Microbiome

RESEARCH Open Access

Disinfection exhibits systematic impacts on the drinking water microbiome

Zihan Dai[1], Maria C. Sevillano-Rivera[2], Szymon T. Calus[1], Q. Melina Bautista-de los Santos[3], A. Murat Eren[4,5], Paul W. J. J. van der Wielen[6,7], Umer Z. Ijaz[1] and Ameet J. Pinto[2*]

Abstract

Limiting microbial growth during drinking water distribution is achieved either by maintaining a disinfectant residual or through nutrient limitation without using a disinfectant. The impact of these contrasting approaches on the drinking water microbiome is not systematically understood. We use genome-resolved metagenomics to compare the structure, metabolic traits, and population genomes of drinking water microbiome samples from bulk drinking water across multiple full-scale disinfected and non-disinfected drinking water systems. Microbial communities cluster at the structural- and functional potential-level based on the presence/absence of a disinfectant residual. Disinfectant residual alone explained 17 and 6.5% of the variance in structure and functional potential of the drinking water microbiome, respectively, despite including multiple drinking water systems with variable source waters and source water communities and treatment strategies. The drinking water microbiome is structurally and functionally less diverse and variable across disinfected compared to non-disinfected systems. While bacteria were the most abundant domain, archaea and eukaryota were more abundant in non-disinfected and disinfected systems, respectively. Community-level differences in functional potential were driven by enrichment of genes associated with carbon and nitrogen fixation in non-disinfected systems and γ-aminobutyrate metabolism in disinfected systems likely associated with the recycling of amino acids. Genome-level analyses for a subset of phylogenetically-related microorganisms suggests that disinfection selects for microorganisms capable of using fatty acids, presumably from microbial decay products, via the glyoxylate cycle. Overall, we find that disinfection exhibits systematic selective pressures on the drinking water microbiome and may select for microorganisms able to utilize microbial decay products originating from disinfection-inactivated microorganisms.

Keywords: Disinfection, Drinking water microbiome, Selection, Metagenomics

Introduction

Drinking water systems harbor diverse and complex microbial communities in bulk water, biofilms on pipe wall, suspended solids, and in loose deposits [1–5]. While treatment processes at the drinking water treatment plants (DWTPs) shape the microbial community that leaves the DWTP [6–9], multiple factors can influence the structure and function of the drinking water microbiome in the drinking water distribution systems (DWDSs). These factors include, but are not limited to, DWDS size, pipe materials and ages, water age within the DWDS, and similar factors within premises plumbing (PP) in buildings and homes [10–14]. Managing the microbiological quality of drinking water during transport through the DWDS and into the PP is essential for the provision of safe drinking water. Unwanted microbial growth and/or changes in the drinking water microbiome composition during transit through the DWDS and PP are associated with several detrimental outcomes. For instance, this could lead to proliferation of

* Correspondence: a.pinto@northeastern.edu
[2]Department of Civil and Environmental Engineering, Northeastern University, Boston, MA, USA
Full list of author information is available at the end of the article

opportunistic pathogens [15–19] and eukaryotic microbes [14, 16, 20, 21], taste and odor issues [22], and impact infrastructure via corrosion damage [23, 24].

Source-to-tap differences in drinking water systems can range from source water type (e.g., surface, ground, reuse water), process configurations at the DWTP, and heterogeneity and condition of the DWDS and PP, yet globally there are two fundamental approaches for managing the drinking water microbiome during transport to the consumer [25]. The first and most widely used approach involves maintenance of a disinfectant residual (e.g., chlorine) in the DWDS. This is accomplished by ensuring the water leaving the DWTP has a chlorine residual and/or by using booster stations in large complex DWDSs to compensate for disinfectant residual decay [26]. Disinfectant residuals counteract microbial growth through inactivation, thus ensuring stable microbial concentrations during distribution. While disinfectant residuals are effective in managing microbial growth in the DWDS, there are some key issues associated with them. These include esthetic and corrosion related problems [25, 27, 28] but more importantly the formation of harmful disinfection byproducts (DBPs) [29–31], which are also regulated. Further, there is an increasing recognition that the disinfectant residuals may be associated with selection of some opportunistic pathogens [16, 32] and antibiotic resistance genes (ARGs) in drinking water [33–35].

The second approach for managing microbial growth in the DWDS, primarily practiced in parts of Western Europe (e.g., Netherlands, Denmark, and Switzerland), involves distribution of drinking water without any disinfectant residuals [36]. These systems focus on minimizing nutrient availability in the DWDS to limit microbial growth using high-quality source waters and/or multi-barrier treatment. While some of these drinking water systems may also use chlorine or other chlorine compounds (e.g., chlorine dioxide) at the DWTP, they ensure that chlorine is not detectable prior to distribution. The efficacy of this approach is supported by evidence that incidences of microbial contamination and associated waterborne illnesses are comparable to systems that maintain a disinfectant residual [25, 37]. This suggests that with appropriate source water quality management, treatment, and well-maintained infrastructure, drinking water can be safely distributed without disinfectant residuals [25].

Despite reports of comparable biological water quality between systems with and without disinfectant residuals, there are a limited number of studies that have systematically compared the microbial community between these two types of systems. Bautista et al. [38] conducted a meta-analyses study involving collation, curation, and comparison of 16S rRNA gene amplicon sequencing data from previously published datasets. While this study was confounded by methodological differences between datasets being used, the key conclusions were that presence/absence of disinfectant residuals impacts microbial community structure and membership and that systems without disinfectant residuals are more diverse than their disinfected counterparts. Recently, Waak et al. [39] compared biofilms between two drinking water systems, one chloraminated systems and one without a disinfectant residual. Consistent with previous findings, they observed higher cell numbers and higher diversity in the system without disinfectant residual, with higher proportional abundance (proportion of total community) of deleterious microbes (i.e., mycobacteria, nitrifiers, corrosion causing bacteria) in the chloraminated system. Both, Bautista et al. [38] and Waak et al. [39] utilized gene-targeted assays (i.e., 16S rRNA gene) to probe drinking water microbiome composition and its differences. While gene-targeted assays can provide valuable information on microbial community structure and membership information, they do not provide insight into metabolic differences that may drive the observed differences in community structure. Further, gene-targeted assays can be limited by primer-bias and can result in non-detection of microbial community members. Both challenges can be overcome by utilizing metagenomics which can provide insights into structure and functional potential of microbial communities without being bias against or towards specific community members. This comes with the limitation that differences between samples/systems emerging from low-abundance microbes may not be detected as this may require ultra-deep sequencing. Further, it is important to note that current sequencing-based approaches (e.g., 16S rRNA gene amplicon sequencing and metagenomics) only provide relative abundance of taxa or genes which are inherently compositional in nature. Although these microbiome characterization approaches are powerful, they do not capture quantitative differences between microbial communities and this would require complimenting with quantitative molecular assays (e.g., quantitative PCR).

We used metagenome analyses and genome-resolved metagenomics to investigate the potential influence of disinfectant residuals on the drinking water microbiomes by comparing drinking water systems from the UK (with disinfectant residual) and the Netherlands (without disinfectant residual). The goals of this study were (1) to determine the extent to which disinfectant residual shapes the structure and functional potential of the drinking water microbiome, (2) to determine whether the selective pressures of disinfection are conserved across drinking water systems, and (3) to identify metabolic pathways underpinning differences in structure and functional potential of the drinking water microbiome. Addressing these questions across different drinking water systems with inherent system-to-system variability (e.g., source water, water chemistry, treatment process) but one consistent difference—i.e., presence or absence of disinfectant residual—will help highlight disinfection that are conserved and thus generalizable across systems.

Results and discussion

Water quality parameters across disinfected and non-disinfected DWDS

Sampling was conducted in seven DWDSs with disinfectant residual between April–August of 2013 and at five DWDSs without disinfectant residual between October–December 2015. The water chemistry varied between the DWDSs considering they were supplied by different DWTPs, our sampling campaign also captures seasonal differences between locations (Fig. 1) (Table S1). Specifically, water temperatures were higher (~ 5 °C) for the disinfected samples compared to the non-disinfected samples. While the pH, DO, nitrogen species (i.e., ammonium and nitrate), and TOC measurements were not significantly different between disinfected and non-disinfected samples, the measured phosphate and total chlorine concentrations were significantly different (p < 0.05). Specifically, the average total chlorine concentrations in disinfected systems was 0.37 mg Cl_2/l (range 0.1–0.73 mg Cl_2/l) while no disinfectant residuals were measurable in the non-disinfected systems. The average phosphate concentrations were 2.3 mg PO_4^{3-}/l while no phosphate was measurable in non-disinfected samples. Phosphate was higher in the disinfected systems as it is likely to be used for corrosion control [40]. While we were unable to obtain information on source water type (i.e., ground vs surface water) used for production of drinking water supplied to the sampled DWDS, conductivity measurements suggested DWDS in both type systems were supplied by a DWTPs drawing from surface and groundwater sources (Fig. 1).

Summary of metagenomic data set

Metagenomic analyses were used to assess the association between presence/absence of disinfectant residual with the structure and functional potential of the drinking water microbiome. A total of 41 drinking water samples were collected from DWDSs (i.e., chlorine) from the UK (n = 23), while those collected from the Netherlands (n = 18) did not have a disinfectant residual. Quality trimming of raw metagenomic data resulted in the retention of 638 million paired-end reads. Co-assembly for each drinking water system was carried out by combining reads from individual sampling location within each drinking water system (Table 1). De novo co-assembly generated 0.04–1.81 million true scaffolds for each sampling location after discarding scaffolds shorter than 500 bp and contaminant scaffolds (Table 1) with an N50 value ranged from 775 to 3300 bp. The proportion of

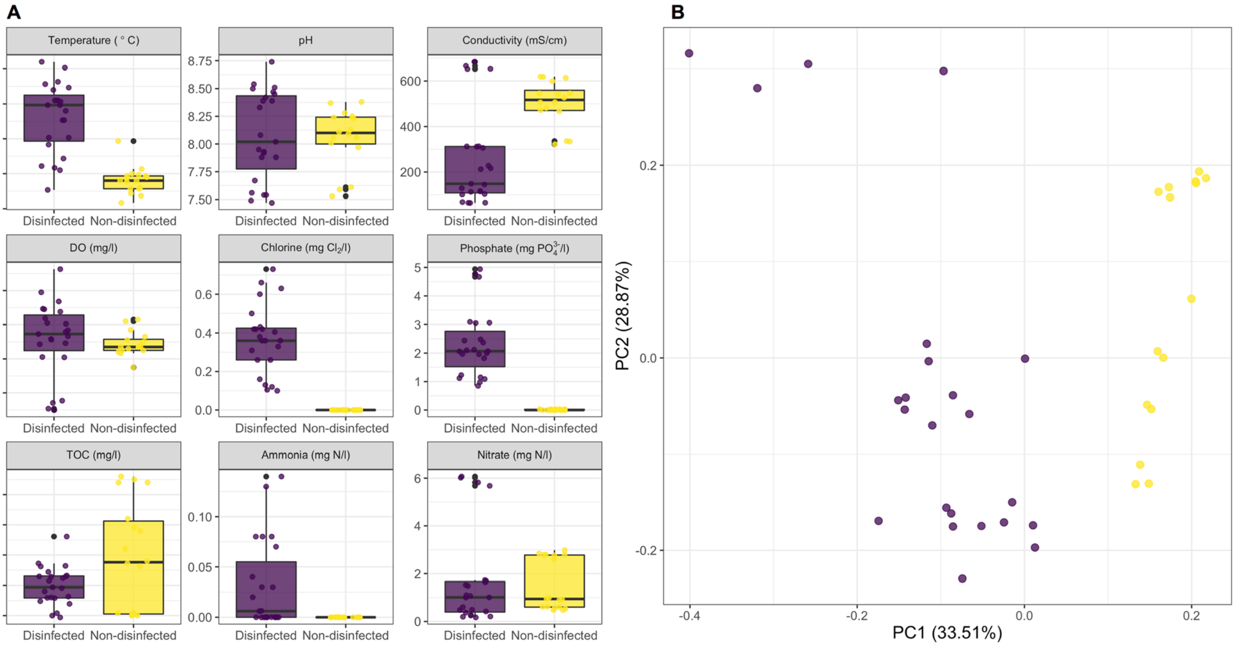

Fig. 1 Summary of water chemistry parameters measured for samples collected from disinfected (purple) and non-disinfected systems (yellow). **b** Principal component analyses using Euclidean distances for measured water chemistry parameters indicates distinct clustering of samples from disinfected and non-disinfected systems

