots of antibiotic resistance genes and mobile genetic elements. *Water Research*, *123*, 468–478.

Haas, B. J., Gevers, D., Earl, A. M., Feldgarden, M., Ward, D. V., Giannoukos, G., et al. (2011). Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Research*, *21*, 494–504.

Haft, D. H., Selengut, J. D., & White, O. (2003). The TIGRFAMs database of protein families. *Nucleic Acids Research*, *31*(1), 371–373.

Handelsman, J. (2004). Metagenomics: Application of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews*, *68*, 669–685.

Harrington, C. T., Lin, E. I., Olson, M. T., & Eshleman, J. R. (2013). Fundamentals of pyrosequencing. *Archives of Pathology and Laboratory Medicine*, *137*, 1296–1303.

Hinrichsen, D. (1998). Feeding a future world. *People and the Planet*, 7(1), 6–9.

Hoff, K. J., Lingner, T., Meinicke, P., & Tech, M. (2009). Orphelia: Predicting genes in metagenomic sequencing reads. *Nucleic Acids Research*, *37*(Web Server issue), W101–W105.

Hugenholtz, P. (2002). Exploring prokaryotic diversity in the genomic era. *Genome Biology Review*, *3*, 0003.1–0003.8.

Hunter, S., Corbett, M., Denise, H., Fraser, M., Gonzalez-Beltran, A., Hunter, C., et al. (2014). EBI metagenomics – a new resource for the analysis and archiving of metagenomic data. *Nucleic Acids Research*, *42*, 600–606.

Hurt, R. A., Qiu, X., Wu, L., Roh, Y., Palumbo, A. V., Tiedje, J. M., et al. (2001). Simultaneous recovery of RNA and DNA from soils and sediments. *Applied and Environmental Microbiology*, *67*, 4495–4503.

Huson, D. H., & Weber, N. (2013). Microbial community analysis using MEGAN. *Methods in Enzymology*, *531*, 465–485.

Hu, Y., Yang, X., Li, J., Lv, N., Liu, F., Wu, J., et al. (2016). The bacterial mobile resistome transfer network connecting the animal and human microbiomes. *Applied and Environmental Microbiology*, *82*, 6672–6681.

Hyatt, D., Chen, G. L., Locascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, *11*, 119.

Inagaki, F., Nunoura, T., Nakagawa, S., Teske, A., Lever, M., Lauer, A., et al. (2006). Biogeographical distribution and diversity of microbes in methane hydrate-bearing deep marine sediments on the Pacific Ocean Margin. *Proceedings of the National Academy of Sciences of the United States of America*, *103*, 2815–2820.

Jorth, P., Turner, K. H., Gumus, P., Nizam, N., Buduneli, N., & Whiteley, M. (2014). Metatranscriptomics of the human oral microbiome during health and disease. *mBio*, *5*(2), e1012–e1014.

Kalmbach, S., Manz, W., & Szewzyk, U. (1997). Isolation of new bacterial species from drinking water biofilms and proof of their *in situ* dominance with highly specific 16S rRNA probes. *Applied and Environmental Microbiology*, *63*, 4164–4170.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., et al. (2008). KEGG for linking genomes to life and the environment. *Nucleic Acids Research*, *36*(Database issue), D480–D484.

Kanehisa, M., Sato, Y., & Morishima, K. (2016). BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *Journal of Molecular Biology*, *428*(4), 726–731.

Kang, D. D., Froula, J., Egan, R., & Wang, Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *Peer Journal*, *3*, e1165.

Kelley, D. R., Liu, B., Delcher, A. L., Pop, M., & Salzberg, S. L. (2012). Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. *Nucleic Acids Research*, *40*, e9.

Kim, Y., Aw, T. G., Teal, T. K., & Rose, J. B. (2015). Metagenomic investigation of viral communities in ballast water. *Environmental Science and Technology*, *49*, 8396–8407.

Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., et al. (2013). Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next generation sequencing-based diversity studies. *Nucleic Acids Research*, *41*, e1.

Kõljalg, U., Nilsson, R. H., Abarenkov, K., Tedersoo, L., Taylor, A. F., Bahram, M., et al. (2013). Towards a unified paradigm for sequence-based identification of fungi. *Molecular Ecology*, *22*(21), 5271–5277.

Konstantinidis, K. T., Ramette, A., & Tiedje, J. M. (2006). The bacterial species definition in the genomic era. *Philosophical Transactions of the Royal Society of London B Biological Sciences*, *361*(1475), 1929–1940.

Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K., & Schloss, P. D. (2013). Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Applied and Environmental Microbiology*, *79*, 5112–5120.

Krause, L., Diaz, N. N., Goesmann, A., Kelley, S., Nattkemper, T. W., Rohwer, F., et al. (2008). Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Research*, *36*(7), 2230–2239.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357.

Lauro, F. M., DeMaere, M. Z., Yau, S., Brown, M. V., Ng, C., Wilkins, D., et al. (2011). An integrative study of a meromictic lake ecosystem in Antarctica. *The ISME Journal*, *5*, 879–895.

Lear, G., Dickie, I., Banks, J. C., Boyer, S., Buckley, H. L., Buckley, T. R., et al. (2018). Methods for the extraction, storage, amplification and sequencing of DNA from environmental samples. *New Zealand Journal of Ecology*, *42*(1), 10.

