Freshwater monitoring by nanopore sequencing

Lara Urban[1§*], Andre Holzer[2§*], J Jotautas Baronas[3], Michael Hall[1], Philipp Braeuninger-Weimer[4], Michael J Scherm[5], Daniel J Kunz[6,7], Surangi N Perera[8], Daniel E Martin-Herranz[1], Edward T Tipper[3], Susannah J Salter[9], and Maximilian R Stammnitz[9*]

[1]*European Bioinformatics Institute, Wellcome Genome Campus, Hinxton CB10 1SD, UK;*

[2]*Department of Plant Sciences, University of Cambridge, Cambridge CB2 3EA, UK;*

[3]*Department of Earth Sciences, University of Cambridge, Cambridge CB2 3EQ, UK;*

[4]*Department of Engineering, University of Cambridge, Cambridge CB3 0FA, UK;*

[5]*Department of Biochemistry, University of Cambridge, Cambridge CB2 1GA, UK;*

[6]*Wellcome Sanger Institute, Wellcome Trust Genome Campus, Hinxton CB10 1SA, UK;*

[7]*Department of Physics, University of Cambridge, Cambridge CB3 0HE, UK;*

[8]*Department of Physiology, Development & Neuroscience, University of Cambridge, Cambridge CB2 3DY, UK;*

[9]*Department of Veterinary Medicine, University of Cambridge, Cambridge CB3 0ES, UK;*

*§ These authors contributed equally * To whom correspondence should be addressed: maxrupsta@gmail.com; andre.holzer.biotech@gmail.com; lara.h.urban@gmail.com*

Key words: *Nanopore sequencing, environmental metagenomics, freshwater ecology, portable bacterial monitoring*

ORCID IDs: *Lara Urban: 0000-0002-5445-9314, Andre Holzer: 0000-0003-2439-6364, J Jotautas Baronas: 0000-0002-4027-3965, Michael Hall: 0000-0003-3683-6208, Philipp Braeuninger-Weimer: 0000-0001-8677-1647, Michael J Scherm: 0000-0002-3289-9159, Daniel J Kunz: 0000-0003-3597-6591, Surangi N Perera: 0000-0003-4827-9242, Daniel E Martin-Herranz: 0000-0002-2285-3317, Edward T Tipper: 0000-0003-3540-3558, Susannah J Salter: 0000-0003-3898-8504, Maximilian R Stammnitz: 0000-0002-1704-9199*

ABSTRACT

While traditional microbiological freshwater tests focus on the detection of specific bacterial indicator species including pathogens, direct tracing of all aquatic DNA through metagenomics poses a profound alternative. Yet, *in situ* metagenomic water surveys face substantial challenges in cost and logistics. Here we present a simple, fast, inexpensive and remotely accessible freshwater diagnostics workflow centred around the portable nanopore sequencing technology. Using defined compositions and spatiotemporal microbiota from surface water of an example river in Cambridge (UK), we provide optimised experimental and bioinformatics guidelines, including a benchmark with twelve taxonomic classification tools for nanopore sequences. We find that nanopore metagenomics can depict the hydrological core microbiome and fine temporal gradients in line with complementary physicochemical measurements. In a public health context, these data feature relevant sewage signals and pathogen maps at species level resolution. We anticipate that this framework will gather momentum for new environmental monitoring initiatives using portable devices.

INTRODUCTION

The global assurance of safe drinking water and basic sanitation has been recognised as a United Nations Millennium Development Goal (Bartram, Lewis, Lenton, & Wright, 2005), particularly in light of the pressures of rising urbanisation, agricultural intensification and climate change (Haddeland et al., 2014; Schewe et al., 2014). Waterborne diseases represent a particular global threat, with zoonotic diseases such as typhoid fever, cholera or leptospirosis resulting in hundreds of thousands of deaths each year (Prüss-Üstün, Kay, Fewtrell, & Bartram, 2002; Prüss-Üstün et al., 2019).

To control for risks of infection by waterborne diseases, microbial assessments can be conducted. While traditional microbial tests focus on the isolation of specific bacterial indicator organisms through selective media outgrowth in a diagnostic laboratory, this cultivation process is all too often time consuming, infrastructure-dependent and lacks behind in automatisation (Salazar & Sunagawa, 2017; Tringe & Rubin, 2005). Environmental metagenomics, the direct tracing of DNA from environmental samples, constitutes a less organism-tailored, data-driven monitoring alternative. Such approaches have been demonstrated to provide robust measurements of relative taxonomic species composition as well as functional diversity in a variety of environmental contexts (Almeida et al., 2019; Bahram et al., 2018; Sunagawa et al., 2015), and overcome enrichment and resolution biases common to culturing (Salazar & Sunagawa, 2017; Tringe & Rubin, 2005). However, they usually depend on

expensive stationary equipment, specialised operational training and substantial time lags between fieldwork, sample preparation, raw data generation and access. Combined, there is an increasing demand for freshwater monitoring frameworks that unite the advantages of metagenomic workflows with high cost effectiveness, fast technology deployability and data transparency (Gardy & Loman, 2018).

In recent years, these challenges have been revisited with the prospect of mobile DNA analysis. The main driver of this is the 'portable' MinION device from Oxford Nanopore Technologies (ONT), which enables real-time DNA sequencing using nanopores (Jain, Olsen, Paten, & Akeson, 2016). Nanopore read lengths can be comparably long, currently up to ~$2*10^6$ bases (Payne, Holmes, Rakyan, & Loose, 2018), which is enabled by continuous electrical sensing of sequential nucleotides along single DNA strands. In connection with a laptop or cloud access for the translation of raw voltage signal into nucleotides, nanopore sequencing can be used to rapidly monitor long DNA sequences in remote locations. Although there are still common concerns about the technology's base-level accuracy, mobile MinION setups have already been transformative for real-time tracing and rapid data sharing during bacterial and viral pathogen outbreaks (Chan et al., 2020; Faria et al., 2018; Faria et al., 2017; Kafetzopoulou et al., 2019; Quick et al., 2015; Quick et al., 2016). In the context of freshwater analysis, a MinION whole-genome shotgun sequencing protocol has already been leveraged for a comparative study of 11 rivers (Reddington et al., 2020). This report highlights key challenges which emerge in serial monitoring scenarios of a relatively low-input DNA substrate (freshwater), for example large sampling volumes (2-4 litres) and data redundancies arising from small shotgun fragments (mean <4 kbp). We reasoned that targeted DNA amplification may be a suitable means to bypass these bottlenecks and assess river microbiomes with nanopore sequencing.

Here we report a simple, inexpensive workflow to assess and monitor microbial freshwater ecosystems with targeted nanopore DNA sequencing. Our benchmarking study involves the design and optimisation of essential experimental steps for multiplexed MinION usage in the context of local environments, together with an evaluation of computational methods for the bacterial classification of nanopore sequencing reads from metagenomic libraries. To showcase the resolution of sequencing-based aquatic monitoring in a spatiotemporal setting, we combine DNA analyses with physicochemical measurements of surface water samples collected at nine locations within a confined ~12 kilometre reach of the River Cam passing through the city of Cambridge (UK) in April, June and August 2018.

RESULTS

Experimental design and computational workflows

Using a bespoke workflow, nanopore full-length (V1-V9) 16S ribosomal RNA (rRNA) gene sequencing was performed on all location-barcoded freshwater samples at each of the three time points (Figure 1; Supplementary Table 1; Material and Methods). River isolates were multiplexed with negative controls (deionised water) and mock community controls composed of eight bacterial species in known mixture proportions.

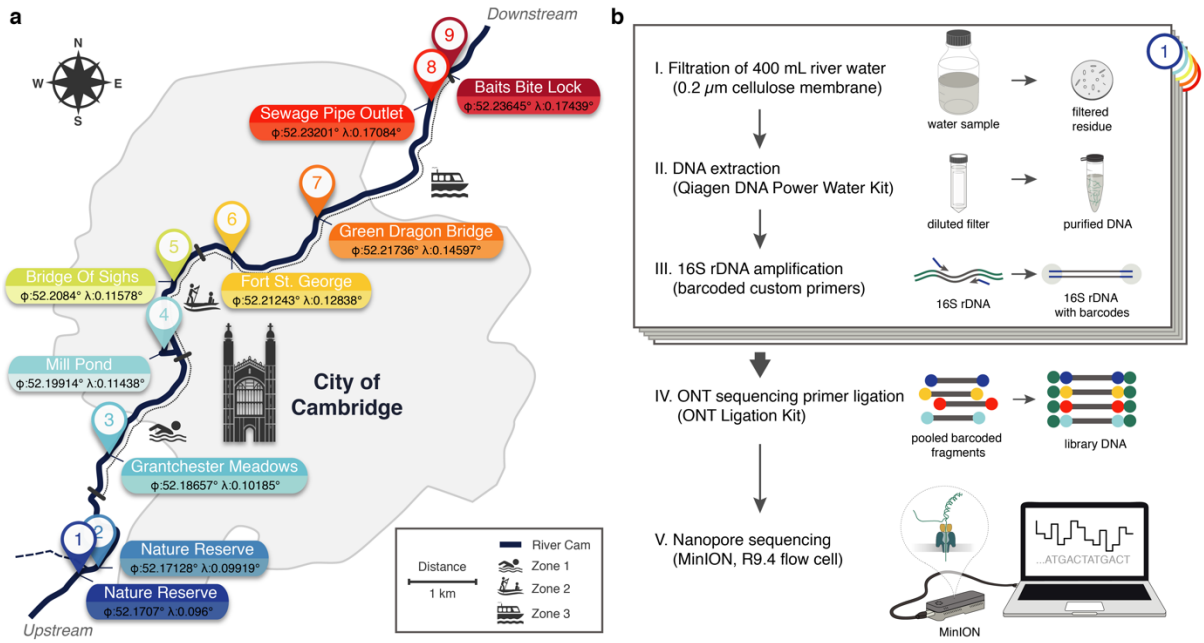


Figure 1: Freshwater microbiome study design and experimental setup. (a) Schematic map of Cambridge (UK), illustrating sampling locations (colour-coded) along the River Cam. Geographic coordinates of latitude and longitude are expressed as decimal fractions according to the global positioning system. (b) Laboratory workflow to monitor bacterial communities from freshwater samples using nanopore sequencing (Material and Methods).

To obtain valid taxonomic assignments from freshwater sequencing profiles using nanopore sequencing, twelve different classification tools were compared through several performance metrics (Figure 2; Material and Methods). Our comparison included established classifiers such as RDP (Wang, Garrity, Tiedje, & Cole, 2007), Kraken (Wood & Salzberg, 2014) and Centrifuge (Kim, Song, Breitwieser, & Salzberg, 2016), as well as more recently developed methods optimised for higher sequencing error rates such as IDTAXA (Murali, Bhargava, & Wright, 2018) and Minimap2 (Li, 2018). Root mean square errors (RMSE) between observed and expected bacteria of the mock community differed slightly across all classifiers. An *Enterobacteriaceae* overrepresentation was observed across all replicates and classification methods, pointing towards a consistent *Escherichia coli* amplification bias potentially caused by skewed taxonomic specificities of the selected 16S primer pair 27f and

1492r (Frank et al., 2008). Robust quantifications were obtained by Minimap2 alignments against the SILVA v.132 database (Quast et al., 2013), for which 99.68 % of classified reads aligned to the expected mock community taxa (mean sequencing accuracy 92.08 %). Minimap2 classifications reached the second lowest RMSE (excluding *Enterobacteriaceae*), and relative quantifications were highly consistent between mock community replicates. Benchmarking of the classification tools on one aquatic sample further confirmed Minimap2's reliable performance in a complex bacterial community, although other tools such as SPINGO (Allard, Ryan, Jeffery, & Claesson, 2015), MAPseq (Matias Rodrigues, Schmidt, Tackmann, & von Mering, 2017), or IDTAXA (Murali et al., 2018) also produced highly concordant results despite variations in speed and memory usage (data not shown).

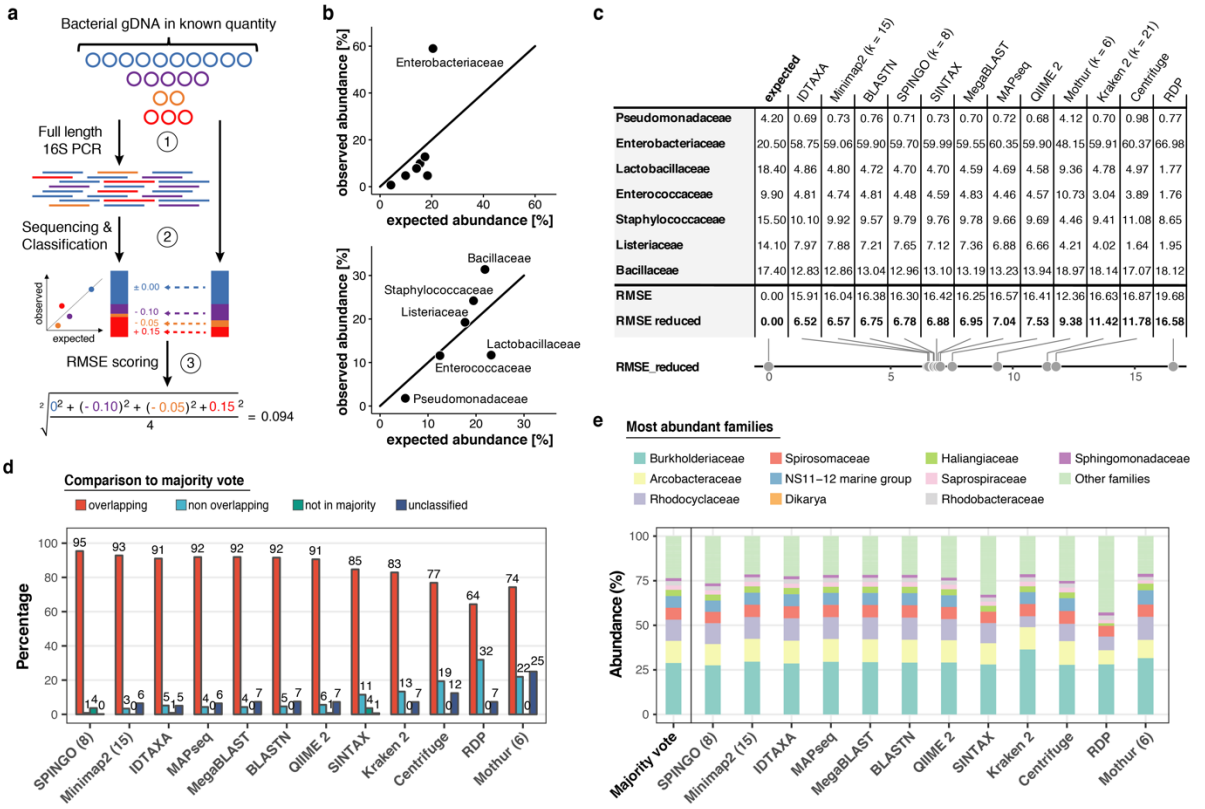


Figure 2: Benchmarking of classification tools with nanopore full-length 16S sequences. (a) Schematic of mock community quantification performance testing. (b) Observed vs. expected read fraction of bacterial families present in 10,000 nanopore reads randomly drawn from mock community sequencing data. Example representation of Minimap2 (kmer length 15) quantifications with (upper) and without (lower) *Enterobacteriaceae* (Material and Methods). (c) Mock community classification output summary for twelve classification tools tested against the same 10,000 reads. Root mean squared errors observed and expected bacterial read fractions are provided with (RMSE) and without *Enterobacteriaceae* (RMSE reduced). (d) Classification output summary for 10,000 reads randomly drawn from an example freshwater sample (Material and Methods). 'Overlapping' fractions (red) represent agreements of a classification tool with the majority of tested methods on the same reads, while 'non-overlapping' fractions (light blue) represent disagreements. Dark green sets highlight rare taxon assignments not featured in any of the 10,000 majority classifications, while dark blue bars show unclassified read fractions. (e) Top 10 represented bacterial taxon families across all classifiers based on the 10,000 aquatic reads used in (d).

