Freshwater monitoring by nanopore sequencing

Authors: 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 depicts both, the hydrological core microbiome and fine temporal gradients in line with complementary physicochemical measurements. Using reference-based sequence clustering, 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[1], particularly in light of the pressures of rising urbanisation, agricultural intensification and climate change[2,3]. These trends enforce an increasing demand for freshwater monitoring frameworks that combine cost effectiveness, fast technology deployability and data transparency[4]. Environmental metagenomics, the tracing of organisms present in a substrate through high-throughput DNA sequencing, yields informative measures of relative taxonomic species occurrence and functional diversity[5]. Microbial metagenomics studies overcome enrichment biases common to traditional culturing approaches[5]; however, they usually depend on expensive and stationary equipment, highly specialised operational training and substantial time lags between fieldwork, sample preparation, raw data generation and access.

In recent years, these challenges have been revisited with the prospect of 'portable' DNA analysis. The main driver of this is the smartphone-sized MinION device from Oxford Nanopore Technologies (ONT), which enables real-time DNA sequencing using nanopores[6]. Nanopore read lengths can be comparably long (currently up to ~$2*10^6$ bases[7]), 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 proven powerful for real-time tracing and open data sharing during bacterial and viral pathogen outbreaks[8-13].

Here we report a simple, inexpensive workflow to assess microbial freshwater ecosystems with nanopore DNA sequencing. Our benchmark 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

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 1a). Samples were complemented with a negative control (deionised water) and a mock community control composed of eight bacterial species in known mixture proportions (Methods).

To obtain valid taxonomic assignments from freshwater sequencing profiles using nanopore sequencing, twelve different classification tools were compared through several performance metrics (Extended Data Figure 1; Methods). 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)[14]. Robust quantifications were obtained by Minimap2[15] alignments against the SILVA v.132 database[16], 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 more complex bacterial community,

although other tools such as SPINGO[17], MAPseq[18], or IDTAXA[19] also produced highly concordant results despite variations in processing speed and memory usage (data not shown).

Diversity analysis and river core microbiome

Using Minimap2 classifications within our bioinformatics consensus workflow (Extended Data Figure 2; 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 2a; Supplementary Table 2). Overall, 55.9 % (n = 4,644,194) of raw reads could be taxonomically assigned to the family level (Figure 2b). 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, $R = 0.814$, $p = 2.1*10^{-4}$; Extended Data Figure 3), 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 2b; Extended Data Figure 3): microbial family alpha diversity ranged between 0.46 (June-6) and 0.92 (April-7) (Simpson index), indicating partially 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 2c). 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 2d). 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. 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 2c). Principal component analysis (PCA) (Figure 3a; Extended Data Figure 4) revealed that the strongest differential abundances along the chronological axis of variation (PC3) derived from the higher abundance of *Carnobacteriaceae* in April (Figure 3b). This family is known for its occurrence in waters with low temperature[20].

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 (Extended Data Figures 5 and 6, respectively; Supplementary Table 1c). 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 (Extended Data Figure 5; 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. 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[21] (Extended Data Figure 5; Methods).

Maps of potential bacterial pathogens at species-level resolution

In line with these physicochemical trends, we 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[22,23], and also 13 wastewater-associated[24] bacterial genera (Supplementary Table 3). Of these, 21 potentially pathogenic and eight wastewater-associated genera were detected across all of the river samples (Figure 3c; 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). The most prolific candidate pathogen genus observed was *Arcobacter*, which features multiple species implicated in acute gastrointestinal infections[25].

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; Extended Data Figure 4a-b). This was characterised by an unusual dominance of *Caedibacteraceae*, *Halomonadaceae* and others (Extended Data Figure 4c). Isolate April-8 also showed a highly distinct bacterial composition, with some families nearly exclusively occurring in this sample (outlier analysis, 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*[24,26] (Figure 3c).

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, *Pseudomonas*, *Legionella* and *Salmonella* (Extended Data Figure 7; Methods). While *Legionella* and *Salmonella* diversities only presented negligible levels of known harmful species, a cluster of sequencing reads in downstream sections indicated a low abundance of the opportunistic, environmental pathogen *Pseudomonas aeruginosa* (Extended Data Figure 7).