Lee, J. E., Lee, S., Sung, J., & Ko, G. (2011). Analysis of human and animal fecal microbiota for microbial source tracking. *The ISME Journal*, *5*(2), 362–365.

Leimena, M. M., Ramiro-Garcia, J., Davids, M., van den Bogert, B., Smidt, H., Smid, E. J., et al. (2013). A comprehensive metatranscriptome analysis pipeline and its validation using human small intestine microbiota datasets. *BMC Genomics*, *14*(1), 530.

Lekunberri, I., Balcázar, J. L., & Borrego, C. M. (2018). Metagenomic exploration reveals a marked change in the river resistome and mobilome after treated wastewater discharges. *Environmental Pollution*, *234*, 538–542.

Leung, H. C., Yiu, S. M., Yang, B., Peng, Y., Wang, Y., Liu, Z., et al. (2011). A robust and accurate binning algorithm for metagenomic sequences with arbitrary species abundance ratio. *Bioinformatics*, *27*(11), 1489–1495.

Lever, M. A., Torti, A., Eickenbusch, P., Michaud, A. B., Santi-Temkiv, T., & Jorgensen, B. B. (2015). A modular method for the extraction of DNA and RNA, and the separation of DNA pools from diverse environmental sample types. *Frontiers in Microbiology*, *6*, 476.

Li, Z., Hobson, P., An, W., Burch, M. D., House, J., & Yang, M. (2012). Earthy odour compounds production and loss in three cyanobacterial cultures. *Water Research*, *46*, 5165–5173.

Li, R., Li, Y., Kristiansen, K., & Wang, J. (2008). SOAP: Short oligonucleotide alignment program. *Bioinformatics*, *24*(5), 713–714.

Li, P.-E., Lo, C.-C., Anderson, J. J., Davenport, K. W., Bishop-Lilly, K. A., Xu, Y., et al. (2017). Enabling the democratization of the genomics revolution with a fully integrated web-based bioinformatics platform. *Nucleic Acids Research*, *45*, 67–80.

Liu, Z., DeSantis, T. Z., Andersen, G. L., & Knight, R. (2008). Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Research*, *36*, e120.

Liu, B., Gibbons, T., Ghodsi, M., Treangen, T., & Pop, M. (2011). Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. *BMC Genomics*, *12*(Suppl. 2), S4.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., et al. (2012a). Comparison of next-generation sequencing systems. *Journal of Biomedicine and Biotechnology*, *2012*, 251364.

Liu, B., & Pop, M. (2009). ARDB-antibiotic resistance genes database. *Nucleic Acids Research*, *37*, D443–D447.

Liu, J. L., & Wong, M. H. (2013). Pharmaceuticals and personal care products (PPCPs): A review on environmental contamination in China. *Environment International*, *59*, 208–224.

Liu, M., Zhang, Y., Yang, M., Tian, Z., Ren, L., & Zhang, S. (2012b). Abundance and distribution of tetracycline resistance genes and mobile elements in an oxytetracycline production wastewater treatment system. *Environmental Science and Technology*, *46*, 7551–7557.

Li, P., Yang, S. F., Lv, B. B., Zhao, K., Lin, M. F., Zhou, S., et al. (2015b). Comparison of extraction methods of total microbial DNA from freshwater. *Genetics and Molecular Research*, *14*(1), 730–738.

Li, B., Yang, Y., Ma, L., Ju, F., Guo, F., Tiedje, J. M., et al. (2015a). Metagenomic and network analysis reveal wide distribution and co-occurrence of environmental antibiotic resistance genes. *The ISME Journal*, *9*, 2490–2502.

Li, N., Zhang, L., Li, F., Wang, Y., Zhu, Y., Kang, H., et al. (2011). Metagenome of microorganisms associated with the toxic Cyanobacteria *Microcystis aeruginosa* analyzed using the 454-sequencing platform. *Chinese Journal of Oceanology and Limnology*, *29*, 505–513.

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., et al. (2010). De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research*, *20*, 265–272.

Loman, N. J., Misra, R. V., Dallman, T. J., Constantinidou, C., Gharbia, S. E., Wain, J., et al. (2012). Performance comparison of benchtop high throughput sequencing platforms. *Nature Biotechnology*, *30*, 434–439.

Lowe, T. M., & Eddy, S. R. (1997). tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, *25*(5), 955–964.

Lozupone, C. A., & Knight, R. (2007). Global patterns in bacterial activity. *Proceedings of the National Academy of Sciences of the United States of America*, *104*, 11436–11440.

Lundberg, D. S., Lebeis, S. L., Paredes, S. H., Yourstone, S., Gehring, J., Malfatti, S., et al. (2012). Defining the core *Arabidopsis thaliana* root microbiome. *Nature*, *488*, 86–90.

Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics*, *9*, 387–402.

Mardis, E. R. (2013). Next-generation sequencing platforms. *Annual Review of Analytical Chemistry*, *6*, 287–303.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, *43*, 376–380.

Markowitz, V. M., Chen, I. M., Palaniappan, K., Chu, K., Szeto, E., Pillay, M., et al. (2014). IMG 4 version of the integrated microbial genomes comparative analysis system. *Nucleic Acids Research*, *42*(Database issue), D560–D567.