Diversity analysis and river core microbiome

Using Minimap2 classifications within our bioinformatics consensus workflow (Supplementary Figure 1; Material and Methods), we then inspected sequencing profiles of three independent MinION runs for a total of 30 river DNA isolates and six controls. This yielded ~8.3 million sequences with exclusive barcode assignments (Figure 3a; Supplementary Table 2). Overall, 55.9 % (n = 4,644,194) of raw reads could be taxonomically assigned to the family level (Figure 3b). To account for variations in sample sequencing depth, rarefaction with a cut-off at 37,000 reads was applied to all samples. While preserving ~90 % of the original family level taxon richness (Mantel test, $R = 0.814$, $p = 2.1*10^{-4}$; Supplementary Figure 2), this conservative thresholding resulted in the exclusion of 14 samples, mostly from the June time point, for subsequent high-resolution analyses. The 16 remaining surface water samples revealed moderate levels of microbial heterogeneity (Figure 3b; Supplementary Figure 2): microbial family alpha diversity ranged between 0.46 (June-6) and 0.92 (April-7) (Simpson index), indicating low-level evenness with a few taxonomic families that account for the majority of the metagenomic signal.

Hierarchical clustering of taxon profiles showed a dominant core microbiome across all aquatic samples (clusters C2 and C4, Figure 3c). The most common bacterial families observed were *Burkholderiaceae* (40.0 %), *Spirosomaceae* (17.7 %), and NS11-12 marine group (12.5 %), followed by *Arcobacteraceae* (4.8 %), *Sphingomonadaceae* (2.9 %) and *Rhodobacteraceae* (2.5 %) (Figure 3d). Members of these families are commonly associated with aquatic environments; for example, *Burkholderiaceae* reads mostly originate from genera such as *Limnohabitans*, *Rhodoferax* or *Aquabacterium*, which validates the suitability of this nanopore metagenomics workflow. Hierarchical clustering additionally showed that two biological replicates collected at the same location and time point (April samples 9.1 and 9.2), grouped with high concordance; this indicates that spatiotemporal trends are discernible even within a highly localised context.

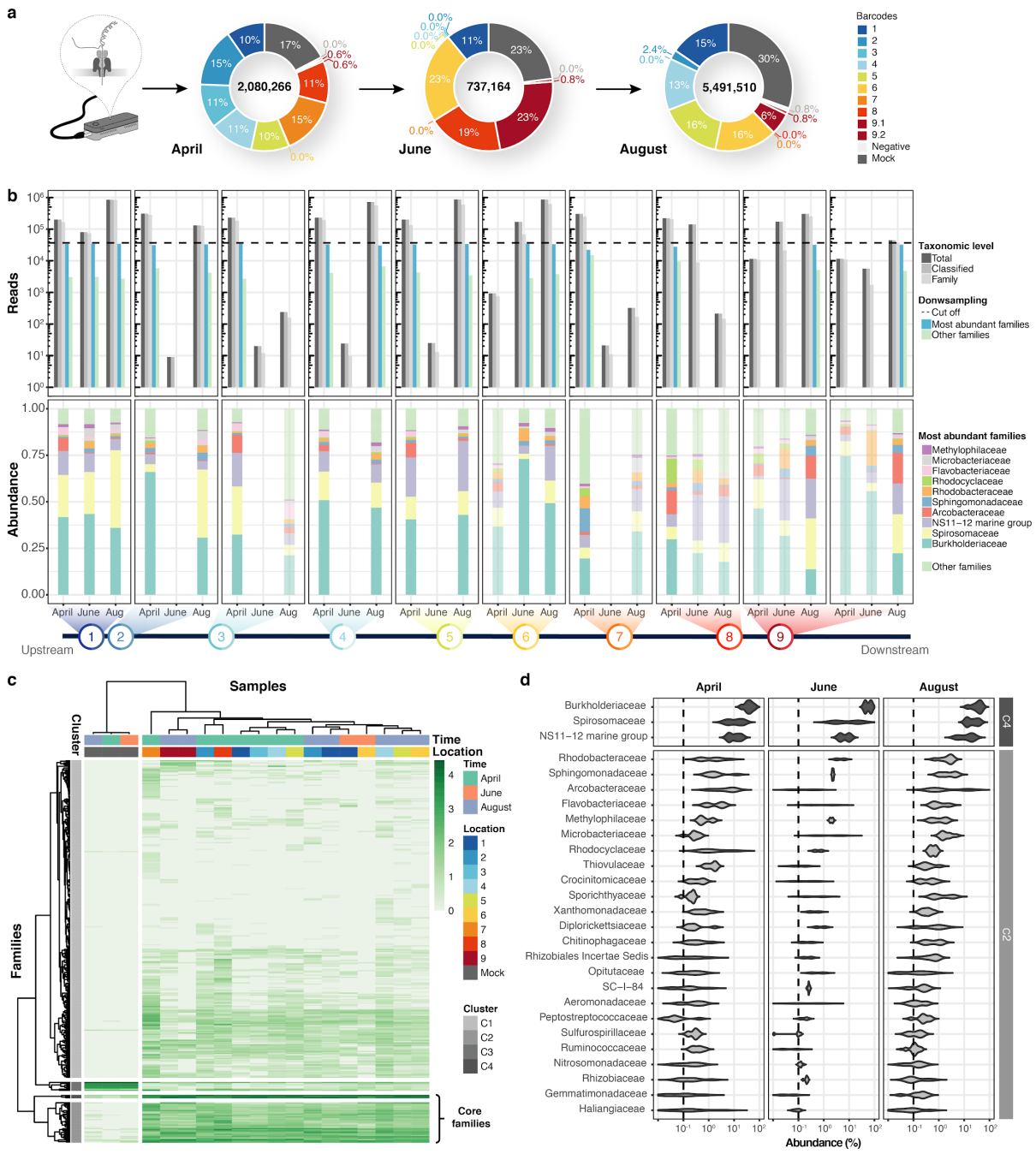


Figure 3: Bacterial diversity of the River Cam. (a) Nanopore sequencing output summary. Values in the centre of the pie charts depict total numbers of classified nanopore sequences per time point. Percentages illustrate representational fractions of locations and control barcodes (negative control and mock community). (b) Read depth and bacterial classification summary. Upper bar plot shows the total number of reads, and the number of reads classified to any taxonomic level, to at least bacterial family level, to the ten most abundant bacterial families across all samples, or to other families. Rarefaction cut-off displayed at 37,000 reads (dashed line). Lower bar plot features fractions of the ten most abundant bacterial families across the samples with more than 100 reads. Colours in bars for samples with less than 37,000 reads are set to transparent. (c) Hierarchical clustering of bacterial family abundances across freshwater samples after rarefaction, together with the mock community control. Four major clusters of bacterial families occur, with two of these (C2 and C4) corresponding to the core microbiome of ubiquitously abundant families, one (C3) corresponding to the main mock community families and one (C1) corresponding to the majority of rare accessory taxa. (d) Detailed river core microbiome. Violin plots ($\log_{10}$ scale of relative abundance [%] across all samples, $n_{April} = 7$, $n_{June} = 2$, $n_{August} = 7$) summarise fractional representation of bacterial families from clusters C2 and C4, sorted by median total abundance. Vertical dashed line depicts 0.1 % proportion.

Besides the dominant core microbiome, microbial profiles showed a marked arrangement of time dependence, with water samples from April grouping more distantly to those from June and August (Figure 3c). Principal component analysis (PCA) (Figure 4a; Supplementary Figure 3) revealed that the strongest differential abundances along the chronological axis of variation (PC3) derived from the higher abundance of *Carnobacteriaceae* in April (Figure 4b). This bacterial family is known for its occurrence in waters with low temperature (Lawson & Caldwell, 2014).

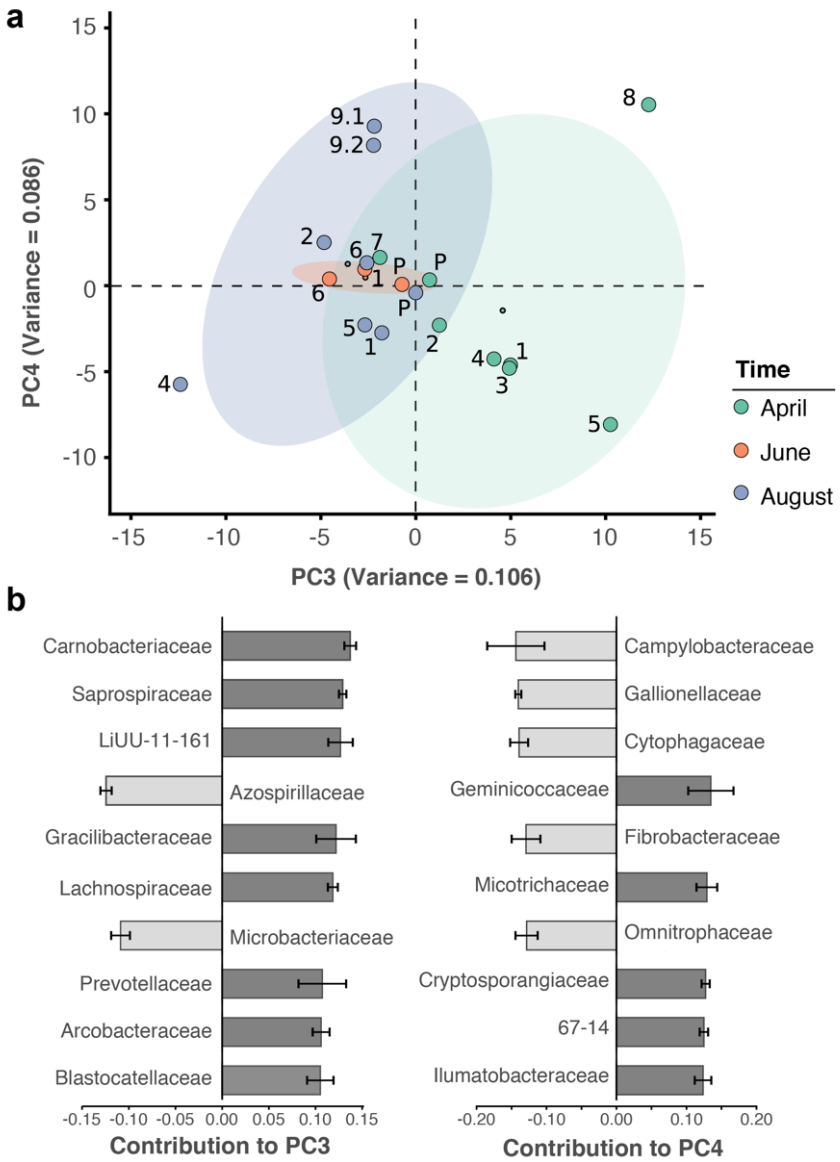


Figure 4: Spatiotemporal axes of taxonomic diversity in the River Cam. (a) PCA of bacterial composition across locations, indicating community dissimilarities along the main time (PC3) and spatial (PC4) axes of variation; dots coloured according to time points. (b) Contribution of individual bacterial families to the PCs in (a). Error bars represent the standard deviation of these families across four independent rarefactions.

Hydrochemistry and seasonal profile of the River Cam

While a seasonal difference in bacterial composition can be expected due to increasing water temperatures in the summer months, additional changes may have also been caused by alterations in river hydrochemistry and flow rate (Figure 5a; Supplementary Figure 4; Supplementary Table 1). To assess this effect in detail, we measured the pH and a range of major and trace cations in all river water samples using inductively coupled plasma-optical emission spectroscopy (ICP-OES), as well as major anions using ion chromatography (Material and Methods). As with the bacterial composition dynamics, we observed significant temporal variation in water chemistry, superimposed on a spatial gradient of generally increasing sodium and chloride concentrations along the river reach (Figure 5b-c). This spatially consistent effect is likely attributed to wastewater and agricultural discharge inputs in and around Cambridge city. A comparison of the major element chemistry in the River Cam transect with the world's 60 largest rivers further corroborates the likely impact of anthropogenic pollution in this fluvial ecosystem (Gaillardet, Dupré, Louvat, & Allègre, 1999) (Figure 5d; Material and Methods).

Maps of potential bacterial pathogens at species-level resolution

The River Cam has been notorious for causing bacterial infections such as leptospirosis amongst its stakeholders such as swimmers, rowers and houseboat owners. In line with the physicochemical trends, we therefore next determined the spatiotemporal enrichment of potentially functionally important bacterial taxa through nanopore sequencing. We retrieved 55 potentially pathogenic bacterial genera through careful integration of species known to affect human health (Jin et al., 2018; Wattam et al., 2017), and also 13 wastewater-associated bacterial genera (Wu et al., 2019) (Supplementary Table 3). Of these, 21 potentially pathogenic and eight wastewater-associated genera were detected across all of the river samples (Figure 6; Material and Methods). Many of these signals were stronger downstream of urban sections, within the mooring zone for recreational and residential barges (location 7; Figure 1a) and in the vicinity of sewage outflow from a nearby wastewater treatment plant (location 8; Figure 1a). The most prolific candidate pathogen genus observed was *Arcobacter*, which features multiple species implicated in acute gastrointestinal infections (Kayman et al., 2012).

