AEM
Journals.ASM.org

Metagenomic Analyses of Drinking Water Receiving Different Disinfection Treatments

Vicente Gomez-Alvarez, Randy P. Revetta, and Jorge W. Santo Domingo

U.S. Environmental Protection Agency, Office of Research and Development, Cincinnati, Ohio, USA

A metagenome-based approach was used to assess the taxonomic affiliation and function potential of microbial populations in free-chlorine-treated (CHL) and monochloramine-treated (CHM) drinking water (DW). In all, 362,640 (averaging 544 bp) and 155,593 (averaging 554 bp) pyrosequencing reads were analyzed for the CHL and CHM samples, respectively. Most annotated proteins were found to be of bacterial origin, although eukaryotic, archaeal, and viral proteins were also identified. Differences in community structure and function were noted. Most notably, *Legionella*-like genes were more abundant in the CHL samples while mycobacterial genes were more abundant in CHM samples. Genes associated with multiple disinfectant mechanisms were identified in both communities. Moreover, sequences linked to virulence factors, such as antibiotic resistance mechanisms, were observed in both microbial communities. This study provides new insights into the genetic network and potential biological processes associated with the molecular microbial ecology of DW microbial communities.

Free chlorine is used as the primary disinfectant in most drinking water (DW) distribution systems (DWDS). However, chlorine disinfection promotes the formation of disinfectant byproducts (DBPs), and as a result, many water utilities are considering changing to monochloramine to ensure regulatory compliance of targeted DBPs (3). While both disinfection strategies aim at mitigating the presence of pathogens, they do not completely eradicate the growth of microorganisms in DWDS. Indeed, diverse microbial communities have been shown to inhabit DWDS. The latter has been documented by using a variety of culture-based assays (10) and culture-independent approaches, such as 16S rRNA gene sequence analysis using Sanger chemistry (37) and pyrosequencing (21). Additionally, fluorescence *in situ* hybridization targeting the 16S rRNA gene has been used to detect active bacteria in DW biofilms (53). Most of the previous approaches are limited in scope. For example, culture-based techniques are biased toward a small fraction of the inhabiting microbiota. On the other hand, most studies using DNA-based approaches have targeted phylogenetic genes, which provide limited information on the public health relevance of the microbial groups detected.

Metagenome-based approaches offer a more comprehensive view of the genetic complexity of natural and engineered microbial communities, allowing us to better assess the microbial taxonomic diversity and metabolic potential within any given community (23). The number of metagenomic studies has increased in recent years because of the availability of next-generation sequencing technologies. The discoveries have ranged from novel photosynthetic pathways to genetic pathways that are important in host-microbial interactions, findings that would have been difficult to obtain by using conventional phylogenetic analyses (15). Comparison of different metagenomes has further enhanced our understanding of processes unique to some microbiomes and provided the genetic information needed to track multiple populations performing a variety of functions. Interestingly, in spite of the public health relevance of DW, very little information is available on DW metagenomes. For example, Schmeisser et al. (42) sequenced 5,000 random clones of DW biofilm from a cosmid library. Of these clones, 2,200 were identified as putative proteins, and over half of them were characterized into nine functional groups, many of which are associated with proteobacteria such as *Rhizobium*, *Pseudomonas*, and *Escherichia*. However, this study was performed with samples from a nonchlorinated DWDS and therefore some of the information might be specific to such systems. Moreover, due to the relatively low number of sequences analyzed, no information on genes linked to microorganisms associated with public health concerns was obtained.

To further understand the DW microbial network of systems receiving disinfection treatments, we analyzed pyrosequencing data of metagenomes from free-chlorine-treated (CHL) and monochloramine-treated (CHM) DW samples. With approximately 2,000,000 reads combined and, on average, 550 bp per read, to our knowledge, thus far, this represents the largest DW metagenomic survey ever performed.

MATERIALS AND METHODS

Sample collection. DW samples ($n = 2$) were collected from two distribution system simulators (DSS). One sample was collected from a free-chlorine system, while the second sample was collected from a chloramine-amended system (see Fig. S1 in the supplemental material). The DSS are currently in operation at the Environmental Protection Agency Test and Evaluation Facility in Cincinnati, OH. The CHL DW sample was obtained from the main flow of a 16.5-m-long by 23.6-cm pipe loop fed with municipal DW following treatment that employs flocculation and settling with pH adjustment followed by sand filtration, granular activated carbon, and chlorination with final discharge to the DWDS (see Fig. S1A in the supplemental material). The CHM DW sample was obtained from the main flow of a 23-m-long 23.6-cm pipe loop fed with the same municipal DW but amended with ammonia to yield a 2-mg liter^{-1} monochloramine residual (see Fig. S1B in the supplemental material). System

Received 28 March 2012 Accepted 16 June 2012
Published ahead of print 22 June 2012
Address correspondence to Jorge W. Santo Domingo, santodomingo.jorge@epa.gov.
Supplemental material for this article may be found at http://aem.asm.org/.

doi:10.1128/AEM.01018-12

properties and water quality characteristics are presented in Table S1 in the supplemental material. Microbial biomass was concentrated from 100 liters to approximately 250 ml retentate by ultrafiltration as previously described (38), with the minor modification that the ultrafiltration unit ran in a continuous mode connected directly to the pipe loop of the DSS (see Fig. S2 in the supplemental material).