Maruyama, A., Nakamura, K., Yamanaka, H., Kondoh, M., & Minamoto, T. (2014). The release rate of environmental DNA from juvenile and adult fish. *PLoS One*, *9*, e114639.

Maxam, A. M., & Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, *74*, 560–564.

McArthur, A. G., Waglechner, N., Nizam, F., Yan, A., Azad, M. A., Baylay, A. J., et al. (2013). The comprehensive antibiotic resistance database. *Antimicrobial Agents and Chemotherapy*, *57*, 3348–3357.

McCliment, E. A., Voglesonger, K. M., O'Day, P. A., Dunn, E. E., Holloway, J. R., & Cary, S. C. (2006). Colonization of nascent, deep sea hydrothermal vents by a novel archaeal and nanoarchaeal assemblage. *Environmental Microbiology*, *8*, 114–125.

McHardy, A. C., Martin, H. G., Tsirigos, A., Hugenholtz, P., & Rigoutsos, I. (2007). Accurate phylogenetic classification of variable-length DNA fragments. *Nature Methods*, *4*(1), 63–72.

McMahon, K. (2015). Metagenomics 2.0. *Environmental Microbiology Reports*, 7, 38–39.

Mohamed, Y. M., Ghazy, M. A., Sayed, A., Ouf1, A., El-Dorry, H., & Siam, R. (2013). Isolation and characterization of a heavy metal-resistant, thermophilic esterase from a Red Sea Brine Pool. *Scientific Reports*, *3*.

Mohiuddin, M., & Schellhorn, H. E. (2015). Spatial and temporal dynamics of virus occurrence in two freshwater lakes captured through metagenomic analysis. *Frontiers in Microbiology*, *6*, 960.

Monzoorul Haque, M., Ghosh, T. S., Komanduri, D., & Mande, S. S. (2009). SOrt-items: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. *Bioinformatics*, *25*(14), 1722–1730.

Moorthie, S., Mattocks, C. J., & Wright, C. F. (2011). Review of massively parallel DNA sequencing technologies. *The HUGO Journal*, *5*, 1–12.

Moran, M. A., Satinsky, B., Gifford, S. M., Luo, H., Rivers, A., Chan, L. K., et al. (2013). Sizing up metatranscriptomics. *The ISME Journal*, 7, 237–243.

Morrison, J. M., Baker, K. D., Zamor, R. M., Nikolai, S., Elshahed, M. S., & Youssef, N. H. (2017). Spatiotemporal analysis of microbial community dynamics during seasonal stratification events in a freshwater lake (Grand Lake, OK, USA). *PLoS One*, *12*(5), e0177488.

Mou, X., Lu, X., Jacob, J., Sun, S., & Heath, R. (2013). Metagenomic identification of bacterioplankton taxa and pathways involved in microcystin degradation in lakeErie. *PLoS One*, *8*, e61890.

Nagarajan, N., & Pop, M. (2013). Sequence assembly demystified. *Nature Reviews Genetics*, *14*, 157–167.

Namiki, T., Hachiya, T., Tanaka, H., & Sakakibara, Y. (2012). MetaVelvet: An extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Research*, *40*(20), e155.

Newby, D. T., Marlowe, E. M., & Maier, R. M. (2009). Nucleic acid-based methods of analysis. In R. M. Maier, I. L. Pepper, & C. P. Gerba (Eds.), *Environmental microbiology* (2nd ed.) (p. 243). Burlington): Academic Press. 28.

Newton, R. J., Jones, S. E., Eiler, A., McMahon, K. D., & Bertilsson, S. (2011). A guide to the natural history of freshwater lake bacteria. *Microbiology and Molecular Biology Reviews*, *75*, 14–49.

Ng, C., DeMaere, M. Z., Williams, T. J., Lauro, F. M., Raftery, M., Gibson, J. A., et al. (2010). Metaproteogenomic analysis of a dominant green sulfur bacterium from Ace Lake, Antarctica. *The ISME Journal*, *4*, 1002–1019.

Nichols, D., Lewis, K., Orjala, J., Mo, S., Ortenberg, R., O'Connor, P., et al. (2008). Short peptide induces an "uncultivable" microorganism to grow *in vitro*. *Applied and Environmental Microbiology*, *74*, 4889–4897.

Noguchi, H., Park, J., & Takagi, T. (2006). MetaGene: Prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Research*, *34*(19), 5623–5630.

Ocepek, M., Pate, M., Kušar, D., Hubad, B., Avberšek, J., Logar, K., et al. (2011). Comparison of DNA extraction methods to detect *Salmonella* spp. in tap water. *Slovenian Veterinary Research*, *48*, 93–98.

Ogram, A., Sayler, G. S., & Barkay, T. (1987). The extraction and purification of microbial DNA from sediments. *Journal of Microbiological Methods*, 7, 57–66.

Oh, S., Caro-Quintero, A., Tsementzi, D., DeLeon-Rodriguez, N., Luo, C., Poretsky, R., et al. (2011). Metagenomic insights into the evolution, function, and complexity of the planktonic microbial community of lake Lanier, a temperate freshwater ecosystem. *Applied and Environmental Microbiology*, 77, 6000–6011.

