Mini Review: Metagenomics as a tool to monitor reclaimed water quality

Pei-Ying Hong[1*], David Mantilla-Calderon[1], Changzhi Wang[1]

1. Water Desalination and Reuse Center, Division of Biological and Environmental Science and Engineering, King Abdullah University of Science and Technology (KAUST), 23955-6900 Thuwal, Saudi Arabia

*** Corresponding author**

Email: peiying.hong@kaust.edu.sa

Telephone: +966 12 8082218

17 **Abstract**

18 Many biological contaminants are disseminated through water, and their occurrence has potential
19 detrimental impacts on public and environmental health. Conventional monitoring tools rely on cultivation
20 and are not robust in addressing modern water quality concerns. This review proposes metagenomics as
21 a means to provide a rapid, nontargeted assessment of biological contaminants in water. When further
22 coupled with the appropriate methods (e.g., quantitative PCR and flow cytometry) and bioinformatic tools,
23 metagenomics can provide information concerning both the abundance and diversity of biological
24 contaminants in reclaimed waters. Further correlation between the metagenomic-derived data of selected
25 contaminants and the measurable parameters of water quality can also aid in devising strategies to
26 alleviate undesirable water quality. Here, we reviewed metagenomic approaches (i.e., both sequencing
27 platforms and bioinformatic tools) and studies that demonstrated their use for reclaimed water quality
28 monitoring. We also provide recommendations on areas of improvement that will allow metagenomics to
29 significantly impact how the water industry performs reclaimed water quality monitoring in the future.

30

Introduction

Water scarcity in the Middle East and North African regions, as well as in countries such as Singapore, Australia and Maldives, has necessitated the use of reclaimed water to alleviate the depletion of nonrenewable freshwater supplies. Reclaimed water is increasingly used in landscape irrigation to maintain green living spaces and for agricultural irrigation to produce food. Reclaimed water is also injected into aquifers to replenish depleting groundwater and used as energy exchange medium in cooling towers. In some places, reclaimed water further undergoes advanced treatment processes, typically involving reverse osmosis membrane filtration, to become a potable water source. Depending on the intended reuse purpose, different wastewater treatment technologies are used to provide the reclaimed water with quality that abides by either the World Health Organization (WHO) guidelines or standards inspired by United States Environmental Protection Agency (USEPA) and International Organization for Standardization (ISO).

Current regulations stipulated by WHO, USEPA and ISO only require the enumeration of fecal indicators (e.g., total and fecal coliforms) to indicate reclaimed water quality. The standard methods used to determine these fecal indicators can be prone to false-negatives if viable bacteria are stressed or injured. Culture-based methods also require time (typically 24 h to 48 h) for the microbial targets to grow to levels that facilitate enumeration. This process impedes the ability for a rapid response. Furthermore, fecal indicators do not occur at frequencies that correlate well with waterborne pathogens in reclaimed water (1); hence, they cannot predict accurately the presence of pathogens. Considering these limitations, standard methods have become increasingly obsolete in addressing modern water quality concerns, especially because emerging contaminants are found in reclaimed waters intended for agriculture and landscape irrigation and can potentially affect public health. These contaminants include bacterial pathogens (particularly those related to antibiotic-resistant ones), viral pathogens, protozoal hosts for intracellular pathogens, and extracellular DNA (e.g., antibiotic resistance genes) (2, 3). Many of these pathogens are fastidious, slow growing and hard to culture for routine monitoring.

Besides culture-based approaches, molecular methods such as quantitative PCR (qPCR) can determine the presence of pathogens or antibiotic resistance genes. However, qPCR is a targeted

approach that only detects the marker genes that hybridize to the designed primers or probes. Hence, this targeted approach would not provide insights into unknown gene targets that do not have any available primer sets. Given the wide spectrum of contaminants that are present, nontargeted methods that can provide information on both the phylogenetic and functional diversities of emerging contaminants simultaneously would be preferred. Additionally, the method should preferably provide quantitative estimates of those targets-of-interest to facilitate evidence-based decision making. Some of the key questions to be asked when evaluating reclaimed water quality include the following: is the wastewater treatment system functioning well to provide reclaimed water that meets the required quality? Is the reclaimed water biologically stable and would it not change much in its quality along the distribution network? Are contaminants present in the reclaimed water that would affect the environment and consumers' health at the point-of-use? Can we infer the presence of nutrients or chemical contaminants in the reclaimed water based on the presence of some of the microbial contaminants?

In this mini-review, we argue that metagenomics is suitable to address the abovementioned questions, hence facilitating reclaimed water quality monitoring. We derived this proposition based on the following evidence gathered from the current literature: **(i)** advances in sequencing technologies have rapidly decreased the associated costs while increasing the number of raw reads available; **(ii)** the availability of bioinformatic tools to facilitate the analysis of metagenomic data, allowing the collection of massive datasets that reveal the gene and functional diversities in a nontargeted manner; and **(iii)** continuous improvement of both sequencing technologies and analytical tools are shortening the time required to perform metagenomics and analysis. However, its ability to provide quantitative measurements, good accuracy and fine resolution of phylogenetic and functional classification in mixed community samples will need to be further improved to fully address the needs of reclaimed water quality monitoring.

Definition of metagenomics

Metagenomics, being DNA based, can only provide information on who is there (i.e., taxonomic and phylogenetic information) and what is there (i.e., functional gene diversity). Depending on the type of microbial target (i.e., virus, bacteria or protozoa), different sample preparation measures and extraction protocols would have to be used to maximize the yield of DNA from these microorganisms before metagenomics. However, because the microbial populations and genes detected by metagenomics are derived from DNA, they may be from nonviable cells or genes that are not being expressed. This approach contrasts with metatranscriptomics (RNA-based sequencing) or metaproteomics (peptide sequencing), which provide information on which microbial populations are alive and actively transcribing their genes and translating the mRNA to proteins. Metagenomics should not be confused with amplicon-based high-throughput sequencing that typically involves only a targeted gene (e.g., 16S rRNA or 18S rRNA genes) (4-6). Metagenomics should not be confused with whole genome sequencing, which refers to single-genome sequencing. The number of papers related to the keywords “metagenomics” and various types of water matrices demonstrates that metagenomics is more widely utilized for surface waters than for reclaimed water (Figure 1). However, the number of papers related to the use of metagenomics also experienced a fast rate of increment, particularly from the year 2013 onwards with the advent and accessibility of sequencing technologies.

Functional metagenomics

Before metagenomics became more mainstream, the earlier approach involved extracting a large amount (e.g., > 10 µg) of high-molecular-weight DNA from a sample (7), creating DNA fragments using endonucleases and then ligating these DNA fragments into artificial chromosome vectors. The size of these DNA fragments can vary from a few kilobases to as long as more than ten kilobases, depending on the fragment size that can be efficiently inserted into the vector. For instance, phage vectors accept inserts of approximately 15 to 20 kb while that of bacterial artificial chromosomes can range from 150 to 350 kbp (8, 9). After gene insertion, the vectors are transformed into *Escerichia coli*, and individual transformants are expressed and screened for the intended functional traits. Transformants that expressed the intended functional traits are then sequenced to denote the inserted gene identities.

Alternatively, all the transformants can be pooled together and sequenced directly without any prescreening. The depth of information derived from this approach of functional metagenomics is limited by the number of transformants picked for screening and sequencing, but this limitation can be easily resolved using an automated colony picker. However, because it involves cloning and incubating cells before sequencing, this approach is subjected to additional bias during cloning and takes a longer time for completion. Due to the amount of time and effort required, functional metagenomics does not facilitate efficient decision making; hence, it has not been widely used for reclaimed water quality monitoring.