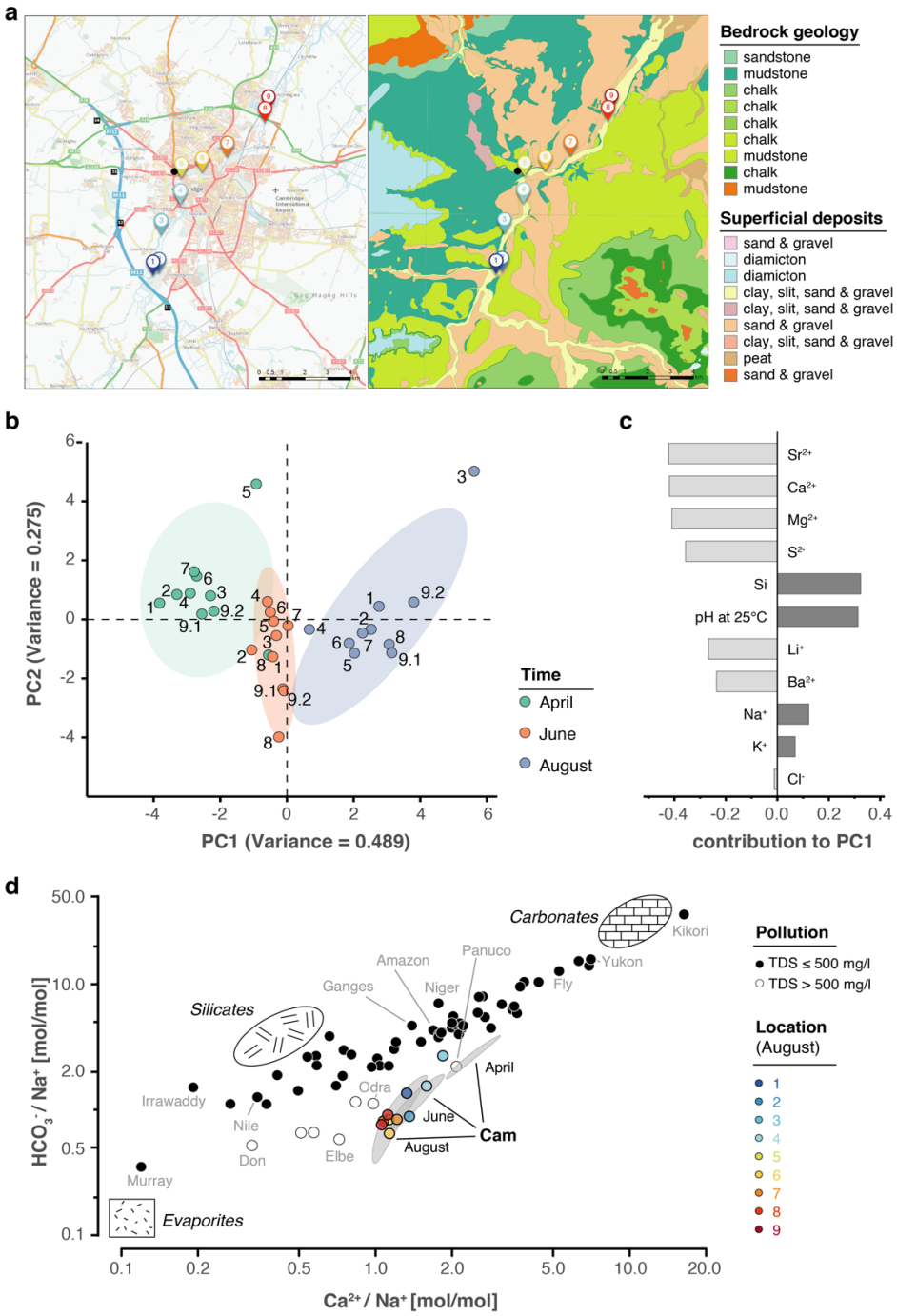


Figure 5: Geological and hydrochemical profile of the River Cam and its basin. (a) Outline of the Cam River catchment surrounding Cambridge (UK), and its corresponding lithology. Overlay of bedrock geology and superficial deposits (British Geological Survey data: DiGMapGB-50, 1:50,000 scale) is shown as visualised by GeoIndex. Bedrock is mostly composed of subtypes of Cretaceous limestone (chalk), gault (clay, sand) and mudstone. Approximate sampling locations are colour-coded as in Figure 1. (b) Principal component analysis of measured pH and 13 inorganic solute concentrations of this study's 30 river surface water samples. PC1 (~49 % variance) displays a strong, continuous temporal shift in hydrochemistry. (c) Parameter contributions to PC1 in (b), highlighting a reduction in water hardness (Ca^{2+}, Mg^{2+}) and increase in pH towards the summer months (June and August). (d) Mixing diagram with Na^{+}-normalised molar ratios, representing inorganic chemistry loads of world's 60 largest rivers; open circles represent polluted rivers with total dissolved solid (TDS) concentrations >500 mg l^{-1}. Cam River ratios are superimposed as ellipses from ten samples per month (50 % confidence, respectively). Separate data points for all samples from August are also shown and colour-coded, indicating the downstream-to-upstream trend of Na^{+} increase (also observed in April and June). End-member signatures show typical chemistry of small rivers draining these lithologies exclusively (carbonate, silicate and evaporite).

In general, much of the taxonomic variation across all samples was caused by sample April-7 (PC1 explains 27.6 % of the overall variance in bacterial composition; Supplementary Figure 3a-b). This was characterised by an unusual dominance of *Caedibacteraceae*, *Halomonadaceae* and others (Supplementary Figure 3c). Isolate April-8 also showed a highly distinct bacterial composition, with some families nearly exclusively occurring in this sample (outlier analysis; Material and Methods). The most predominant bacteria in this sewage pipe outflow are typically found in wastewater sludge or have been shown to contribute to nutrient pollution from effluents of wastewater plants, such as *Haliangiaceae*, *Nitospiraceae*, *Rhodocyclaceae*, and *Saprospiracea* (Nielsen, Saunders, Hansen, Larsen, & Nielsen, 2012; Wu et al., 2019) (Figure 6).

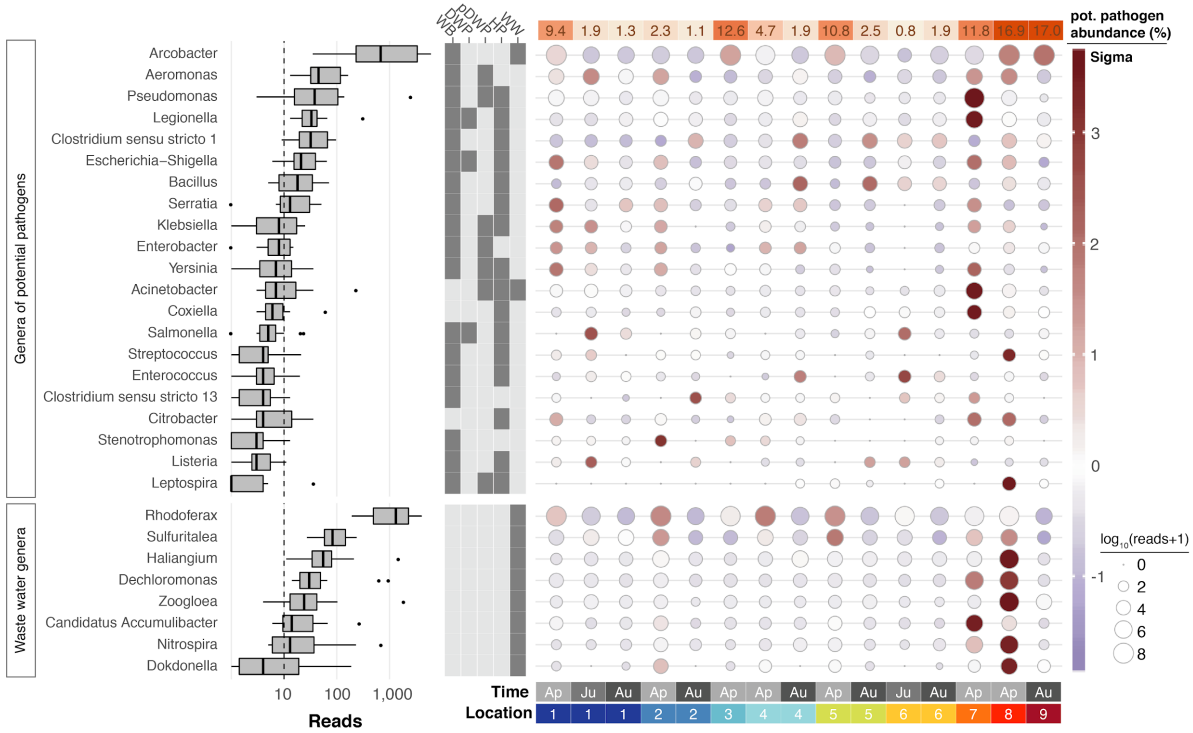


Figure 6: Potentially pathogenic and wastewater treatment related bacteria in the River Cam. The boxplots on the left show the abundance distribution across locations per bacterial genus. Error bars represent Q1 – 1.5*IQR (lower), and Q3 + 1.5*IQR (upper), respectively; Q1: first quartile, Q3: third quartile, IQR: interquartile range. The centre colour-scale table depicts the categorisation of subsets of genera as waterborne bacterial pathogens (WB), drinking water pathogens (DWP), potential drinking water pathogens (pDWP), human pathogens (HP) and core genera from wastewater treatment plants (WW) (dark grey: included, light grey: excluded) (Supplementary Table 3). The right-hand circle plot shows the distribution of bacterial genera across locations of the River Cam. Circle sizes represent overall read size fractions, while circle colours (sigma scheme) represent the standard deviation from the observed mean relative abundance within each genus.

Using multiple sequence alignments between nanopore reads and pathogenic species references, we further resolved the phylogenies of three common potentially pathogenic genera occurring in our river samples, *Legionella, Salmonella* and *Pseudomonas* (Figure 7a-c; Material and Methods). While *Legionella* and *Salmonella*

diversities presented negligible levels of known harmful species, a cluster of reads in downstream sections indicated a low abundance of the opportunistic, environmental pathogen *Pseudomonas aeruginosa* (Figure 7c). We also found significant variations in relative abundances of the *Leptospira* genus, which was recently described to be enriched in wastewater effluents in Germany (Numberger et al., 2019) (Figure 7d).

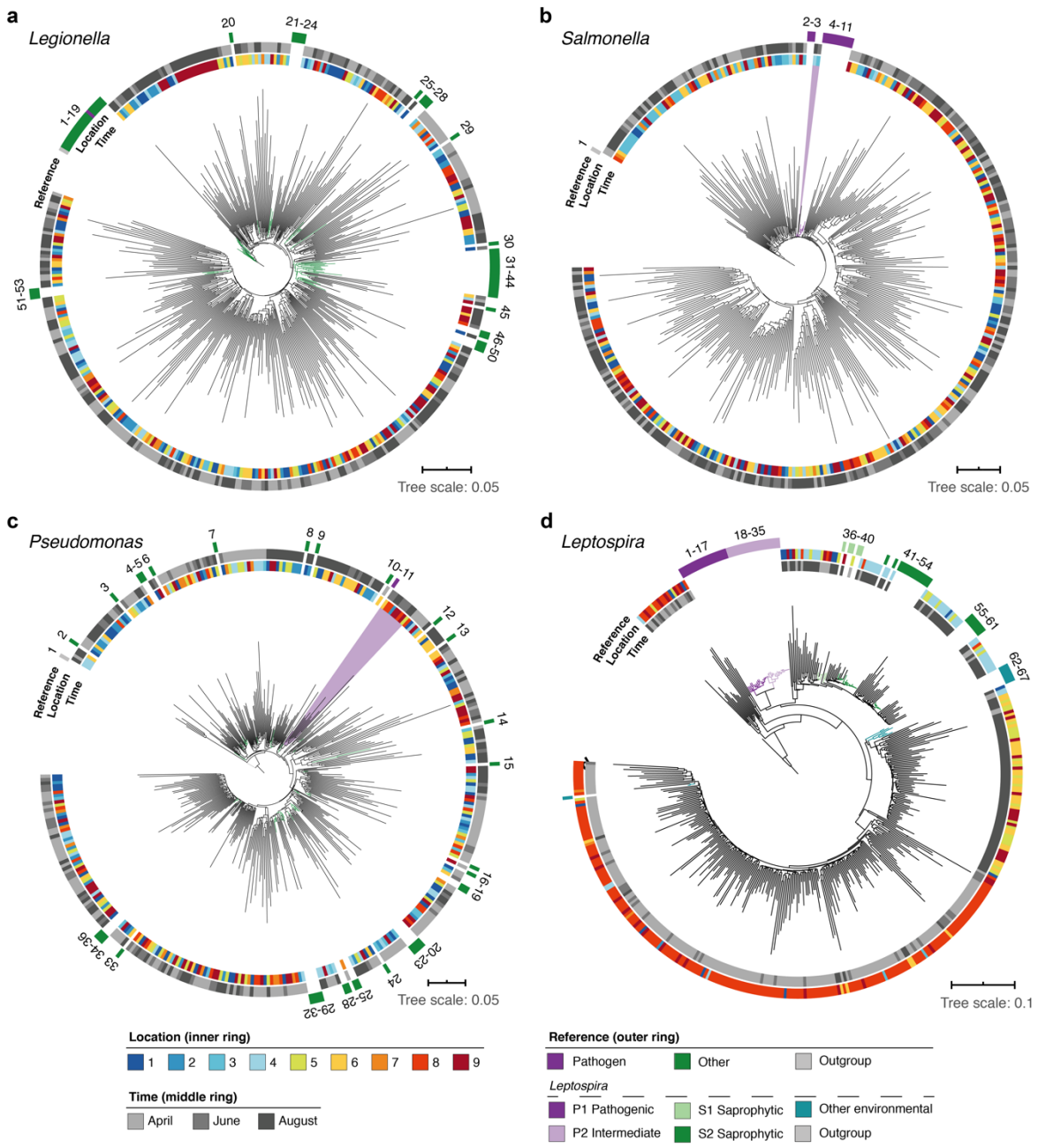


Figure 7: High-resolution phylogenetic clustering of candidate pathogenic genera in the River Cam. Phylogenetic trees illustrating multiple sequence alignments of exemplary River Cam nanopore reads classified as (a) *Legionella*, (b) *Salmonella*, (c) *Pseudomonas*, or (d) Leptospira, together with known reference species sequences ranging from pathogenic to saprophytic taxa within the genes (Supplementary Table 4). Reads highlighted in light violet background display close clustering with pathogenic isolates of (b) *Salmonella spp.*, (c) *Pseudomonas aeruginosa* and (d) *Leptospira spp.*

Indeed, the peak of River Cam *Leptospira* reads falls into an area of increased sewage influx (Figure 6). The *Leptospira* genus contains several potentially pathogenic species capable of causing life-threatening leptospirosis through waterborne infections, however also features close-related saprophytic and 'intermediate' taxa (Vincent et al., 2019; Wynwood et al., 2014). To resolve its complex phylogeny in the River Cam surface, we aligned *Leptospira* reads from all samples together with many reference sequences assigned to pre-classified pathogenic, saprophytic and other environmental *Leptospira* species (Vincent et al., 2019) (Figure 7d; Supplementary Table 4; Material and Methods). Despite the presence of nanopore sequencing errors (Supplementary Figure 5) and correspondingly inflated read divergence, we could pinpoint spatial clusters and a distinctly higher similarity between our amplicons and saprophytic rather than pathogenic *Leptospira* species. These findings were subsequently validated by targeted, *Leptospira* species-specific qPCR (Supplementary Table 5; Material and Methods), confirming that the latest nanopore sequencing quality is high enough to yield indicative results for bacterial monitoring workflows at the species level.