Oulas, A., Pavloudi, C., Polymenakou, P., Pavlopoulos, G. A., Papanikolaou, N., Kotoulas, G., et al. (2015). Metagenomics: Tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. *Bioinformatics and Biology Insights*, 75–88.

Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H.-Y., Cohoon, M., et al. (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Research*, *33*(17), 5691–5702.

Pace, N. R. (1997). A molecular view of microbial diversity and the biosphere. *Science*, *276*, 734–740.

Pareek, C. S., Smoczynski, R., & Tretyn, A. (2011). Sequencing technologies and genome sequencing. *Journal of Applied Genetics*, *52*(4), 413–435.

Paszkiewicz, K., & Studholme, D. J. (2010). De novo assembly of short sequence reads. *Briefings in Bioinformatics*, *11*(5), 457–472.

Pati, A., Heath, L. S., Kyrpides, N. C., & Ivanova, N. (2011). ClaMS: A classifier for metagenomic sequences. *Standards in Genomic Sciences*, *5*(2), 248–253.

Patil, K. R., Roune, L., & McHardy, A. C. (2012). The PhyloPythiaS web server for taxonomic assignment of metagenome sequences. *PLoS One*, 7(6), e38581.

Peano, C., Pietrelli, A., Consolandi, C., Rossi, E., Petiti, L., Tagliabue, L., et al. (2013). An efficient rRNA removal method for RNA sequencing in GC-rich bacteria. *Microbial Informatics and Experimentation*, *3*(1), 1.

Pehrsson, E. C., Forsberg, K. J., Gibson, M. K., Ahmadi, S., & Dantas, G. (2013). Novel resistance functions uncovered using functional metagenomic investigations of resistance reservoirs. *Frontiers in Microbiology*, *4*, 145.

Peng, Y., Leung, H. C., Yiu, S. M., & Chin, F. Y. (2012). IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, *28*(11), 1420–1428.

Penn, K., Wang, J., Fernando, S. C., & Thompson, J. R. (2014). Secondary metabolite gene expression and interplay of bacterial functions in a tropical freshwater cyanobacterial bloom. *The ISME Journal*, *8*, 1866–1878.

Perez-Losada, M., Castro-Nallar, E., Bendall, M. L., Freishtat, R. J., & Crandall, K. A. (2015). Dual transcriptomic profiling of host and microbiota during health and disease in pediatric asthma. *PLoS One*, *10*, e0131819.

Pester, M., Rattei, T., Flechl, S., Grongroft, A., Richter, A., Overmann, J., et al. (2012). amoA-based consensus phylogeny of ammonia-oxidizing archaea and deep sequencing of amoA genes from soils of four different geographic regions. *Environmental Microbiology*, *14*, 525–539.

Pevzner, P. A., Tang, H., & Waterman, M. S. (2001). A Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences of the United States of America*, *98*, 9748–9753.

Pillai, S., Gopalan, V., & Lam, A. K. (2017). Review of sequencing platforms and their applications in phaeochromocytoma and paragangliomas. *Critical Reviews In Oncology-Hematology*, *116*, 58–67.

Pitcher, D. G., Saunders, N. A., & Owen, R. J. (1989). Rapid extraction of bacterial genomic DNA with guanidium thiocyanate. *Letters in Applied Microbiology*, *8*, 151–156.

Poretsky, R., Rodriguez-R, L. M., Luo, C., Tsementzi, D., & Konstantinidis, K. T. (2014). Strengths and Limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial community dynamics. *PLoS One*, *9*(4), e93827.

Powell, S., Forslund, K., Szklarczyk, D., Trachana, K., Roth, A., Huerta-Cepas, J., et al. (2014). EggNOG v4.0: Nested orthology inference across 3686 organisms. *Nucleic Acids Research*, *42*(Database issue), D231–D239.

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., et al. (2012). A tale of three next generation sequencing platforms: Comparison of ion torrent, pacific Biosciences and illumina MiSeq sequencers. *BMC Genomics*, *13*, 341.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2013). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, *41*(Database issue), D590–D596.

Quince, C., Lanzen, A., & Curtis, T. P. (2009). Accurate determination of microbial diversity from 454 pyrosequencing data. *Nature Methods*, *6*, 639–641.

Quince, C., Lanzen, A., Davenport, R. J., & Turnbaugh, P. J. (2011). Removing noise from pyrosequenced amplicons. *BMC Bioinformatics*, *12*, 38.

Rappe, M. S., & Giovannoni, S. J. (2003). The uncultured microbial majority. *Annual Review of Microbiology*, *57*, 369–394.

Rastogi, G., & Sani, R. K. (2011). Molecular techniques to assess microbial community structure, function, and dynamics in the environment. In Ahmad, et al. (Ed.), *Microbes and microbial technology: Agricultural and environmental applications* (pp. 29–58). Springer Science + Business Media, LLC.

Reuter, J. A., Spacek, D., & Snyder, M. P. (2015). High-throughput sequencing technologies. *Molecular Cell*, *58*(4), 586–597.

Rho, M., Tang, H., & Ye, Y. (2010). FragGeneScan: Predicting genes in short and error-prone reads. *Nucleic Acids Research*, *38*(20), e191.