However, the advantage of functional metagenomics is that inserted genes can express their enzymes, and subsequent biochemical characterization of those enzymes may lead to useful products. For example, Song et al. extracted high-molecular-weight DNA from the contents of the rumen, and fragmented these DNAs to sizes ranging from 10 to 50 kbp before creating fosmid libraries. The clones were screened for cellulolytic activity, and those with positive cellulolytic activity were pooled together for DNA extraction and sequencing (10). Further gene annotation revealed a novel glycosyl hydrolase family 5 cellulase gene with endo-β-1,4-glucanase. Although not demonstrated in the study because of its potential application, this approach can potentially result in enzymes that can be applied to disrupt undesirable biofilms (11, 12).

Current sequencing platforms for modern metagenomic approaches

In daily routine monitoring of reclaimed water quality, utilities may assess the biological stability of their reclaimed water. Biological stability is defined as the steady-state concentration of bacterial cells and composition in the water (13). A sudden increase in the concentration of bacterial cells would infer either the growth of microorganisms, influx of microbial contaminants or a failing distribution network, which might detrimentally impact operations and safety at the point-of-use. In addition to monitoring for biologically stable reclaimed water, utilities may also be interested in determining the performances of their treatment processes by tracking log removal values. This can be done by enumerating the concentration of contaminants before and after treatment. Furthermore, to determine the risks associated with pathogens in our reclaimed water supplies, quantitative estimates of pathogens are needed to

facilitate microbial risk assessments. These questions require a timely response and modern metagenomic approaches (also referred to as shotgun sequencing), bypassing the need for cloning and cultivation and showing promise to address these questions.

A succession of sequencing platforms is available, from the now defunct 454 pyrosequencing and Ion Torrent to the current mainstream Illumina, as well as Nanopore and PacBio, which can generate longer reads than Illumina reading chemistries depending on the quality and fragment size of the DNA template. Regardless of the sequencing platform, the main distinguishing feature is the ability to generate a large number of short-length reads (typically 100 to 300 bp per read) per run (Table 1) at costs typically ranging from 1 to 3 thousand dollars per run. Most of these sequencing platforms require significantly lower concentrations of DNA (typically 10 ng to 1 µg of DNA) than the clone-based functional metagenomic approach. The DNA amount required is low because modern sequencing platforms rely on solid phase or emulsion-based PCR to exponentially amplify the gene molecules so that the detection sensitivity can be enhanced. However, this can also introduce amplification bias incurred during PCR (14) and sequencing errors due to low-fidelity polymerase (15). Shotgun sequencing also does not require DNA to be of high molecular weight because the library preparation steps require DNA to be fragmented to approximately 400 bp before ligating with the index adaptors. However, overly fragmenting DNA will also impair the sequencing quality by generating reads with lengths shorter than the norm. Therefore, the optimization of protocols is required to minimize associated error rates and lapses in sequencing quality.

Availability of bioinformatic tools: Genome-centric approach

Sequencing results can be analyzed using either a genome-centric or gene-centric approach. A genome-centric approach relies on assembling the short-length reads into contigs or scaffolds (larger genomic fragments), and further assembling the contigs or scaffolds into draft or complete genomes. Assembly can be performed with supervision, whereby reads are aligned against reference genomes based on sequence similarity. Homologous regions of the individual raw reads are also matched and linked together to form contigs in a de novo manner and then are aligned against reference genomes. Alternatively, the assembly can be performed using an unsupervised approach that relies on

discriminative sequence composition and/or co-abundance of reads (16). The unsupervised approach groups contigs into bin clusters that are further differentiated based on the sequencing coverage. Contigs associated with a particular bin cluster can be retrieved for further de novo assembly to form draft population genomes. Several programs, including MetaBat (17), Concoct (18), and MaxBin (19), facilitate the reconstruction of microbial genomes from a metagenomic dataset. The quality of the genome bins is further assessed using CheckM (20) to derive the percentage of completeness and contamination level. For example, most draft genomes obtained via the unsupervised approach are classified to be of acceptable quality based on a substantial level of completeness (≥ 70%) and low level of contamination (≤ 5%) (20, 21).

A genome-centric approach can potentially be used to identify the presence of pathogens in reclaimed water although not without challenge. Assuming typical reclaimed water may have up to 2000 unique species with an average genome size of 4 Mbp (22), each in equal relative abundance, 8 Gbp of reads would have to be obtained per sample to achieve a 1× sequencing coverage of all genomes in this sample. An ecosystem with an equal distribution of species is unlikely, and a higher likelihood of assembling a genome usually applies to microbial cells that are predominant and, hence, overrepresented in terms of sequencing reads. This phenomenon does not consider that the current sequencing platforms require PCR to amplify gene targets before sequencing, thereby incurring a selective bias against those with a GC-rich genome (and hence achieving a lower sequencing coverage). In most instances, trying to identify a unique genome confidently requires more than 5× sequencing coverage (23). An even higher coverage is needed to discern the genomes arising from multiple pathogenic strains of the same species that may coexist in the same mixed microbial consortium (24). Considering the current throughput reported by Illumina NovaSeq 6000, this would require at least 1 lane in an S2 flowcell per sample to achieve the needed coverage (Table 1). Therefore, it is more likely to obtain only draft genomes from metagenomic data. Draft genome databases are growing rapidly, and any new microbiological resource deposited in a repository available to the community is announced frequently online in the full open-access journal *Microbiology Resource Announcements* published by the American Society for Microbiology. However, many of the draft genomes are contaminated with fragments of sequences from other species (25), and validation of these contigs and draft genomes

remains a key essential step (26). However, no good validation approach exists that can assess the accuracy in the metagenomic assembly unless a pure culture of that microbial target can be isolated and propagated and whole-genome sequencing is performed followed by verification against the data derived from metagenomics.

Assembly of metagenomic data would be more useful to elucidate dominant species present in reclaimed water—for example, nitrifying bacteria or heterotrophs that correlate with the nutrient content of the water—because they are more likely to show a higher sequencing coverage and, hence, more confident assembly results. However, dominant taxa can be elucidated rapidly using amplicon-based sequencing and may not require the use of metagenomics unless functional annotation is required. Although it is assumed that metagenomics may achieve a better resolution and accuracy in taxonomic classifications because more genes associated with the microbial target can be evaluated simultaneously, a recent study suggested the contrary. Tessler et al. analyzed 49 samples from a floodplain system using both 16S rRNA gene-based amplicon and shotgun sequencing (27). The authors demonstrated that amplicon sequencing could assign more reads at the phyla and family levels and could be relatively more robust across both biodiversity and community ecology analyses than metagenomics. This observation can be explained by the possibility that taxonomic resolution derived from metagenomics is detrimentally impacted by the coverage and size of whole-genome databases because, in instances where whole genomes of target species are absent, many of the reads obtained from shotgun sequencing would be mapped as unknown (27). This error would inherently reduce the number of taxonomically applicable reads. Furthermore, horizontal gene transfer is a ubiquitous and rampant phenomenon in microbial ecosystems (28, 29). Because shotgun sequencing assigns taxonomic classifications based on genes across the entire genome, regardless of whether they are core genes, this can lead to incidences of contradictory and inaccurate identifications if those assigned genes were instead horizontally transferred from another microbial species. By contrast, amplicon sequencing only considers one type of gene at a time and, by choosing a core gene (e.g., the 16S rRNA gene), which is rarely transferred horizontally (30) to be sequenced, taxonomical classifications can be assigned more accurately than metagenomics.