DISCUSSION

Using an inexpensive, easily adaptable and scalable framework based on nanopore sequencing, we provide the first spatiotemporal nanopore sequencing atlas of bacterial microbiota along a river reach. Our results suggest that our workflow robustly assesses both, the core microbiome of an example fluvial ecosystem and heterogeneous bacterial compositions in the context of supporting physical (temperature, flow rate) and hydrochemical (pH, inorganic solutes) parameters. We show that the technology's current sequencing accuracy of ~92 % allows for the designation of significant human pathogen community shifts along rural-to-urban river transitions, as illustrated by downstream increases in the abundance of pathogen candidates.

Our assessment of popular bioinformatics workflows for taxonomic classification highlights current challenges with error-prone nanopore sequences. We observed differences in terms of bacterial quantifications, read misclassification rates and consensus agreements between the twelve tested computational methods. In this computational benchmark, using the SILVA v.132 reference database, one of the most balanced performances was achieved by Minimap2 alignments. As nanopore sequencing quality continues to increase through refined pore chemistries, basecalling algorithms and consensus sequencing workflows (Calus, Ijaz, & Pinto, 2018; Karst et al., 2020; Rang, Kloosterman, & de Ridder, 2018), future bacterial taxonomic classifications are likely to improve and advance opportunities for aquatic species discovery.

We show that nanopore amplicon sequencing data can resolve the core microbiome of a freshwater body, as well as its temporal and spatial fluctuations. Besides common freshwater bacteria, we find that the differential abundances of *Carnobacteriaceae* most strongly contribute to seasonal loadings in the River Cam. *Carnobacteriaceae* have been previously associated with cold environments (Lawson & Caldwell, 2014), and we found these to be more abundant in colder April samples (mean 11.3 °C, vs. 15.8 °C in June and 19.1 °C in August). This might help to establish this family as an indicator for bacterial community shifts along with water temperature fluctuations.

Most routine freshwater surveillance frameworks focus on semi-quantitative diagnostics of only a limited number of target taxa, such as pathogenic *Salmonella, Legionella* and faecal coliforms (Ramirez-Castillo et al., 2015; Tan et al., 2015), whereas metagenomics approaches can give a complete and detailed overview of environmental microbial diversity. Beyond nanopore shotgun-sequencing (Reddington et al., 2020), our proof-of-principle analysis highlights that the combination of targeted full-length 16S rRNA gene MinION sequencing is a suitable complement to hydrochemical controls in pinpointing relatively contaminated freshwater sites, some of which in case of the River Cam had been previously highlighted for their pathogen diversity and abundance of antimicrobial resistance genes (Rowe et al., 2017; Rowe et al., 2016). Nanopore amplicon sequencing has here allowed us to reliably distinguish closely related pathogenic and non-pathogenic bacterial species of the common *Legionella*, *Salmonella*, *Pseudomonas* and *Leptospira* genera. For *Leptospira* bacteria, which are of particular interest to the Cam River stakeholder community, we further validated the nanopore sequencing results through the gold standard qPCR workflow of Public Health England (Supplementary Table 5).

A number of experimental intricacies should be addressed towards future nanopore freshwater sequencing studies with our approach, mostly by scrutinising water DNA extraction yields, PCR biases and molar imbalances in barcode multiplexing (Figure 3a; Supplementary Figure 5). Our results show that it would be theoretically feasible to obtain meaningful river microbiota from >100 barcoded samples on a single nanopore flow cell, thereby enabling water monitoring projects involving large collections at costs below £20 per sample (Supplementary Table 6). On the other hand, shotgun nanopore sequencing approaches such as the one employed by Reddington et al. (2020) may bypass pitfalls often observed in amplicon sequencing, namely taxon-specific primer biases (Frank et al., 2008), 16S rDNA copy number fluctuations between species (Darby, Todd, & Herman, 2013) or the

omission of functionally relevant sequence elements. In combination with sampling protocol adjustments, shotgun nanopore sequencing could moreover be used for the serial monitoring of eukaryotic microorganisms and viruses in freshwater ecosystems (Reddington et al., 2020).

Since the commercial launch of the MinION in 2015, a wide set of microbial nanopore sequencing applications in the context rRNA gene (Benitez-Paez, Portune, & Sanz, 2016; Cusco, Catozzi, Vines, Sanchez, & Francino, 2018; Kerkhof, Dillon, Haggblom, & McGuinness, 2017; Krehenwinkel et al., 2019) and shotgun (Leggett et al., 2019; Nicholls, Quick, Tang, & Loman, 2019; Reddington et al., 2020; Stewart et al., 2019) metagenomics have attracted the interest of a growing user community. Two independent case studies have recently provided decomposition analyses of faecal bacterial pathogens in MinION libraries derived from river and spring waters in Montana, USA (Hamner et al., 2019) and Kathmandu Valley, Nepal (Acharya et al., 2019). Although it is to be expected that short-read metagenomics technology continues to provide valuable environmental insights, as illustrated through global cataloguing efforts of ocean (Sunagawa et al., 2015), wastewater (Wu et al., 2019) and soil (Bahram et al., 2018) microbiomes, these traditional platforms remain unfeasible for the monitoring of remote environments – especially in low-resource settings. We reason that the low investment costs, convenience of MinION handling and complementary development of portable DNA purification methods (Boykin et al., 2019; Gowers et al., 2019) will allow for such endeavours to become increasingly accessible to citizens and public health organisations around the world, ultimately democratising the opportunities and benefits of DNA sequencing.

MATERIAL AND METHODS

1.1 Freshwater sampling

We monitored nine distinct locations along a 11.62 km reach of the River Cam, featuring sites upstream, downstream and within the urban belt of the city of Cambridge, UK. Measurements were taken at three time points, in two-month intervals between April and August 2018 (Figure 1; Supplementary Table 1a). To warrant river base flow conditions and minimise rain-derived biases, a minimum dry weather time span of 48h was maintained prior to sampling (Fisher, Newton, Dila, & McLellan, 2015). One litre of surface water was collected in autoclaved DURAN bottles (Thermo Fisher Scientific, Waltham, MA, USA), and cooled to 4 °C within three hours. Two bottles of water were collected consecutively for each time point, serving as biological replicates of location 9 (samples 9.1 and 9.2).

1.2 Physical and chemical metadata

We assessed various chemical, geological and physical properties of the River Cam (Figure 5; Supplementary Figure 4; Supplementary Table 1b-c).

In situ water temperature was measured immediately after sampling. To this end, we linked a DS18B20 digital temperature sensor to a portable custom-built, grid mounted Arduino nano v3.0 system. The pH was later recorded under temperature-controlled laboratory conditions, using a pH edge electrode (HI-11311, Hanna Instruments, Woodsocket, RI, USA).

To assess the dissolved ion concentrations in all collected water samples, we aerated the samples for 30 seconds and filtered them individually through a 0.22 µM pore-sized Millex-GP polyethersulfone syringe filter (MilliporeSigma, Burlington, MA, USA). Samples were then acidified to pH ~2, by adding 20 µL of 7M distilled HNO_3 per 3 mL sample. Inductively coupled plasma-optical emission spectroscopy (ICP-OES, Agilent 5100 SVDV; Agilent Technologies, Santa Clara, CA, USA) was used to analyse the dissolved cations Na^+, K^+, Ca^{2+}, Mg^{2+}, Ba^{2+}, Li^+, as well as Si and SO_4^{2-} (as total S) (Supplementary Table 1b). International water reference materials (SLRS-5 and SPS-SW2) were interspersed with the samples, reproducing certified values within 10 % for all analysed elements. Chloride concentrations were separately measured on 1 mL of non-acidified aliquots of the same samples, using a Dionex ICS-3000 ion chromatograph (Thermo Fisher Scientific, Waltham, MA, USA) (Supplementary Table 1b). Long-term repeat measurements of a USGS natural river water standard T-143 indicated precision of more than 4 % for Cl^-. However, the high Cl^- concentrations of the samples in this study were not fully bracketed by the calibration curve and we therefore assigned a more conservative uncertainty of 10 % to Cl^- concentrations.

High calcium and magnesium concentrations were recorded across all samples, in line with hard groundwater and natural weathering of the Cretaceous limestone bedrock underlying the river catchment (Figure 5a). There are no known evaporite salt deposits in the river catchment, and therefore the high dissolved Na^+, K^+ and Cl^- concentrations in the River Cam are likely derived from anthropogenic inputs (Rose, 2007) (Figure 5c-d). We calculated bicarbonate concentrations through a charge balance equation (concentrations in mol/L):

$$\text{conc}(HCO_3^-) = \text{conc}(Li^+) + \text{conc}(Na^+) + \text{conc}(K^+) + 2*\text{conc}(Mg^{2+}) + 2*\text{conc}(Ca^{2+}) - \text{conc}(Cl^-) - 2*\text{conc}(S^{2-})$$

The total dissolved solid (TDS) concentration across the 30 freshwater samples had a mean of 458 mg/L (range 325 - 605 mg/L) which is relatively high compared to most rivers, due to 1.) substantial solute load in the Chalk groundwater (particularly Ca^{2+}, Mg^{2+}, and HCO_3^-) and 2.) likely anthropogenic contamination (particularly Na^+, Cl^-, and SO_4^{2-}). The TDS range and the major ion signature of the River Cam is similar to other anthropogenically heavily-impacted rivers (Gaillardet et al., 1999), exhibiting enrichment in Na^+ (Figure 5d).

Overall, ion profiles clustered substantially between the three time points, indicating characteristic temporal shifts in water chemistry. PC1 of a PCA on the solute concentrations [µmol/L] shows a strong time effect, separating spring (April) from summer (June, August) samples (Figure 5b). We highlighted the ten most important features (i.e., features with the largest weights) and their contributions to PC1 (Figure 5c).

We integrated sensor data sets on mean daily air temperature, sunshine hours and total rainfall from a public, Cambridge-based weather station (Supplementary Figure 4a-c; Supplementary Table 1c). Similarly, mean gauged daily Cam water discharge [m^3s^{-1}] of the River Cam was retrieved through publicly available records from three upstream gauging stations connected to the UK National River Flow Archive (https://nrfa.ceh.ac.uk/), together with historic measurements from 1968 onwards (Supplementary Figure 4d)

1.3 DNA extraction

Within 24 hours of sampling, 400 mL of refrigerated freshwater from each site was filtered through an individual 0.22 µm pore-sized nitrocellulose filter (MilliporeSigma, Burlington, MA, USA) placed on a Nalgene polysulfone bottle top filtration holder (Thermo Fisher Scientific) at -30 mbar vacuum pressure. Additionally, 400 mL de-ionised (DI) water was also filtered. We then performed DNA extractions with a modified DNeasy PowerWater protocol (Qiagen, Hilden, Germany). Briefly, filters were cut into small slices with sterile scissors and transferred to 2 mL Eppendorf tubes containing lysis beads. Homogenization buffer PW1 was added, and the tubes subjected to ten minutes of vigorous shaking at 30 Hz in a TissueLyser II machine (Qiagen). After subsequent DNA binding and washing steps in accordance with the manufacturer's protocol, elution was done in 50 µL EB. We used Qubit dsDNA HS Assay (Thermo Fisher Scientific) to determine water DNA isolate concentrations (Supplementary Table 2a).

1.4 Bacterial full-length 16S rDNA sequence amplification

DNA extracts from each sampling batch and DI water control were separately amplified with V1-V9 full-length (~1.45 kbp) 16S rRNA gene primers, and respectively multiplexed with an additional sample with a defined bacterial mixture composition of eight species (*Pseudomonas aeruginosa*, *Escherichia coli*, *Salmonella enterica*, *Lactobacillus fermentum*, *Enterococcus faecalis*, *Staphylococcus aureus*, *Listeria monocytogenes*, *Bacillus subtilis*; D6305, Zymo Research, Irvine, CA, USA) (Figure 2), which was previously assessed using nanopore shotgun metagenomics (Nicholls et al., 2019). We used common primer binding sequences 27f and 1492r, both coupled to unique 24 bp barcodes and a nanopore motor protein tether sequence (Supplementary Table 7). Full-length 16S rDNA PCRs were performed with 30.8 µL DI water, 6.0 µL barcoded primer pair (10 µM), 5.0 µL PCR-buffer with $MgCl_2$ (10x), 5.0 µL dNTP mix (10x), 3.0 µL freshwater DNA extract, and 0.2 µL Taq (Qiagen) under the following conditions:

94 °C - 2 minutes

94 °C - 30 seconds, 60 °C - 30 seconds, 72 °C - 45 seconds (35 cycles)

72 °C - 5 minutes

1.5 Nanopore library preparation

Amplicons were purified from reaction mixes with a QIAquick purification kit (Qiagen). Two rounds of alcoholic washing and two additional minutes of drying at room temperature were then performed, prior to elution in 30 µL 10 mM Tris-HCl pH 8.0 with 50 mM NaCl. After concentration measurements with Qubit dsDNA HS, twelve barcoded extracts of a given batch were pooled in equimolar ratios, to approximately 300 ng DNA total (Supplementary Table S2b). We used KAPA Pure Beads (KAPA Biosystems, Wilmington, MA, USA) to concentrate full-length 16S rDNA products in 21 µL DI water. Multiplexed nanopore ligation sequencing libraries were then made by following the SQK-LSK109 protocol (Oxford Nanopore Technologies, Oxford, UK).

1.6 Nanopore sequencing

R9.4 MinION flow cells (Oxford Nanopore Technologies) were loaded with 75 µl of ligation library. The MinION instrument was run for approximately 48 hours, until no further sequencing reads could be collected. Fast5 files were basecalled using Guppy (version 3.15) and output DNA sequence reads with Q>7 were saved as fastq files. Various output metrics per library and barcode are summarised in Supplementary Table 2c.