Table 1 Sequencing and de novo co-assembly statistics for metagenomes from 12 drinking water systems

Drinking water system	Paired-end reads (millions)	Scaffolds (> 500 bp)	True scaffolds	True scaffold assembly size (Mbp)	% Mappedreads	GC content (%)	N50 (bp)	ORFs per Mbp	Coding density
D1	195.73	555493	546375	615.10	99.02	54.66	1131	1403.68	0.48
D2	46.87	38567	36733	53.03	96.24	55.34	2112	1419.84	0.64
D3	17.40	192457	190882	249.69	91.15	57.82	1531	1498.24	0.63
D4	36.01	123852	122486	204.78	93.03	57.57	3300	1316.54	0.60
D5	36.74	227196	225149	269.12	88.73	59.09	1313	1527.12	0.60
D6	17.39	42209	41459	57.23	95.89	59.16	1641	1504.23	0.65
D8	19.4	77973	76996	108.07	95.38	61.07	1751	1475.71	0.68
ND1	45.52	521371	517773	472.02	83.82	53.75	855	1803.21	0.61
ND2	25.98	363819	361304	316.18	75.03	53.44	802	1807.05	0.56
ND3	48.63	667992	663968	562.73	81.63	52.93	775	1838.06	0.60
ND4	17.78	164328	163361	143.22	66.73	56.48	808	1822.84	0.63
ND5	130.92	1812573	1804048	1834.75	93.74	56.38	1005	1672.04	0.60

D disinfected, *ND* non-disinfected, *N50* minimum contig length that account for 50% of the bases, *ORF* open reading frame

quality-trimmed reads mapping back to true scaffolds ranged from 67 to 99% (Table 1) across all samples.

Non-disinfected systems are more diverse than disinfected systems

Non-disinfected systems were significantly ($p < 0.0001$) more diverse compared to systems that maintained a disinfectant residual (Fig. 2a) based on the Nonpareil estimated diversity index [41]. This observation is consistent with previous comparisons of bulk water [42] and biofilm [39] samples from disinfected and non-disinfected systems. Lower diversity in disinfected systems is likely due to stronger selective pressure of the disinfectant residual as compared to that nutrient limitation in non-disinfected systems. As a result of the higher diversity in non-disinfected systems, the metagenomic sequencing for these samples provided significantly lower coverage of the sampled microbial community (Fig. 2b) as compared to systems with a disinfectant residual ($p < 0.0001$).

Microbial community membership and structure is different between disinfected and non-disinfected systems

We used 2872 small-subunit (SSU) rRNA genes (2742 genes > 100 bp) identified on the assembled scaffolds to determine community membership and structure across sampling locations. While bacteria were dominant members of the drinking water microbiome in both types of systems (2C, 2D), the relative abundance of archaea and eukaryota were dependent on the presence/absence of disinfectant residual (Fig. 2c, e). Specifically, the relative abundance of eukaryota was higher in disinfected systems as compared to non-disinfected system (2C), while archaea were ubiquitous across non-disinfected samples (Fig. 2c, e) they were only detected in a single disinfected sample (D2). Non-disinfected systems were taxonomically more diverse, with respect to bacteria and archaea, as compared to disinfected systems. Specifically, a total of 14 bacterial and 6 archaeal phyla were detected in one or more non-disinfected systems that were not detected in any of the disinfected systems. Several of these unique phyla, while not dominant in non-disinfected systems, were detected at relative abundances between 1–5% (e.g., *Nitrospirae*, *Nanoarchaeota*).

The bacterial community was dominated by *Proteobacteria*, in particular *Alphaproteobacteria* and *Gammaproteobacteria*, in both disinfected and non-disinfected systems with *Deltaproteobacteria* being much more prevalent and abundant in non-disinfected systems (Fig. 2d). *Actinobacteria* were more abundant than *Proteobacteria* in two drinking water systems and constituted 44% and 33% of the community in systems D4 and ND1, respectively. Overall, the relative abundance of *Proteobacteria* was higher in disinfected systems, ranging from 28 to 90%, as compared to non-disinfected systems, ranging from 30 to 57%. *Patescibacteria* was the second most abundant phylum across non-disinfected systems, constituting 15 to 29% of the SSU rRNA genes, while they were only detected in one disinfected sample (D2) with a relative abundance of 1%. Within *Patescibacteria*, *Parcubacteria* were the most commonly detected phyla followed by *Microgenomatia* and *Gracilibacteria*.

The observed differences between disinfected and non-disinfected DWDS for bacteria and archaea are largely consistent with previous meta-analyses of amplicon sequencing data from the 16S rRNA gene [42]. In contrast to bacteria and archaea, results from eukaryotes,

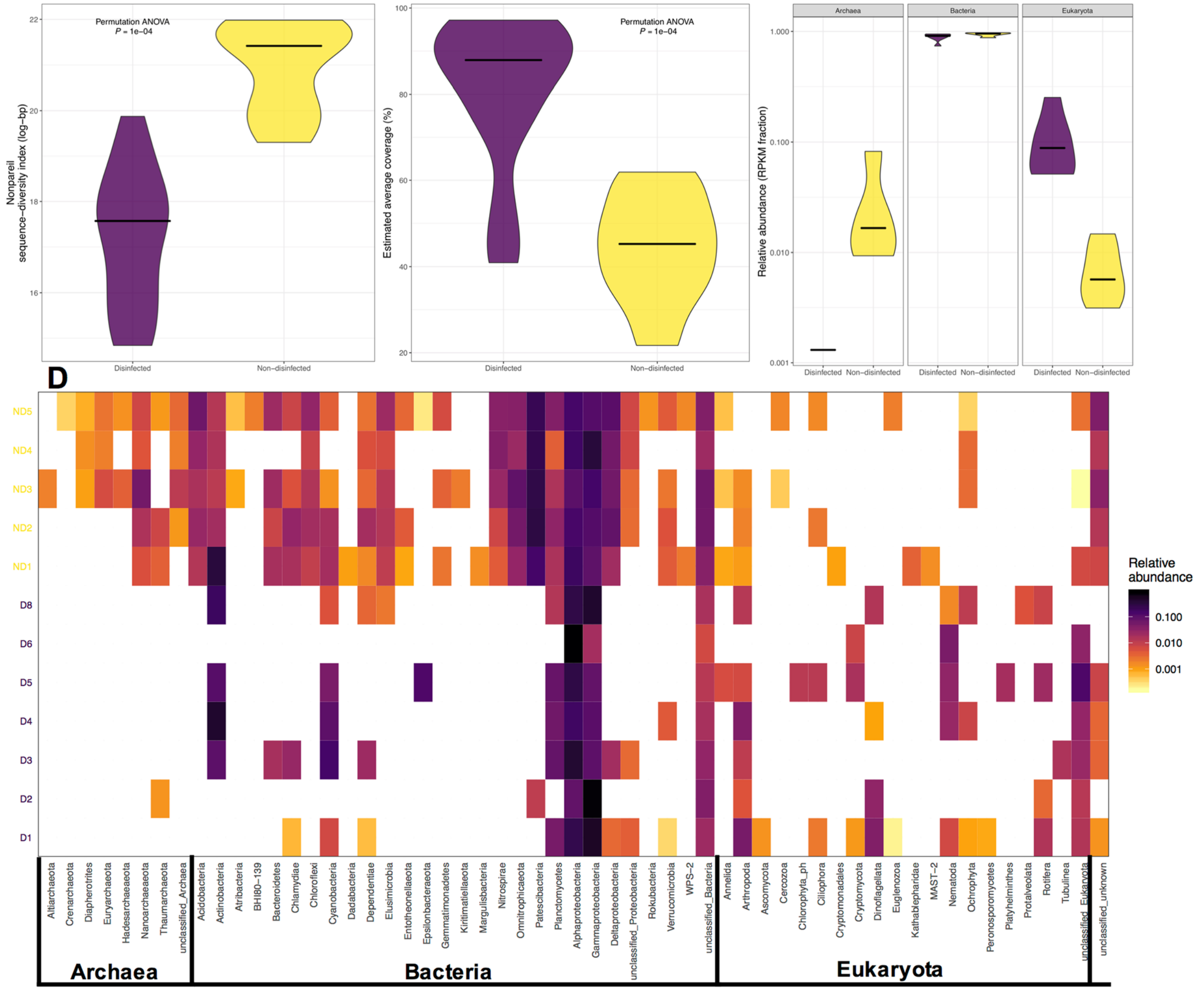

Fig. 2 Comparison of **a** diversity and **b** coverage between disinfected and non-disinfected drinking water systems estimated using Nonpareil. **c** Comparison of relative abundance of bacterial, archaeal, and eukaryotic communities in drinking water systems with and without disinfectant residuals. **d** Log10 transformed relative abundance of different phyla (classes for phylum Proteobacteria) across sampling location for the bacteria, archaea, and eukaryota

which have not been systematically investigated in the drinking water microbiome, were surprising in terms of their higher relative abundance eukaryotic in disinfected as compared to non-disinfected systems. For instance, SSU rRNA genes associated *Nematoda* were detected in nearly every disinfected system, but were not detected in non-disinfected systems. Specifically, SSU rRNA genes from two free-living nematode genera, i.e., *Araeolaimida* and *Monhysterida*, were detected in five of the eight disinfected systems. Similarly, SSU rRNA genes from the phylum *Rotifera* were only detected in disinfected systems and were largely associated with the monogonont rotifers within the genera *Ploimida*. While the relative abundance of scaffolds determined to be of eukaryotic in origin was higher in disinfected compared to non-disinfected systems, this does not mean that eukaryotes were proportionally larger part of the drinking water microbiome in disinfected compared to the non-disinfected systems. Genome sizes of picoeukaryotic microbes can be orders of magnitude larger than that of bacteria and archaea and vary significantly between picoeukaryotes themselves. Further, the higher overall diversity and lower sequencing coverage (Fig. 1) could also have resulted in under sampling of the eukaryotic community in non-disinfected systems.

Drinking water systems cluster at the nucleotide level based on presence/absence of disinfectant residuals

Samples (for read-based analyses) and drinking water systems (for scaffold-based analyses) clustered with each other based on the presence/absence of disinfectant residuals (Fig. 3a, b) based on Mash distance estimates.