Applied and Environmental Microbiology

Availability of bioinformatics tools: Gene-centric approach

Considering the limitations of the genome-centric approach, the alternative gene-centric approach can be used to analyze metagenomic data derived for reclaimed water quality monitoring. For this approach, the raw reads are input into classifier or profiler programs to map both the phylogenetic and functional profiles of the sample data. For example, an interactive toolbox such as MEGAN (31) provides taxonomic analysis by mapping reads against the NCBI or Silva database. MEGAN also provides functional analysis using various protein databases (e.g., SEED and KEGG). Free public resources such as MG-RAST (32) provide taxonomic and functional analyses similar to MEGAN. Additionally, it serves as a public depository for metagenomic data where users interested in meta-data analysis can download open access metagenomic datasets for further data-mining. Despite its ease of use, functional analysis on both MG-RAST and MEGAN tend to provide only the classification of proteins at the functional class level (e.g., proteins related to biosynthesis, degradation, folding, processing and modification) (33) and does not facilitate downstream scientific inquiry on the annotated genes that are related to each of these functional classes.

In addition to MEGAN and MG-RAST, in recent years, an increasing number of classification tools (Table 2) have been developed (34). However, the databases associated with each classification method may differ. Some classifiers match DNA sequences obtained from metagenomics to DNA databases, while others match DNA sequences to protein or marker gene databases. To exemplify, the common databases include Silva, Ribosomal database project (RDP), Greengenes and NCBI for taxonomic classification, or databases such as FOAM and PFAM for protein sequences. Depending on these databases, the numbers of taxon or functional genes classified back as output data can differ (35, 36). Ye et al. evaluated the different classifier methods and noted a wide variation in the total species abundance obtained by the different classifiers that have their associated default databases for the same sample. However, if a common database is constructed and used across the different classifier methods, the variation in the total species identified becomes lower (35). Likewise, the antibiotic resistance gene prediction potential (including the ability to annotate correctly the number of antibiotic resistance genes and associated classes) differs depending on the type of antibiotic resistance databases (e.g., ARDB,

247 ARG-miner, CARD, and SARG) (37). These observations are worth noting because companies (e.g.,
248 CosmosID, DNAsense, and BaseClear) are now providing metagenomic and bioinformatic services for
249 the generated data, making it particularly convenient for users without any experience handling large
250 datasets to utilize metagenomics as a routine monitoring tool. However, most of these companies use
251 their in-house-developed databases for genome-centric or gene-centric analysis of metagenomic data,
252 making protocol standardization and cross comparison of results particularly challenging. Therefore, each
253 method or company can provide classification results that differ, which would not facilitate interlab
254 comparisons.

255 For some classification methods, particularly those that come with relatively large databases, time
256 is needed to install and build the databases in local servers for first-time users. We performed an analysis
257 to determine the time needed to classify a dataset of approximately 890k sequences and found that,
258 depending on the method, the time can range from 2 min to 2 h using a one-node CPU and 200 Gb RAM
259 (Table 2). With advances in computing power, the time needed to analyze the full metagenomic dataset is
260 likely to shorten. However, there is a likelihood that most of the reads in environmental surveys of
261 reclaimed water can result in being unclassified or unable to identify confidently at the species/strains
262 level with the profiling methods (38). The collation of large genomic databases remains in its early stages
263 compared with well-curated 16S rRNA gene databases such as RDP, Silva, and Greengenes, particularly
264 for viruses and eukaryotes for which the completeness of the existing databases may not be as well
265 developed as that for bacteria (39). Furthermore, the classification results derived from shotgun
266 sequencing reads, particularly those that are present in a relative abundance <0.1%, are likely to
267 represent false-positive identification (35). Therefore, a bottleneck lies in collating well-curated databases
268 to facilitate our abilities to generate meaningful data related to phylogenetic identification from
269 metagenomic data.

270 In addition to classification for the phylogenetic identities of the microbial community, several
271 databases are available to identify antibiotic resistance genes (ARG), metal resistance genes (MRG) and
272 virulence factors (Table 2). Once the reads are classified accordingly, the mapped reads across a
273 constant can, in theory, be normalized either as **(i)** the number of target sequences per million sequence

reads (i.e., counts per million, CPM), **(ii)** the number of target sequences per number of marker genes (e.g., the 16S rRNA gene), **(iii)** the number of target sequences per cell number (40), **(iv)** RPKM (reads per kilobase per million mapped reads) (41) or FPKM (fragments per kilobase per million mapped reads, analogous to RPKM and used especially in paired-end shotgun sequencing reads). CPM is usually more commonly used than RPKM or FPKM. Regardless of the normalization step used, such normalization is required to obtain relative abundance from metagenomics that can be used in comparative analysis. Relative abundances can also be used for correlation against meta-data (e.g., water quality data or operational data). For example, Hendriksen and coworkers utilized metagenomics to monitor the occurrence and diversity of ARGs in urban sewage collected from 79 sites in 60 different countries. They expressed the number of reads assigned to ARGs per kilobase per million fragments (FPKM) across the different geographical regions and found that Africa and South America have higher median ARG reads than Asia, Europe, Middle East, North America and Oceania. They further correlated these relative abundance values with World Bank variables (e.g., extent of open defecation practices, life expectancy, infection and malnutrition rates) and observed a strong correlation between the ARG relative gene abundance and socioeconomic, health and environmental factors (42). This corroborates the conclusion from another study demonstrating a strong correlation between antimicrobial resistance indices (obtained through nonmetagenomic methods) and improving sanitation and good governance (43).

Alternatively, multivariate analysis can also be performed using the relative abundance of all identified taxons/genes across the different samples. Changes in the alpha diversity (a quantitative measure of community diversity) of these marker genes identified from metagenomics can also be performed although a need exists to discern between technical variability (natural changes to a treatment due to the stochastic nature of the system) vs. biological results (made in response to the treatment) (44). Such analysis was demonstrated in a recent study that monitored the surface water quality at multiple locations in Haiti postearthquake. The authors determined that the relative abundance of bacteria was differentiated based on the sampling locations, but the Chao1 alpha diversity was not significantly different among the sampling sites. The authors further determined the relative abundance of marker genes associated with known waterborne pathogens such as *E. coli* O157:H7 and *Vibrio cholerae*, as well

as the presence of phages associated with these pathogens in some of the sample replicates, indicating potential breaches in sanitation infrastructure after the earthquake (45).

Concerning the genome-centric and gene-centric approaches, we ask readers to also refer to a recent review paper by Lal Gupta et al., in which the authors illustrated a workflow to determine the scope and distribution of resistomes in complex environments using both a read-based profiling approach and a de novo assembly based profiling approach on metagenomic data (37). The workflow suggested by Lal Gupta et al. can potentially be applied to determine the classifications of both taxonomy and other functional genes such as metal resistance genes and virulence factors (Table 2). A wide suite of tools for assembly and annotation is available, and each may generate different results. Choosing the most appropriate or accurate metagenomic tool to facilitate reclaimed water quality monitoring is not easy because most of the existing tools utilize databases that are not initially developed for this sample type. Several studies were conducted to identify accurate tools for general environmental shotgun sequencing data, with one recent study concluding that k-mer-based approaches (e.g., Kraken) may outperform other tools in terms of accuracy (46) and speed (Table 2). Regardless of which metagenomic analytical pipeline is chosen to use for reclaimed water quality, the same pipeline should be used consistently across all samples to facilitate comparison.