1.7 Leptospira validation

In collaboration with Public Health England, raw water DNA isolates of the River Cam from each location and time point were subjected to the UK reference service for leptospiral testing (Supplementary Table 5). This test is based on quantitative real-time PCR (qPCR) of 16S rDNA and *LipL32*, implemented as a TaqMan assay for the detection and differentiation of pathogenic and non-pathogenic *Leptospira* spp. from human serum. Briefly, the assay consists of a two-component PCR; the first component is a duplex assay that targets the gene encoding the outer membrane lipoprotein *LipL32*, which is reported to be strongly associated with the pathogenic phenotype. The second reaction is a triplex assay targeting a well conserved region within the 16S rRNA gene (*rrn*) in *Leptospira* spp. Three different genomic variations correlate with pathogenic (PATH probe), intermediate (i.e., those with uncertain pathogenicity in humans; INTER probe) and non-pathogenic *Leptospira* spp. (ENVIRO probe), respectively.

2. DNA sequence processing workflow

The described data processing and read classification steps were implemented using the Snakemake workflow management system (Köster & Rahmann, 2012) and are available on Github - together with all necessary downstream analysis scripts to reproduce the results of this manuscript (https://github.com/d-j-k/puntseq).

2.1 Read data processing

Reads were demultiplexed and adapters trimmed using Porechop (version 0.2.4, https://github.com/rrwick/porechop). The only non-default parameter set was '--check_reads' (to 50,000), to increase the subset of reads to search for adapter sets. Next, we removed all reads shorter than 1.4 kbp and longer than 1.6 kbp with Nanofilt (version 2.5.0, https://github.com/wdecoster/nanofilt).

We assessed read statistics including quality scores and read lengths using NanoStat (version 1.1.2, https://github.com/wdecoster/nanostat), and used Pistis (https://github.com/mbhall88/pistis) to create quality control plots. This allowed us to assess GC content and Phred quality score distributions, which appeared consistent across and within our reads. Overall, we obtained 2,080,266 reads for April, 737,164 for June, and 5,491,510 for August, with a mean read quality of 10.0 (Supplementary Table 2c).

2.2 Benchmarking of bacterial taxonomic classifiers using nanopore reads

We used twelve different computational tools for bacterial full-length 16S rDNA sequencing read classification (section 2.2.1).:

Tool	Version	Commands
BLASTN (Altschul, Gish, Miller, Myers, & Lipman, 1990; Camacho et al., 2009)	v.2.9.0+	# build database makeblastdb -in silva.fna -parse_seqids -blastdb_version 5 -title "2019-08-24_SILVA_BLASTdatabase" -dbtype nucl # run BLASTN blastn -db silva.fna -query Cam16S.fa -out Cam16S.out -outfmt '6'
Centrifuge (Kim et al., 2016)	v.1.0.4	# build database centrifuge -x centrifuge_16s_database -U Cam16S.fa --threads config["centrifuge_16s"]["threads"] --report-file Cam16S_report.tsv -S Cam16S.tab --met-stderr centrifuge-kreport -x centrifuge_16s_database Cam16S.tab {input} > Cam16S.kreport
IDTAXA (Murali et al., 2018)	Implemented in R *DECIPHER* v.2.10.2 (Wright, 2016)	load(“SILVA_SSU_r132_March2018.RData”) IdTaxa(Cam16S.fa, trainingSet, strand = "both", threshold = 0)
Kraken 2 (Wood, Lu, & Langmead, 2019; Wood & Salzberg, 2014)	v.2.0.7	# build database kraken2 --db kraken2_16s_database --output Cam16S,out --report Cam16S.kreport --gzip-compressed --threads 1 Cam16S.fa
MAPseq (Matias Rodrigues et al., 2017)	v.1.2.3	mapseq Cam16S.fa silva_ref.fa > Cam16S.mseq
MegaBLAST (Camacho et al., 2009; Morgulis et al., 2008)	v.2.9.0+	# build database makeblastdb -in silva.fna -parse_seqids -blastdb_version 5 -title "2019-08-24_SILVA_BLASTdatabase" -dbtype nucl # run megaBLAST blastn -task "megablast" -db silva.fna -query Cam16S.fa -out Cam16S.out -outfmt '6'
Minimap2 (Li, 2018)	v.2.13-r852-dirty	minimap2 -k 15 -d silva_k15.mmi silva.fna minimap2 -ax map-ont -L silva_k15.mmi Cam16S.fa > Cam16S.sam
Mothur (Schloss et al., 2009)	v.1.43.0	align.seqs(candidate=Cam16S.fa, template=mothur.silva.nr_v132.align, processors=1, ksize=6, align=needleman)
QIIME 2 (Bolyen et al., 2019)	v.2019.7	# classification using classify-consensus-blast qiime feature-classifier classify-consensus-blast --i-query Cam16S.qza --i-reference-reads silva.qza --i-reference-

		taxonomy silva_tax.qza --o-classification Cam16S.qza --output-dir /Qiime2/Cam16S_blastn
RDP (Wang et al., 2007)	Implemented in R *DADA2* v.1.12.1 (Callahan et al., 2016)	assignTaxonomy(seqs = Cam16S.fa, refFasta = silva_nr_v132_train_set.fa.gz", tryRC = T, outputBootstraps=T,minBoot=0)
SINTAX (Robert C. Edgar, 2016)	Implemented in VSEARCH v.2.13.3 (Rognes, Flouri, Nichols, Quince, & Mahe, 2016)	vsearch -makeudb_usearch silva_tax.fa -output silva_tax.udb vsearch -sintax Cam16S.fa -db silva_tax.udb -tabbedout Cam16S.sintax -strand both -sintax_cutoff 0.5
SPINGO (Allard et al., 2015)	v.1.3	spindex -k 8 -p 1 -d silva_spingo_orig.fa spingo -d silva_spingo_orig.fa -k 8 -a -i Cam16S.fa > Cam16S.spingo

2.2.1 Datasets

We used nanopore sequencing data from our mock community and freshwater amplicons for benchmarking the classification tools. We therefore subsampled (a) 10,000 reads from each of the three mock community sequencing replicates (section 1.4), and (b) 10,000 reads from an aquatic sample (April-8; three random draws served as replicates). We then used the above twelve classification tools to classify these reads against the same database, SILVA v.132 (Quast et al., 2013) (Figure 2).

2.2.2 Comparison of mock community classifications

For the mock community classification benchmark, we assessed the number of unclassified reads, misclassified reads (i.e. sequences not assigned to any of the seven bacterial families), and the root mean squared error (RMSE) between observed and expected taxon abundance of the seven bacterial families. Following the detection of a strong bias towards the *Enterobacteriaceae* family across all classification tools, we also analysed RMSE values after exclusion of this family (Figure 2b-c).

2.2.3 Comparison of river community classifications

For the aquatic sample, the number of unclassified reads were counted prior to monitoring the performance of each classification tool in comparison with a consensus classification, which we defined as majority vote across classifications from all computational workflows. We observed stable results across all three draws of 10,000

reads from the same dataset (data not shown), indicating a robust representation of the performance of each classifier.

2.2.4 Overall classification benchmark

Minimap2 performed second best at classifying the mock community (lowest RMSE), while also delivering freshwater bacterial profiles in line with the majority vote of other classification tools (Figure 2d-e), in addition to providing rapid speed (data not shown). Yet, the application of this software to our entire dataset caused insufficient memory errors (at ~150 Gb RAM with kmer length 12), likely due to major sequence redundancies within the SILVA v.132 reference fasta file. Therefore, to run each of our full samples within a reasonable memory limit of 50 Gb, it was necessary to reduce the number of threads to 1, raise the kmer size ('-k') to 15 and set the minibatch size ('-K') to 25M (i.e., the number of query bases that are processed at any time), prolonging the runtime of several samples to ~three days.

2.3 Bacterial analyses

2.3.1 General workflow

After applying Minimap2 to the processed reads as explained above (section 2.2.4), we processed the resulting SAM files by firstly excluding all header rows starting with the '@' sign and then transforming the sets of read IDs, SILVA IDs, and alignment scores to tsv files of unique read-bacteria assignments either on the bacterial genus or family level. All reads that could not be assigned to the genus or family level were discarded, respectively. In the case of a read assignment to multiple taxa with the same alignment score, we determined the lowest taxonomic level in which these multiple taxa would be included. If this level was above the genus or family level, respectively, we discarded the read.

2.3.2 Estimating the level of misclassifications and DNA contaminants

Across three independent sequencing replicates of the same linear bacterial community standard (section 2.2.1), we found that the fraction of reads assigned to unexpected genus level taxa lies at ~1 % when using the Minimap2 classifier and the SILVA v.132 database.

Raw quantified DNA, PCR amplicons and sequencing read counts were considerably less abundant in DI water negative controls, as compared to actual freshwater specimens (Supplementary Table 2a). Only the negative control of the most prolific flow cell run (August 2018) passed the relatively high confidence threshold of 37,000 sequencing reads on the family level (Figure 3b; Supplementary Figure 2; section 2.4). Further inspection of these negative control reads revealed that their metagenomic profile closely mimicked the taxonomic classification profiles of river samples within the same sequencing batch, in addition to low-level kit contaminants like alphaproteobacteria of the *Bradyrhizobium* and *Methylobacterium* genus (Salter et al., 2014) which were otherwise nearly completely absent in any of the true aquatic isolates (Supplementary Table 8).

2.3.3 Determination of nanopore sequencing accuracy

Minimap2 alignments against mock community taxa were used to determine the mean read-wise nanopore sequencing accuracy for this study (92.08 %), as determined by the formula:

accuracy = 1 - (read mismatch length ÷ read alignment length)

These values were calculated for each of all eight species against each sequencing replicate, using the samtools (v.1.3.1) stats function (Li et al., 2009).

2.4 Rarefaction and high-confidence samples

Sample-specific rarefaction curves were generated by successive subsampling of sequencing reads classified by Minimap2 against the SILVA v.132 database (section 2.2.1). For broader comparative data investigations, we chose to only retain samples that passed a conservative minimum threshold of 37,000 reads. Family and genus-level species richness was hence mostly kept at ~90 % of the original values, in accordance with stable evenness profiles across a series of 100 bootstrap replicates (Supplementary Figure 2; section 2.4.1). Although we mainly present a single example rarefied dataset within this manuscript, we repeated each analysis, including PCAs, hierarchical clustering and Mantel tests, based on additional rarefied datasets to assess the stability of all results.

2.4.1 Mantel test

We performed Mantel tests (using scikit-bio version 0.5.1) to compare rarefied datasets with the full dataset. We therefore compared the Euclidean distance based on Z-standardised bacterial genera between all samples with more than 37,000 reads (two-sided test, 99,999 permutations). This resulted in a Pearson correlation of 0.814 (p

= $2.1*10^{-4}$) for our main rarefied dataset (results of the Mantel test applied to the remaining three other rarefied datasets: R = 0.819 and p = $1.0*10^{-4}$, R = 0.828 and p = $8.0*10^{-5}$, R = 0.815 and p = $1.4*10^{-4}$, respectively). Results of the Mantel tests applied to the genus-level bacterial classifications were also similar for all four subsampled datasets (R = 0.847 and p = $1.0*10^{-5}$, R = 0.863 and p = $1.0*10^{-5}$, R = 0.851 and p = $1*10^{-5}$, R = 0.856 and p = $1.0*10^{-5}$).

2.5 Meta-level bacterial community analyses

All classification assessment steps and summary statistics were performed in R or python (https://github.com/d-j-k/puntseq). We used the python package 'scikit-bio' for the calculation of the Simpson index and the Shannon's diversity as well as equitability index.

2.6 Data processing for hierarchical clustering, principal component and outlier analyses

Rarefied read count data was subjected to a $\log_{10}(x+1)$ transformation before hierarchical clustering using the complete linkage method. For PCA analyses, rarefied read count data was subjected to $\log_{10}(x+1)$ and Z-transformations. Negative control samples were removed. Mock community samples were initially removed to then be re-aligned to the eigenspace determined by the aquatic samples. We provide PCA visualisations of the main principal components (PCs explaining >10 % variance, respectively). For each of these relevant PCs, we further highlight the ten most important features (i.e., taxa with the largest weights) and their contributions to the PCs in barplots. For detecting outlier bacterial families per sample, we chose bacteria which were 1.) identified by more than 500 reads and 2.) which were at least five times more abundant in any single sample than in the mean of all samples combined.

2.7 Pathogen candidate assessments

A list of 55 known bacterial pathogenic genera, spanning 37 families, was compiled for targeted sequence testing. This was done through the careful integration of curated databases and online sources, foremost using PATRIC (Wattam et al., 2017) and data on known waterborne pathogens (Jin et al., 2018) (Supplementary Table 3a). Additionally, we integrated known genera from a large wastewater reference collection (Wu et al., 2019) (Supplementary Table 3b).

To identify if DNA reads assigned to *Leptospiraceae* were more similar to sequence reads of previously identified pathogenic, intermediate or environmental *Leptospira* species, we built a neighbour-joining tree of *Leptospiraceae* reads classified in our samples data, together with sequences from reference databases (Figure 3d; species names and NCBI accession numbers in clockwise rotation around the tree in Supplementary Table 4d). We matched the orientation of our reads, and then aligned them with 68 *Leptospira* reference sequences and the *Leptonema illini* reference sequence (DSM 21528 strain 3055) as outgroup. We then built a neighbour-joining tree using Muscle v.3.8.31 (R. C. Edgar, 2004), excluding three reads in the 'Other Environmental' clade that had extreme branch lengths >0.2. The reference sequences were annotated as pathogenic and saprophytic clades P1, P2, S1, S2 as recently described (Vincent et al., 2019). Additional published river water *Leptospira* that did not fall within these clades were included as 'Other Environmental' (Ganoza et al., 2006). Similarly, we constructed phylogenies for the *Legionella*, *Salmonella* and *Pseudomonas* genus, using established full-length 16S reference species sequences from NCBI (Supplementary Table 4a-c).

3. Total project cost

This study was designed to enable freshwater microbiome monitoring in budget-constrained research environments. Although we had access to basic infrastructure such as pipettes, a PCR and TissueLyser II machine, as well a high-performance laptop, we wish to highlight that the total sequencing consumable costs were held below £4,000 (Supplementary Table 6a). Here, individual costs ranged at ~£75 per sample (Supplementary Table 6b). With the current MinION flow cell price of £720, we estimate that per-sample costs could be further reduced to as low as ~£15 when barcoding and pooling ~100 samples in the same sequencing run (Supplementary Table 6c). Assuming near-equimolar amplicon pooling, flow cells with an output of ~5,000,000 reads can yield well over 37,000 sequences per sample and thereby surpass this conservative threshold applied here for comparative river microbiota analyses.