Riesenfeld, C. S., Goodman, R. M., & Handelsman, J. (2004). Uncultured soil bacteria are a reservoir of new antibiotic resistance genes. *Environmental Microbiology*, *6*, 981–989.

Rivera, I. N. G., Lipp, E. K., Gil, A., Choopun, N., Huq, A., & Colwell, R. R. (2003). Method of DNA extraction and application of multiplex polymerase chain reaction to detect toxigenic *Vibrio cholera* O1 and O139 from aquatic ecosystems. *Environmental Microbiology*, *5*, 599–606.

Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., et al. (2010). De novo assembly and analysis of RNA-seq data. *Nature Methods*, 7(11), 909–912.

Rohwer, F., & Edwards, R. (2002). The phage proteomic tree: A genome-based taxonomy for phage. *Journal of Bacteriology*, *184*, 4529–4535.

Romero, J. L., Grande Burgos, M. J., Pérez-Pulido, R., Gálvez, A., & Lucas, R. (2017). Resistance to antibiotics, biocides, preservatives and metals in bacteria isolated fromsea-foods: Co-selection of strains resistant or tolerant to different classes of compounds. *Frontiers in Microbiology*, *8*, e1650.

Roose-Amsaleg, C., & Laverman, A. M. (2016). Do antibiotics have environmental side-effects? Impact of synthetic antibiotics on biogeochemical processes. *Environmental Science and Pollution Research*, *23*, 4000–4012.

Roumpeka, D. D., Wallace, R. J., Escalettes, F., Fotheringham, I., & Watson, M. (2017). A review of bioinformatics tools for bio-Prospecting from metagenomic sequence data. *Frontiers in Genetics*, *8*, 23.

Roux, S., Enault, F., Robin, A., Ravet, V., Personnic, S., Theil, S., et al. (2012). Assessing the diversity and specificity of two freshwater viral communities through metagenomics. *PLoS One*, 7, e33641.

Rusch, D. B., Halpern, A. L., Sutton, G., Heidelberg, K. B., Williamson, S., Yooseph, S., et al. (2007). The sorcerer II global ocean sampling expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biology*, *5*, e77.

Salam, L. B., Ilori, M. O., Amund, O. O., LiiMien, Y., & Nojiri, H. (2018). Characterization of bacterial community structure in a hydrocarbon-contaminated tropical African soil. *Environmental Technology*, *39*(7), 939–951.

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74, 5463–5467.

Santas, A. J., Persaud, T., Wolfe, B. A., & Bauman, J. M. (2013). Non-invasive method for a state-wide survey of eastern hellbenders *Cryptobrachus alleganiensis* using environmental DNA. *International Journal of Zoology*, *2013*(1), 1–6.

Sato, K., & Sakakibara, Y. (2015). MetaVelvet-sl: An extension of the Velvet assembler to a de novo metagenomic assembler utilizing supervised learning. *DNA Research*, *22*, 69–77.

Schadt, E. E., Turner, S., & Kasarskis, A. (2010). A window into third-generation sequencing. *Human Molecular Genetics*, *19*, 227–240.

Schloss, P. D. (2010). The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. *PLoS Computational Biology*, *6*, e10000844.

Schloss, P. D., Gevers, D., & Westcott, S. L. (2011). Reducing the effects of PCR amplification and sequencing artefacts on 16S rRNA-based studies. *PLoS One*, *6*, e27310.

Schloss, P. D., & Handelsman, J. (2005). Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Applied and Environmental Microbiology*, *71*(3), 1501–1506.

Schloss, P. D., & Handelsman, J. (2006a). Introducing SONS, a tool for operational taxonomic unit-based comparisons of microbial community memberships and structures. *Applied and Environmental Microbiology*, *72*(10), 6773–6779.

Schloss, P., & Handelsman, J. (2006b). Introducing TreeClimber, a test to compare microbial community structures. *Applied and Environmental Microbiology*, *72*(4), 2379 2379.

Schulz, M. H., Zerbino, D. R., Vingron, M., & Birney, E. (2012). Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, *28*(8), 1086–1092.

Seaman, T. (2014). Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068–2069.

Seiler, C., & Berendonk, T. (2012). Heavy metal driven co-selection of antibiotic resistance in soil and water bodies impacted by agriculture and aquaculture. *Frontiers in Microbiology*, *3*, e399.

Shade, A., Read, J. S., Welkie, D. G., Kratz, T. K., Wu, C. H., & McMahon, K. D. (2011). Resistance, resilience, recovery: Aquatic bacterial dynamics after water column disturbance. *Environmental Microbiology*, *13*, 2752–2767.

Shade, A., Read, J. S., Youngblut, N. D., Fierer, N., Knight, R., Kratz, T. K., et al. (2012). Lake microbial communities area resilient after a whole-ecosystem disturbance. *The ISME Journal*, *6*, 2153–2167.

Sharpton, T. J. (2014). An introduction to the analysis of shotgun metagenomic data. *Frontiers of Plant Science*, *5*, 209.

Sigee, D. C. (2005). *Freshwater microbiology: Biodiversity and dynamic interactions of microorganisms in the freshwater environment*. West Sussex PO19 8SQ, England: John Wiley & Sons Ltd.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J., & Birol, I. (2009). ABySS: A parallel assembler for short read sequence data. *Genome Research*, *19*(6), 1117–1123.