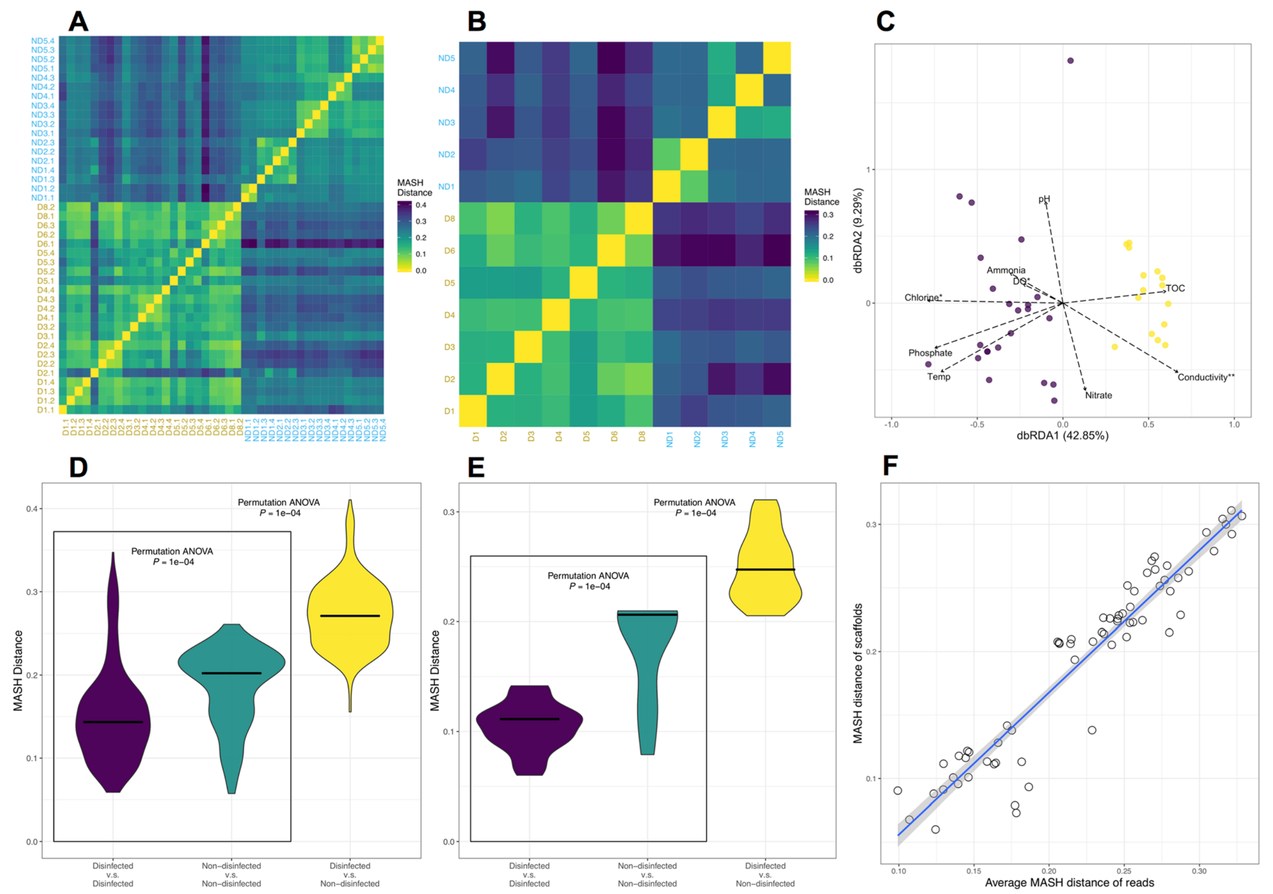

Fig. 3 Comparison of nucleotide composition using paired reads each from each sample and true scaffolds in each drinking water system according to Mash distance. **a**, **b** Heatmaps based on pairwise Mash distances of reads and scaffolds. Heatmaps are colored according to Mash distance; yellow denotes a distance of 0. Labels on x- and y-axis are colored according to disinfection strategies. **c** MDS clustering of read-based Mash distances between samples with vectors representing water chemistry/environmental parameters implemented using dbRDA. **d**, **e** Violin plots indicating the distribution of pairwise Mash distances of reads and scaffolds. Plots are colored according to the system type for which pairwise comparisons were conducted. Purple denotes comparisons between disinfected samples, yellow denotes comparisons between non-disinfected samples, and green denotes comparisons between disinfected and non-disinfected samples. **f** Correlation between average Mash distances of reads across samples and Mash distances of scaffolds across sampling locations

We further evaluated the significance and explanatory power of measured water chemistry parameters in explaining the observed clustering between disinfected and non-disinfected systems. To do this, we initially performed BioEnv analyses to identify water chemistry parameters and their combinations that were highly correlated with observed Mash distances between samples (Table S2). This identified chlorine as being strongly correlated with the Mash distances between samples ($R = 0.54$, $p < 0.001$) while the maximum correlation between water chemistry and Mash distances was observed for a combination of chlorine, phosphate, and TOC ($R = 0.62$, $p < 0.001$). We subsequently utilized dbRDA to independently determine the environmental/water chemistry variables most significantly associated with Mash distances between samples. While chlorine was identified as a significant variable ($p < 0.01$), dbRDA identified conductivity ($p < 0.001$) and DO ($p < 0.01$) as significant variables (Table S3). Finally, variance partitioning analysis was used to determine the proportion of variance in the Mash distance matrices explained by individual and combination of variables identified as significant by dbRDA (Table S4). This resulted in chlorine, conductivity, and DO individually explaining ~ 17%, 12%, and 1% of the variance in the Mash distance matrix, with ~ 60% of the variance unexplained by these three variables.

We further compared the distribution of Mash distances between drinking water metagenomes within disinfected, within non-disinfected, and between disinfected and non-disinfected systems. Mash distances between drinking water metagenomes from disinfected systems were significantly different ($p < 0.0001$) and exhibited a lower mean for disinfected as compared to non-disinfected systems. Further, the pairwise Mash distances between disinfected and non-disinfected systems were significantly different and higher from those estimated within each category (i.e., disinfected or non-disinfected). This was consistent for both read- and scaffold-based analyses (Fig. 3d, e). Finally, the average pairwise Mash distances estimated using reads (i.e., between samples) and scaffolds (i.e., between DWDSs) were highly correlated (Pearson's $R = 0.95$, $P < 0.05$) (Fig. 3c), indicating the de novo assembly process did not result in loss of information on factors driving the differences between disinfected and non-disinfected systems.

These analyses provide a few key insights. First, Mash distance-based (both read and scaffold based) clustering of samples occurs depending on presence and absence of disinfectant residual suggests that the microbial communities are more similar within each group (i.e., disinfected and non-disinfected) and dissimilar between the two groups (i.e., disinfected vs non-disinfected). Second, while disinfected and non-disinfected samples cluster distinctly from each other, disinfected systems exhibit lower nucleotide-level heterogeneity as compared to their non-disinfected systems indicating that the factors governing microbial community in disinfected systems likely impose stronger selective pressures on the microbial community as compared to those in non-disinfected systems. Third, non-disinfected systems exhibit greater diversity not only within a system (Fig. 2) but also across systems as compared to disinfected systems. Despite the strong correlation between pairwise Mash distances of reads and scaffolds (Fig. 3f), the median Mash distances for pairwise comparison of samples within each type of system (i.e., disinfected and non-disinfected) is higher for the scaffold-based analyses as compared to the read-based analyses. This is likely from the omission of low abundance microorganisms during de novo assembly and thus suggests that composition of medium-to-high abundance organisms is likely to be more variable between non-disinfected systems as compared to disinfected systems.

Finally, while the water chemistry and environmental parameters between disinfected and non-disinfected systems were distinct (Fig. 1b), the parameters that most strongly correlated with Mash distances between samples were limited to a combination of chlorine, phosphate, and TOC for BioEnv analyses and chlorine, conductivity, and DO based on dbRDA. Both independent exploratory analyses consistently identified chlorine presence/absence and concentration as one of the key drivers of the difference in microbial communities across the samples. Further, variance partition analysis indicated that ~ 17% of the variance in the Mash distance matrix was driven exclusively by chlorine; this makes chlorine the most important parameter measured as part of this study in terms of differentiating between drinking water metagenomes. The significance of phosphate determined by BioEnv analyses is likely because chlorine and phosphate concentrations are inherently associated due to common use of the latter for corrosion control in DWDSs that maintain a chlorine residual [40]. Further, while it is unlikely that DO (identified as significant by dbRDA) directly affects microbial community composition (all DO concentrations were near or greater than saturation), it is possible that this may reflect the use of advanced oxidation process (e.g., ozonation) during drinking water treatment. Similarly, conductivity (identified as significant by dbRDA) is unlikely to directly influence the microbial community, but rather this may reflect the source water type and treatment processes being used for drinking water production. Specifically, source water derived from groundwater sources or from reservoirs under the influence of groundwater typically have much higher conductivities than those that rely on surface water supply. Similarly, chemicals used for softening and coagulation/flocculation processes may influence water conductivity. Thus, we speculate that the influence of conductivity may serve as a surrogate for a combination of source water and treatment process. These analyses clearly identify chlorine as one of the major measured parameters driving the Mash distances between samples, followed by conductivity (a potential surrogate for source water and treatment process). Further, the fact that the major proportion of the variance remains unexplained suggests that additional aspects such as treatment process configuration, DWDS characteristics, and other water chemistry parameters which were not characterized/measured as part of this study also likely play a strong role in differentiating between microbial communities in disinfected and non-disinfected drinking water systems.

Protein coding sequences cluster based on the presence/absence of disinfectant residuals

A total of 8 million protein coding sequences were predicted and translated from true scaffolds, of which approximately 17 to 27% were annotated against KEGG database (Table S5). Consistent with the nucleotide-level analyses, samples clustered based on the presence and absence of disinfectant residual (Fig. 4a, b, c) rather than by DWDS. Further, BioEnv analyses identified the combination of chlorine, phosphate, and ammonia as being strongly and significantly correlated ($R = 0.392$, $p < 0.001$) with Bray-Curtis distances between samples estimated using abundance (i.e., RPKM) of KOs (Table S6). Similar to nucleotide-based analyses, chlorine presence/absence and concentration was the measured parameter more strongly and significantly associated with differences in functional potential between samples at the

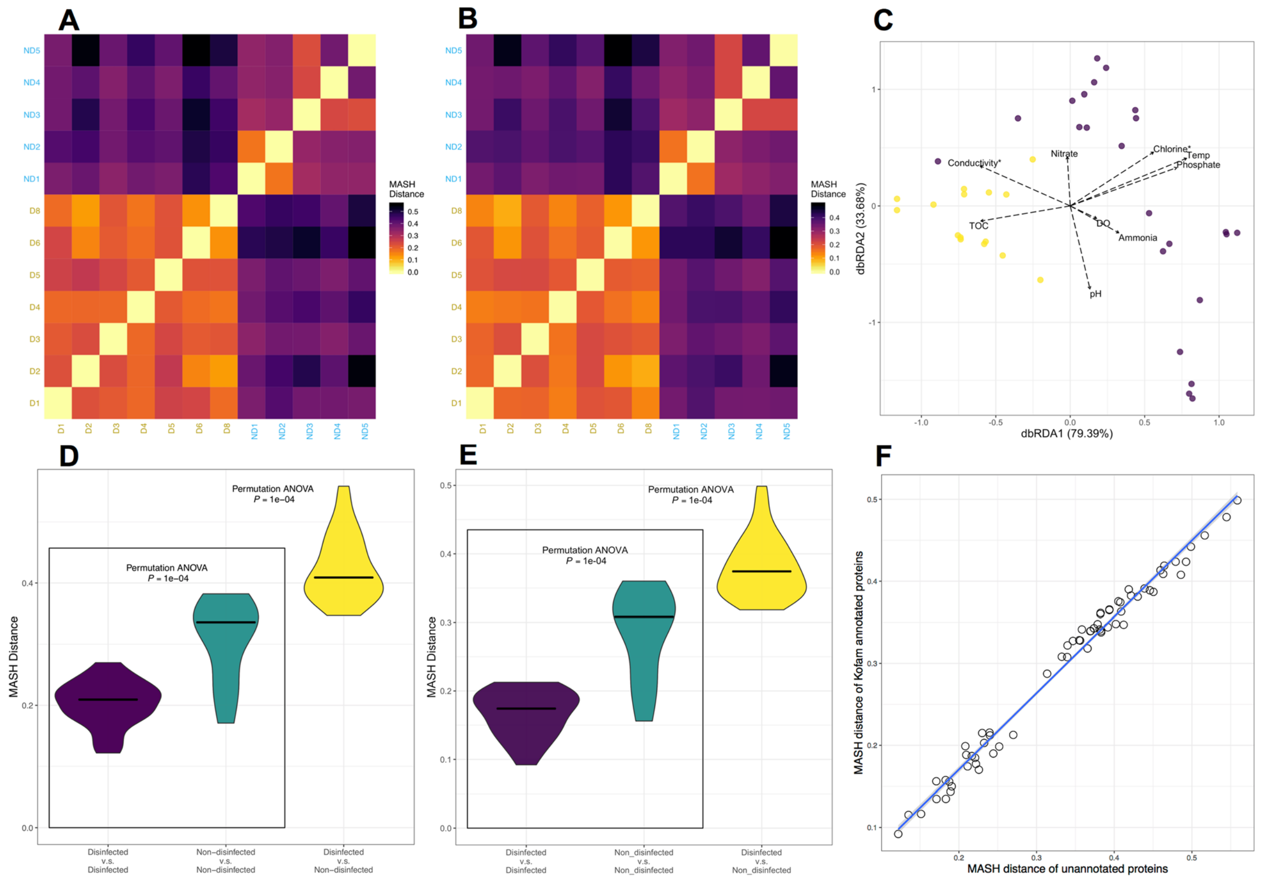

Fig. 4 Comparison of functional potential among all and KEGG protein-coding amino acid sequences across sampling locations. This analysis estimates dissimilarity in amino acid composition of samples, similar to the nucleotide composition analyses presented earlier. **a**, **b** Heatmaps based on pairwise Mash distances of all protein-coding sequences and Bray-Curtis distances using KO. Heatmaps are colored according to Mash/Bray-Curtis distance; yellow denotes a distance of 0. Labels on x- and y-axis are colored according to disinfection strategies; dark golden denotes samples with chlorine, while blue denotes samples without disinfectant residuals. **c** MDS clustering of using Bray-Curtis distances using KO abundances between samples with vectors representing water chemistry/environmental parameters implemented using dbRDA. Violin plots indicating the distribution of pairwise **d** Mash distances of all and **e** Bray-Curtis distances KEGG-annotated proteins. Crossbars indicate the median value of Mash distances. **f** Correlation between pairwise Mash distances estimated using all and Bray-Curtis distances for KEGG-annotated protein coding sequences