ACKNOWLEDGEMENTS

We thank Meltem Gürel, Christian Schwall, Jack Monahan, Eirini Vamva, Astrid Wendler, Ben Wagstaff, Elliot Brooks, Jennifer Pratscher, Rob Field, David Seilly, Mervyn Greaves, Tim Brooks, Daniel Bailey, Jenny Molloy, Michal Filus, Aleix Lafita, Oana Stroe, Abigail Wood, Paul Saary, Jane Clarke, Fiona Gilsenan and her family, Zamin Iqbal, Rob Finn, Alex Greenwood, Daniela Numberger, Julian Parkhill, Simon Frost, Sam Stubbs, Mark Holmes, Alicja Dabrowska, Alex Patto, Adrien Leger, Kim Judge, Alina Ham, Heather Martinez, Gemma

Gambrill, Víctor de Lorenzo, David Sargan, Lisa Schmunk, Amanda Clare and Alejandro de Miquel Bleier for helpful comments and assistance with this project. We thank Lilo and Manfred Fuchs from the Fuchs Fund for supporting LU's conference participation and presentation.

FUNDING

This study was funded by the OpenPlant Fund (BBSRC BB/L014130/1) and the University of Cambridge RCUK Catalyst Seed Fund. LU, MH and DEMH were funded by an EMBL PhD Fellowship. LU's Fellowship was financed by the European Union's Horizon2020 research and innovation programme (grant agreement number N635290). AH and MRS received Gates Cambridge Trust PhD scholarships. DJK was supported by the Wellcome Trust under grants 203828/Z/16/A and 203828/Z/16/Z. MJS was funded through the Oliver Gatty Studentship. SNP was funded by Wellcome Ph.D. Studentship 102453/Z/13/Z. JJB and ETT acknowledge NERC standard grant NE/P011659/1.

AUTHOR CONTRIBUTIONS

LU, AH, JJB, PBW, MJS, DJK, ETT and MRS designed the research; PBW, DJK, DEMH and MRS acquired project funding; LU, AH, PBW, MJS, SNP, DJK, DEMH and MRS collected river samples; LU, AH, JJB, PBW, MJS, SNP, DJK and MRS performed the experiments; LU, AH, JJB, MH, SJS, and MRS analysed the data; LU, AH and MRS wrote the paper with input from all co-authors.

COMPETING INTERESTS

All authors of this manuscript declare no competing interest.

MATERIALS AND CORRESPONDENCE

Correspondence and requests for materials should be addressed to Maximilian Stammnitz (maxrupsta@gmail.com), or to Andre Holzer (andre.holzer.biotech@gmail.com) and Lara Urban (lara.h.urban@gmail.com).

DATA AVAILABILITY

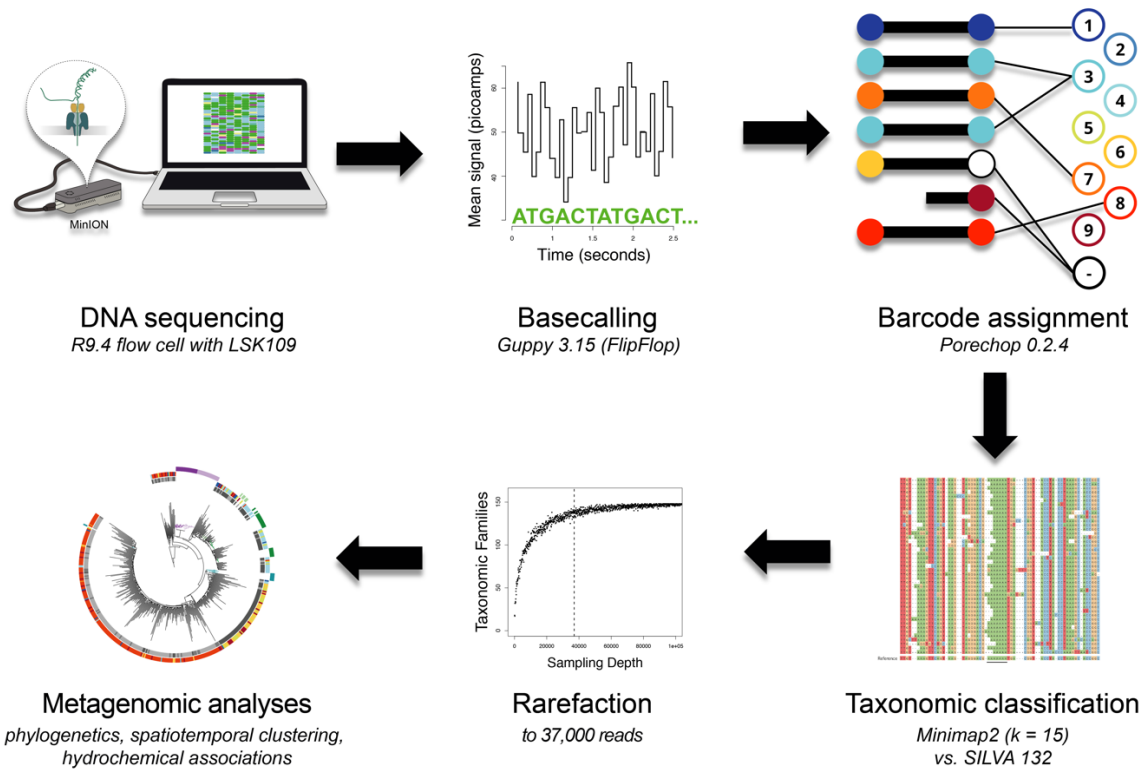
Sequencing datasets generated and analysed during this study are available from the European Nucleotide Archive, project accession PRJEB34900 (https://www.ebi.ac.uk/ena/data/view/PRJEB34900). The following

figures of this manuscript are based on this data: Figures 2, 3, 4, 6, 7, Supplementary Figures 2, 3, 5. Environmental measurements are available from public repositories, https://www.cl.cam.ac.uk/research/dtg/weather/ and https://nrfa.ceh.ac.uk/. The following figure of this manuscript are based on this data: Figure 5 and Supplementary Figure 4. There are no restrictions on data availability.

CODE AVAILABILITY

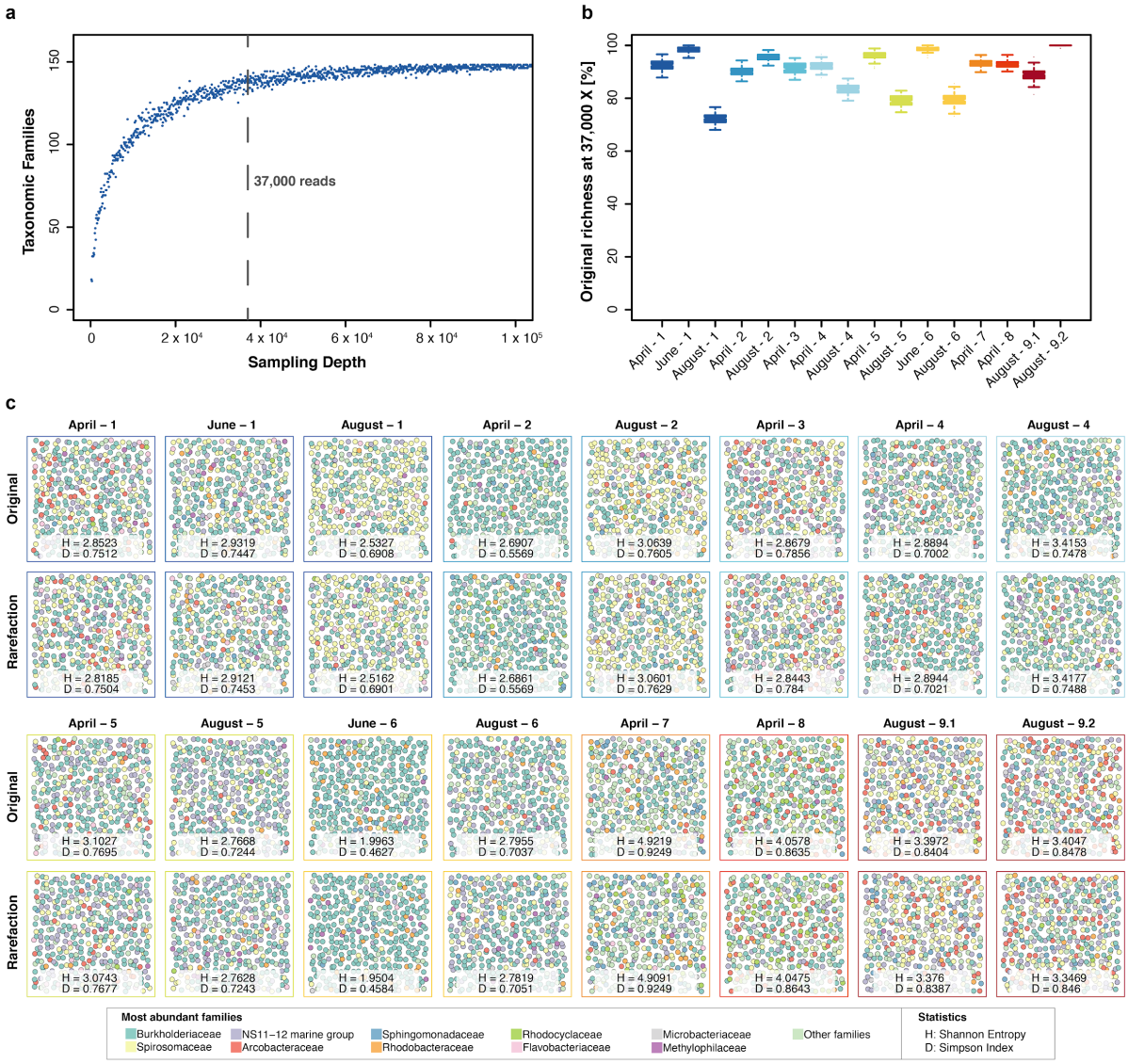
Our Github repository (https://github.com/d-j-k/puntseq/) features a Snakemake framework that integrates all data pre-processing steps, and a Singularity that contains all necessary software (https://github.com/d-j-k/puntseq/tree/master/analysis/). We further provide complete and rarefied SILVA 132 classifications from runs of Minimap2 (https://github.com/d-j-k/puntseq/tree/master/minimap2_classifications), which can be directly used as an input for reproducible downstream analyses.

SUPPLEMENTARY FIGURES

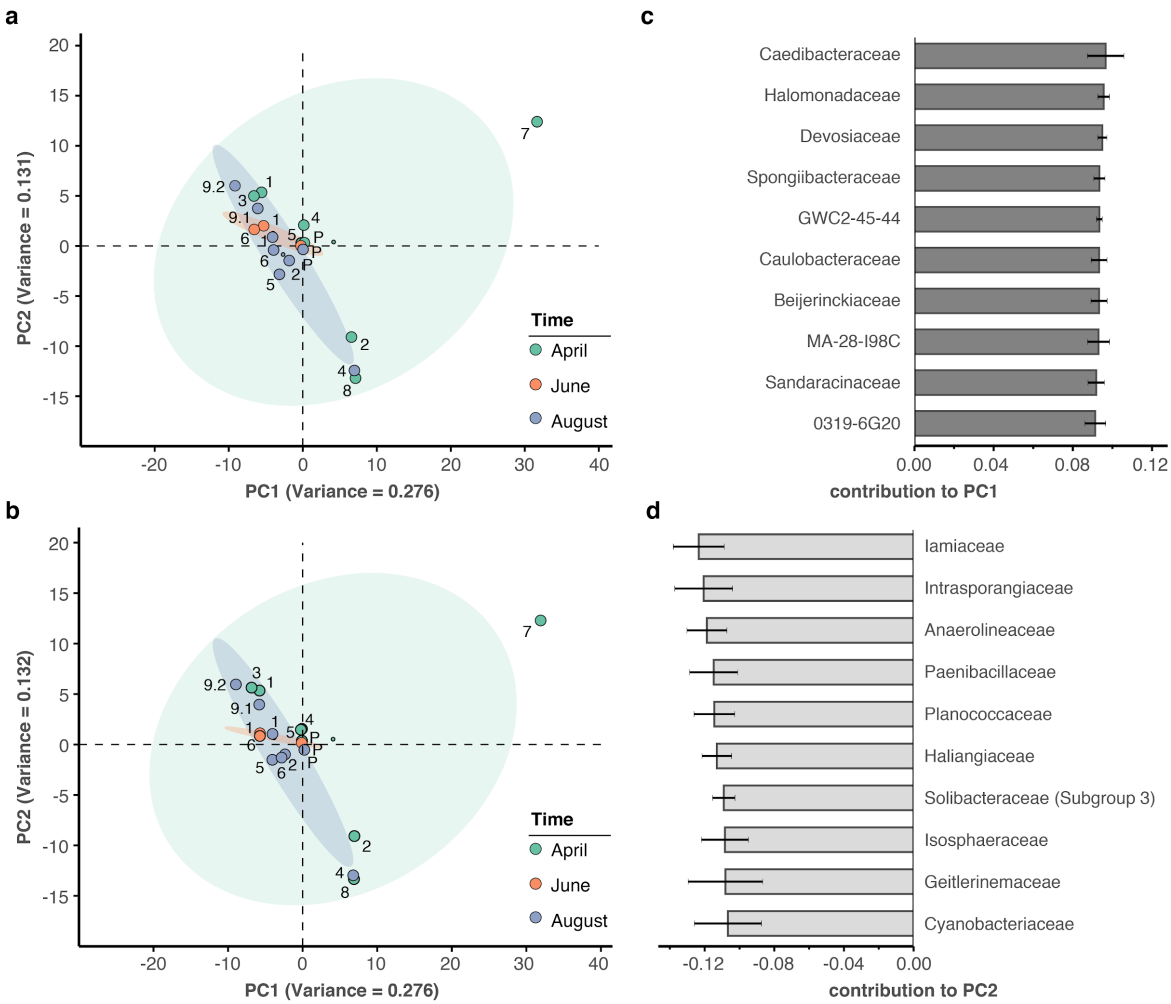


Supplementary Figure 1: Bioinformatics consensus workflow. Essential data processing steps, from nanopore sequencing to spatiotemporal bacterial composition analysis (Material and Methods). After full-length 16S rDNA sequencing with the MinION (R9.4 flow cell), local basecalling of the raw fast5 files was performed using Guppy (Wick, Judd, & Holt, 2019). Output fastq files were filtered for length and quality (Material and Methods), and reads assigned to their location barcode using Porechop. We then used Minimap2 (k = 15) and the SILVA v.132