Molecular procedures. The retentate was filtered onto 0.2-μm polycarbonate membranes (GE Osmonics, Minnetonka, MN), and total DNA was extracted by using the UltraClean Soil DNA kit in accordance with the manufacturer's instructions (MoBio Laboratories Inc., Solana Beach, CA). The DNA yield of each sample was approximately 3 ng/μl. This yield did not meet the minimal amount required for pyrosequencing (5 μg). As a result, DNA extracts were subjected to random genome amplification to further increase the DNA yield (44). Previous studies concluded that whole-genome amplification biases were minimal and applied equally to all samples, and any bias would have been annulled during comparative study (9, 47). Briefly, to generate fragments 300 to 600 bp in length, genomic DNA was mechanically sheared for 15 s using a 60 Sonic Dismembrator (Fisher Scientific) and a sonication setting of 3. The resuspended fragments were incubated at 95°C for 5 min with K9-DNA primer (18). The mixture was placed on ice for 5 min and incubated with 50 U of DNA polymerase I large Klenow fragment (New England BioLabs, Ipswich, MA) for 3.5 h at room temperature. The reaction was stopped by heating for 10 min at 75°C. Randomly labeled Klenow extension products were purified, and PCR amplifications were performed in triplicate 100-μl reaction mixtures containing 10 ng of DNA, 1× PCR buffer, 2.5 mM each deoxynucleoside triphosphate, 1% acetamide, 0.625 U of *Ex Taq* (Invitrogen, Carlsbad, CA), and 0.2 μM K9-PCR primer (18) under the following conditions: 28 cycles of 94°C for 40 s, 53°C for 1 min, and 72°C for 30 s with an extension step of 72°C for 1.5 min. PCR products were purified with the QIAquick PCR product cleanup kit (Qiagen, Valencia, CA) and pooled to generate a minimum of 5 μg of DNA (500 ng/μl), which was then used as the template for shotgun pyrosequencing.

Metagenome libraries were generated with the 454 Life Sciences GS-FLX Titanium platform. In all, 1,024,242 and 849,349 reads for the CHL and CHM metagenomes were generated in this study, respectively (see Table S2 in the supplemental material). Quality control filters, an internal tool in the MG-RAST v3.0 pipeline (31), excluded 606,145 and 579,244 reads from further analyses, respectively. Prior to annotation, 35% of the CHL and 18% of the CHM metagenomes were identified as clusters of artificially replicated sequences and removed using a dereplication pipeline tool (17) (http://microbiomes.msu.edu/replicates). Filter parameters included a cutoff value of 0.9, no length difference requirement, and an initial base pair match of 3 bp (17). The metagenomes generated in this study are freely available from the SEED platform at the MG-RAST website (projects 4470954.3 and 4470937.3).

Metagenome analyses. In all, 362,640 and 155,593 reads averaging 544 and 554 bp for the CHL and CHM metagenomes were used in the metagenomic analyses, respectively (see Table S2 in the supplemental material). Approximately 45% of our reads were annotated (e-value cutoff of $1e^{-05}$) with an assigned function or a specific gene by either the MG-RAST v3.0 pipeline (31) (http://metagenomics.anl.gov) or the RAMMCAP pipeline (25) (http://camera.calit2.net) (see Table S2 in the supplemental material). The MG-RAST v3.0 pipeline analysis included phylogenetic comparisons and functional annotations against the SEED database (31). The RAMMCAP pipeline (i.e., CAMERA) assigned functions by comparison to the Pfam, TIGRfam, and clusters of orthologous groups (COG) databases (25). Prior to quantification, reads were normalized against the total number of hits in their respective databases (e.g., COG, Pfam) as described previously (2). Comparable average genome size permitted us to quantitatively compare the metagenomic data (13). The ChaoI estimators of COG richness were computed with the software SPADE v2.1 (5) (http://chao.stat.nthu.edu.tw) by using the number of individual COGs per unique COG function. The statistical significance of differences between metagenome profiles was calculated on the basis of Fisher's exact test with corrected q values (Storey's false discovery rate [FDR] multiple testing correction approach) using the STAMP v2.0 software package (34).

Nonmetric multidimensional scaling (NMDS) and cluster analysis (CA) based on the relative-abundance data were used to identify the relationships among the community structures of DW metagenomes and 22 publicly available metagenomes, covering a wide variety of habitats (see Table S3 in the supplemental material). Direct comparison of selected metagenomes was performed on the MG-RAST server (31) to avoid potential bias that can be introduced by using different protocols implemented in other annotation pipelines. Prior to quantitative characterization, counts were normalized (relative abundance) against the total number of hits in the respective database and transformed [$\log(x + 1)$]. The comparison was assessed by the Bray-Curtis similarity coefficient of the transformed data using the PAST v2.03 software (20). This estimator compares the structures by accounting for the abundance distributions of attributes. Cluster dendrograms were generated by the unweighted-pair group method using average linkages with the MEGA v5.03 software (49) and using 1,000 replicates to develop bootstrap confidence values. When comparing metagenomes, it is likely that some biases are introduced because of the different protocols implemented for each research project. However, a previous study concluded that there is no evidence of technical bias (e.g., DNA extraction, sequencing protocol, or random genome amplification) for multivariable analysis (47).

Metabolic pathways and taxonomic assignments of functional genes. The entire metabolic pathway for the CHL and CHM samples was annotated using the SEED database and visualized using the KEGG Mapper, an internal tool in the MG-RAST server (31). Sequences assigned to functional groups (e.g., virulence factors and nitrogen) were identified and retrieved from MG-RAST and RAMMCAP output files (see metagenome analysis section). BLASTX analyses were conducted against the NCBI nonredundant protein sequence (nr) database using the CAMERA 2.0 server (48). Assignment and comparison of taxonomic groups and tree representation of the NCBI taxonomy were performed using MEGAN v4.6.3.1 (22).

16S rRNA sequence analysis. Bacterial 16S rRNA gene sequences were aligned using the mothur v1.24.1 software (41) (http://www.mothur.org). Classification and identification of nearest-neighbor sequences were performed using the Classifier tool (Ribosomal Database Project II release 10.26) (7) and BLASTn (1), respectively. Phylogenetic trees were constructed from the alignments based on the maximum-likelihood method and calculated using the Tamura-Nei model (50). The MEGA v5.03 software (49) was used to build trees using 500 replicates to develop bootstrap confidence values.