single parameter level (R = 0.382, p < 0.001). In contrast to nucleotide-based analyses, conductivity and chlorine were the only two variables identified as significantly associated with Bray-Curtis distances between samples estimated using the relative abundance of KO's in samples using dbRDA (Table S7). Variance partitioning indicated that both conductivity and chlorine individually explained approximately 6.5% of the variance in Bray-Curtis distance matrix estimated using KO abundance (Table S8). A comparison of the pairwise Mash distances within each group (i.e., disinfected, non-disinfected) and between them indicated that the diversity in functional potential was significantly different for both predicted protein coding-sequences and KEGG annotated proteins (p < 0.0001). The median value of Mash distances between the non-disinfected samples was greater than that for disinfected samples (Fig. 4d, e), and the differences in Mash distances between the two groups was larger than the distances within each group. And finally, despite the fact that only 17–27% of predicted proteins were annotated against the KEGG database, the Mash distances between metagenomes estimated using all predicted protein coding sequences and those that were annotated against the KEGG database were highly correlated (Pearson's R≈ 1.00, p < 0.05) (Fig. 4f), suggesting that focusing on annotated proteins does not result in significant loss of information while performing direct comparisons between samples from disinfected and non-disinfected systems.

These analyses based on protein coding sequencing provide several key insights. First, clustering of samples into disinfected and non-disinfected groups is consistent for both community composition (i.e., read-based nucleotide composition analyses) and functional potential, irrespective of the use of all predicted ORF's and KEGG annotated protein sequences. Non-disinfected systems are significantly more heterogeneous across systems as compared to their disinfected counterparts. This suggests that selection pressures exerted within disinfected systems are not only evident at community structure/membership (Fig. 3), but also evident at the community functional potential level. Further, consistent with microbial community composition, chlorine was also identified as one of the key measured parameters driving differences between samples based on functional potential using both BioEnv and dbRDA analyses. In contrast to TOC which was included in the BioEnv parameter combination for microbial community composition level analyses, ammonia was identified as part of the combination at the functional potential level. While the exact reason behind this difference cannot be ascertained in this study, this may likely be associated with the fact that non-disinfected systems are severely nitrogen limited as compared to disinfected systems, while both types of systems were likely not carbon limited (Table S1). Similar to the nucleotide level analyses, both conductivity and chlorine were identified as significantly ($p < 0.01$) associated with differences between samples, with variance partitioning analysis allocating an equal amount of variation to both parameters (Table S8). As speculated above, if conductivity is considered a signal for source water and treatment process type, then the impact of these two parameters on the functional potential of the microbial community is relatively similar to that of presence/absence of the disinfectant residual. Finally, the residuals from the variance partitioning analysis were noticeably larger (84%) for functional potential analyses as compared to the microbial community composition (60%), suggesting that the impact of unmeasured/uncharacterized factors/parameters on the microbial community functional potential was significantly larger than their impact on community composition. While it cannot be ruled out, it is unlikely that the higher fraction of unexplained variation was due to only a proportion of ORFs being annotated; this is because Mash distances estimated using only KEGG-annotated ORFs were highly correlated with those estimated using all predicted ORFs using suggesting little to minimal loss of discriminatory power while using only annotated proteins.

Differentially abundant metabolic modules are consistent with microbial growth control strategies

A total of 7281 KOs were identified in all samples with 5922 remaining post-filtering based on scaffold coverage (> 1x) and frequency of KO detection in each drinking water system (detected more than once). The 5922 KO's were further categorized into 540 KEGG modules and upon further filtering to remove KEGG modules with no more than one missing block and greater than equal to 50% completion, a total of 208 KEGG modules were retained. Of these, a total of 57 KEGG modules exhibited significantly differential abundance between disinfected and non-disinfected samples (p value < 0.005) (Table S9, S10). Modules associated with ribosomal synthesis, ribonucleotide biosynthesis, and RNA polymerase were ignored from further consideration. Similarly, modules most likely associated with plant metabolism (e.g., Crassulacean acid metabolism) were also ignored. This resulted in 29 and 22 KEGG modules that were more abundant in non-disinfected system and disinfected systems, respectively. These included modules associated with energy metabolism (disinfected, i.e., D = 2; non-disinfected, i.e., ND = 5), carbohydrate and lipid metabolism (D = 11, ND = 10), nucleotide and amino acid metabolism (D = 5, ND = 13), and secondary metabolism (D = 4, ND = 1).

Metabolic modules associated with polyamine biosynthesis, aromatics degradation, terpenoid biosynthesis, and fatty acid metabolism were significantly enriched in disinfected systems. Specifically, metabolic pathways associated with benzene (M00548) and benzoate (M00551) degradation to catechol and methyl catechol were highly enriched in disinfected systems. Further, eukaryota-associated metabolic modules such as terpenoid backbone biosynthesis (M00367) and modules associated with peroxisomal beta-oxidation of very long chain fatty acids (M00861) are likely to be enriched in the disinfected systems due to the higher relative abundance of eukaryota in samples collected from disinfected as compared to non-disinfected systems respectively. Further, modules related to γ-aminobutyrate (GABA) metabolism (M00136, M00027) were enriched in disinfected systems. The GABA shunt pathway converts glutamate to GABA using glutamate decarboxylase (GAD), followed by the reversible conversion from α-ketoglutarate to succinate semialdehyde (SSA) through the activity of GABA transaminase (GABA-AT), and finally succinate is formed by succinate semialdehyde dehydrogenase (SSDH) activity. In contrast, the key metabolic modules enriched in non-disinfected systems were associated with carbon fixation and methane metabolism (M00377, M00620, and M00422) and nitrogen fixation (M00175) (Table S10). The differentially abundant carbon fixation modules included the Wood-Ljungdahl pathway, acetyl-CoA pathway, and the incomplete reductive citrate cycle. These pathways can fix carbon dioxide to produce acetyl-CoA which can then be converted to other necessary biosynthetic intermediates of the carbon metabolism [43, 44].

The enrichment of carbon and nitrogen fixation modules in non-disinfected systems is consistent with nutrient limitation as the strategy for microbial growth control in non-disinfected drinking water systems. While the measured total organic carbon concentrations in non-disinfected systems did not indicate carbon-limited conditions, DWTP's supplying water to non-disinfected DWDSs typically achieve far superior levels of removal of assimilable organic carbon (AOC) [28]. Similarly, the nitrogen availability in the form of ammonia was consistently zero for non-disinfected systems compared to disinfected systems which have residual ammonia concentrations ranged from 0.01–0.15 mg/l of ammonia-nitrogen. In contrast, the enrichment of KEGG modules associated with GABA metabolism in disinfected systems suggests the potential importance of stress protection and the utilization of microbial decay products. Previous studies have shown that GABA metabolism is associated with bacterial survival under various types of environmental stresses, including oxidative stress, acidic stress, and osmotic stress [45–48]. Meanwhile, GABA can also play a significant role in the nitrogen metabolism of bacteria. For instance, putrescine formed due to the breakdown of amino acids potentially from decaying biomass, can be converted to GABA (M00136) and finally metabolized via the GABA shunt pathway [47]. The enrichment of GABA metabolism in disinfected systems may thus be associated with greater protection against disinfectant stress and by allowing access to decay products from inactivated cells.

Average genome size differences between disinfected and non-disinfected system vary between read-based and MAG-based analyses

We further investigated differences in genome sizes between disinfected and non-disinfected systems. Genome sizes can be indicative of the metabolic capacity of microorganisms [49] and thus provide insights into whether the presence/absence of disinfectants selects for organisms with larger or smaller metabolic repertoire [50] in comparison to organisms detected in non-disinfected systems. Average genome size estimates from disinfected systems were significantly larger than those from non-disinfected systems based on MicrobeCensus estimates using entire metagenomic data (Fig. 5a); this was consistent even when reads mapping to phyla

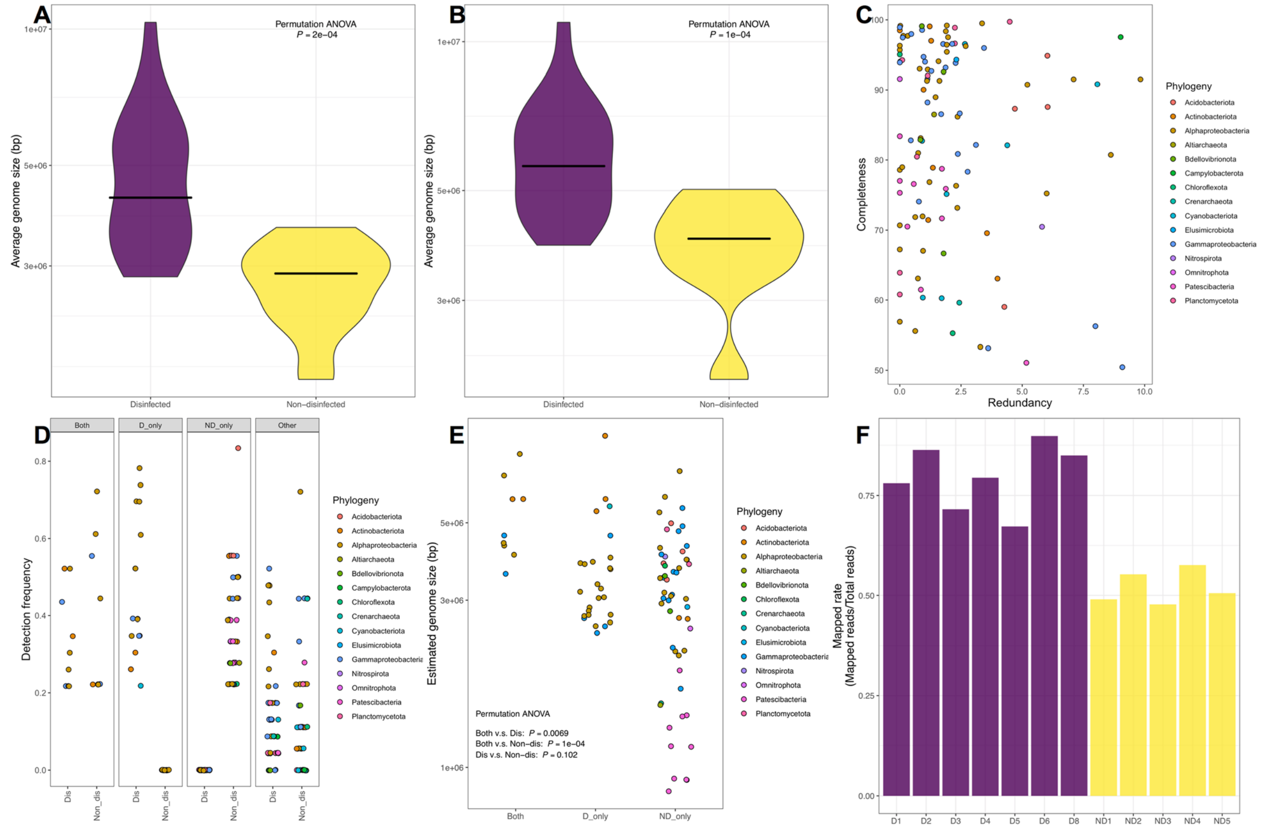

Fig. 5 Violin plots indicating the genome size estimated by MicrobeCensus **a** before and **b** after Patescibacteria removal suggest average genome sizes in disinfected systems are larger than those in non-disinfected systems. **c** The 115 MAGs assembled with > 50% completeness and < 10% redundancy were categorized into **d** four groups based on their detection frequency in disinfected and non-disinfected systems. **e** While the estimated genome sizes of MAGs in D-only, ND-only, and both categories were not significantly different, the ND-only category consisted of large number of smaller genomes. **f** Bar plot indicating the proportion of reads mapped to 115 genomes across samples. Purple and yellow denotes samples from systems with and without a disinfectant residual, respectively