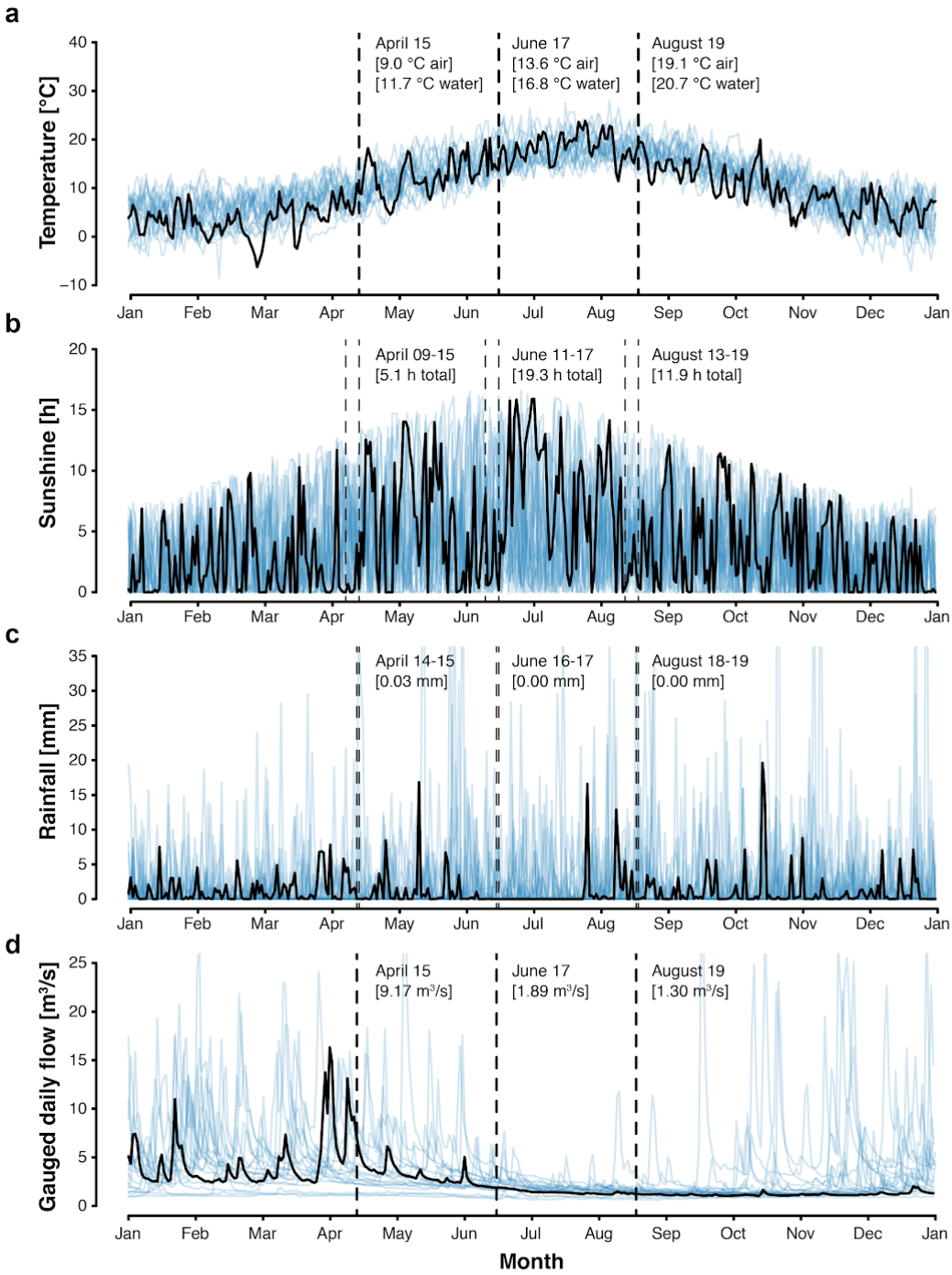
database for taxonomic classifications. Rarefaction reduced each sample to the same number of reads (37,000), allowing for a robust comparison of bacterial composition across samples in various downstream analyses.



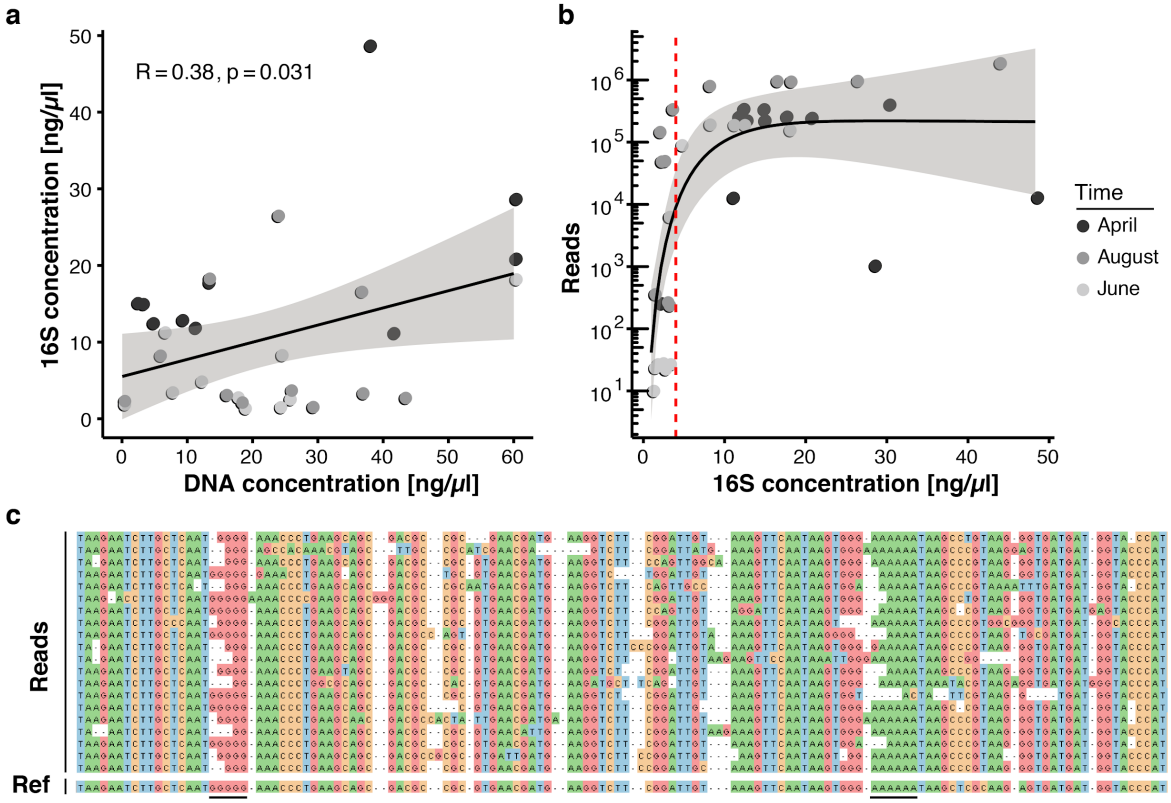
Supplementary Figure 2: Impact of rarefaction on diversity estimation. (a) Example rarefaction curve for bacterial family classifications of the 'April-1' sample. The chosen cut-off preserves most (~90 %) of the original family taxon richness (vertical line). (b) Difference between original and rarefied family richness at 37,000 reads across all freshwater sequencing runs with quantitative sequencing outputs above the chosen cut-off. Boxplots feature 100 independent rarefactions per sample. Error bars represent Q1 – 1.5*IQR (lower), and Q3 + 1.5*IQR (upper), respectively. (c) Diversity visualisation of the ten most abundant bacterial families across all samples with sequencing outputs >37,000 reads, through 400 'unordered bubbles'. Taxonomic proportions and colours are in accordance with Figure 2b. Shannon (H) and Simpson (D) indices for all samples indicate marginal differences between pairs of original and rarefied sets.



Supplementary Figure 3: Principal component analysis of river bacterial family compositions. (a-b) PCA with two independent rarefaction sets to 37,000 reads in all freshwater sequencing samples. Numbers and coloured dots indicate locations for each time point. The first and second principal components (PC1 and PC2, combined variance: ~41 %) robustly capture outlier samples 'April-7' along PC1 and 'April-2', 'August-4' and 'April-8' along PC2. (c-d) Fractional loads of the ten bacterial families most strongly contributing to changes along PC1 (c) and along PC2 (d). Error bars represent standard deviation of these families to the respective PC across four independent rarefactions. Subsequent principal components (PC3 and PC4) are less outlier-driven and depict spatial and temporal metagenomic trends within the River Cam.



Supplementary Figure 4: Cambridge weather and River Cam flow rate. (a) Daily air temperature [°C], (b) daily sunshine [hours], and (c) daily rainfall [mm] of Cambridge in 2018 (black trend line) vs. 1998-2017 (blue background trend lines). (d) Cam River gauged daily flow [m^3s^{-1}] in 2018 (black trend line) vs. 1968-2017 (blue background trend lines). Data was compiled from public repositories https://www.cl.cam.ac.uk/research/dtg/weather/ and https://nrfa.ceh.ac.uk/. Gauged daily flow measurements at Jesus Lock, Cambridge (between sampling locations 5 and 6; NRFA #33016) were discontinued in 1983. Yet, contemporary flow rates can be modelled with high accuracy (Pearson's R = 0.9, R^2 = 0.8) through linear data integration of three upstream stations already in operation since before 1983: Rhee at Wimpole (NRFA #33027, 70.2 % model weight), Granta at Stapleford (NRFA #33053, 19.6 % model weight) and Cam at Dernford (NRFA #33024, 10.3 % model weight).



Supplementary Figure 5: Key challenges of freshwater monitoring with nanopore sequencing. (a-b) Correlation analysis between DNA extraction yield, 16S amplification yield and raw sequencing output (Supplementary Table 2). (a) DNA concentrations (x-axis) obtained from 30 freshwater samples after extraction with the DNeasy PowerWater Kit (Material and Methods) are compared against the DNA concentration of the same samples after full-length 16S PCR amplification (y-axis), as measured by Qubit dsDNA HS. Fitted linear model displays the 95 % confidence interval (R = Pearson correlation coefficient). (b) The DNA concentration obtained for each sample after full-length 16S PCR amplification (x-axis) is compared against the final number of demultiplexed nanopore sequencing reads. Logarithmic fit with 95 % confidence interval indicates that samples with a minimum input concentration measurement of ~5 ng/µl yielded sequencing outputs sufficient to pass the rarefaction threshold of 37,000 reads. (c) Multiple sequence alignment of an example set of related nanopore 16S sequences, displaying increased indel rates at homopolymer reference sites (underlined); the mean sequencing error rate for this study lies at 7.92 %.

SUPPLEMENTARY TABLE LEGENDS

Table S1: Summary of samples and metadata. (a) Sampling locations. (b) Environmental metadata by sample. (c) Environmental metadata by time point.

Table S2: Summary of raw DNA, amplicon and sequencing yields. (a) Water DNA extraction yields. (b) Full-length 16S PCR amplicon extraction yields. (c) Nanopore sequencing read metrics.

Table S3: Summary of pathogen and wastewater bacterial genera tested. (a-b) List of pathogen (a) and wastewater (b) candidate bacterial genera.

Table S4: Summary of reference sequences for high-resolution pathogen mapping. (a-d) References and NCBI accessions for *Legionella* (a), *Salmonella* (b), *Pseudomonas* (c) and *Leptospira* (d).

Table S5: Summary of multi-species *Leptospira* quantifications by Taqman qPCR.

Table S6: Summary of project costs. (a) Basic sequencing workflow cost estimate. (b) Cost estimate per sample, based on a 12-plex MinION sequencing run. (c) Projected cost estimate per sample, based on a 100-plex MinION sequencing run.

Table S7: Summary of full-length 16S primer sequences (5' - 3').

Table S8: Summary of negative controls. (a-c) Relative classification output per sample (%), sorted by negative control abundances in April (a), June (b) and August (c).

REFERENCES

Acharya, K., Khanal, S., Pantha, K., Amatya, N., Davenport, R. J., & Werner, D. (2019). A comparative assessment of conventional and molecular methods, including MinION nanopore sequencing, for surveying water quality. *Sci Rep, 9*(1), 15726. doi:10.1038/s41598-019-51997-x

Allard, G., Ryan, F. J., Jeffery, I. B., & Claesson, M. J. (2015). SPINGO: a rapid species-classifier for microbial amplicon sequences. *BMC Bioinformatics, 16*, 324. doi:10.1186/s12859-015-0747-1

Almeida, A., Mitchell, A. L., Boland, M., Forster, S. C., Gloor, G. B., Tarkowska, A., . . . Finn, R. D. (2019). A new genomic blueprint of the human gut microbiota. *Nature, 568*(7753), 499-504. doi:10.1038/s41586-019-0965-1

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology, 215*(3), 403-410. doi:10.1016/S0022-2836(05)80360-2

Bahram, M., Hildebrand, F., Forslund, S. K., Anderson, J. L., Soudzilovskaia, N. A., Bodegom, P. M., . . . Bork, P. (2018). Structure and function of the global topsoil microbiome. *Nature, 560*(7717), 233-237. doi:10.1038/s41586-018-0386-6

Bartram, J., Lewis, K., Lenton, R., & Wright, A. (2005). Focusing on improved water and sanitation for health. *The Lancet, 365*(9461), 810-812. doi:10.1016/s0140-6736(05)17991-4

Benitez-Paez, A., Portune, K. J., & Sanz, Y. (2016). Species-level resolution of 16S rRNA gene amplicons sequenced through the MinION portable nanopore sequencer. *Gigascience, 5*, 4. doi:10.1186/s13742-016-0111-z

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., . . . Caporaso, J. G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol, 37*(8), 852-857. doi:10.1038/s41587-019-0209-9

Boykin, L. M., Sseruwagi, P., Alicai, T., Ateka, E., Mohammed, I. U., Stanton, J. L., . . . Ndunguru, J. (2019). Tree Lab: Portable genomics for Early Detection of Plant Viruses and Pests in Sub-Saharan Africa. *Genes (Basel), 10*(9). doi:10.3390/genes10090632

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods, 13*(7), 581-583. doi:10.1038/nmeth.3869

Calus, S. T., Ijaz, U. Z., & Pinto, A. J. (2018). NanoAmpli-Seq: a workflow for amplicon sequencing for mixed microbial communities on the nanopore sequencing platform. *Gigascience, 7*(12). doi:10.1093/gigascience/giy140

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics, 10*, 421. doi:10.1186/1471-2105-10-421

Chan, J. F.-W., Yuan, S., Kok, K.-H., To, K. K.-W., Chu, H., Yang, J., . . . Yuen, K.-Y. (2020). A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *The Lancet*. doi:10.1016/s0140-6736(20)30154-9

Cusco, A., Catozzi, C., Vines, J., Sanchez, A., & Francino, O. (2018). Microbiota profiling with long amplicons using Nanopore sequencing: full-length 16S rRNA gene and the 16S-ITS-23S of the rrn operon. *F1000Res, 7*, 1755. doi:10.12688/f1000research.16817.2