We also found significant variations in relative abundances of the *Leptospira* genus, which was recently described to be enriched in wastewater effluents in Germany[27]. Indeed, the peak of River Cam *Leptospira* reads falls into an area of increased sewage influx (Figure 3c). The *Leptospira* genus contains several potentially pathogenic species capable of causing life-threatening leptospirosis through waterborne infections[28], however also features close-related saprophytic and 'intermediate' taxa[29]. To resolve its complex phylogeny in the River Cam surface, we aligned *Leptospira* reads from all samples together with various reference sequences assigned to pre-classified pathogenic, saprophytic and other environmental *Leptospira* species[29] (Figure 3d; Supplementary Table 4a; Methods). Despite the presence of nanopore sequencing errors (Extended Data Figure 8) 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, 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, we provide the first spatiotemporal nanopore sequencing atlas of bacterial microbiota along a river reach. Beyond the core microbiome of an example fluvial ecosystem, our results suggest that it is possible to robustly assess the heterogeneity in accessory bacterial composition in the context of supporting physical (temperature, flow rate) and hydrochemical (pH, inorganic solutes) parameters. We show that the technology's current 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.

Furthermore, 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[30-32], 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[20], 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[33,34]. Our proof-of-principle analysis highlights that the combination of full-length 16S rRNA gene amplification and nanopore sequencing can complement hydrochemical controls in pinpointing relatively contaminated freshwater sites, some of which had been previously highlighted for their pathogen diversity and abundance of antimicrobial resistance genes[35,36]. Nanopore sequencing here allowed for the reliable distinction of closely related pathogenic and non-pathogenic bacterial species of the common *Salmonella*, *Legionella*, *Pseudomonas* and *Leptospira* genera. For *Leptospira* bacteria, we further validated 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, mostly by scrutinising water DNA extraction yields, PCR biases and molar imbalances in barcode multiplexing

(Figure 2a; Extended Data Figure 8). Yet, 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). Barcoded shotgun nanopore sequencing protocols may pose a viable alternative strategy to bypass pitfalls often observed in amplicon-based workflows, namely taxon-specific primer biases[14], 16S rDNA copy number fluctuations between species[37] and the omission of functionally relevant sequence elements. In combination with sampling protocol adjustments, shotgun nanopore sequencing could moreover be used for the monitoring of eukaryotic microorganisms and viruses in freshwater ecosystems.

Since the commercial launch of the MinION in 2015, a wide set of nanopore sequencing applications like rRNA gene[38-41] and shotgun[42-45] metagenomics have attracted the interest of a growing user community; indeed, two independent case studies have recently provided decomposition analyses of faecal bacterial pathogens in MinION libraries derived from river and spring waters in Crow Agency (Montana, USA)[46] and Kathmandu Valley (Nepal)[47]. 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[48], wastewater[24] and soil[49] microbiomes, many traditional platforms are unfeasible for monitoring remote environments - especially in low-resource settings. We reason that the low investment costs, the convenience of MinION handling and complementary development of portable DNA purification methods[50] will allow for such endeavours to become increasingly accessible to citizens and public health organisations around the world, ultimately democratising the opportunities, open sharing and benefits of DNA sequencing.

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, Nick Loman, 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, DEMH, 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, DK, DEMH, 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 correspondance*:* 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)