known to have smaller genomes (e.g., *Patescibacteria*) were selectively removed from the data set (Fig. 5b). This suggests that microorganisms in disinfected systems may be metabolically more diverse than their counterparts from non-disinfected systems. Nonetheless, these results were not consistent when compared with estimated genome sizes of MAGs recovered as part of this study. Specifically, we recovered a total of 115 dereplicated MAGs with completeness > 50% and redundancy < 10% (Table S11). These 115 MAGS were binned into four categories based on the detection or non-detection in disinfected samples. Specifically, MAGs were binned in the four groups (i.e., both, D-only, ND-only, and other) based on genome coverage and detection frequency criteria outlined in the "Materials and methods" section (see "MAG-level analyses" section). This resulted in 9, 16, 41, and 49 MAGs categorized as both, D-only, ND-only, and other (Fig. 5c, d) (Table S11). In contrast to read-based estimates of average genome size, MAG-based genome size estimates were not significantly different between the three key categories (Both = 4.4 ± 0.77Mbp, D-only = 3.22 ± 0.81Mbp, ND-only = 3.48 ± 1.22Mbp) (Fig. 5e). Yet, the ND-only category consisted of several smaller genomes (n = 17) compared to the D category. The lack of genome size differences between disinfected and non-disinfected samples based on MAG-based analyses compared to metagenome-level read-based analyses may be due to the proportion of read-based data represented by the MAGs. Specifically, while 60–90% of the reads from disinfected systems mapped to the 115 MAGs with the mapping rate from non-disinfected systems averaging around 50% (Fig. 5f). Thus, it is likely that the metagenomic assembly and binning process may have resulted in the suboptimal recovery of smaller genomes from non-disinfected sample which eliminates the signal in genome size differences observed at the metagenome level.

Metabolic capacities differ between metagenome assembled genomes from disinfected and non-disinfected systems

The clustering of MAGs (Fig. 6a) based on presence/absence of KEGG metabolic modules was largely driven by phylogenetic placement of MAGs, rather than their classifications into groups based on the detection frequencies in disinfected and non-disinfected systems (Fig. 6b). Further, there was insufficient representation of MAGs from D-only/ND-only categories across all phylogenetic clusters (e.g., at the species or genus level) to allow for direct comparisons of metabolic potential of closely related MAGs exclusively frequent in disinfected and non-disinfected systems. Nonetheless, there were seven and five high quality (completeness > 90%, redundancy < 10%) alphaproteobacterial MAGs that were exclusively frequent in disinfected (average detection frequency in disinfected = 55%) and non-disinfected systems (average detection frequency in non-disinfected = 29%) (Fig. 6a). Thus, we focused metabolic module comparisons between these 12 MAGs only. We evaluated differences in metabolic capacity of these MAGs by (1) considering all KEGG modules ≥ 75% complete within MAGs to be present in them and (2) all modules present in more than half of the high-quality MAGs within each category to be present within each category (Fig. 6c, Table S12). We subsequently confirmed the presence/absence of genes within key metabolic modules using KO-level annotation for these 12 MAGs.

The metabolic module associated with the glyoxylate cycle (M00012) was present in 86% of the MAGs in the D-only category while being only partially complete in most of the ND-only MAGs. Specifically, isocitrate lyase (*aceA:* K01637) and malate synthase (*aceB*: K01638), two key genes involved in the glyoxylate cycle, were present in 40% and 100% of the MAGs from D-only, respectively and both genes were absent in all ND-only MAGs included in this analysis. The glyoxylate shunt is associated with the use of non-carbohydrate carbon sources (i.e., via gluconeogenesis), such as break down products from lipids, fatty acids etc. [51]. The likely benefit of the glyoxylate shunt and associated use of lipids and fatty acids as carbon source is further supported by the fact that KEGG module associated with propionyl-coA metabolism (M00741) was complete in 6/7 as compared to 2/5 MAGs from the D-only and ND-only categories. This metabolic module is associated with the conversion of propionyl-coA, a toxic byproduct of fatty and amino acid degradation, to succinyl-coA. High biomass turnover rates, due to disinfectant-induced microbial inactivation, may result in resource pools enriched in microbial decay products thus allowing a significant advantage for microorganisms capable of necrotrophic growth [52] aided by the glyoxylate cycle. Thus, it is feasible that the ability to utilize microbial decay products may provide a distinct advantage to microorganisms inhabiting disinfected DWDSs.

The glyoxylate shunt may provide additional benefits for microorganisms subject to disinfectant stress via enhanced fitness to oxidative stress [51] and enhanced persistence when challenged with other chemical stressors (e.g., antibiotics) [53]. In contrast to module level analyses at the metagenome level where carbon fixation capacity was significantly more abundant in non-disinfected as compared to disinfected systems, the alphaproteobacterial MAGs from D-only systems harbored the capacity for carbon fixation via the Calvin-Benson-Bassham cycle (M00165, M00166, M00167) while this capacity was mostly absent from MAGs in the ND-only category. Nonetheless, these MAG-based

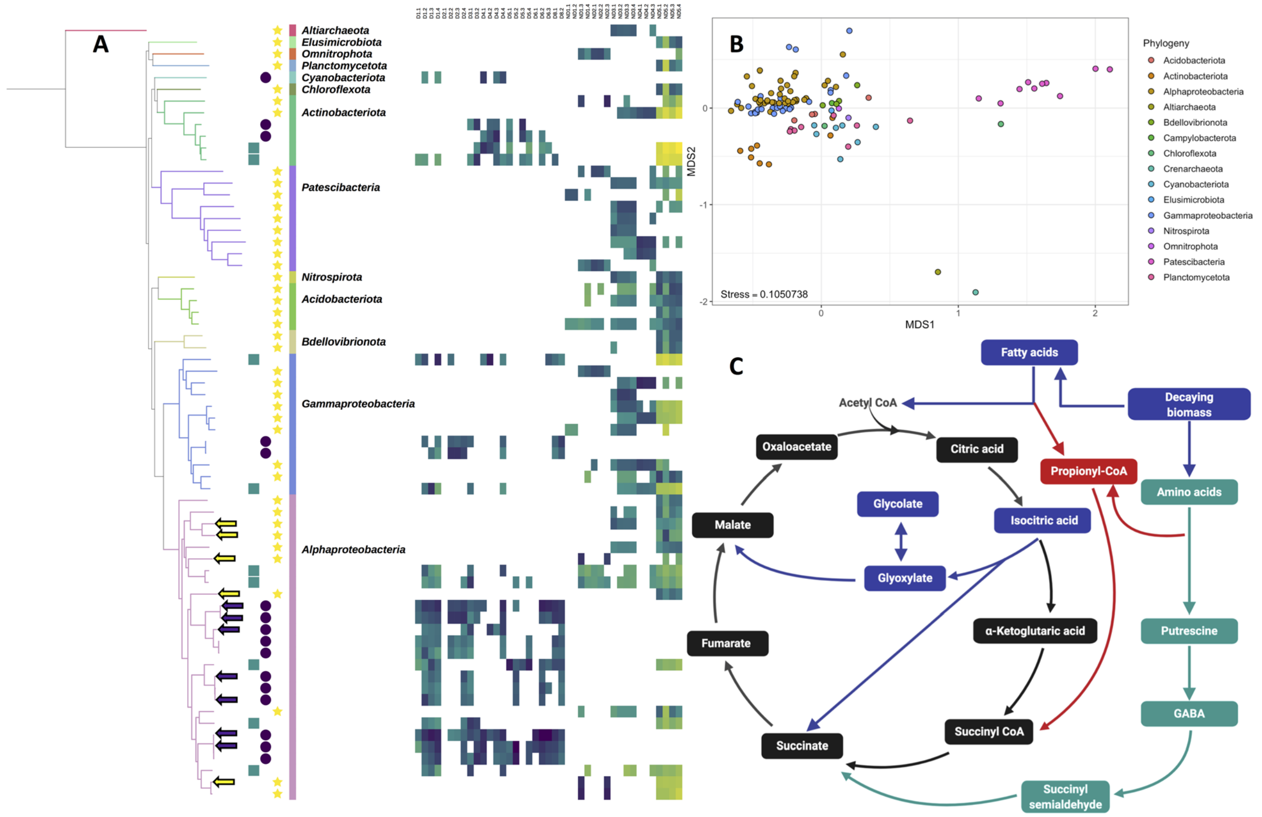

Fig. 6 a Phylogenomic tree of 66 MAGs classified as D-only (purple circles), ND-only (yellow stars), and both (teal squares) constructed using 48 ribosomal proteins along and their relative abundance (RPKM) in the samples collected from disinfected and non-disinfected systems. RPKM's for MAGs are only reported for samples where 25% of the nucleotides in a MAG were covered by at least one read. **b** Clustering of all MAGs based on their clustering metabolic potential (i.e., completeness of KEGG modules) was primarily drive by phylogeny. **c** The metabolic modules identified as differentially abundant in disinfected systems using metagenome level analyses are shown using teal arrows and squares and those more prevalent in high quality alphaproteobacterial MAGs from D-only (purple arrows—**a**) compared to those from ND-only category (yellow arrows—a) are shown using blue arrows and boxes, while red arrows and boxes indicate modules identified as more prevalent in D-only systems using both metagenome and MAG level analyses

analyses are limited in phylogenetic scope and do not weigh the importance of MAGs to their respective systems based on their relative abundance. Hence, we suggest that metagenome-level analyses should take precedence over findings at the MAG level when they conflict. While the glyoxylate shunt was not identified as significantly enriched in the disinfected systems at the metagenome level analyses, the GABA shunt (metagenome level analyses) and glyoxylate shunt (MAG level analyses) may both be involved in the use of non-carbohydrate carbon sources suggesting that reuse of microbial decay products may indeed be a key bacterial trait that allows for persistence in disinfected drinking water systems. Further lending support to this is that that propionyl-coA metabolism was identified as significantly enriched in disinfected systems compared to non-disinfected systems using both metagenome-level and MAG-level analyses. Interestingly, only one metabolic module was identified as being more than twice as prevalent in alphaproteobacterial MAGs from ND-only systems compared to those from D-only systems (i.e., M00156: cbb3-type Cytochrome C oxidase). The greater metabolic capacity of alphaproteobacterial D-only MAGs compared to ND-only MAGs was also confirmed at the KO-level by evaluating the presence/absence of KO's in the D-only and ND-only category MAGs. Specifically, while only 8 KOs were twice or more as prevalent in ND-only MAGs compared to D-only MAGs, the total KOs that were twice or more as prevalent in D-only MAGs was 109. This supports the conclusion that metabolic repertoire of alphaproteobacterial D-only MAGs is significantly larger than that of ND-only MAGs. Notable among the genes that were twice as frequent in D-only MAGs compared ND-only MAGs included those involve in SOS-response-mediated mutagenesis involving translesion synthesis (i.e., *imuA*: K14160, *imuB*: K14161, and *dnaE2*: K14162) [54], glyoxylate reductase (*gyaR*: K00015) which may be likely involved in regulating glyoxylate concentrations, and vitamin B12 transporter (*btuB*: K16092). SOS response is typically activated in response to significant cellular accumulation of damaged DNA [55] and *imuA* and *imuB* co-expression with *dnaE2* has been shown to be responsive to UV damage [54]. Thus, the higher prevalence of SOS response-related genes in D-only MAGs may be associated with the DNA damage caused by disinfectants. Further, the ability to synthesize vitamin B12, an essential co-factor, is limited to certain bacteria and archaea and thus the ability to uptake vitamin B12 from the environment is essential for growth [56]. The higher abundance of vitamin B12 transporters is consistent with metagenome level observations that the microbial community in disinfected systems rely more on scavenging from the environment as compared to non-disinfected systems.