Darby, B. J., Todd, T. C., & Herman, M. A. (2013). High-throughput amplicon sequencing of rRNA genes requires a copy number correction to accurately reflect the effects of management practices on soil nematode community structure. *Mol Ecol, 22*(21), 5456-5471. doi:10.1111/mec.12480

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res, 32*(5), 1792-1797. doi:10.1093/nar/gkh340

Edgar, R. C. (2016). SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS sequences. *bioRxiv*, 074161. doi:10.1101/074161

Faria, N. R., Kraemer, M. U. G., Hill, S. C., Goes de Jesus, J., Aguiar, R. S., Iani, F. C. M., . . . Pybus, O. G. (2018). Genomic and epidemiological monitoring of yellow fever virus transmission potential. *Science, 361*(6405), 894. doi:10.1126/science.aat7115

Faria, N. R., Quick, J., Claro, I. M., Theze, J., de Jesus, J. G., Giovanetti, M., . . . Pybus, O. G. (2017). Establishment and cryptic transmission of Zika virus in Brazil and the Americas. *Nature, 546*(7658), 406-410. doi:10.1038/nature22401

Fisher, J. C., Newton, R. J., Dila, D. K., & McLellan, S. L. (2015). Urban microbial ecology of a freshwater estuary of Lake Michigan. *Elementa (Wash D C), 3*. doi:10.12952/journal.elementa.000064

Frank, J. A., Reich, C. I., Sharma, S., Weisbaum, J. S., Wilson, B. A., & Olsen, G. J. (2008). Critical evaluation of two primers commonly used for amplification of bacterial 16S rRNA genes. *Appl Environ Microbiol, 74*(8), 2461-2470. doi:10.1128/AEM.02272-07

Gaillardet, J., Dupré, B., Louvat, P., & Allègre, C. J. (1999). Global silicate weathering and CO2 consumption rates deduced from the chemistry of large rivers. *Chemical Geology, 159*(1), 3-30. doi:10.1016/S0009-2541(99)00031-5

Ganoza, C. A., Matthias, M. A., Collins-Richards, D., Brouwer, K. C., Cunningham, C. B., Segura, E. R., . . . Vinetz, J. M. (2006). Determining risk for severe leptospirosis by molecular analysis of environmental surface waters for pathogenic Leptospira. *PLoS Med, 3*(8), e308. doi:10.1371/journal.pmed.0030308

Gardy, J. L., & Loman, N. J. (2018). Towards a genomics-informed, real-time, global pathogen surveillance system. *Nat Rev Genet, 19*(1), 9-20. doi:10.1038/nrg.2017.88

Gowers, G. F., Vince, O., Charles, J. H., Klarenberg, I., Ellis, T., & Edwards, A. (2019). Entirely Off-Grid and Solar-Powered DNA Sequencing of Microbial Communities during an Ice Cap Traverse Expedition. *Genes (Basel), 10*(11). doi:10.3390/genes10110902

Haddeland, I., Heinke, J., Biemans, H., Eisner, S., Florke, M., Hanasaki, N., . . . Wisser, D. (2014). Global water resources affected by human interventions and climate change. *Proc Natl Acad Sci U S A, 111*(9), 3251-3256. doi:10.1073/pnas.1222475110

Hamner, S., Brown, B. L., Hasan, N. A., Franklin, M. J., Doyle, J., Eggers, M. J., . . . Ford, T. E. (2019). Metagenomic Profiling of Microbial Pathogens in the Little Bighorn River, Montana. *Int J Environ Res Public Health, 16*(7). doi:10.3390/ijerph16071097

Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol, 17*(1), 239. doi:10.1186/s13059-016-1103-0

Jin, D., Kong, X., Cui, B., Jin, S., Xie, Y., Wang, X., & Deng, Y. (2018). Bacterial communities and potential waterborne pathogens within the typical urban surface waters. *Sci Rep, 8*(1), 13368. doi:10.1038/s41598-018-31706-w

Kafetzopoulou, L. E., Pullan, S. T., Lemey, P., Suchard, M. A., Ehichioya, D. U., Pahlmann, M., . . . Duraffour, S. (2019). Metagenomic sequencing at the epicenter of the Nigeria 2018 Lassa fever outbreak. *Science, 363*(6422), 74. doi:10.1126/science.aau9343

Karst, S. M., Ziels, R. M., Kirkegaard, R. H., Sørensen, E. A., McDonald, D., Zhu, Q., . . . Albertsen, M. (2020). Enabling high-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. *bioRxiv*, 645903. doi:10.1101/645903

Kayman, T., Abay, S., Hizlisoy, H., Atabay, H. I., Diker, K. S., & Aydin, F. (2012). Emerging pathogen Arcobacter spp. in acute gastroenteritis: molecular identification, antibiotic susceptibilities and genotyping of the isolated arcobacters. *J Med Microbiol, 61*(Pt 10), 1439-1444. doi:10.1099/jmm.0.044594-0

Kerkhof, L. J., Dillon, K. P., Haggblom, M. M., & McGuinness, L. R. (2017). Profiling bacterial communities by MinION sequencing of ribosomal operons. *Microbiome, 5*(1), 116. doi:10.1186/s40168-017-0336-9

Kim, D., Song, L., Breitwieser, F. P., & Salzberg, S. L. (2016). Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res, 26*(12), 1721-1729. doi:10.1101/gr.210641.116

Köster, J., & Rahmann, S. (2012). Snakemake--a scalable bioinformatics workflow engine. *Bioinformatics, 28*(19), 2520-2522. doi:10.1093/bioinformatics/bts480

Krehenwinkel, H., Pomerantz, A., Henderson, J. B., Kennedy, S. R., Lim, J. Y., Swamy, V., . . . Prost, S. (2019). Nanopore sequencing of long ribosomal DNA amplicons enables portable and simple biodiversity assessments with high phylogenetic resolution across broad taxonomic scale. *Gigascience, 8*(5). doi:10.1093/gigascience/giz006

Lawson, P. A., & Caldwell, M. E. (2014). The Family Carnobacteriaceae. In (pp. 19-65). Berlin, Heidelberg: Springer Berlin Heidelberg.

Leggett, R. M., Alcon-Giner, C., Heavens, D., Caim, S., Brook, T. C., Kujawska, M., . . . Clark, M. D. (2019). Rapid MinION profiling of preterm microbiota and antimicrobial-resistant pathogens. *Nat Microbiol*. doi:10.1038/s41564-019-0626-z

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics, 34*(18), 3094-3100. doi:10.1093/bioinformatics/bty191

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079. doi:10.1093/bioinformatics/btp352

Matias Rodrigues, J. F., Schmidt, T. S. B., Tackmann, J., & von Mering, C. (2017). MAPseq: highly efficient k-mer search with confidence estimates, for rRNA sequence analysis. *Bioinformatics, 33*(23), 3808-3810. doi:10.1093/bioinformatics/btx517

Morgulis, A., Coulouris, G., Raytselis, Y., Madden, T. L., Agarwala, R., & Schaffer, A. A. (2008). Database indexing for production MegaBLAST searches. *Bioinformatics, 24*(16), 1757-1764. doi:10.1093/bioinformatics/btn322

Murali, A., Bhargava, A., & Wright, E. S. (2018). IDTAXA: a novel approach for accurate taxonomic classification of microbiome sequences. *Microbiome, 6*(1), 140. doi:10.1186/s40168-018-0521-5

Nicholls, S. M., Quick, J. C., Tang, S., & Loman, N. J. (2019). Ultra-deep, long-read nanopore sequencing of mock microbial community standards. *Gigascience, 8*(5). doi:10.1093/gigascience/giz043

Nielsen, P. H., Saunders, A. M., Hansen, A. A., Larsen, P., & Nielsen, J. L. (2012). Microbial communities involved in enhanced biological phosphorus removal from wastewater--a model system in environmental biotechnology. *Curr Opin Biotechnol, 23*(3), 452-459. doi:10.1016/j.copbio.2011.11.027

Numberger, D., Ganzert, L., Zoccarato, L., Muhldorfer, K., Sauer, S., Grossart, H. P., & Greenwood, A. D. (2019). Characterization of bacterial communities in wastewater with enhanced taxonomic resolution by full-length 16S rRNA sequencing. *Sci Rep, 9*(1), 9673. doi:10.1038/s41598-019-46015-z

Payne, A., Holmes, N., Rakyan, V., & Loose, M. (2018). Whale watching with BulkVis: A graphical viewer for Oxford Nanopore bulk fast5 files. *bioRxiv*. doi:10.1101/312256

Prüss-Üstün, A., Kay, D., Fewtrell, L., & Bartram, J. (2002). Estimating the burden of disease from water, sanitation, and hygiene at a global level. *Environmental Health Perspectives, 110*(5), 537-542. doi:10.1289/ehp.110-1240845

Prüss-Üstün, A., Wolf, J., Bartram, J., Clasen, T., Cumming, O., Freeman, M. C., . . . Johnston, R. (2019). Burden of disease from inadequate water, sanitation and hygiene for selected adverse health outcomes: An updated analysis with a focus on low- and middle-income countries. *Int J Hyg Environ Health, 222*(5), 765-777. doi:10.1016/j.ijheh.2019.05.004

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., . . . Glockner, F. O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res, 41*(Database issue), D590-596. doi:10.1093/nar/gks1219

Quick, J., Ashton, P., Calus, S., Chatt, C., Gossain, S., Hawker, J., . . . Loman, N. J. (2015). Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella. *Genome Biol, 16*, 114. doi:10.1186/s13059-015-0677-2

Quick, J., Loman, N. J., Duraffour, S., Simpson, J. T., Severi, E., Cowley, L., . . . Carroll, M. W. (2016). Real-time, portable genome sequencing for Ebola surveillance. *Nature, 530*(7589), 228-232. doi:10.1038/nature16996

Ramirez-Castillo, F. Y., Loera-Muro, A., Jacques, M., Garneau, P., Avelar-Gonzalez, F. J., Harel, J., & Guerrero-Barrera, A. L. (2015). Waterborne pathogens: detection methods and challenges. *Pathogens, 4*(2), 307-334. doi:10.3390/pathogens4020307

Rang, F. J., Kloosterman, W. P., & de Ridder, J. (2018). From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biol, 19*(1), 90. doi:10.1186/s13059-018-1462-9

Reddington, K., Eccles, D., O'Grady, J., Drown, D. M., Hansen, L. H., Nielsen, T. K., . . . Brown, B. L. (2020). Metagenomic analysis of planktonic riverine microbial consortia using nanopore sequencing reveals insight into river microbe taxonomy and function. *Gigascience, 9*(6). doi:10.1093/gigascience/giaa053

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahe, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ, 4*, e2584. doi:10.7717/peerj.2584

Rose, S. (2007). The effects of urbanization on the hydrochemistry of base flow within the Chattahoochee River Basin (Georgia, USA). *Journal of Hydrology, 341*(1-2), 42-54. doi:10.1016/j.jhydrol.2007.04.019

Rowe, W., Baker-Austin, C., Verner-Jeffreys, D. W., Ryan, J. J., Micallef, C., Maskell, D. J., & Pearce, G. P. (2017). Overexpression of antibiotic resistance genes in hospital effluents over time. *J Antimicrob Chemother, 72*(6), 1617-1623. doi:10.1093/jac/dkx017

Rowe, W., Verner-Jeffreys, D. W., Baker-Austin, C., Ryan, J. J., Maskell, D. J., & Pearce, G. P. (2016). Comparative metagenomics reveals a diverse range of antimicrobial resistance genes in effluents entering a river catchment. *Water Sci Technol, 73*(7), 1541-1549. doi:10.2166/wst.2015.634

Salazar, G., & Sunagawa, S. (2017). Marine microbial diversity. *Curr Biol, 27*(11), R489-R494. doi:10.1016/j.cub.2017.01.017

Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., . . . Walker, A. W. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology, 12*(1), 87. doi:10.1186/s12915-014-0087-z

Schewe, J., Heinke, J., Gerten, D., Haddeland, I., Arnell, N. W., Clark, D. B., . . . Kabat, P. (2014). Multimodel assessment of water scarcity under climate change. *Proc Natl Acad Sci U S A, 111*(9), 3245-3250. doi:10.1073/pnas.1222460110

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., . . . Weber, C. F. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol, 75*(23), 7537-7541. doi:10.1128/AEM.01541-09

Stewart, R. D., Auffret, M. D., Warr, A., Walker, A. W., Roehe, R., & Watson, M. (2019). Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. *Nat Biotechnol, 37*(8), 953-961. doi:10.1038/s41587-019-0202-3

Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., . . . Bork, P. (2015). Structure and function of the global ocean microbiome. *Science, 348*(6237).

Tan, B., Ng, C., Nshimyimana, J. P., Loh, L. L., Gin, K. Y., & Thompson, J. R. (2015). Next-generation sequencing (NGS) for assessment of microbial water quality: current progress, challenges, and future opportunities. *Front Microbiol, 6*, 1027. doi:10.3389/fmicb.2015.01027

Tringe, S. G., & Rubin, E. M. (2005). Metagenomics: DNA sequencing of environmental samples. *Nat Rev Genet, 6*(11), 805-814. doi:10.1038/nrg1709

Vincent, A. T., Schiettekatte, O., Goarant, C., Neela, V. K., Bernet, E., Thibeaux, R., . . . Picardeau, M. (2019). Revisiting the taxonomy and evolution of pathogenicity of the genus Leptospira through the prism of genomics. *PLoS Negl Trop Dis, 13*(5), e0007270. doi:10.1371/journal.pntd.0007270

Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol, 73*(16), 5261-5267. doi:10.1128/AEM.00062-07

Wattam, A. R., Davis, J. J., Assaf, R., Boisvert, S., Brettin, T., Bun, C., . . . Stevens, R. L. (2017). Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. *Nucleic Acids Res, 45*(D1), D535-D542. doi:10.1093/nar/gkw1017

Wick, R. R., Judd, L. M., & Holt, K. E. (2019). Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol, 20*(1), 129. doi:10.1186/s13059-019-1727-y

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biol, 20*(1), 257. doi:10.1186/s13059-019-1891-0

Wood, D. E., & Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology, 15*(3), R46. doi:10.1186/gb-2014-15-3-r46

Wright, E. S. (2016). Using DECIPHER v2.0 to Analyze Big Biological Sequence Data in R. *The R Journal, 8*(1), 352-359.

Wu, L., Ning, D., Zhang, B., Li, Y., Zhang, P., Shan, X., . . . Zhou, J. (2019). Global diversity and biogeography of bacterial communities in wastewater treatment plants. *Nat Microbiol, 4*(7), 1183-1195. doi:10.1038/s41564-019-0426-5

Wynwood, S. J., Graham, G. C., Weier, S. L., Collet, T. A., McKay, D. B., & Craig, S. B. (2014). Leptospirosis from water sources. *Pathogens and Global Health, 108*(7), 334-338. doi:10.1179/2047773214Y.0000000156