RESULTS AND DISCUSSION

Comparison of metagenomes. Diversity analyses showed that the DW microbiome is as functionally complex as the distal gut microbiome (ChaoI, ~2,900 COGs), but less diverse than wastewater biofilms (4,122 COGs), whale fall (3,332 COGs), soil (3,394 COGs), and Sargasso Sea samples (3,714 COGs) (16). On the other hand, when the metabolism profiles of DW samples were compared to other metagenomes, the DW samples were similar to soil, freshwater, wastewater and marine metagenomes (Fig. 1A; see Fig. S3A in the supplemental material). Metabolic pathways (i.e., MG-RAST subsystems) common among these environments included nitrogen, potassium, and phosphorous metabolism; motility; metabolism of aromatic compounds; and stress response subsystems. Overall, these data suggest that the diversity of the DW samples is comparable to that of many natural communities and therefore should not be considered a simple system composed of a few bacterial groups with limited functional capabilities. The comparison analyses also suggest that DW sources play an important role in the overall composition of finished DW. The metab-

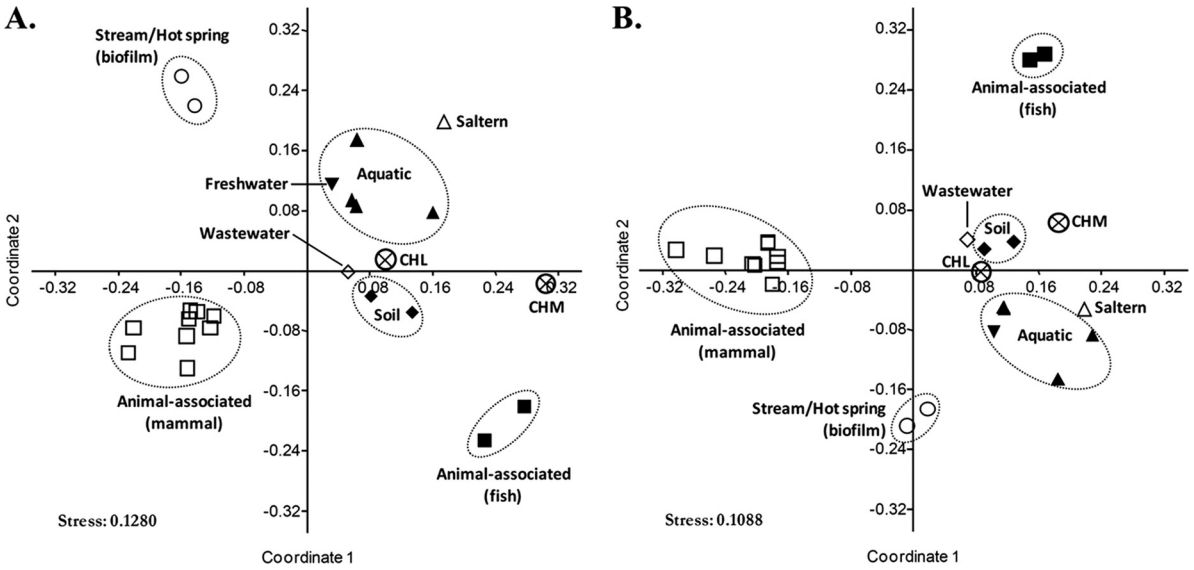

FIG 1 NMDS ordination plot of terrestrial, aquatic, and animal-associated metagenomes based on the relative distribution of 983 functions (i.e., level 3) (A) and taxonomic identification at the genus level of annotated proteins identified by MG-RAST v2.1 (B). Cell division, DNA, RNA, and protein synthesis functions were excluded because they are shared by all bacteria. DW treatments: CHL and CHM. For details of metagenomes, see Table S3 in the supplemental material.

olism of DW metagenomes was less similar to that of animal-associated metagenomes (Fig. 1A; see Fig. S3A in the supplemental material). These gut microbiomes are enriched for a variety of functions involved in pathogenesis, including cell wall and capsule synthesis, phages, prophages, transposable, and plasmids, dormancy and sporulation, iron acquisition and metabolism, and amino acid subsystems. Both NMDS and CA on Bray-Curtis distance (based on taxonomic assignments of annotated proteins) show essentially similar clustering in both community structure and community membership (Fig. 1B; see Fig. S3B in the supplemental material), which is in agreement with previous studies based on 16S rRNA sequences (4). Habitats that are taxonomically similar to DW share a relatively high abundance of some bacterial groups compared to animal-associated metagenomes, specifically for members of the *Alphaproteobacteria* (24% versus 8%, respectively), *Betaproteobacteria* (12% versus 6%), and *Actinobacteria* (11% versus 3%). In contrast, *Bacteroidia* (21% versus 2%), *Clostridia* (13% versus 6%), and *Epsilonproteobacteria* (9% versus 1%) were proportionally more abundant in the animal-associated metagenomes. In general, our results highlight the value of using taxonomic composition based on annotated proteins to ascertain differences and similarities among environmental microbiomes (Fig. 1; see Fig. S3 in the supplemental material).

DSS microbial composition. Differences in community structure between the DW samples were noted when annotated proteins were assigned taxonomic affiliations (Fig. 2; see Fig. S4 in the

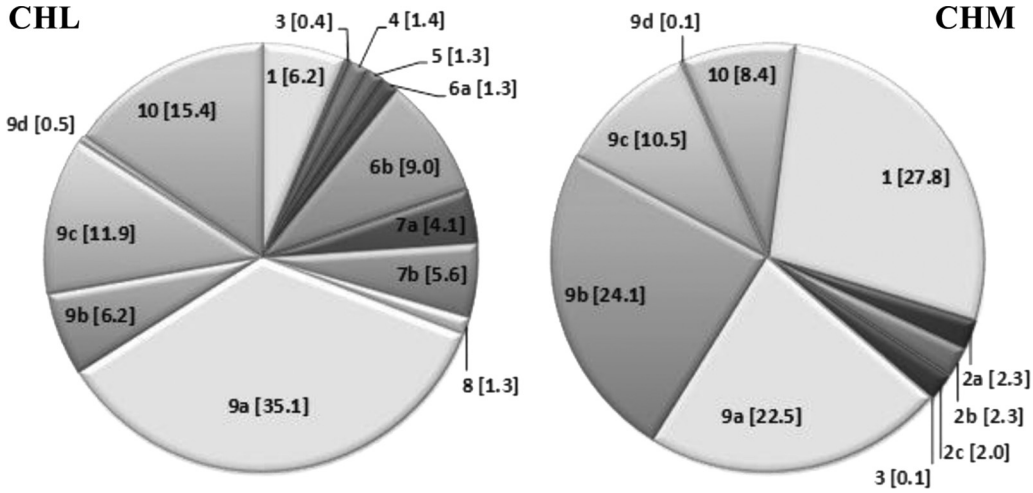

FIG 2 Distribution of members of the *Bacteria* domain as determined by taxonomic identification of annotated proteins (at the class level). Each number in brackets is the percentage of the total number of sequences in each group. *Bacteria* domain: 1, *Actinobacteria*; 2a, *Cytophaga*; 2b, *Flavobacteria*; 2c, *Sphingobacteria*; 3, *Chlamydiae*; 4, *Chlorobia*; 5, *Chloroflexi*; 6a, *Gloeobacteria*; 6b, *Cyanobacteria*; 7a, *Bacilli*; 7b, *Clostridia*; 8, *Planctomycetacia*; 9a, *Alphaproteobacteria*; 9b, *Betaproteobacteria*; 9c, *Deltaproteobacteria*; 9d, *Gammaproteobacteria*; 10, other classes each representing <1%.