Conclusions

To our knowledge, this is the first study to provide metagenomic insights into differences in structure and functional potential of drinking water microbiomes across full-scale drinking water systems that rely on disinfection (i.e., disinfected) or nutrient limitation (i.e., non-disinfected) to manage microbial growth. Understanding the microbial implications of these two microbial growth control strategies is essential to not only develop a better understanding of ecological and metabolic traits guiding community level processes in these system but is also critical for providing a community-level context to the microbiological safety in either type of drinking water system. In this study, we show that disinfection exhibits consistent, systematic, and significant association with drinking water microbiome at the membership, structure, and functional potential at the metagenome and MAG levels, irrespective of the drinking water system under consideration (e.g., source water type, treatment process, etc.). In doing so, we also identify key metabolic traits associated with carbon and nitrogen metabolism that are over represented in bacteria in disinfected systems compared to non-disinfected systems. This suggests that the influence and efficacy of disinfection on the drinking water microbiome may not simply be associated with differential disinfection resistance [57], but may also expand to other metabolic traits that include the use of carbon and nitrogen sources made available via microbial inactivation and its regulation.

Materials and methods

Sample collection and processing

Drinking water samples were collected from 12 drinking water systems in the Netherlands (n = 5) between October to December 2013 (non-disinfected, i.e., ND) and the UK (n = 7) between April to August in 2015 (disinfected, i.e., D) where chlorine was the residual disinfectant. Samples were collected at two to four locations in each DWDS which resulted in 23 D and 18 ND samples. A total 15 L of water was filtered through three sterile Sterivex filters with 0.22 μm pore size polyethersulfone membrane (EMD Millipore^{TM} SVGP01050) using a peristaltic pump (Watson-Marlow 323S/D) to harvest microbial cells. Immediately after filtration, the membranes were removed aseptically from the Sterivex cartridge, cut into pieces and then transferred to Lysing Matrix E tubes. The membranes were stored at 4 °C for 24 h or less before being transported to the laboratory and stored at − 80 °C. Further details of sample treatments and preservation are described in Sevillano-Rivera et al. [35], along with detailed description of chemical analyses. Briefly, Orion 5 Star Meter (Thermo Fisher Scientific, Waltham, MA) was used to measure temperature, pH, and conductivity, and dissolved oxygen, while total chlorine and phosphate were determined on-site using DR 2800 VIS Spectrophotometer (Hach Lange, the UK) and EPA-approved HACH kits. Nitrogen species were measured according to standard method, 4500-NH3-F for ammonia, 4500-NO2-B for nitrite, and 4500-NO3-B for nitrate respectively in laboratory [58], while total organic carbon (TOC) was determined using Shimadzu TOC-LCPH Analyzer (Shimadzu, Kyoto, Japan).

DNA extractions

The total genomic DNA was extracted directly from filter membranes using Maxwell16 DNA extraction system (Promega) and LEV DNA kit (AS1290, Promega, Madison, WI, USA). The filters with collected biomass in lysing matrix E tubes were incubated with 300 μL of lysing buffer and 30 μL of Proteinase K and incubated at 56 °C. A total of 500 μL of chloroform:isoamyl alcohol (24:1, pH 8.0) was added to the tube, vortexed and this was followed by bead beating for 40 s at 6 m/s using a FastPrep 24 instrument (MP Biomedicals, Santa Ana, CA, USA), and centrifugation at 14,000*g* for 10 min. The bead beating and centrifugation steps were repeated twice more with transfer of supernatant to clean tube followed by replacement of the aqueous phase with fresh lysing buffer. The aqueous phase was then subject to DNA purification using the Maxwell LEV DNA kit. The extracted DNA was quantified using Qubit HS dsDNA assay with Qubit 2.0 Fluorometer (Life Technologies, UK). Negative controls consisting of reagent blanks (no input material) and filter blanks (filter membranes from unused Sterivex filters) were processed identically as the samples for DNA extraction. Genomic DNA extracted from mock community, consisting of 10 organisms, detailed previously [35], was spiked into negative controls extracted (n = 8) from the reagent and filter blanks. These negative controls were also included in following library preparation and high-throughput sequencing (see below).

Library preparation and Illumina sequencing

Sequencing libraries were prepared using the Nextera XT DNA Sample Preparation Kit (Illumina Inc.). All DNA extracts (including negative controls) were cleaned up with HighPrep PCR magnetic beads (MagBio Inc.) to remove short fragments after library preparation and quantified with qPCR according to Illumina guidelines. All libraries were pooled together in equimolar proportion and pooled library was quantified with Qubit HS dsDNA assay and further concentrated using HighPrep PCR magnetic beads (MagBio Inc). Metagenomic sequencing on prepared libraries were performed on four lanes of Illumina HiSEQ 2500 flow cell (2 × 250-bp read length, Rapid Run Mode) at University of Liverpool Centre for Genomic Research (Liverpool, UK).

Metagenomic read-based analyses

The FASTQ files were trimmed using Cutadapt v1.2.1 (Martin 2014) with a "-O 3" flag, and Sickle v1.200 (Joshi and Fass 2011) using a threshold of window quality score (≥ 20) and read length after trimming (≥ 10 bp). A further trimming was applied using Trimmomatic v0.35 [59] to remove any remaining Illumina Nextera adaptors and trim reads according to quality score with a 4-base wide sliding window and a minimum average quality score of 20 and singlet reads were excluded in downstream analyses as well. To estimate metagenome diversity and coverage for each sample, Nonpareil 3.0 [41] was used in kmer mode on the quality-filtered reads. Diversity and coverage information for each metagenome was estimated using command "Nonpareil.set()" in R package "Nonpareil". MicrobeCensus [60] was used on quality-trimmed reads to estimate average genome size across samples with flag "-n 100000000" for all samples. To eliminate the potential effects of bacteria with small genomes (i.e., *Patescibacteria*) on average genome size estimations, pre-processed reads were mapped against 12 *Patescibacteria* metagenome assembled genomes (MAGs) from this study (see below) and 1037 *Patescibacteria* genomes from GTDB-tk [61]. The reads mapped in proper pair to *Patescibacteria* were removed using samtools ("-F2" flag). MicrobeCensus was used again to estimate average genome size using the same parameters.

Metagenome assembly and mapping

Filtered pair-ended reads were then pooled from each drinking water system for co-assembly, which resulted in 12 paired-end FASTQ files for co-assembly, including seven from disinfected (Dis) and five from non-disinfected (NonDis) systems. De novo co-assembly was performed using MetaSPAdes v3.10.1 [62] with recommended k-values for 2 × 250 bp reads (21,33,55,77,99, 127). All scaffolds shorter than 500 bp were discarded and UniVec_Core build 10.0 (National Center for Biotechnology Information 2016) was used for contamination vector screening and any scaffold with a significant hit to the UniVec database was removed. Reads from each sample were then mapped back to the filtered scaffolds using BWA-MEM v0.7.12 with default settings [63].

To eliminate the scaffolds that may have originated from sample or post-processing contamination, reads from negative controls were first mapped back to mock community genomes using BWA-MEM v0.7.12 [63], and all reads not mapped in proper pair were extracted using samtools v1.3.1 (Li et al. 2009) with "-f2" flag and were considered "contaminant reads." Sample reads (S), contaminant reads (C), and negative control reads (NC) were mapped back to filtered scaffolds in each co-assembly. Properly-paired mapped reads were extracted using samtools v1.3.1 with "-f2" flag from the BAM files. Relative abundance and normalized coverage deviation of each scaffold was calculated using reads from samples and those identified as contaminant reads in negative controls:

$$\text{Relative abundance}_S = \frac{\text{Scaffold coverage}_S}{\sum_{i=1}^{n} \text{Scaffold coverage}_S}$$

$$\text{Relative abundance}_C = \frac{\text{Scaffold coverage}_C}{\sum_{i=1}^{n} \text{Scaffold coverage}_{NC}}$$

$$\text{Normalized coverage deviation} = \frac{\text{Standard deviation of scaffold coverage}}{\text{Average scaffold coverage}}$$

To distinguish true scaffolds from contamination, relative abundance (RA) and normalized coverage deviation (NCD) estimated using sample reads (S) and contaminant reads (C) were compared for all scaffolds:

$$\text{Scaffold} = \begin{cases} \text{True scaffold,} & \text{if} \quad \begin{array}{r} RA_C = 0 \\ RA_S > RA_C \text{ and } NCD_S < NCD_C \end{array} \\ \text{Contaminant scaffold,} & \text{if} \quad \begin{array}{r} RA_S = 0 \\ RA_C > RA_S \text{ and } NCD_C < NCD_S \end{array} \end{cases}$$

True scaffolds, the scaffolds with higher RA and lower NCD in samples compared to negative controls, were kept for downstream analyses while contaminant scaffolds were excluded from all further analyses.

Nucleotide and protein composition analyses

MASH v1.1.1 [64] was used to estimate the dissimilarity between samples using quality-filtered reads (with "-r" and "-m 2" flags) and dissimilarity between drinking water systems using true scaffolds with the sketch size of 100,000. Prodigal v2.6.3 [65] was used to identify open reading frames (ORFs) in the true scaffolds and translate ORFs to protein-coding amino acid sequences. Following prediction and translation, HMMER v3.1b2 [66] was used to annotate ORFs against the Pfam database v31.0 [67] with a maximum e-value of $1e-5$ and curated bit score thresholds (the gathering thresholds). Subsequently, MASH distances were calculated between drinking water metagenomes using predicted ORFs, as well as Pfam annotated proteins with the sketch size of 100,000 and "-a" flag.

Taxonomic classification and phylogenetic analyses

The program "cmsearch" was implemented in Infernal v1.1.2 [68] to search scaffolds against SSU rRNA covariance models (CMs) for bacteria, archaea, and eukaryota; these are default models used by SSU-ALIGN v0.1 [69] using HMM-only approach and only significant hits were considered. The results were filtered according to length (≥ 100 bp alignment) and e-value ($<1e-5$). SSU rRNA sequences detected in contaminant scaffolds were removed and if more than one SSU gene sequence was found on a single scaffold, only the longest SSU gene sequence was retained. Relative abundance of each SSU gene sequence was calculated for each sampling location as follows:

$$\mathrm{RPKM}^{i}_{SSU} = \frac{\text{Scaffold coverage}^{i}}{\sum_{i=1}^{n} \text{SSU containing Scaffold coverage per Mb}^{i} \times \text{Scaffold length per kb}^{i}}$$

$$\text{Relative abundance}^{i}_{SSU} = \frac{\mathrm{RPKM}^{i}_{SSU}}{\sum_{i=1}^{n} \text{RPKM of scaffold containing SSU gene}^{i}}$$

SSU rRNA gene sequences were classified using Mothur v1.33.3 (Schloss et al. 2009) with SILVA database [70] (Release 132) with a minimum confidence threshold of 80%.

Annotation and comparison of functional orthologies and modules between samples

The protein-coding sequences were searched against KOfam, a HMM profile database for KEGG orthology [71] with predefined score thresholds using KofamScan [72]. Only KEGG orthologies (KO) identified on scaffolds with (> 1x) coverage for each sample and those detected more than once across samples within a single drinking water system were retained for further analyses. Average read count for each KO was calculated using scaffold coverage, average length of reads mapped, and total number of reads mapped to each scaffold in a sample using above equations. To assess functions at KEGG module level, BRITE hierarchy file was retrieved from KEGG website, and KO's were categorized into KEGG modules. The abundance of KEGG module in each sample was calculated using the median abundance of the detected KEGG orthologies within each module. The completeness of each KEGG module was calculated using "KO2MODULEclusters2.py".

Metagenome binning and refining

Anvi'o (versions: v2.2.2, v2.4.0, v4 and v5.1) [73] was used for metagenome binning and refining. Briefly, CONCOCT [74] integrated in Anvi'o was used to cluster scaffolds (longer than 2500 bp) into metagenome bins using tetra-nucleotide composition and coverage information across all samples within each metagenomic co-assembly. The "merge" method of CheckM v1.0.7 [75] was used to identify the bins that may emerge from the same microbial population but may have been separated during automated binning process. Following merging of compatible bins, RefineM v0.0.21 [76] was used to automatically refine bins according to genomic properties (i.e., the mean GC content, tetra-nucleotide signature, and coverage) and taxonomic classification. The completeness and redundancy of each refined bin was estimated using CheckM based on collections of lineage specific single-copy genes resulting in a total of 154 bins with greater than > 50% completeness. Among these bins, 130 bins had a redundancy of < 10% redundancy, while 24 bins are with > 10% redundancy. Further manual curation of these bins was performed using Anvi'o, resulting in 156 curated metagenome assembled genomes (MAGs). The 156 MAGs were de-replicated using dRep v2.2.2 [77] and MAGS with > 10% redundancy were discarded which resulted in 115 dereplicated MAGs with completeness > 50% and redundancy < 10%. All raw sequencing data and dereplicated MAGs are available on NCBI at BioProject number PRJNA533545. The dereplicated MAGs are also available in figshare at the following url: https://doi.org/10.6084/m9.figshare.11833269.