Improving pathogen detection capabilities

In cases of public health outbreaks that may be due to the use of reclaimed water, a need exists to promptly identify the causative microbial agent. However, the current state of metagenomics may not be well poised to facilitate a rapid decision-making process because sample preparation, sequencing and bioinformatic analysis can assume a considerable amount of time. If the time needed for DNA extraction, library preparation and sequencing is considered, the whole procedure would have taken approximately 39 to 55 h using an Illumina sequencing platform. This process can be sped up using newer sequencing platforms such as Nanopore sequencing platforms but would still take approximately 18 h to complete the entire preparation and sequencing (47). For example, Nanopore MinION required approximately 24 h to determine the presence of Ebola virus in a human clinical specimen (48). However, viral genomes are

small (< 1 Mbp) and do not represent the average genome size of bacterial or protozoal pathogens. Hence, the time needed to draft the complete genome of large prokaryotic or eukaryotic cells would be significantly longer. Alternatively, instead of focusing on complete genomes, the draft genomes of bacterial isolates can be obtained through metagenomics. The reads were merged and assembled to obtain longer contigs or draft genomes before mapping against bacterial pathogen databases. The contigs can then be identified for marker genes associated with the pathogenic species at a certain threshold confidence level. Using this approach, bacterial pathogens such as *Bacillus anthracis*, *Klebsiella pneumoniae* and nontuberculosis mycobacteria were detected in the effluent of a wastewater treatment plant that only utilized a conventional activated sludge tank to decontaminate the wastewater (49). This approach can be further sped up to provide a preliminary analysis of the functional traits within 6 h of sequencing on Nanopore platforms (47).

Although metagenomics has demonstrated the huge potential to reveal novel insights into gene functions, identifying pathogens through assembly may be challenging because waterborne pathogens are generally present in low abundances and would theoretically show up with very low read counts. For example, with approximately 4 Gb per library, pathogenic *E. coli* that tested positive using culture-based methods were not detected by metagenomics (50). Brute-force ultradeep sequencing can be performed to obtain high read coverage of those rare taxa, but this approach can be costly. In recent years, within the field of clinical diagnostics, attempts have been made to identify pathogens using a scoring system after metagenomics. To do so, sequences are obtained from both background controls and test samples before alignment and identification against a curated database (e.g., nucleotide or protein databases of NCBI). The number of reads that aligned positively to a known hit (e.g., target X) in the database is determined first in the background/control samples. This would generate a mean number of reads assigned to X along with the standard deviation that is present in the background/control samples. Subsequently, the number of reads assigned to X in a separate test sample can also be obtained. A z-score can then be obtained (51), and one can then denote which targets demonstrate the highest z-score regardless of its raw abundance counts and, hence, presumably is the causative agent for clinical infection. This or similar scoring approaches have been tested for clinical diagnostics, where samples are derived from blood, urine or biopsy (51-54). However, no demonstration of this approach has yet been

made on reclaimed water samples because it may be technically challenging to do so given the more diverse microbial community in reclaimed water than infected clinical specimens.

Improving semiquantitative capabilities of metagenomics

Most of the studies expressed marker genes in relative abundance, calculated using the following equation:

$$\text{Relative abundance of marker gene x} = \frac{Number\ of\ sequences\ assigned\ to\ gene\ x}{Total\ number\ of\ sequences}$$

However, this calculation does not consider the reference sequence length and how it would impact match hits (55). For example, in the SNC-ARDB database for ARGs, reference sequences can range from 186 to 4728 bp. The number of reads signifying marker gene x that maps positively to a reference sequence of 186 bp may be different from the number of reads that maps positively to a reference sequence of 4728 bp when using the same criteria of ≥ 90% sequence identity and alignment length of ≥ 25 amino acids. Hence, Li et al. demonstrated a correction factor that normalizes the number of sequences assigned to gene x by the reference sequence length (55). In the same study, they further expressed the reads in a similar way to that obtained from quantitative PCR, whereby they normalized the marker gene x results with the total number of reads that matched against the 16S rRNA reference sequence.

Relative abundance of marker gene x normalized against 16S rRNA genes=

$$\frac{(Number\ of\ sequences\ assigned\ to\ gene\ x)(Length\ of\ reads/Length\ of\ gene\ x\ reference\ sequence)}{(Number\ of\ sequences\ assigned\ to\ 16S)(Length\ of\ reads/Length\ of\ 16S\ reference\ sequence)}$$

Alternatively, metagenomic data can also be assessed for the average coverage of a set of approximately 30 essential single-copy marker genes that were found in nearly all Bacteria and Archaea (40, 56). Because these are single-copy marker genes, the average number of these gene counts can be interpreted to be similar to the number of bacterial and archaeal cell numbers. Subsequently, that value

can be used as a normalization factor to determine the number of reads of marker gene x per prokaryote cell.

The abovementioned methods used for metagenomic datasets can only achieve information on the relative abundance and cannot provide quantitative measurements in terms of the contaminant number per liter of reclaimed water. The latter set of values is usually needed for quantitative microbial risk assessment (QMRA). A possible way to overcome this challenge would be to couple flow cytometry with metagenomics on the same sample. For example, the total cell counts can be first estimated by enumerating them with nucleic acid stains and flow cytometry. This would generate a value associated with the number of cells per L. This value can then be multiplied by the normalized marker gene x count per prokaryote cell obtained via metagenomics to derive the marker gene x count per L. However, even with these estimated values, the dose-response models and transmission probability associated with emerging contaminants such as ARB or ARG are still unavailable to facilitate QMRA although recent efforts have been made to introduce dose-response models that incorporate stochastic death dynamics between ARB and antibiotic-susceptible bacteria (57), hence allowing the consideration of ARB in existing dose-response models.

Applications of metagenomics to monitor reclaimed water quality

Metagenomics is commonly used to conduct a baseline characterization of the diversity and relative abundance of contaminants that are present in reclaimed water. For example, Chopyk et al. collected water samples from tidal brackish rivers, freshwater ponds and creeks and water reclamation facilities and proceeded to process these samples for shotgun metagenomics (58). The samples were evaluated for taxonomic and functional differences. Although no apparent differences were found in the overall phylogenetic distribution of the microbial community among the samples, the diversity of ARGs in at least one of the reclaimed water samples was higher than that in the other water samples. This outlier trend may be an anomaly arising from the small sample size or a potential breach in the treatment process.

406 In addition to ARGs, the diversity of viruses that are present in reclaimed water can also be
407 elucidated by metagenomics. Most of the assigned reads obtained from metagenomics were determined
408 to be bacteriophages assigned to the families Myoviridae, Podoviridae and Siphoviridae (59-61). By
409 contrast, human enteric viruses account for < 1% of the total sequences obtained from treated effluent
410 postmembrane filtration (59). By matching against databases designed to annotate viral sequences (e.g.,
411 MetaVir), viruses of potential public health relevance and belonging to the families Herpesvirales,
412 Adenoviridae, Polyomaviridae and Parvodiridae are detectable in the postmembrane filtrated effluents
413 (59). Coincidentally, Polyomaviridae was also detected in the postmembrane chlorinated effluent sampled
414 from a WWTP at another location (61). These earlier studies use a gene-centric approach to identify the
415 marker genes associated with potential viruses at the family level and, hence, cannot describe the viral
416 pathogens at the species level. Additionally, most of the detected human enteric viruses are double-
417 stranded DNA viruses and not single-stranded RNA viruses that would need to be first recovered through
418 RNA extraction and transcribed to obtain cDNA before proceeding with shotgun metagenomic
419 sequencing.