supplemental material). Specifically, *Mycobacterium* (*Actinobacteria*), *Acidovorax* (*Betaproteobacteria*), *Burkholderia* (*Betaproteobacteria*), *Pseudomonas* (*Gammaproteobacteria*), and *Dechloromonas* (*Betaproteobacteria*) were dominant in the CHM water, while *Caulobacter* (*Alphaproteobacteria*), *Rhodopseudomonas* (*Alphaproteobacteria*), *Synechococcus* (*Cyanobacteria*), *Bradyrhizobium* (*Alphaproteobacteria*), and *Pseudomonas* (*Gammaproteobacteria*) were the most abundant members in chlorinated water (see Table S4 in the supplemental material). These results further suggest that disinfectants can play a role in the microbial composition of DWDS (33). Differences in community structure were observed when the analysis was performed using 16S rRNA sequences recovered from the DW metagenome libraries (see Fig. S5 in the supplemental material). Some of the DW metagenome 16S rRNA sequences were related to sequences retrieved from 16S rRNA gene clone libraries generated with water samples taken from the same distribution system, including unclassified alphaproteobacteria detected in CHL and CHM water samples (see Fig. S5 in the supplemental material) (37, 54).

In this study, we did not remove free associated DNA or dead cells and therefore it is possible that some of the sequences identified are associated with active/live cells, as well as with dead microbial populations. While DWDS are relatively harsh environments because of oligotrophic conditions and the presence of disinfectants, there is evidence that some bacteria can survive in distribution systems. For example, in a study based on RNA as the target used to generate 16S rRNA gene clone libraries, Revetta et al. (37) showed that proteobacteria were the dominant members within the "active" fraction of the community. Similar groups were identified as dominant in this metagenomic study when the sequences of both 16S rRNA genes and functional genes were analyzed. This was the case not only for the dominant bacterial groups but also for phylogenetic groups such as cyanobacteria and clostridia, providing further evidence that many of these populations might also be active. Future studies using DNA-binding agents such as propidium monoazide (32) and alternate methods such as metatranscriptomics (45) will be useful in further identifying active bacteria in DWDS.

The majority of the annotated proteins were related to the *Bacteria* domain, although eukaryotic, archaeal, and viral proteins were also identified (see Table S2 in the supplemental material). It should be noted that less than 45% of our sequences were annotated with an assigned function using the MG-RAST (SEED database) and CAMERA (Pfam, TIGRfam, and COG databases) pipelines (see Table S2 in the supplemental material). Bacterial genomes dominate these databases, and therefore it is possible that a fraction of these unclassified sequences is associated with genes of yet-to-be-annotated eukaryotic and prokaryotic genomes. Consequently, future studies should focus on sequencing of the genomes of different microbial groups inhabiting DWDS and use a combination of biochemical and molecular assays to confirm predicted gene functions.

Public health implications. Several sequences retrieved from the metagenomes were associated with bacterial groups and genes with potential public health relevance. For example, cyanobacterium-like sequences were found in both types of water samples although they were more abundant in chlorinated water (Table 1). *Synechococcus* was the most frequently identified cyanobacterium. Cyanobacterial sequences affiliated with *Anabaena*, *Gloeobacter*, *Microcystis*, and *Nostoc* were also present. Cyanobacteria are one of the most important groups of toxin-producing aquatic bacteria. Species within the cyanobacterial genera identified in this study have been reported to produce various secondary metabolites, including microcystins, cytotoxins, and neurotoxins (52). Cyanobacterial populations have also been detected in 16S rRNA gene clone libraries of chlorine-treated DW (37). It should be noted that the correct classification of cyanobacteria using 16S rRNA sequencing analyses is cumbersome because of their close similarity to chloroplast rRNA gene sequences (7, 30). Hence, the relative abundance of cyanobacterium-like proteins in this study further supports earlier observations suggesting that cyanobacteria can be inhabitants of DWDS. Moreover, the fact that cyanobacterial sequences have been detected in clone libraries generated with RNA extracts suggests that some species are capable of withstanding harsh environmental conditions (38).

TABLE 1 Distribution of *Mycobacterium* spp., *Legionella* spp., cyanobacteria, and protozoa[a] as determined by taxonomic identification of annotated proteins

Domain and organism	% of total in:	
	CHL water	CHM water
Bacteria		
Mycobacterium	1.29	19.65
Legionella	0.31	0.09
Cyanobacterium	9.18	0.88
Eukaryota		
Amoeba	0.03	<0.001
Ciliate	<0.002	<0.002
Slime mold	0.98	0.02

[a] Members of these protozoan groups are known to harbor *Legionella* spp. (24).