MAG-level analyses

Taxonomy assignment of MAGs was performed using GTDB-tk v0.1.3 [61] with the flag "classify_wf". Genome sizes of MAGs were estimated by multiplying the number of nucleotides in the MAG with the inverse of the CheckM estimated completeness. The MAGs were annotated using the HMM profile database for KEGG orthology with predefined score thresholds using KofamScan [72]. The KO's for each MAG were then categorized into modules based on BRITE hierarchy file retrieved from KEGG [71], and the completeness of KEGG modules in each genome was calculated using script "KO2MODULEclusters2.py". Anvi'o was used to extract a collection of 48 single-copy ribosomal proteins [78] from each MAG using "anvi-get-sequences-for-hmm-hits" with a maximum number of missing ribosomal proteins of 40. Subsequently, a phylogenetic tree was reconstructed using concatenated alignment of ribosomal proteins sequences using FastTree v2.1.7 [79]. Interactive Tree Of Life (iTOL) v4 (Letunic and Bork 2007) was used to visualize the phylogenetic tree.

Program "Union" in EMBOSS v6.6.0.0 [80] was used to concatenate all scaffolds in each MAG into a single sequence. Reads from all samples were cross-mapped to all MAGs using BWA-MEM v0.7.12 with default settings and proportion of each nucleotide in MAG covered by at least 1x coverage was determined using BEDtools [81]. A MAG was considered detected in a sample if ≥ 25% of its bases were covered by at least one read from the corresponding sample. This approach was used to determine whether MAGs were detected in all the samples. Further, the MAGs were binned into four categories based on their detection/non-detection within samples. Specifically, MAGs were divided into "D-only" if there were detected in ≥20% of the samples from the disinfected systems and not detected in any samples from the non-disinfected systems, "ND-only" if there were detected in ≥ 20% of the samples from the non-disinfected systems and not detected in any samples from the disinfected systems, "both" if there were detected in ≥ 20% of disinfected and non-disinfected systems, while the remaining MAGs were classified in the "other" category. Subsequently, reads from all samples were cross-mapped back to all the MAGs using BBMap v38.24 [82] with a minimum identity of 90%, and "ambiguous = best" and "pairedonly = t" flags. After filtering for detection (see above), reads per kilobase of per million reads (RPKM) for each MAG and each sample were calculated using the equation:

$$\mathrm{RPKM} = \frac{\text{Number of reads mapped to MAG}}{\text{Total number of reads in sample per Million} \times \text{MAG length in kbp}}$$

Statistics

Differences between disinfected and non-disinfected systems for (1) Mash distances distributions and (2) relative abundances were determined using Permutational ANOVA, and Pearson's correlations between pairwise mash distances were estimated in R. BioEnv in "sinkr" (https://github.com/menugget/sinkr), and "vegan" [83] packages were used to identify environmental parameters (i.e., water chemistry) and their combinations that explain the differences in the structure (i.e., Mash distances between samples estimated using reads) and functional potential (i.e., Bray Curtis distance estimated between samples using KO abundance (i.e., RPKM). BioEnv permutes through 2^n−1 possible combination of selected environmental parameters, 511 combinations in this case, and selects the combinations of scaled environmental variables which capture maximum correlation between dissimilarities of community datasets water chemistry and microbial community structure or functional potential. While, BioEnv analyses identified combination of variables that are highly correlated with differences in microbial community structure of functional potential, it does not identify the proportion of variance in microbial community structure of functional potential explained by individual variables or their combination. To this end, we used distance-based redundancy analysis (dbRDA) to perform constrained ordinations on community structure and functional potential to bypass the limitation of usual RDA and CCA, which can only use Euclidean distance measure. Function dbrda() from "vegan" package was used with pairwise Mash distances calculated between samples estimated using reads based Mash distance and Bray-Curtis distances based on KO RPKM to investigate relationships between the environmental variables and community data on both nucleotide composition and KO level. The function varpart() in the vegan package was used to determine the fraction of variation captured parameters identified as significantly associated with read-based Mash and KO relative abundance-based Bray-Curtis distance matrices. DeSeq2 package v1.18.1 [84] was used to identify differentially abundant KEGG modules between disinfected and non-disinfected systems by only considering KEGG modules with a maximum of one block missing and equal to or greater than 50% complete. The median scaffold-length normalized read count of KO's within each module was used in DESeq2 analyses with a maximum adjusted *p* value of 0.005.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s40168-020-00813-0.

Additional file 1: Table S1. Summary of water quality parameters measured for the samples collected as part of this study. (*NM = not measured). **Table S2.** BioEnv analyses in vegan package to determine the subset of variables significantly correlated with community similarities. This determines the Spearman's correlation between Euclidean distances of scaled environmental variables with the Mash distances estimated using metagenomic reads. **Table S3.** Distance based Redundancy Analysis using Mash distance matrix generated using pairwise Mash distances between samples estimated using metagenomic reads. **Table S4.** Variance Partition analysis using water chemistry/environmental parameters identified as significant being significantly associated with read-based Mash distances by dbRDA analyses. **Table S5.** Number of predicted open reading frames (ORFs) for each metagenome co-assembly and number annotated against the KEGG database using Kofamscan. **Table S6.** BioEnv analyses in vegan package to determine the subset of variables significantly correlated with similarities in functional potential of community estimates using KEGG annotation. This determines the Spearman's correlation between Euclidean distances of scaled environmental variables with Bray Curtis distance matrix generated from RPKM of KO detected in samples. **Table S7.** Distance based Redundancy Analysis using pairwise Bray-Curtis distances between samples estimated using from RPKM of KO detected in samples. **Table S8.** Variance Partition analysis using water chemistry/environmental parameters identified as significant being significantly associated with KO Bray-Curtis distance matrix by dbRDA analyses. **Table S9.** Summary of modules that were significantly higher abundance in disinfected systems as compared to non-disinfected systems. **Table S10.** Summary of modules that were significantly higher abundance in non-disinfected systems as compared to disinfected systems.

Additional file 2: Table S11. Summary statistics for metagenome assembled genomes (MAGs) extracted from the metagenome assemblies. Metagenome assembled genomes (MAG) were finalized after dereplication using dRep (https://github.com/MrOlm/drep) using all MAGs assembled in this study. As a result, the name assigned to a MAG does not represent sampling location it was assembled from. MAGs were assigned taxonomy using the Genome taxonomy database (GTDBTK: https://gtdb.ecogenomic.org/) version 0.1.3. The completeness and redundancy of MAGs was estimated using CheckM (https://github.com/Ecogenomics/CheckM/wiki) version 1.0.7. Only MAGs >50% completeness and <10% redundancy were included in the study. The genome statistics were estimated using Prokka. The coding density was estimated by dividing the cumulative length of coding sequences (CDS) divided by the length of the MAG. The MAGs were assigned four categories, "D-only", "ND-only", "Both", and "Other". "D-only" was assigned to MAGs detected in >20% of disinfected samples and not detected in non-disinfected samples. "ND-only" was assigned to MAGs detected in <20% of disinfected samples and not detected of non-disinfected samples. "Both" was assigned to MAGs that were detected in >20% of disinfected and non-disinfected samples. "Other" was assigned to MAGs that did not fall in either of the above three classes. (see excel spreadsheet).

Additional file 3: Table S12. KEGG modules and their completeness estimates within each MAG assembled as part of this study.

Acknowledgements

The authors are grateful to Prof. Karthik Anantharaman (University of Wisconsin Madison) for providing helpful critiques of the manuscript. The authors would also like to thank Dr. Jacob Price (University of Wisconsin Madison) for helpful discussions on statistical considerations for data analyses.

Availability of data and material

All raw and processed data is available on NCBI and affiliated with bioproject number PRJNA533545

Authors' contributions

ZD, MSR, and AJP designed the experiments and sample collection scheme. MSR, QMB, STC, and AJP performed sample collection and processing. ZD, MSR, UIZ, AME, and AJP performed data analyses and interpretation. PvD facilitated sampling in Netherlands. All authors contributed to data interpretation and writing of the manuscript. The authors read and approved the final manuscript.

Funding

ZD was supported by the Lord Kelvin Adam Smith Scholarship at the University of Glasgow. MS was supported by the College of Engineering at Northeastern University. This work was funded by Engineering and Physical Science Research Council (EP/M016811/1) and the National Science Foundation (NSF-CBET 1749530). UZI is supported by NERC Independent Research Fellowship (NERC NE/L011956/1).

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Infrastructure and Environment Research Division, School of Engineering, University of Glasgow, G12 8LT, Glasgow, UK. [2]Department of Civil and Environmental Engineering, Northeastern University, Boston, MA, USA. [3]Department of Civil and Environmental Engineering, University of Michigan, Ann Arbor, Michigan, USA. [4]Department of Medicine, University of Chicago, Chicago, IL, USA. [5]Josephine Bay Paul Center, Marine Biological Laboratory, Woods Hole, MA, USA. [6]KWR Watercycle Research Institute, Nieuwegein, Netherlands. [7]Laboratory of Microbiology, Wageningen University, Wageningen, Netherlands.

Received: 22 November 2019 Accepted: 25 February 2020
Published online: 20 March 2020

References

1. LeChevallier MW, Welch NJ, Smith DB. Full-scale studies of factors related to coliform regrowth in drinking water. Applied and Environmental Microbiology. 1996;62:2201–11.
2. Liu G, Bakker GL, Li S, Vreeburg JHG, Verberk JQJC, Medema GJ, Liu WT, Van Dijk JC. Pyrosequencing reveals bacterial communities in unchlorinated drinking water distribution system: an integral study of bulk water, suspended solids, loose deposits, and pipe wall biofilm. Environmental Science & Technology. 2014;48:5467–76.
3. Liu G, Zhang Y, van der Mark E, Magic-Knezev A, Pinto A, van den Bogert B, Liu W, van der Meer W, Medema G. Assessing the origin of bacteria in tap water and distribution system in an unchlorinated drinking water system by SourceTracker using microbial community fingerprints. Water Research. 2018;138:86–96.
4. Berry D, Xi C, Raskin L. Microbial ecology of drinking water distribution systems. Current Opinion in Biotechnology. 2006;17:297–302.
5. Proctor CR, Hammes F. Drinking water microbiology—from measurement to management. Current Opinion in Biotechnology. 2015;33:87–94.
6. Chao Y, Ma L, Yang Y, Ju F, Zhang X-X, Wu W-M, Zhang T. Metagenomic analysis reveals significant changes of microbial compositions and protective functions during drinking water treatment. Scientific Reports. 2013;3:3550.
7. Roeselers G, Coolen J, van der Wielen PWJJ, Jaspers MC, Atsma A, de Graaf B, Schuren F. Microbial biogeography of drinking water: patterns in phylogenetic diversity across space and time. Environ Microbiol. 2015;17:2505–14.
8. Pinto AJ, Xi C, Raskin L. Bacterial community structure in the drinking water microbiome is governed by filtration processes. Environmental Science & Technology. 2012;46:8851–9.
9. Lautenschlager K, Hwang C, Ling F, Liu WT, Boon N. K??ster O, Egli T, Hammes F: Abundance and composition of indigenous bacterial communities in a multi-step biofiltration-based drinking water treatment plant. Water Research. 2014;62:40–52.
10. Pinto AJ, Schroeder J, Lunn M. Sloan W. Raskin L: Spatial-temporal survey and occupancy-abundance modeling to predict bacterial community dynamics in the drinking water microbiome. mBio. 2014;5:e01135–14.
11. Baron JL, Vikram A, Duda S, Stout JE, Bibby K. Shift in the microbial ecology of a hospital hot water system following the introduction of an on-site monochloramine disinfection system. PLoS ONE. 2014;9.
12. Proctor CR, Gächter M, Kötzsch S, Rölli F, Sigrist R, Walser J-C, Hammes F. Biofilms in shower hoses – choice of pipe material influences bacterial growth and communities. Environmental Science: Water Research & Technology. 2016;2:670–82.
13. Jia S, Shi P, Hu Q, Li B, Zhang T, Zhang XX. Bacterial community shift drives antibiotic resistance promotion during drinking water chlorination. Environmental Science & Technology. 2015;49:12271–9.
14. Wang H, Masters S, Edwards MA, Falkinham JO, Pruden A. Effect of disinfectant, water age, and pipe materials on bacterial and eukaryotic community structure in drinking water biofilm. Environmental Science & Technology. 2014;48:1426–35.
15. Prest EI, Hammes F, van Loosdrecht MCM, Vrouwenvelder JS. Biological stability of drinking water: controlling factors, methods, and challenges. Frontiers in Microbiology. 2016;7.
16. Wang H, Pryor MA, Edwards MA, Falkinham JO, Pruden A. Effect of GAC pre-treatment and disinfectant on microbial community structure and opportunistic pathogen occurrence. Water Research. 2013;47:5760–72.
17. Liu G, Verberk JQJC, Van Dijk JC. Bacteriology of drinking water distribution systems: an integral and multidimensional review. Applied Microbiology and Biotechnology. 2013;97:9265–76.
18. van der Wielen PWJJ, van der Kooij D. Nontuberculous mycobacteria, fungi, and opportunistic pathogens in unchlorinated drinking water in the Netherlands. Applied and Environmental Microbiology. 2013;79:825.