Skvortsov, T., de Leeuwe, C., Quinn, J. P., McGrath, J. W., Allen, C. C. R., McElarney, Y., et al. (2016). Metagenomic characterization of the viral community of Lough Neagh, the largest freshwater lake in Ireland. *PLoS One*, *11*(2), e0150361.

Smalla, K., Cresswell, N., Mendonca-Hagler, L. C., Wolters, A., & van Elsas, J. D. (1993). Rapid DNA extraction protocol from soil for polymerase chain reaction-mediated amplification. *Journal of Applied Bacteriology*, *74*, 78–85.

Soergel, D. A. W., Dey, N., Knight, R., & Brenner, S. E. (2012). Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences. *The ISME Journal*, *6*, 1440–1444.

Sogin, M. L., Morrison, H. G., Huber, J. A., Welch, D. M., Huse, S. M., Neal, P. R., et al. (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences of the United States of America*, *103*, 12115–12120.

Sommer, M. O. A., Dantas, G., & Church, G. M. (2009). Functional characterization of the antibiotic resistance reservoir in the human microflora. *Science*, *325*, 1128–1131.

Soo, R. M., Wood, S. A., Grzymski, J. J., McDonald, I. R., & Cary, S. C. (2009). Microbial biodiversity of thermophilic communities in hot mineral soils of Tramway Ridge, Mount Erebus, Antarctica. *Environmental Microbiology*, *11*, 715–728.

Sorek, R., & Cossart, P. (2010). Prokaryotic transcriptomics: A new view on regulation, physiology and pathogenicity. *Nature Reviews Genetics*, *11*, 9–16.

Sorek, R., Zhu, Y. W., Creevey, C. J., Francino, M. P., Bork, P., & Rubin, E. M. (2007). Genome-wide experimental determination of barriers to horizontal gene transfer. *Science*, *318*, 1449–1452.

Stahl, D. A., Flesher, B., Mansfield, H. R., & Montgomery, L. (1988). Use of phylogenetically based hybridization probes for studies of ruminal microbial ecology. *Applied and Environmental Microbiology*, *54*, 1079–1084.

Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., et al. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, *353*, 78–82.

Staley, C., Gould, T. J., Wang, P., Phillips, J., Cotner, J. B., & Sadowsky, M. J. (2014). Core functional traits of bacterial communities in the Upper Mississippi River show limited variation in response to land cover. *Frontiers in Microbiology*, *5*, 414.

Steffen, M. M., Dearth, S. P., Dill, B. D., Li, Z., Larsen, K. M., Campagna, S. R., et al. (2014). Nutrients drive transcriptional changes that maintain metabolic homeostasis but alter genome architecture in Microcystis. *The ISME Journal*, *8*, 2080–2092.

Steffen, M. M., Li, Z., Effler, T. C., Hauser, L. J., Boyer, G. L., & Wilhelm, S. W. (2012). Comparative metagenomics of toxic freshwater cyanobacteria bloom communities on two continents. *PLoS One*, 7, e44002.

Steffen, M. M., Belisle, B. S., Watson, S. B., Boyer, G. L., Bourbonniere, R. A., & Wilhelm, S. W. (2015). Metatranscriptomic evidence for co-occuring top-down and bottom-up controls on toxic cyanobacterial communities. *Applied and Environmental Microbiology*, *81*, 3268–3276.

Steven, B., McCann, S., & Ward, N. L. (2012). Pyrosequencing of plastid 23S rRNA genes reveals diverse and dynamic cyanobacterial and algal populations in two eutrophic lakes. *FEMS Microbiology Ecology*, *82*, 607–615.

Stokes, H. W., & Gillings, M. R. (2011). Gene flow, mobile genetic elements and the recruitment of antibiotic resistance genes into Gram-negative pathogens. *FEMS Microbiology Reviews*, *35*, 790–819.

Streit, W. R., & Schmitz, R. A. (2004). Metagenomics- key to the uncultured microbes. *Current Opinion in Microbiology*, 7, 492–498.

Su, Z., Ning, B., Fang, H., Hong, H., Perkins, R., Tong, W., et al. (2011). Next-generation sequencing and its applications in molecular diagnostics. *Expert Review of Molecular Diagnostics*, *11*, 333–343.

Su, X., Xu, J., & Ning, K. (2012). Parallel-meta: Efficient metagenomic data analysis based on high-performance computation. *BMC Systems Biology*, *6*(Suppl. 1), S16.

Tanaka, T., Sakai, R., Kobayashi, R., Hatakeyama, K., & Matsunaga, T. (2009). Contributions of phosphate to DNA adsorption/desorption behaviors on aminosilane-modified magnetic nanoparticles. *Langmuir*, *25*, 2956–2961.

Tan, B. F., Ng, C., Nshimyimana, J. P., Loh, L. L., Gin, K. Y. -H., & Thompson, J. R. (2015). Next-generation sequencing (NGS) for assessment of microbial water quality: Current progress, challenges, and future opportunities. *Frontiers in Microbiology*, *6*, 1027.