Metagenome analyses also confirmed the ubiquity of mycobacteria in CHM DWDS (Table 1; see Table S4 in the supplemental material). Other microbial diversity studies have documented the presence of mycobacteria in DWDS (43). Among the mycobacterial sequences identified, nontuberculous mycobacteria (NTM) were relatively abundant, specifically, *Mycobacterium mucogenicum* (see Fig. S5 in the supplemental material). Members of the NTM group are considered ubiquitous in the environment and potentially pathogenic to individuals with predisposing conditions (51). A possible survival mechanism of mycobacteria in CHM water is the production of exopolysaccharides (EPS), as they can protect cells from direct exposure to disinfectants (6). Indeed, we detected a high number of sequences associated with *Mycobacterium*-related EPS biosynthesis in CHM water, while in the CHL sample, fewer EPS sequences were detected and most belong to the *Alphaproteobacteria* class (Table 2 and Fig. 3A). EPS biosynthesis is a tightly regulated and energy-intensive process that is responsible for adhesion to surfaces and cohesion in a biofilm and promotes biofilm formation (11, 19). However, it is unknown if the mode of action of chloramine may select for this survival mechanism. A group of virulence factors involved in *Mycobacterium* intracellular parasitism were also identified in the CHM sample, including the mammalian cell entry (MCE) and phospholipid ABC transporter (*yrbE*) proteins. MCE and *yrbE*-encoded proteins confer on *Mycobacterium* spp. the ability to invade and survive inside host cells (39). Most of the sequences were associated with members of the *M. avium* complex, the *M. chelonae* group, *M. smegmatis*, and *M. vanbaalenii* (Fig. 3B and C). The CHM

TABLE 2 Distribution of annotated proteins associated with disinfectant and antibiotic resistance mechanisms

Mechanism	Gene	% of total in: CHL water	% of total in: CHM water	*P* value
Disinfectant resistance				
Glutathione protection				
Glutathione synthetase	*gshB*	0.006	0.007	NS[b]
Glutathione reductase	*gorA*	0.085	0.042	<0.001
Nonspecific DNA-binding protein	*dps*	0.026	0.070	<0.001
OxyR system				
Hydrogen peroxide-inducible regulator	*oxyR*	0.009	0.038	<0.001
Peroxidase/catalase	*katG*	0.134	0.152	NS
Alkyl hydroperoxide reductase protein	*ahpF*	ND[a]	0.004	NS
Glutaredoxin reductase	*grxA*	0.013	0.015	NS
Thioredoxin reductase	*trxB*	0.582	0.095	<0.001
SoxRS system				
Redox-sensitive transcriptional regulator	*soxR*	0.001	0.004	NS
Manganese superoxide dismutase	*sodA*	0.019	ND	NS
RpoS regulated genes				
RNA polymerase sigma factor	*rpoS*	0.012	0.004	NS
Cu-Zn superoxide dismutase	*sodC*	0.002	ND	NS
Exo-DNase III	*xthA*	0.080	0.074	NS
EPS (biofilms)				
EPS transport protein		0.001	0.134	<0.001
Capsular EPS		0.013	0.195	<0.001
Antibiotic resistance				
Beta-lactamase	*bla*	0.076	0.130	<0.01
Sulfonamides, dihydropteroate synthase	*sul*	0.014	0.006	NS
Tetracycline, small GTP-binding protein[c]		0.119	0.065	<0.01
Multidrug efflux pump and transporter				
Multidrug resistance protein (pumps)		0.040	0.114	<0.01
Small multidrug resistance protein	*smr*	0.006	0.002	NS
RND[d] family membrane fusion protein	*mfp*	0.323	0.145	<0.001
Multidrug efflux transporter	*mexF*	0.037	0.025	NS
RND	*rnd*	0.402	0.467	NS
Multidrug efflux pump	*acrBDF*	1.885	1.751	NS
Membrane protein	*marC*	0.010	0.010	NS

[a] ND, not detected.
[b] NS, not significant.
[c] Related to TetM.
[d] RND, resistance-nodulation-cell division.

sample contained mycobacterial genes associated with the production of β-lactamases (Fig. 3D), which is compatible with the intrinsic resistance of several mycobacterial species to β-lactam antibiotics (12).

The metagenomic data suggest that *Mycobacterium* spp. were in lower abundance in the CHL water sample. In contrast, a higher number of sequences associated with *Legionella* and protozoan species known to harbor *Legionella* were retrieved from the CHL water sample (Table 1). The propagation of *Legionella*, a potential pathogen in DWDS, is facilitated by the interactions with protozoa in biofilms, such as amoeba, ciliates, and slime molds (24). Our analysis identified virulence factors and resistance functions associated with *Legionella* spp., such as *dotL* and *icmE* (part of the type IV secretion system); which are required for intracellular growth (56); glutathione peroxidase and reductase, which are involved in the oxidative stress response (26); *helA*, which provides resistance to heavy metals (29); metallo-β-lactamases enzymes involved in the breakdown of antibiotics (14); and *emrA*, a multidrug efflux system (27). A possible explanation for the lower abundance of *Legionella* in the CHM sample may relate to the effectiveness of monochloramine at reducing the amoebal hosts within DW biofilms (55).

Metabolic potential and resistance mechanisms. The multivariable analysis results obtained from whole-metagenome libraries highlighted the potential ecological niche differences within DW distribution systems (Fig. 2; see Fig. S4 and Table S4 in the supplemental material). For example, a greater abundance of reads associated with ammonium transport, nitrification, and denitrification were identified in the CHM sample (Fisher's exact test, $q = 0.05$) (see Fig. S6 in the supplemental material). Overall,