420 The abovementioned studies characterized the microbial contaminants that are present in
421 reclaimed water collected at the end of the wastewater treatment process. This sampling point is typically
422 defined as the point-of-entry before the reclaimed water is transported or distributed to the point-of-use.
423 Because reclaimed water typically still contains organic carbon and other essential nutrients that can
424 support microbial regrowth, the reclaimed water quality can potentially change within the distribution
425 network depending on factors such as the residual disinfectant concentration, hydraulic retention time,
426 distance of network and so on. To determine changes in water quality and, hence, infer the extent of the
427 biological stability of reclaimed waters, metagenomics can be used to characterize the microbial
428 community in the reclaimed water at the point-of-use and compare against that at the point-of-entry.
429 Garner et al. determined that, in four of their studied reclaimed water distribution networks, a decrease
430 was observed in the relative abundance and diversity of ARGs from the point-of-entry to point-of-use.
431 However, the relative abundance of certain ARGs correlates with the concentration of biological dissolved
432 organic carbon, suggesting the need to limit the amount of organic carbon in the distribution systems (62).
433 Similarly, Zaouri et al. utilized a metagenomic approach to simultaneously monitor the taxonomic profiles

of bacterial and viral communities, as well as the antibiotic resistome in aquifers that were recharged with treated wastewater (63). By comparing the upstream controls, the authors determined that bacterial families such as Planctomycetes are present at a higher relative abundance in recharged aquifers, likely because of the higher organic carbon content in these waters upon exposure to treated wastewater. This observation reiterates the earlier observation that organic carbon can change the microbial community, likely because of microbial regrowth. Additionally, Zaouri et al. observed that the viral family Picornaviridae is present in higher relative abundance in recharged aquifers compared with the controls (63), suggesting potential dissemination of the human enteric viruses at the point-of-use due to reclaimed water.

Collectively, these studies demonstrate the use of metagenomics to **(i)** identify microbial populations and functional genes in water matrices, **(ii)** compare samples for the reclaimed water quality at either a temporal or spatial scale, **(iii)** and correlate data from metagenomics to other meta-data (e.g., organic content, residual disinfectant, and temperature) to determine which variable to control to alleviate unwanted detrimental changes in reclaimed water quality.

Perspectives

Metagenomics provides a nontargeted approach to simultaneously examine both phylogenetic and functional profiles associated with the water matrices. However, to revamp the way water industry is monitoring reclaimed water quality, continued development in metagenomics is needed in the following areas:

- **Improvement in databases**: there should be a continuous effort to perform whole-genome sequencing of a wide consortium of biological pathogens relevant to reclaimed water, particularly viruses and protozoa. These assembled genomes should be made available in public depositories for further curation of databases, which would improve the resolution of future information we can obtain from metagenomics.

- **Standardized protocols for data analysis**: similar to other methods that are endorsed by regulatory agencies for water quality monitoring, shotgun sequencing protocols and bioinformatic pipelines should also be standardized so that metagenomic data can be benchmarked against regulatory standards and cross-compared across different laboratories.
- **Developing bioinformatic tools to identify rare taxa (e.g., low-abundance pathogens)**: While brute-force ultradeep sequencing can help in identifying rare taxa, this incurs a cost that can add up significantly if routinely adopted for reclaimed water quality monitoring. The huge amount of data would also need more time for analysis to be completed. To circumvent this bottleneck, rapid bioinformatic tools need to be developed to identify low-abundance pathogens and samples with poor water quality. Potential tools include the z-scoring system already demonstrated for clinical samples, but would need to be fine tuned for reclaimed water quality monitoring, and data mining or a machine learning algorithm to identify trends and outliers that can isolate aberrations in reclaimed water quality.
- **Conduct more studies to demonstrate the use of metagenomics in reclaimed water quality monitoring**: Developing metagenomics as a toolkit to denote and predict water quality would require more studies to provide a representative sample size that can identify which biomarkers correlate with certain measurable water quality data (e.g., pH, residual chlorine, and organic carbon concentration). Current studies mainly focus on monitoring reclaimed water for the microbial community and ARGs. Other functional genes, such as mobile genetic elements, virulence factors and metal resistance genes, also play an equal role as ARGs in affecting potential safety concerns when reusing the waters and should also be evaluated in future studies.

Conclusions

The advent of next-generation sequencing technologies and faster computing capabilities and the availability of databases have facilitated the use of metagenomics for reclaimed water quality monitoring. Metagenomics can determine changes in both the phylogenetic and functional diversities of emerging

contaminants in a nontargeted manner. Such information can be used to elucidate the removal efficiency achieved by wastewater treatment technologies and to monitor changes in reclaimed water quality over a distribution network. The data derived from metagenomics are semiquantitative (i.e., in terms of relative abundance). However, when complemented with other tools—for example, flow cytometry and quantitative PCR—an estimated abundance data can be derived, although more studies are required to facilitate the use of these data in risk assessment or for comparison against regulatory limits.

Acknowledgments

The authors would like to acknowledge funding support from the KAUST Competitive Research Grant URF/1/3407-01-01 awarded to PYH. This review was written based on the content presented by PYH at the 2018 Singapore International Water Week. The authors declare no conflict of interest.

References

1. Harwood VJ, Levine AD, Scott TM, Chivukula V, Lukasik J, Farrah SR, Rose JB. 2005. Validity of the Indicator Organism Paradigm for Pathogen Reduction in Reclaimed Water and Public Health Protection. Applied and Environmental Microbiology 71:3163.
2. Leclerc H, Schwartzbrod L, Dei-Cas E. 2002. Microbial agents associated with waterborne diseases. Crit Rev Microbiol 28:371-409.
3. Sanganyado E, Gwenzi W. 2019. Antibiotic resistance in drinking water systems: Occurrence, removal, and human health risks. Sci Total Environ 669:785-797.
4. Lluch J, Servant F, Paisse S, Valle C, Valiere S, Kuchly C, Vilchez G, Donnadieu C, Courtney M, Burcelin R, Amar J, Bouchez O, Lelouvier B. 2015. The Characterization of Novel Tissue Microbiota Using an Optimized 16S Metagenomic Sequencing Pipeline. PLoS One 10:e0142334.
5. Mohd Shaufi MA, Sieo CC, Chong CW, Gan HM, Ho YW. 2015. Deciphering chicken gut microbial dynamics based on high-throughput 16S rRNA metagenomics analyses. Gut Pathogens 7:4.
6. Moreno Y, Moreno-Mesonero L, Amoros I, Perez R, Morillo JA, Alonso JL. 2018. Multiple identification of most important waterborne protozoa in surface water used for irrigation purposes by 18S rRNA amplicon-based metagenomics. Int J Hyg Environ Health 221:102-111.
7. Rondon MR, August PR, Bettermann AD, Brady SF, Grossman TH, Liles MR, Loiacono KA, Lynch BA, MacNeil IA, Minor C, Tiong CL, Gilman M, Osburne MS, Clardy J, Handelsman J, Goodman RM. 2000. Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. Appl Environ Microbiol 66:2541-7.
8. Ish-Horowicz D, Burke JF. 1981. Rapid and efficient cosmid cloning. Nucleic Acids Res 9:2989-98.
9. Stone NE, Fan JB, Willour V, Pennacchio LA, Warrington JA, Hu A, de la Chapelle A, Lehesjoki AE, Cox DR, Myers RM. 1996. Construction of a 750-kb bacterial clone contig and restriction map in the region of human chromosome 21 containing the progressive myoclonus epilepsy gene. Genome Res 6:218-25.
10. Song Y-H, Lee K-T, Baek J-Y, Kim M-J, Kwon M-R, Kim Y-J, Park M-R, Ko H, Lee J-S, Kim K-S. 2017. Isolation and characterization of a novel endo-β-1,4-glucanase from a metagenomic library of the black-goat rumen. Brazilian Journal of Microbiology 48:801-808.
11. Chai Z, Wang J, Tao S, Mou H. 2014. Application of bacteriophage-borne enzyme combined with chlorine dioxide on controlling bacterial biofilm. LWT - Food Science and Technology 59:1159-1165.
12. Orgaz B, Neufeld RJ, SanJose C. 2007. Single-step biofilm removal with delayed release encapsulated Pronase mixed with soluble enzymes. Enzyme and Microbial Technology 40:1045-1051.
13. Lautenschlager K, Hwang C, Liu WT, Boon N, Koster O, Vrouwenvelder H, Egli T, Hammes F. 2013. A microbiology-based multi-parametric approach towards assessing biological stability in drinking water distribution networks. Water Res 47:3015-25.
14. Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A. 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol 12:R18.
15. Brandariz-Fontes C, Camacho-Sanchez M, Vila C, Vega-Pla JL, Rico C, Leonard JA. 2015. Effect of the enzyme and PCR conditions on the quality of high-throughput DNA sequencing results. Sci Rep 5:8056.
16. Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. 2013. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat Biotechnol 31:533-8.