19. Garner E, McLain J, Bowers J, Engelthaler DM, Edwards MA, Pruden A. Microbial ecology and water chemistry impact regrowth of opportunistic pathogens in full-scale reclaimed water distribution systems. Environmental Science & Technology. 2018;52:9056–68.
20. van Lieverloo JHM, Hoogenboezem W, Veenendaal G, van der Kooij D. Variability of invertebrate abundance in drinking water distribution systems in the Netherlands in relation to biostability and sediment volumes. Water Research. 2012;46:4918–32.
21. Christensen SCB, Nissen E, Arvin E, Albrechtsen H-J. Distribution of Asellus aquaticus and microinvertebrates in a non-chlorinated drinking water supply system – effects of pipe material and sedimentation. Water Research. 2011;45:3215–24.
22. Otten TG, Graham JL, Harris TD, Dreher TW. Elucidation of taste- and odor-producing bacteria and toxigenic cyanobacteria in a Midwestern drinking water supply reservoir by shotgun metagenomic analysis. Applied and Environmental Microbiology. 2016;82:5410–20.
23. Beech IB, Sunner J. Biocorrosion: towards understanding interactions between biofilms and metals. Current Opinion in Biotechnology. 2004; 15:181–6.
24. Zhang Y, Griffin A, Edwards M. Nitrification in premise plumbing: role of phosphate, pH and pipe corrosion. Environmental Science & Technology. 2008;42:4280–4.
25. Rosario-Ortiz F, Rose J, Speight V, Gunten Uv, Schnoor J: How do you like your tap water? Science (80-) 2016, 351:912–914.
26. Potgieter S, Pinto A, Sigudu M, du Preez H, Ncube E, Venter S. Long-term spatial and temporal microbial community dynamics in a large-scale drinking water distribution system with multiple disinfectant regimes. Water Research. 2018;139:406–19.
27. Kooij Dvd, Wielen PWJJvd, Rosso D, Shaw A, Borchardt D, Ibisch R, Apgar D, Witherspoon J, Toro DMd, Paquin PR, et al: Microbial Growth in Drinking Water Supplies. IWA Publishing; 2013.
28. van der Kooij D, van der Wielen PWJJ. Microbial growth in drinking-water supplies: problems, causes, control and research needs: IWA Publishing; 2013.
29. Richardson SD. Disinfection by-products and other emerging contaminants in drinking water. TrAC Trends in Analytical Chemistry. 2003;22:666–84.
30. Sedlak DL, von Gunten U. The Chlorine Dilemma. Science. 2011;331:42–3.
31. Li X-F, Mitch WA. Drinking water disinfection byproducts (DBPs) and human health effects: multidisciplinary challenges and opportunities. Environmental Science & Technology. 2018;52:1681–9.
32. Falkinham JO, Pruden A, Edwards M. Opportunistic premise plumbing pathogens: increasingly important pathogens in drinking water. Pathogens. 2015;4:373–86.
33. Zhang H, Chang F, Shi P, Ye L, Zhou Q, Pan Y, Li A. Antibiotic resistome alteration by different disinfection strategies in a full-scale drinking water treatment plant deciphered by metagenomic assembly. Environmental Science & Technology. 2019;53:2141–50.
34. Shi P, Jia S, Zhang XX, Zhang T, Cheng S, Li A. Metagenomic insights into chlorination effects on microbial antibiotic resistance in drinking water. Water Research. 2013;47:111–20.
35. Sevillano M, Dai Z, Calus S, Santos QMB-dl, Eren AM, Wielen PWJJvd, Ijaz UZ, Pinto AJ: Disinfectant residuals in drinking water systems select for mycobacterial populations with intrinsic antimicrobial resistance. bioRxiv 2019:675561.
36. Bertelli C, Courtois S, Rosikiewicz M, Piriou P, Aeby S, Robert S, Loret J-F, Greub G. Reduced chlorine in drinking water distribution systems impacts bacterial biodiversity in biofilms. Frontiers in Microbiology. 2018;9.
37. Hambsch B, Böckle K, van Lieverloo JHM. Incidence of faecal contaminations in chlorinated and non-chlorinated distribution systems of neighbouring European countries. Journal of Water and Health. 2007; 5:119–30.
38. Bautista-de los Santos QM, Schroeder JL, Blakemore O, Moses J, Haffey M, Sloan W, Pinto AJ. The impact of sampling, PCR, and sequencing replication on discerning changes in drinking water bacterial community over diurnal time-scales. Water Research. 2016;90:216–24.
39. Waak MB, Hozalski RM, Hallé C, LaPara TM. Comparison of the microbiomes of two drinking water distribution systems—with and without residual chloramine disinfection. Microbiome. 2019;7:87.
40. Volk C, Dundore E, Schiermann J, LeChevallier M. Practical evaluation of iron corrosion control in a drinking water distribution system. Water Research. 2000;34:1967–74.
41. Rodriguez-R LM, Gunturu S, Tiedje JM, Cole JR, Konstantinidis KT: Nonpareil 3: Fast estimation of metagenomic coverage and sequence diversity. mSystems 2018, 3.
42. Santos QMB-dl, L. Schroeder J, C. Sevillano-Rivera M, Sungthong R, Z. Ijaz U, T. Sloan W, J. Pinto A: Emerging investigators series: microbial communities in full-scale drinking water distribution systems – a meta-analysis. Environmental Science: Water Research & Technology 2016, 2:631-644.
43. Berg IA. Ecological aspects of the distribution of different autotrophic CO2 fixation pathways. Appl Environ Microbiol. 2011;77:1925–36.
44. Caspi R, Billington R, Fulcher CA, Keseler IM, Kothari A, Krummenacker M, Latendresse M, Midford PE, Ong Q, Ong WK, et al. The MetaCyc database of metabolic pathways and enzymes. Nucleic Acids Research. 2018;46:D633–9.
45. Coleman ST, Fang TK, Rovinsky SA, Turano FJ, Moye-Rowley WS. Expression of a glutamate decarboxylase homologue is required for normal oxidative stress tolerance in Saccharomyces cerevisiae. Journal of Biological Chemistry. 2001;276:244–50.
46. Feehily C, O'Byrne CP, Karatzas KAG. Functional γ-Aminobutyrate shunt in Listeria monocytogenes: role in acid tolerance and succinate biosynthesis. Applied and Environmental Microbiology. 2013;79:74–80.
47. Feehily C. Karatzas KaG: Role of glutamate metabolism in bacterial responses towards acid and other stresses. Journal of Applied Microbiology. 2013;114:11–24.
48. Metzner M, Germer J, Hengge R. Multiple stress signal integration in the regulation of the complex σS-dependent csiD-ygaF-gabDTP operon in Escherichia coli. Molecular Microbiology. 2004;51:799–811.
49. Barberán A, Ramirez KS, Leff JW, Bradford MA, Wall DH, Fierer N. Why are some microbes more ubiquitous than others? Predicting the habitat breadth of soil bacteria. Ecology Letters. 2014;17:794–802.
50. Guieysse B, Wuertz S. Metabolically versatile large-genome prokaryotes. Current Opinion in Biotechnology. 2012;23:467–73.
51. Ahn S, Jung J, Jang I-A, Madsen EL, Park W. Role of glyoxylate shunt in oxidative stress response. Journal of Biological Chemistry. 2016;291:11928–38.
52. Chatzigiannidou I. Props R. Boon N: Drinking water bacterial communities exhibit specific and selective necrotrophic growth. npj Clean Water. 2018;1:1–4.
53. Dolan SK, Welch M. The glyoxylate shunt, 60 years on. Annual Review of Microbiology. 2018;72:309–30.
54. Galhardo RS, Rocha RP, Marques MV, Menck CFM. An SOS-regulated operon involved in damage-inducible mutagenesis in Caulobacter crescentus. Nucleic Acids Research. 2005;33:2603–14.
55. Baharoglu Z, Mazel D. SOS, the formidable strategy of bacteria against aggressions. FEMS Microbiology Reviews. 2014;38:1126–45.
56. Heal KR, Qin W, Ribalet F, Bertagnolli AD, Coyote-Maestas W, Hmelo LR, Moffett JW, Devol AH, Armbrust EV, Stahl DA, Ingalls AE. Two distinct pools of B12 analogs reveal community interdependencies in the ocean. Proceedings of the National Academy of Sciences of the United States of America. 2017;114:364–9.
57. Chiao T-H, Clancy TM, Pinto A, Xi C, Raskin L. Differential resistance of drinking water bacterial populations to monochloramine disinfection. Environmental Science & Technology. 2014;48:4038–47.
58. Clesceri LS, Greenberg AE, Eaton AD. Standard methods for the examination of water and wastewater, 20th Edition: APHA American Public Health Association; 1998.
59. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.
60. Nayfach S, Pollard KS. Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. Genome Biology. 2015;16:51.
61. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil PA, Hugenholtz P. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nature Biotechnology. 2018;36:996.
62. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. MetaSPAdes: A new versatile metagenomic assembler. Genome Res. 2017;27:824–34.
63. Li H: Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:13033997 [q-bio] 2013.
64. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016;17:132.
65. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: Prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119.

66. Eddy SR. Accelerated profile HMM searches. PLoS Comput Biol. 2011;7:e1002195.
67. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, et al. The Pfam protein families database in 2019. Nucleic Acids Res. 2019;47:D427–32.
68. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 2013;29:2933–5.
69. Nawrocki EP: Structural RNA homology search and alignment using covariance models. PhD Thesis. Washington University in St. Louis, 2009.
70. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2012;41:D590–6.
71. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Research. 2016;44:D457–62.
72. Aramaki T, Blanc-Mathieu R, Endo H, Ohkubo K, Kanehisa M, Goto S, Ogata H: KofamKOALA: KEGG ortholog assignment based on profile HMM and adaptive score threshold. bioRxiv 2019:602110.
73. Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, Delmont TO. Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ. 2015;3:e1319.
74. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson AF, Quince C. Binning metagenomic contigs by coverage and composition. Nature Methods. 2014;11:1144–6.
75. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Research. 2015;25:1043–55.
76. Parks DH, Rinke C, Chuvochina M, Chaumeil PA, Woodcroft BJ, Evans PN, Hugenholtz P, Tyson GW. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol. 2017;2:1533–42.
77. Olm MR, Brown CT, Brooks B, Banfield JF. DRep: A tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J. 2017;11:2864–8.
78. Campbell BJ, Yu L, Heidelberg JF, Kirchman DL. Activity of abundant and rare bacteria in a coastal ocean. Proceedings of the National Academy of Sciences of USA. 2011;108:12776–81.
79. Price MN, Dehal PS, Arkin AP. FastTree 2 - Approximately maximum-likelihood trees for large alignments. PLoS One. 2010;5:e9490.
80. Rice P, Longden L, Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics. 2000;16:276–7.
81. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
82. BBMap short-read aligner, and other bioinformatics tools. [http://sourceforge.net/projects/bbmap/].
83. Oksanen Blanchet FG, Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., Wagner, H., J: Vegan: community ecology package. 2013.
84. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014;15:550.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.