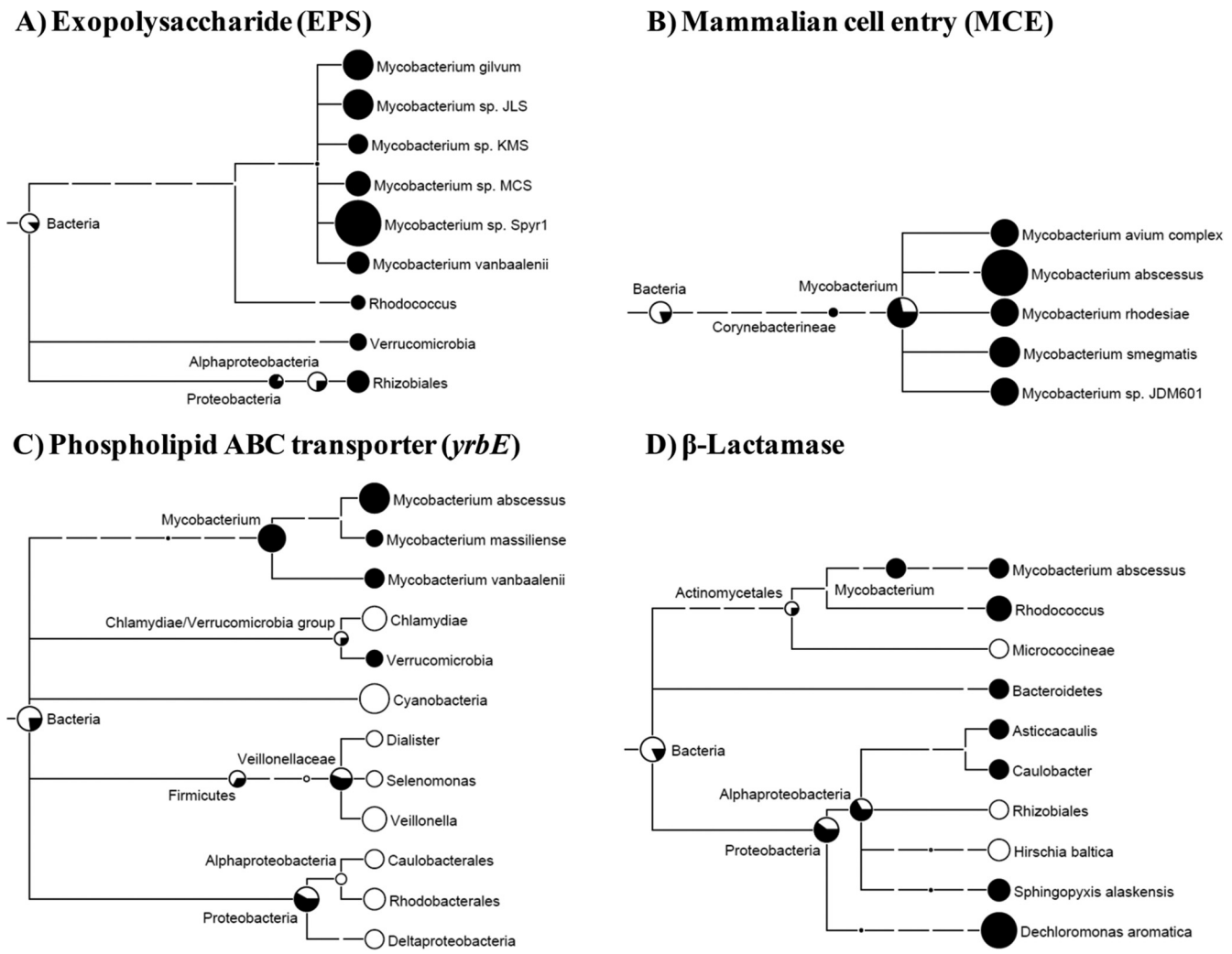

FIG 3 Relative abundance of taxonomic groups based on MEGAN analysis of protein families associated with virulence factors (i.e., pathogenicity) from two DW treatment metagenomes. Each circle is scaled logarithmically to represent the number of reads assigned to each taxonomic group. DW treatments: CHL (white) and CHM (black).

a significant diversity of genes involved in the nitrogen cycle was identified in this study (see Fig. S7 in the supplemental material). In addition, these results indicate the availability of ammonia byproducts during chloramine formation (see Table S1 in the supplemental material) that promote the growth of nitrifying and denitrifying bacteria in CHM distribution systems (36). Bacterial nitrification in DW distribution systems has been associated with loss of disinfection residual, resulting in water quality degradation (8). Among the nitrifying bacteria identified in the CHM sample were members of the genus *Nitrospira* (see Fig. S7 in the supplemental material).

To supplement the results of the analysis based on the SEED subsystems of MG-RAST (see Table S2 in the supplemental material), the data sets were further annotated with KEGG Mapper. Global gene annotation displayed a functionally complex system with a high number of pathways detected in both metagenomes (see Fig. S8 in the supplemental material). The difference between the two treatments is attributable mostly to changes in the distribution of functional genes (Table 2; see Fig. S6 in the supplemental material) combined with the taxonomic affiliation of the genes (Fig. 3; see Fig. S7 in the supplemental material).

Bacterial resistance to disinfectants and antibiotic resistance mechanisms in DW can have significant impacts on human health, as well as serious economic consequences (6). Several disinfectant and antibiotic resistance mechanisms were identified in both libraries, including genes associated with the prevention and repair of radical-induced damage, inactivation and exclusion (i.e., efflux), and bacterial aggregation in biofilms (Table 2). Both DW environments exhibit conditions (see Table S1 in the supplemental material) that are favorable for the establishment of distinct communities harboring virulence factors. In fact, the dominant CHM members were associated with various species of *Mycobacterium*, while CHL is composed predominantly of various members of the *Alphaproteobacteria* (e.g., *Caulobacterales*, *Rhodobacterales*), *Cyanobacteria*, *Deltaproteobacteria*, and *Firmicutes* (Fig. 3). The detection of disinfectant resistance mechanisms suggests that both communities may experience oxidative stress and that different mechanisms may be needed for effective protection against

disinfectants at the community level. While overall the occurrence of most antibiotic resistance mechanisms is not preferential to one particular treatment (Table 2), a few protein groups associated with oxidant defense mechanisms were more abundant in a particular treatment. For example, sequences related to glutathione reductase (*gorA*) and thioredoxin reductase (*trxB*) were more abundant in the CHL sample, while the nonspecific DNA-binding protein expressed in starved cells (*dps*) was more abundant in the CHM metagenomes (Table 2). Specifically, the *gorA* reductase serves as a repair mechanism for the oxidized form of glutathione, a thiol group in many Gram-negative bacteria (46), and the induction of *trxB* reduces disulfide bonds in proteins damaged by oxygen radicals (40). In fact, species associated with the *gorA*-encoded reductase were identified as Gram-negative members of the families *Caulobacteraceae* and *Rhodobacteraceae*. The *dps* gene has also been implicated in the protection of DNA from oxidation damage (28).