17. Kang DD, Froula J, Egan R, Wang Z. 2015. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ 3:e1165.
18. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijas UZ, Loman NJ, Andersson AF, Quince C. 2013. CONCOCT: Clustering cONtigs on COerage and ComposiTion. Preprint at https://arxiv.org/abs/1312.4038v1.
19. Wu YW, Simmons BA, Singer SW. 2016. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics 32:605-607.
20. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25:1043-55.
21. Haroon MF, Thompson LR, Stingl U. 2016. Draft Genome Sequence of Uncultured SAR324 Bacterium lautmerah10, Binned from a Red Sea Metagenome. Genome Announc 4.
22. Al-Jassim N, Ansari MI, Harb M, Hong PY. 2015. Removal of bacterial contaminants and antibiotic resistance genes by conventional wastewater treatment processes in Saudi Arabia: Is the treated wastewater safe to reuse for agricultural irrigation? Water Res 73:277-90.
23. Sczyrba A, Hofmann P, Belmann P, Koslicki D, Janssen S, Droge J, Gregor I, Majda S, Fiedler J, Dahms E, Bremges A, Fritz A, Garrido-Oter R, Jorgensen TS, Shapiro N, Blood PD, Gurevich A, Bai Y, Turaev D, DeMaere MZ, Chikhi R, Nagarajan N, Quince C, Meyer F, Balvociute M, Hansen LH, Sorensen SJ, Chia BKH, Denis B, Froula JL, Wang Z, Egan R, Don Kang D, Cook JJ, Deltel C, Beckstette M, Lemaitre C, Peterlongo P, Rizk G, Lavenier D, Wu YW, Singer SW, Jain C, Strous M, Klingenberg H, Meinicke P, Barton MD, Lingner T, Lin HH, Liao YC, et al. 2017. Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software. Nat Methods 14:1063-1071.
24. Sczyrba A, Hofmann P, Belmann P, Koslicki D, Janssen S, Droge J, Gregor I, Majda S, Fiedler J, Dahms E, Bremges A, Fritz A, Garrido-Oter R, Jorgensen TS, Shapiro N, Blood PD, Gurevich A, Bai Y, Turaev D, DeMaere MZ, Chikhi R, Nagarajan N, Quince C, Meyer F, Balvociute M, Hansen LH, Sorensen SJ, Chia BKH, Denis B, Froula JL, Wang Z, Egan R, Kang DD, Cook JJ, Deltel C, Beckstette M, Lemaitre C, Peterlongo P, Rizk G, Lavenier D, Wu YW, Singer SW, Jain C, Strous M, Klingenberg H, Meinicke P, Barton MD, Lingner T, Lin HH, Liao YC, et al. 2017. Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software. Nature Methods 14:1063-+.
25. Breitwieser FP, Lu J, Salzberg SL. 2017. A review of methods and databases for metagenomic classification and assembly. Briefings in Bioinformatics 20:1125-1136.
26. Gao F. 2018. Recent developments of software and database in microbial genomics and functional genomics. Briefings in Bioinformatics 20:732-734.
27. Tessler M, Neumann JS, Afshinnekoo E, Pineda M, Hersch R, Velho LFM, Segovia BT, Lansac-Toha FA, Lemke M, DeSalle R, Mason CE, Brugler MR. 2017. Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing. Scientific Reports 7:6589.
28. Lawrence JG. 1999. Gene transfer, speciation, and the evolution of bacterial genomes. Current Opinion in Microbiology 2:519-523.
29. Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature 405:299-304.
30. Vernikos G, Medini D, Riley DR, Tettelin H. 2015. Ten years of pan-genome analyses. Current Opinion in Microbiology 23:148-154.
31. Huson DH, Auch AF, Qi J, Schuster SC. 2007. MEGAN analysis of metagenomic data. Genome Research 17:377-386.