Tatusov, R. L., Galperin, M. Y., Natale, D. A., & Koonin, E. V. (2000). The COG database: A tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research*, *28*(1), 33–36.

Teeling, H., Waldmann, J., Lombardot, T., Bauer, M., & Glockner, F. O. (2004). TETRA: A web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics*, *5*, 163.

Thomas, T., Gilbert, J., & Meyer, F. (2012). Metagenomics – a guide from sampling to data analysis. *Microbial Informatics and Experimentation*, *2*(1), 3.

Thomsen, P. F., Kielgast, J., Iversen, L. L., Wiuf, C., Rasmussen, M., Gilbert, M. T., et al. (2012). Monitoring endangered freshwater biodiversity using environmental DNA. *Molecular Ecology*, *21*, 2565–2573.

Tranvik, L., Downing, J. A., Cotner, J. B., Loiselle, S. A., Striegl, R. G., Ballatore, T. J., et al. (2009). Lakes and reservoirs as regulators of carbon cycling and climate. *Limnology and Oceanography*, *54*, 2298–2314.

Treangen, T. J., Koren, S., Sommer, D. D., Liu, B., Astrovskaya, I., Ondov, B., et al. (2013). MetAMOS: A modular and open source metagenomic assembly and analysis pipeline. *Genome Biology*, *14*(1), R2.

Tremblay, J., Singh, K., Fern, A., Kirton, E. S., He, S., Woyke, T., et al. (2015). Primer and platform effects on 16S rRNA tag sequencing. *Frontiers in Microbiology*, *6*, 771.

Tringe, S. G., von Mering, C., Kobayashi, A., Salamov, A. A., Chen, K., Chang, H. W., et al. (2005). Comparative metagenomics of microbial communities. *Science*, *308*, 554–557.

Tsai, Y. L., & Olsen, B. H. (1991). Rapid method for direct extraction of DNA from soil and sediments. *Applied and Environmental Microbiology*, *57*, 1070–1074.

Tseng, C. H., Chiang, P. W., Shiah, F. K., Chen, Y. L., Liou, J. R., Hsu, T. C., et al. (2013). Microbial and viral metagenomes of a subtropical freshwater reservoir subject to climatic disturbances. *The ISME Journal*, 7, 2374–2386.

Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., et al. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, *428*, 37–43.

Uyaguari-Diaz, M. I., Chan, M., Chaban, B. L., Croxen, M. A., Finke, J. K., Hill, J. E., et al. (2016). A comprehensive method for amplicon-based and metagenomic characterization of viruses, bacteria, and eukaryotes in freshwater samples. *Microbiome*, *4*, 20.

Välitalo, P., Kruglova, A., Mikola, A., & Vahala, R. (2017). Toxicological impacts of antibiotics-son aquatic micro-organisms: A mini-review. *International Journal of Hygiene and Environmental Health*, *220*, 558–569.

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., et al. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, *304*, 66–74.

Vorosmarty, C. J., McIntyre, P. B., Gessner, M. O., Dudgeon, D., Prusevich, A., Green, P., et al. (2010). Global threats to human water security and river biodiversity. *Nature*, *467*(7315), 555–561.

Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). A native Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, *73*, 5261–5267.

Wang, Y., Ran, S., Man, B., & Yang, G. (2011). Ethanol induces condensation of single DNA molecules. *Soft Matter*, *7*, 4425–4434.

Wang, J., Wang, J., Zhao, Z., Chen, J., Lu, H., Liu, G., et al. (2017). PAHs accelerate the propagation of antibiotic resistance genes in coastal water microbial community. *Environmental Pollution*, *231*, 1145–1152.

Weinstock, G. M. (2012). Genomic approaches to studying the human microbiota. *Nature*, *489*, 250–256.

Wilcox, T. M., McKelvey, K. S., Young, M. K., Sepulveda, A. J., Shepard, B. B., Jane, S. F., et al. (2016). Understanding environmental DNA detection probabilities: A case study using a stream dwelling char *Salvelinus fontinalis*. *Biological Conservation*, *194*, 209–216.

Wright, J. J., Lee, S., Zaikova, E., Walsh, D. A., & Hallam, S. J. (2009). DNA extraction from 0.22 μM sterivex filters and cesium chloride density gradient centrifugation. *Journal of Visualized Experiments*, 1352.

Wurzbacher, C., Fuchs, A., Attermeyer, K., Frindte, K., Grossart, H. P., Hupfer, M., et al. (2017). Shifts among eukaryote, bacteria, and archaea define the vertical organization of a lake sediment. *Microbiome*, *5*(1), 41.

Wu, C. H., Sercu, B., Van de Werfhorst, L. C., Wong, J., DeSantis, T. Z., Brodie, E. L., et al. (2010). Characterization of coastal urban watershed bacterial communities leads to alternative community-based indicators. *PLoS One*, *5*(6), e11285.

Wu, Y., Yu, C. P., Yue, M., Liu, S. P., & Yang, X. Y. (2016). Occurrence of selected PPCPs and sulfonamide resistance genes associated with heavy metals pollution in surface sediments from Chao Lake, China. *Environmental Earth Sciences*, *75*, 43.