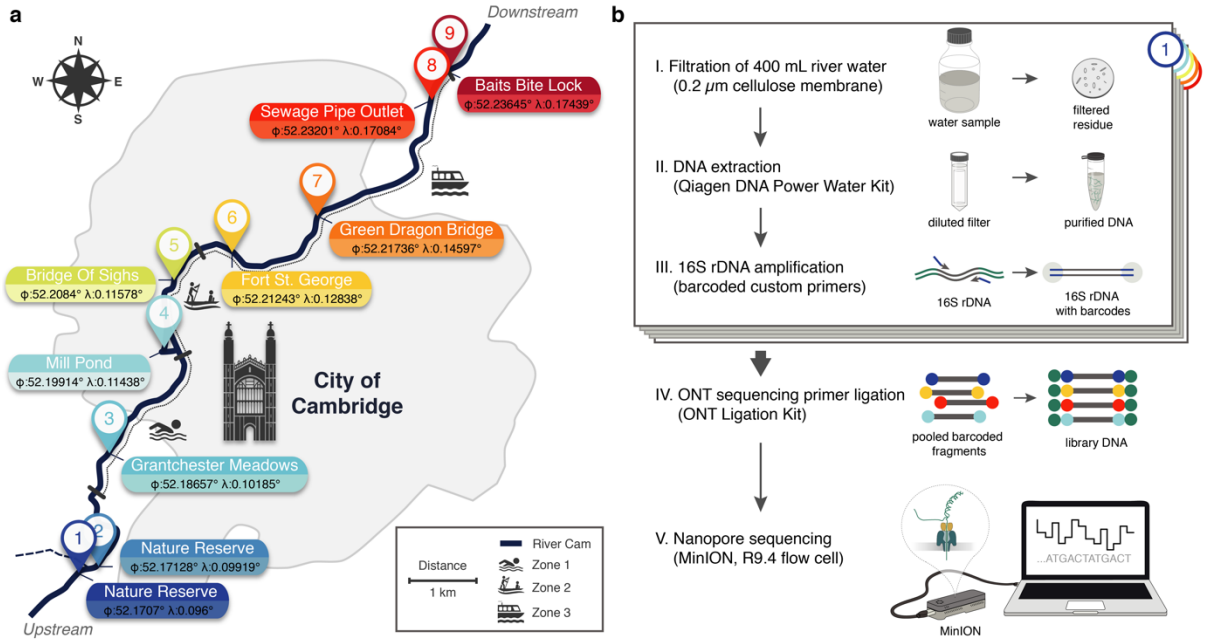


Figure 1: Freshwater microbiome study design and experimental workflow. (a) Schematic map of Cambridge (UK) illustrating sampling locations (colour-coded) along the Cam River. Geographic coordinates of latitude and longitude are expressed as decimal fractions according to the global positioning system.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