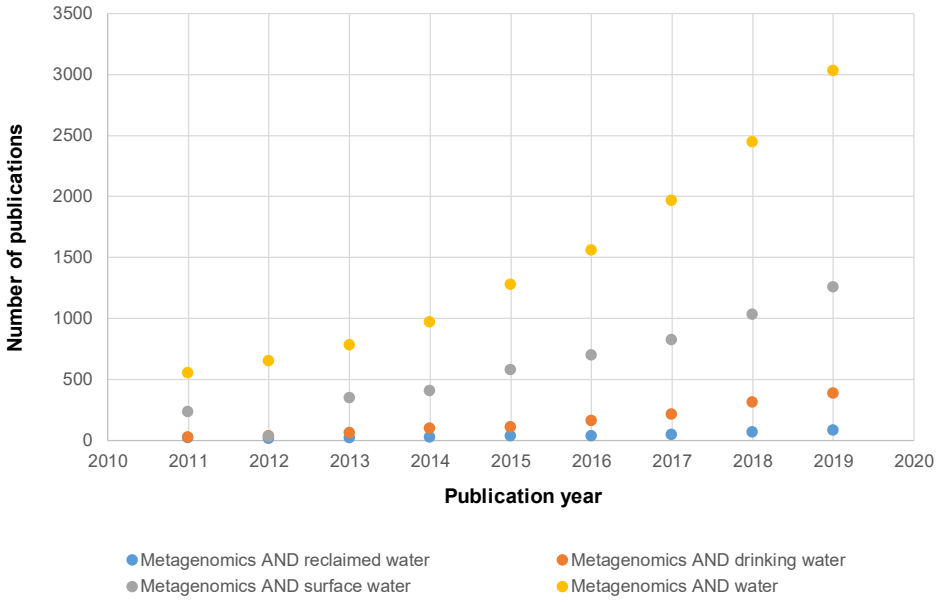
32. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, Wilkening J, Edwards RA. 2008. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9:386.
33. Meyer F, Bagchi S, Chaterji S, Gerlach W, Grama A, Harrison T, Paczian T, Trimble WL, Wilke A. 2017. MG-RAST version 4—lessons learned from a decade of low-budget ultra-high-throughput metagenome analysis. Briefings in Bioinformatics 20:1151-1159.
34. McIntyre ABR, Ounit R, Afshinnekoo E, Prill RJ, Henaff E, Alexander N, Minot SS, Danko D, Foox J, Ahsanuddin S, Tighe S, Hasan NA, Subramanian P, Moffat K, Levy S, Lonardi S, Greenfield N, Colwell RR, Rosen GL, Mason CE. 2017. Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. Genome Biology 18.
35. Ye SH, Siddle KJ, Park DJ, Sabeti PC. 2019. Benchmarking Metagenomics Tools for Taxonomic Classification. Cell 178:779-794.
36. Balvociute M, Huson DH. 2017. SILVA, RDP, Greengenes, NCBI and OTT - how do these taxonomies compare? Bmc Genomics 18.
37. Lal Gupta C, Kumar Tiwari R, Cytryn E. 2020. Platforms for elucidating antibiotic resistance in single genomes and complex metagenomes. Environment International 138:105667.
38. Kim Y, Aw TG, Teal TK, Rose JB. 2015. Metagenomic Investigation of Viral Communities in Ballast Water. Environ Sci Technol 49:8396-407.
39. Hull NM, Ling F, Pinto AJ, Albertsen M, Jang HG, Hong PY, Konstantinidis KT, LeChevallier M, Colwell RR, Liu WT. 2019. Drinking Water Microbiome Project: Is it Time? Trends Microbiol 27:670-677.
40. Yin X, Jiang XT, Chai B, Li L, Yang Y, Cole JR, Tiedje JM, Zhang T. 2018. ARGs-OAP v2.0 with an expanded SARG database and Hidden Markov Models for enhancement characterization and quantification of antibiotic resistance genes in environmental metagenomes. Bioinformatics 34:2263-2270.
41. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods 5:621-628.
42. Hendriksen RS, Munk P, Njage P, van Bunnik B, McNally L, Lukjancenko O, Roder T, Nieuwenhuijse D, Pedersen SK, Kjeldgaard J, Kaas RS, Clausen P, Vogt JK, Leekitcharoenphon P, van de Schans MGM, Zuidema T, de Roda Husman AM, Rasmussen S, Petersen B, Global Sewage Surveillance project c, Amid C, Cochrane G, Sicheritz-Ponten T, Schmitt H, Alvarez JRM, Aidara-Kane A, Pamp SJ, Lund O, Hald T, Woolhouse M, Koopmans MP, Vigre H, Petersen TN, Aarestrup FM. 2019. Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. Nat Commun 10:1124.
43. Collignon P, Beggs JJ, Walsh TR, Gandra S, Laxminarayan R. 2018. Anthropological and socioeconomic factors contributing to global antimicrobial resistance: a univariate and multivariable analysis. Lancet Planet Health 2:e398-e405.
44. Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, Gonzalez A, Kosciolek T, McCall LI, McDonald D, Melnik AV, Morton JT, Navas J, Quinn RA, Sanders JG, Swafford AD, Thompson LR, Tripathi A, Xu ZJZ, Zaneveld JR, Zhu QY, Caporaso JG, Dorrestein PC. 2018. Best practices for analysing microbiomes. Nature Reviews Microbiology 16:410-422.
45. Roy MA, Arnaud JM, Jasmin PM, Hamner S, Hasan NA, Colwell RR, Ford TE. 2018. A Metagenomic Approach to Evaluating Surface Water Quality in Haiti. Int J Environ Res Public Health 15.
46. Gardner PP, Watson RJ, Morgan XC, Draper JL, Finn RD, Morales SE, Stott MB. 2019. Identifying accurate metagenome and amplicon software via a meta-analysis of sequence to taxonomy benchmarking studies. PeerJ 7:e6160-e6160.

47. Quick J, Ashton P, Calus S, Chatt C, Gossain S, Hawker J, Nair S, Neal K, Nye K, Peters T, De Pinna E, Robinson E, Struthers K, Webber M, Catto A, Dallman TJ, Hawkey P, Loman NJ. 2015. Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella. Genome Biol 16:114.
48. Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, Bore JA, Koundouno R, Dudas G, Mikhail A, Ouedraogo N, Afrough B, Bah A, Baum JH, Becker-Ziaja B, Boettcher JP, Cabeza-Cabrerizo M, Camino-Sanchez A, Carter LL, Doerrbecker J, Enkirch T, Dorival IGG, Hetzelt N, Hinzmann J, Holm T, Kafetzopoulou LE, Koropogui M, Kosgey A, Kuisma E, Logue CH, Mazzarelli A, Meisel S, Mertens M, Michel J, Ngabo D, Nitzsche K, Pallash E, Patrono LV, Portmann J, Repits JG, Rickett NY, Sachse A, Singethan K, Vitoriano I, Yemanaberhan RL, Zekeng EG, Trina R, Bello A, Sall AA, Faye O, et al. 2016. Real-time, portable genome sequencing for Ebola surveillance. Nature 530:228-232.
49. Li B, Ju F, Cai L, Zhang T. 2015. Profile and Fate of Bacterial Pathogens in Sewage Treatment Plants Revealed by High-Throughput Metagenomic Approach. Environ Sci Technol 49:10492-502.
50. Suttner B, Johnston ER, Orellana LH, Rodriguez-R LM, Hatt JK, Carychao D, Carter MQ, Cooley MB, Konstantinidis KT. 2020. Metagenomics as a Public Health Risk Assessment Tool in a Study of Natural Creek Sediments Influenced by Agricultural and Livestock Runoff: Potential and Limitations. Applied and Environmental Microbiology 86:e02525-19.
51. Wilson MR, O'Donovan BD, Gelfand JM. 2018. Chronic Meningitis Investigated via Metagenomic Next-Generation Sequencing (vol 75, pg 947, 2018). Jama Neurology 75:1028-1028.
52. Doan T, Wilson MR, Crawford ED, Chow ED, Khan LM, Knopp KA, O'Donovan BD, Xia DX, Hacker JK, Stewart JM, Gonzales JA, Acharya NR, DeRisi JL. 2016. Illuminating uveitis: metagenomic deep sequencing identifies common and rare pathogens. Genome Medicine 8:90.
53. Grumaz S, Stevens P, Grumaz C, Decker SO, Weigand MA, Hofer S, Brenner T, von Haeseler A, Sohn K. 2016. Next-generation sequencing diagnostics of bacteremia in septic patients. Genome Medicine 8:73.
54. Schmidt K, Mwaigwisya S, Crossman LC, Doumith M, Munroe D, Pires C, Khan AM, Woodford N, Saunders NJ, Wain J, O'Grady J, Livermore DM. 2017. Identification of bacterial pathogens and antimicrobial resistance directly from clinical urines by nanopore-based metagenomic sequencing. Journal of Antimicrobial Chemotherapy 72:104-114.
55. Li B, Yang Y, Ma L, Ju F, Guo F, Tiedje JM, Zhang T. 2015. Metagenomic and network analysis reveal wide distribution and co-occurrence of environmental antibiotic resistance genes. ISME J 9:2490-502.
56. Nayfach S, Pollard KS. 2015. Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. Genome Biol 16:51.
57. Chandrasekaran S, Jiang SC. 2019. A dose response model for quantifying the infection risk of antibiotic-resistant bacteria. Sci Rep 9:17093.
58. Chopyk J, Nasko DJ, Allard S, Bui A, Treangen T, Pop M, Mongodin EF, Sapkota AR. 2020. Comparative metagenomic analysis of microbial taxonomic and functional variations in untreated surface and reclaimed waters used in irrigation applications. Water Res 169:115250.
59. O'Brien E, Munir M, Marsh T, Heran M, Lesage G, Tarabara VV, Xagoraraki I. 2017. Diversity of DNA viruses in effluents of membrane bioreactors in Traverse City, MI (USA) and La Grande Motte (France). Water Res 111:338-345.
60. Rosario K, Nilsson C, Lim YW, Ruan Y, Breitbart M. 2009. Metagenomic analysis of viruses in reclaimed water. Environ Microbiol 11:2806-20.