Conclusions. Our study is the first to apply next-generation sequencing techniques to characterize the composition and functional diversity of DW bacterial populations in relation to the disinfection strategy applied. Taxonomic profiles based on bacterial functional genes were in general agreement with the previous description of DWDS in which proteobacteria and actinobacteria (to a lesser extent) are predominant members of the DWDS. The impact of lateral gene transfer on the profiles predicted in this study deserves future consideration, particularly as many (but not all) virulence factors might exhibit a higher rate of horizontal gene transfer than other functional genes. However, the taxonomic profiles developed in this study were based on a wide array of genes, for most of which the rate of horizontal transfer is arguably relatively lower, albeit unknown. Since taxonomic calls were made at the class level, this reduces potential assignment errors. In our study, the high abundance of virulence factors annotated for many bacterial groups correlated with the relative abundance of housekeeping genes and other functional genes, as was the case for mycobacterial sequences, suggesting that the probability that these genes came from mycobacterial species is high. It should be noted that horizontal gene transfer is more prominent among closely related taxa as "genome sequence similarity and GC content similarity are strong barriers to lateral gene transfer in prokaryotes" (35).

While the sequencing depth in this study compensated in part for the limited number of samples that were analyzed ($n = 2$), additional metagenomic surveys are needed in order to better understand the total microbial genetic potential of these systems. However, through randomization procedures (e.g., Fisher's exact test with Storey's FDR multiple testing correction approach), we found some statistically distinct functional groups in each of the water samples. By identifying such groups and the genes associated with them, their potential role in bacterial survival of disinfectant treatment can be studied by using multiple genetic assays. The data from this study suggest that disinfection treatments exert an effect on the overall microbial composition and function of DWDS. Furthermore, the data provide examples of multiple resistance mechanisms in DW microbial communities. While the activation, regulation, and selection of a particular mechanism in these treatments remain speculative, the results provide a metagenomic insight into the functional diversity and potential role of specific populations in water distribution systems. The metagenomic analysis further confirmed that DW distribution systems are a potential reservoir of hygienically relevant microorganisms (i.e., those harboring virulence factors). Additional metagenomic approaches will help us develop a more comprehensive understanding of the microbial ecology of DWDS relative to disinfection regimens. Such information is critical to the design of effective management practices and subsequently helps to safeguard human health.

ACKNOWLEDGMENTS

We thank Claudine Curioso and Regina Lamendella for assistance in sample processing. Keith Kelty, David Wahman, and Kayla Quinter provided analytical support.

Although this report was approved for publication by the United States Environmental Protection Agency (USEPA), any opinions expressed in it are ours. They do not necessarily reflect the official positions and policies of the USEPA. Any mention of products or trade names does not constitute endorsement or recommendation for use.

REFERENCES

1. **Altschul SF, et al.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**:3389–3402.
2. **Beszteri B, Temperton B, Frickenhaus S, Giovannoni SJ.** 2010. Average genome size: a potential source of bias in comparative metagenomics. ISME J. **4**:1075–1077.
3. **Bougeard CM, Goslan EH, Jefferson B, Parsons SA.** 2010. Comparison of the disinfection by-product formation potential of treated waters exposed to chlorine and monochloramine. Water Res. **44**:729–740.
4. **Caporaso JG, et al.** 2011. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc. Natl. Acad. Sci. U. S. A. **108**:4516–4522.
5. **Chao A, Shen TJ.** 2005. Program SPADE (Species Prediction and Diversity Estimation) v2.1. Program and user's guide. http://chao.stat.nthu.edu.tw/softwareCE.html.
6. **Chapman JS.** 2003. Disinfectant resistance mechanisms, cross-resistance, and co-resistance. Int. Biodeterior. Biodegradation **51**:271–276.
7. **Cole JR, et al.** 2009. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res. **37**(Database issue):D141–D145.
8. **Cunliffe DA.** 1991. Bacterial nitrification in chloraminated water supplies. Appl. Environ. Microbiol. **57**:3399–3402.
9. **Edwards RA, et al.** 2006. Using pyrosequencing to shed light on deep mine microbial ecology. BMC Genomics **7**:57.
10. **Figueras MJ, Borrego JJ.** 2010. New perspectives in monitoring drinking water microbial quality. Int. J. Environ. Res. Public Health **7**:4179–4202.
11. **Flemming HC, Wingender J.** 2010. The biofilm matrix. Nat. Rev. Microbiol. **8**:623–633.
12. **Flores AR, Parsons LM, Pavelka MS, Jr.** 2005. Genetic analysis of the beta-lactamases of *Mycobacterium tuberculosis* and *Mycobacterium smegmatis* and susceptibility to β-lactam antibiotics. Microbiology **151**:521–532.
13. **Frank JA, Sørensen SJ.** 2011. Quantitative metagenomic analyses based on average genome size normalization. Appl. Environ. Microbiol. **77**:2513–2521.
14. **Garau G, Di Guilmi AM, Hall BG.** 2005. Structure-based phylogeny of the metallo-beta-lactamases. Antimicrob. Agents Chemother. **49**:2778–2784.
15. **Gilbert JA, Dupont CL.** 2011. Microbial metagenomics: beyond the genome. Ann. Rev. Mar. Sci. **3**:347–371.
16. **Gill SR, et al.** 2006. Metagenomic analysis of the human distal gut microbiome. Science **312**:1355–1359.
17. **Gomez-Alvarez V, Teal TK, Schmidt TM.** 2009. Systematic artifacts in metagenomes from complex microbial communities. ISME J. **3**:1314–1317.
18. **Grothues D, Cantor CR, Smith CL.** 1993. PCR amplification of megabase DNA with tagged random primers (T-PCR). Nucleic Acids Res. **21**:1321–1322.
19. **Hall-Stoodley L, Costerton JW, Stoodley P.** 2004. Bacterial biofilms: from the natural environment to infectious diseases. Nat. Rev. Microbiol. **2**:95–108.