Wylie, K. M., Truty, R. M., Sharpton, T. J., Mihindukulasuriya, K. A., Zhou, Y., Gao, H., et al. (2012). Novel bacterial taxa in the human microbiome. *PLoS One*, *7*, e35294.

Xia, H., Li, T., Deng, F., & Hu, Z. (2013). Freshwater cyanophages. *Virologica Sinica*, *28*, 253–259.

Xiong, X., Frank, D. N., Robertson, C. E., Hung, S. S., Markle, J., Canty, A. J., et al. (2012). Generation and analysis of a mouse intestinal metatranscriptome through illumina based RNA-sequencing. *PLoS One*, *7*(4), e36009.

Yamanaka, H., Motozawa, H., Tsuji, S., Miyazawa, R. C., Takahara, T., & Minamoto, T. (2016). On-site filtration of water samples for environmental DNA analysis to avoid DNA degradation during transportation. *Ecological Research*, *31*, 963–967.

Yang, Y., Liu, W., Xu, C., Wei, B., & Wang, J. (2017). Antibiotic resistance genes in lakes from middle and lower reaches of the yangtze river, China: Effect of land use and sediment characteristics. *Chemosphere*, *178*, 19–25.

Yang, Y., & Wang, W. (2018). Benzyldimethyldodecyl ammonium chloride shifts the proliferation of functional genes and microbial community in natural water from eutrophic lake. *Environmental Pollution*, *236*, 355–365.

Yang, Y., Xu, C., Cao, X., Lin, H., & Wang, J. (2017). Antibiotic resistance genes in surface water of eutrophic urban lakes are related to heavy metals, antibiotics, lake morphology and anthropic impact. *Ecotoxicology*, *26*, 831–840.

Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominquez-Bello, M. G., Contreras, M., et al. (2012). Human gut microbiome viewed across age and geography. *Nature*, *486*, 222–227.

Yilmaz, P., Kottmann, R., Field, D., Knight, R., Cole, J. R., Amaral-Zettler, L., et al. (2011). Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nature Biotechnology*, *29*(5), 415–420.

Yin, Q., Yue, D. M., Peng, Y. K., Liu, Y., & Xiao, L. (2013). Occurrence and distribution of antibiotic-resistant bacteria and transfer of resistance genes in Lake Taihu. *Microbes and Environments*, *28*, 479–486.

Yost, S., Duran-Pinedo, A. E., Teles, R., Krishnan, K., & Frias-Lopez, J. (2015). Functional signatures of oral dysbiosis during periodontitis progression revealed by microbial metatranscriptome analysis. *Genome Medicine*, 7(1), 27.

Youssef, N., Sheik, C. S., Krumholz, L. R., Najar, F. Z., Roe, B. A., Elshahed, M. S., et al. (2009). Comparison of species richness estimates obtained using nearly complete fragments and simulated pyrosequencing-generated fragments in 16S rRNA gene-based environmental surveys. *Applied and Environmental Microbiology*, *75*, 5227–5236.

Zerbino, D. R., & Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, *18*(5), 821–829.

Zhang, W., Sturm, B. S. M., Knapp, C. W., & Graham, D. W. (2009). Accumulation of tetracycline resistance genes in aquatic biofilms due to periodic waste loadings from swine lagoons. *Environmental Science and Technology*, *43*, 7643–7650.

Zhang, S. H., Xu, W. T., & Shi, H. (2012). Comparison of the extractions of DNA and study on the biotechnological detection of *Pseudomonas aeruginosa* in river. *Science and Technology of Food Industry*, *33*, 375–379.

Zhao, X., Yang, L., Yu, Z., Peng, N., Xiao, L., Yin, D., et al. (2008). Characterization of depth-related microbial communities in lake sediment by denaturing gradient gel electrophoresis of amplified 16S rRNA fragments. *Journal of Environmental Sciences*, *20*, 224–230.

Zheng, H., & Wu, H. (2010). Short prokaryotic DNA fragment binning using a hierarchical classifier based on linear discriminant analysis and principal component analysis. *Journal of Bioinformatics and Computational Biology*, *8*(6), 995–1011.

Zhou, J., Bruns, M. A., & Tiedje, J. M. (1996). DNA recovery from soils of diverse composition. *Applied and Environmental Microbiology*, *62*, 316–322.

Zhou, J., He, Z., Yang, Y., Deng, Y., Tringe, S. G., & Alvarez-Cohen, L. (2015). High throughput metagenomic technologies for complex microbial community analysis: Open and closed formats. *mBio*, *6*(1), e02288 14.

Zhou, Z. C., Zheng, J., Wei, Y. Y., Chen, T., Dahlgren, R. A., Shang, X., et al. (2017). Antibiotic resistance genes in an urban river as impacted by bacterial community and physicochemical parameters. *Environmental Science and Pollution Research*, *24*, 23753–23762.

Zhu, Y. G., Johnson, T. A., Su, J. Q., Qiao, M., Guo, G. X., Stedtfeld, R. D., et al. (2013). Diverse and abundant antibiotic resistance genes in Chinese swine farms. *Proceedings of the National Academy of Sciences*, *110*, 3435–3440.

Zhu, W., Lomsadze, A., & Borodovsky, M. (2010). Ab initio gene identification in metagenomic sequences. *Nucleic Acids Research*, *38*(12), e132.