61. Jumat MR, Hasan NA, Subramanian P, Heberling C, Colwell RR, Hong PY. 2017. Membrane Bioreactor-Based Wastewater Treatment Plant in Saudi Arabia: Reduction of Viral Diversity, Load, and Infectious Capacity. Water 9.
62. Garner E, Chen C, Xia K, Bowers J, Engelthaler DM, McLain J, Edwards MA, Pruden A. 2018. Metagenomic Characterization of Antibiotic Resistance Genes in Full-Scale Reclaimed Water Distribution Systems and Corresponding Potable Systems. Environ Sci Technol 52:6113-6125.
63. Zaouri N, Jumat MR, Cheema T, Hong PY. 2020. Metagenomics-based evaluation of groundwater microbial profiles in response to treated wastewater discharge. Environ Res 180:108835.
64. UIUC. Roy J Carver Biotechnology Center: DNA services https://biotech.illinois.edu/htdna/samplesubmission. Accessed 16 December 2019.
65. Nanopore. Products overview. https://nanoporetech.com/products/comparison. Accessed 16 December 2019.
66. DNAlink. PacBio (Sequel II/Sequel/ RS II). https://www.dnalinkseqlab.com/pacbio-sequel-rsii/ Accessed 16 December 2019.
67. Wood DE, Lu J, Langmead B. 2019. Improved metagenomic analysis with Kraken 2. Genome Biology 20:257.
68. Menzel P, Ng KL, Krogh A. 2016. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat Commun 7:11257.
69. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. 2012. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods 9:811-4.
70. Milanese A, Mende DR, Paoli L, Salazar G, Ruscheweyh HJ, Cuenca M, Hingamp P, Alves R, Costea PI, Coelho LP, Schmidt TSB, Almeida A, Mitchell AL, Finn RD, Huerta-Cepas J, Bork P, Zeller G, Sunagawa S. 2019. Microbial abundance, activity and population genomic profiling with mOTUs2. Nat Commun 10:1014.
71. Alcock BP, Raphenya AR, Lau TTY, Tsang KK, Bouchard M, Edalatmand A, Huynh W, Nguyen A-LV, Cheng AA, Liu S, Min SY, Miroshnichenko A, Tran H-K, Werfalli RE, Nasir JA, Oloni M, Speicher DJ, Florescu A, Singh B, Faltyn M, Hernandez-Koutoucheva A, Sharma AN, Bordeleau E, Pawlowski AC, Zubyk HL, Dooley D, Griffiths E, Maguire F, Winsor GL, Beiko RG, Brinkman FSL, Hsiao WWL, Domselaar GV, McArthur AG. 2019. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic Acids Research doi:10.1093/nar/gkz935.
72. Pal C, Bengtsson-Palme J, Rensing C, Kristiansson E, Larsson DG. 2014. BacMet: antibacterial biocide and metal resistance genes database. Nucleic Acids Res 42:D737-43.
73. Liu B, Zheng D, Jin Q, Chen L, Yang J. 2019. VFDB 2019: a comparative pathogenomic platform with an interactive web interface. Nucleic Acids Res 47:D687-D692.
74. Yoon SH, Park YK, Kim JF. 2015. PAIDB v2.0: exploration and analysis of pathogenicity and resistance islands. Nucleic Acids Res 43:D624-30.
75. Brettin T, Davis JJ, Disz T, Edwards RA, Gerdes S, Olsen GJ, Olson R, Overbeek R, Parrello B, Pusch GD, Shukla M, Thomason JA, 3rd, Stevens R, Vonstein V, Wattam AR, Xia F. 2015. RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. Sci Rep 5:8365.
76. Leplae R, Hebrant A, Wodak SJ, Toussaint A. 2004. ACLAME: a CLAssification of Mobile genetic Elements. Nucleic Acids Res 32:D45-9.

727

728



729

730 **Figure 1.** Number of publications associated with each keyword set and retrieved from Scopus from the

731 years 2011 to 2019.

732

733

734 **Table 1.** Current sequencing platforms and their average read length and throughput reported by either
735 manufacturers or selected service laboratories

	Directional reads	Read length (bp)	Throughput per lane	Ref.
NovaSeq 6000				
SP flowcell	Single reads	100	400–500 million	(64)
	Paired reads	2 × 150 or 2 × 250	800 million	
S1 flowcell	Single reads	100	800 million	
	Paired reads	2 × 100 or 2 × 150	1.5 billion	
S2 flowcell	Single reads	100	1.5 billion	
S4 flowcell	Paired reads	2 × 150	5–6 billion	
HiSeq 4000				
8 lane flowcell	Single reads	50–150	300–400 million	(64)
	Paired reads	50–150	650–800 million	
HiSeq 2500				
Rapid V2 flowcell	Single reads	50–260	150–200 million	(64)
	Paired reads	50–260	220–400 million	
MiSeq				
V3 flowcell	Paired reads	300	10–30 million	(64)
V2 flowcell	Paired reads	250	6–20 million	
V2 nano flowcell	Paired reads	250	500 thousand–2 million	
Flongle	Single reads	Dependent on the quality and fragment size of the DNA template	2 Gbp	(65)
MinION Mk and GridION Mk			50 Gbp	
PromethION			220 Gbp	
Sequel	Single reads	Dependent on the quality and fragment size of the DNA template, but reportedly > 1000	500 thousand	(66)
Sequel II	Single reads	Dependent on the quality and fragment size of the DNA template, but reportedly > 1000	4 million	

736

737 **Table 2.** Tools and databases available for phylogenetic and marker gene identification. * denotes the time required to generate the classification results from a test dataset generated
738 from Illumina HiSeq4000 paired-end sequencing. The dataset contains approximately 7 million trimmed paired reads of average 150 bp, 600 Mb fastq.gz file.

Phylogenetic identification	**Version used for this review**	**Database type**	**Target collection**	**Database size**	**Latest update**	**Time required to build database at first use**	**Time required to generate classification results ***	**Ref**
Kraken2	v2.0.8-beta	DNA	Refseq bacteria	103 Gb	2019	15 h	3 min	(67)
MiniKraken2	v2	DNA	Refseq bacteria, archaea, virus and the GRCh38 human genome dataset	8 Gb	2019	Not required	2 min	
Kaiju	v1.7.2	Protein	Eukaryotes, bacteria, viral genomes	97 Gb	2019	4 h	2 h	(68)
MetaPhlAn2	0c3ed7b7718b	Marker genes	Eukaryotes, bacteria, archaea, virus	1 Gb	2018	Not required	2 h	(69)
mOTUs2	v2.5.1	Marker genes	Eukaryotes, bacteria, archaea	1.5 Gb	2018	Not required	40 min	(70)
Marker gene identification	**Latest version at point of writing**	**Target collection**			**Database size**		**Latest update**	**Ref**
CARD	v3.0.4	Antibiotic resistance genes (ARGs)			2602 genes		2019	(71)
SARG	v2.0	ARGs			12307 genes		2018	(40)
BacMet	v2.0	Antibacterial biocide and metal resistance genes			753 genes (experimentally confirmed); 155512 genes (predicted)		2018	(72)
VFDB	Refreshed weekly	Virulence factors			3220 genes (experimentally confirmed); 28587 genes (predicted)		2019	(73)
PAIDB	v2.0	Pathogenicity islands (PAIs) and antimicrobial resistance islands (REIs)			223 PAIs with 1331 genes; 88 REIs with 108 genes		2015	(74)
PATRIC	v3.5.43	Virulence factors (VFs) and ARGs			130963 VFs, 257681 ARGs		2019	(75)
ACLAME	0.4	Mobile genetic elements (MGEs)			122154 proteins from 2326 MGEs		2009	(76)

739

740

741