20. **Hammer Ø, Harper DAT, Ryan PD.** 2001. PAST: paleontological statistics software package for education and data analysis. Palaeontol. Electron. **4**:1–9.
21. **Hong P-Y, et al.** 2010. Pyrosequencing analysis of bacterial biofilm communities in water meters of a drinking water distribution system. Appl. Environ. Microbiol. **76**:5631–5635.
22. **Huson DH, Mitra S, Ruscheweyh H-J, Weber N, Schuster SC.** 2011. Integrative analysis of environmental sequences using MEGAN 4. Genome Res. **21**:1552–1560.
23. **Lamendella R, Santo Domingo JW, Ghosh S, Martinson J, Oerther DB.** 2011. Comparative fecal metagenomics unveils unique functional capacity of the swine gut. BMC Microbiol. **11**:103. doi:10.1186/1471-2180-11-103.
24. **Lau HY, Ashbolt NJ.** 2009. The role of biofilms and protozoa in *Legionella* pathogenesis: implications for drinking water. J. Appl. Microbiol. **107**:368–378.
25. **Li W.** 2009. Analysis and comparison of very large metagenomes with fast clustering and functional annotation. BMC Bioinformatics **10**:359–367.
26. **Locksley RM, Jacobs RF, Wilson CB, Weaver WM, Klebanoff SJ.** 1982. Susceptibility of *Legionella pneumophila* to oxygen-dependent microbicidal systems. J. Immunol. **129**:2192–2197.
27. **Lomovskaya O, Lewis K.** 1992. *emr*, an *Escherichia coli* locus for multidrug resistance. Proc. Natl. Acad. Sci. U. S. A. **89**:8938–8942.
28. **Martinez A, Kolter R.** 1997. Protection of DNA during oxidative stress by the nonspecific DNA-binding protein Dps. J. Bacteriol. **179**:5188–5194.
29. **McClain MS, Hurley MC, Brieland JK, Engleberg NC.** 1996. The *Legionella pneumophila hel* locus encodes intracellularly induced homologs of heavy-metal ion transporters of *Alcaligenes* spp. Infect. Immun. **64**:1532–1540.
30. **McDonald D, et al.** 2012. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J. **6**:610–618.
31. **Meyer F, et al.** 2008. The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics **9**:386–394.
32. **Nocker A, Sossa-Fernandez P, Burr MD, Camper AK.** 2007. Use of propidium monoazide for live/dead distinction in microbial ecology. Appl. Environ. Microbiol. **73**:5111–5117.
33. **Norton CD, LeChevallier MW.** 2000. A pilot study of bacteriological population changes through potable water treatment and distribution. Appl. Environ. Microbiol. **66**:268–276.
34. **Parks DH, Beiko RG.** 2010. Identifying biologically relevant differences between metagenomic communities. Bioinformatics **26**:715–721.
35. **Popa O, Hazkani-Covo E, Landan G, Martin W, Dagan T.** 2011. Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. Genome Res. **21**:599–609.
36. **Regan JM, Harrington GW, Baribeau H, De Leon R, Noguera DR.** 2003. Diversity of nitrifying bacteria in full-scale chloraminated distribution systems. Water Res. **37**:197–205.
37. **Revetta RP, Matlib RS, Santo Domingo JW.** 2011. 16S rRNA gene sequence analysis of drinking water using RNA and DNA extracts as targets for clone library development. Curr. Microbiol. **63**:50–59.
38. **Revetta RP, Pemberton A, Lamendella R, Iker B, Santo Domingo JW.** 2010. Identification of bacterial populations in drinking water using 16S rRNA-based sequence analyses. Water Res. **44**:1353–1360.
39. **Ripoll F, et al.** 2009. Nonmycobacterial virulence genes in the genome of the emerging pathogen *Mycobacterium abscessus*. PLoS One **4**:e5660. doi:10.1371/journal.pone.0005660.
40. **Russel M, Model P.** 1988. Sequence of thioredoxin reductase from *Escherichia coli*. Relationship to other flavoprotein disulfide oxidoreductases. J. Biol. Chem. **263**:9015–9019.
41. **Schloss PD, et al.** 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. **75**:7537–7541.
42. **Schmeisser C, et al.** 2003. Metagenome survey of biofilms in drinking-water networks. Appl. Environ. Microbiol. **69**:7298–7309.
43. **September SM, Brözel VS, Venter SN.** 2004. Diversity of nontuberculoid *Mycobacterium* species in biofilms of urban and semiurban drinking water distribution systems. Appl. Environ. Microbiol. **70**:7571–7573.
44. **Shanks OC, Santo Domingo JW, Graham JE.** 2006. Use of competitive DNA hybridization to identify differences in the genomes of bacteria. J. Microbiol. Methods **66**:321–330.
45. **Shi Y, Tyson GW, DeLong EF.** 2009. Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column. Nature **459**:266–269.
46. **Smirnova GV, Krasnykh TA, Oktyabrsky ON.** 2001. Role of glutathione in the response of *Escherichia coli* to osmotic stress. Biochemistry (Mosc) **66**:973–978.
47. **Smith RJ, et al.** 2012. Metagenomic comparison of microbial communities inhabiting confined and unconfined aquifer ecosystems. Environ. Microbiol. **14**:240–253.
48. **Sun S, et al.** 2011. Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. Nucleic Acids Res. **39**(Database issue):D546–D551.
49. **Tamura K, et al.** 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. **28**:2731–2739.
50. **Tamura Nei KM, Kumar S.** 2004. Prospects for inferring very large phylogenies by using the neighbor-joining method. Proc. Natl. Acad. Sci. U. S. A. **101**:11030–11035.
51. **Vaerewijck MJ, Huys G, Palomino JC, Swings J, Portaels F.** 2005. Mycobacteria in drinking water distribution systems: ecology and significance for human health. FEMS Microbiol. Rev. **29**:911–934.
52. **Valério E, Chaves S, Tenreiro R.** 2010. Diversity and impact of prokaryotic toxins on aquatic environments: a review. Toxins **2**:2359–2410.
53. **Williams MM, Braun-Howland EB.** 2003. Growth of *Escherichia coli* in model distribution system biofilms exposed to hypochlorous acid or monochloramine. Appl. Environ. Microbiol. **69**:5463–5471.
54. **Williams MM, Santo Domingo JW, Meckes MC, Kelty CA, Rochon HS.** 2004. Phylogenetic diversity of drinking water bacteria in a distribution system simulator. J. Appl. Microbiol. **96**:954–964.
55. **Zhang W, DiGiano FA.** 2002. Comparison of bacterial regrowth in distribution systems using free chlorine and chloramine: a statistical study of causative factors. Water Res. **36**:1469–1482.
56. **Zink SD, Pedersen L, Cianciotto NP, Abu-Kwaik Y.** 2002. The Dot/Icm type IV secretion system of *Legionella pneumophila* is essential for the induction of apoptosis in human macrophages. Infect. Immun. **70**:1657–1663.