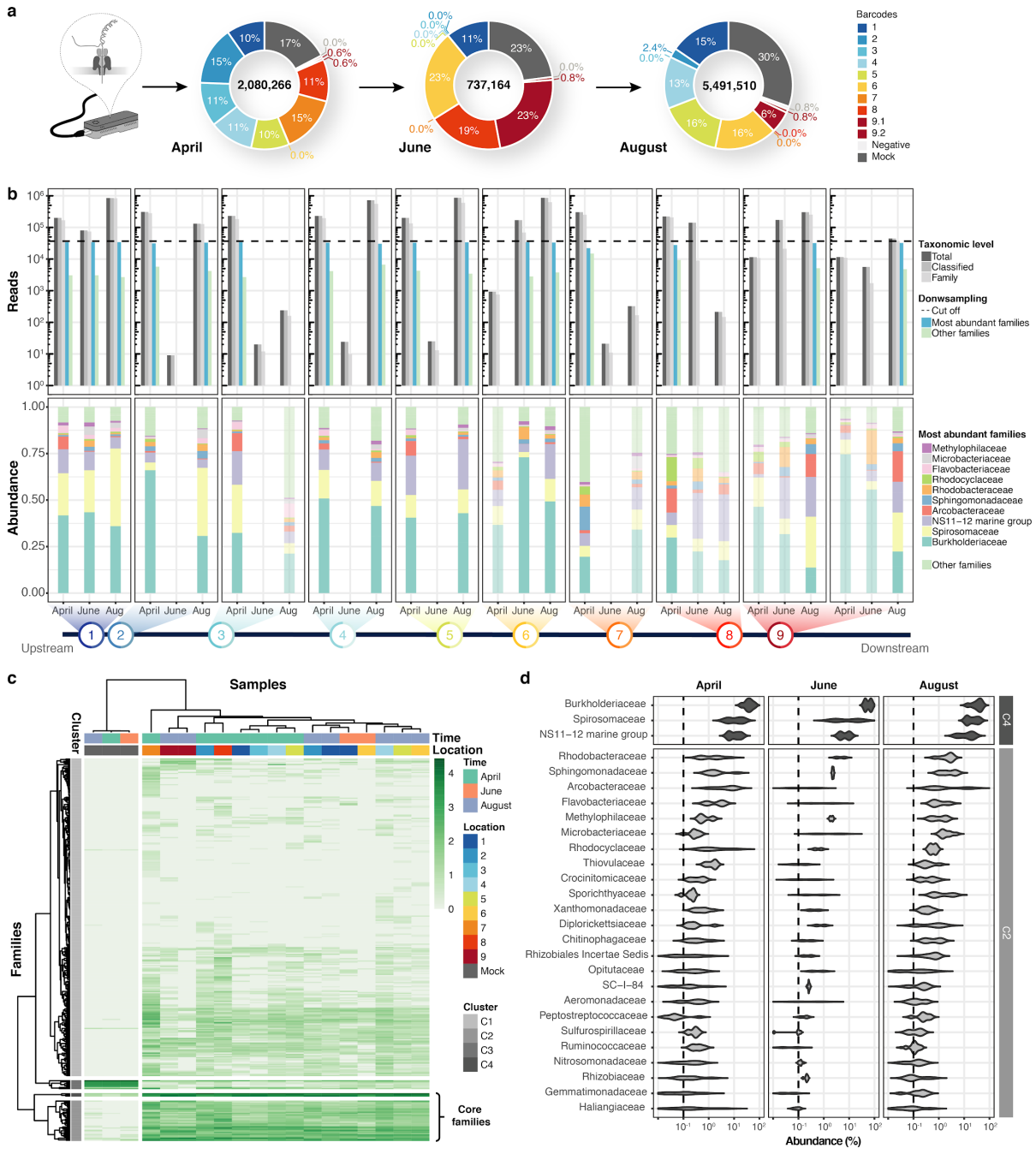


Figure 2: Bacterial diversity of the River Cam. (a) Nanopore sequencing output summary. Values in the centre of the pie charts depict total numbers of classified nanopore sequences per time point. Percentages illustrate representational fractions of locations and control barcodes (negative control and 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.

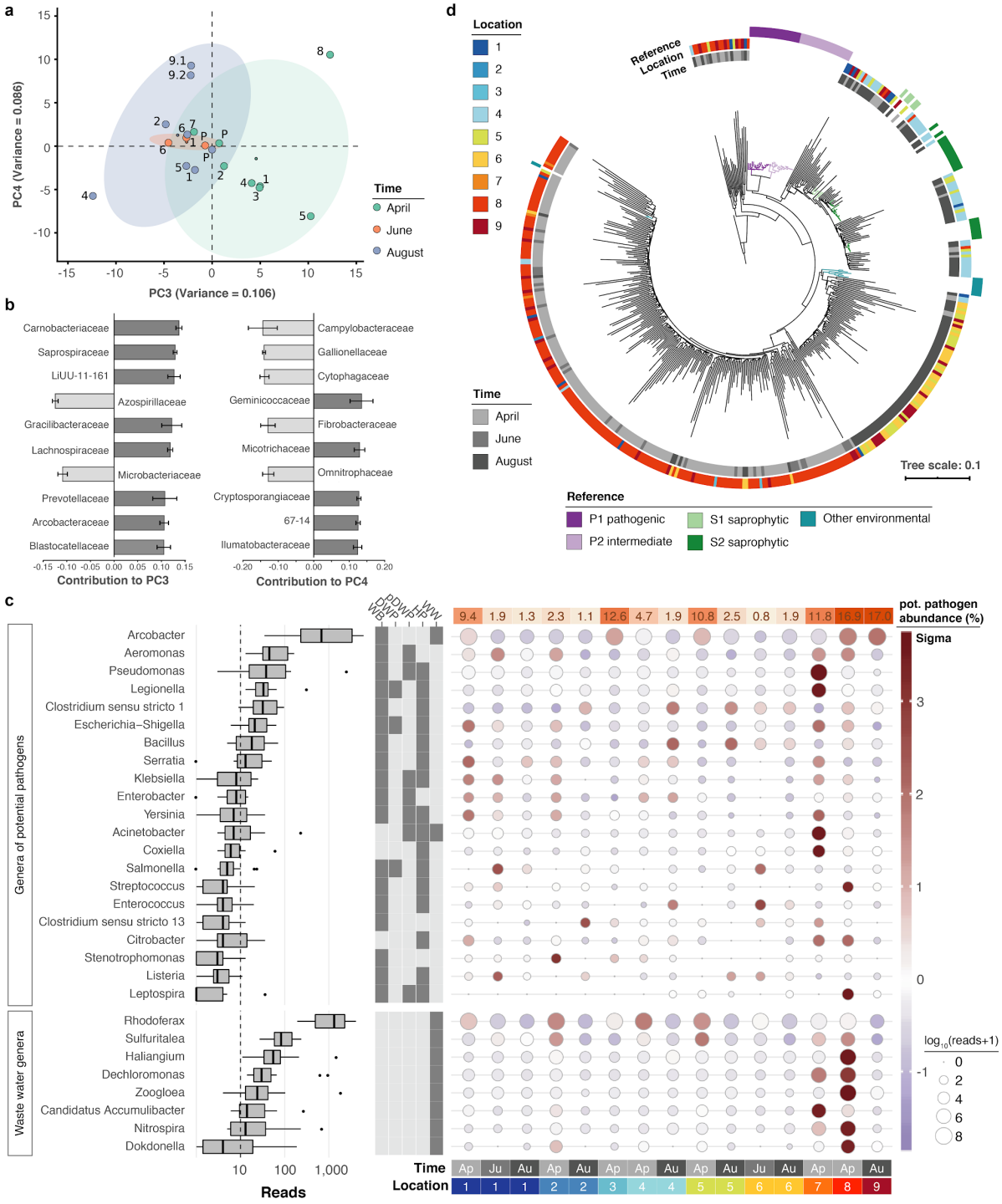
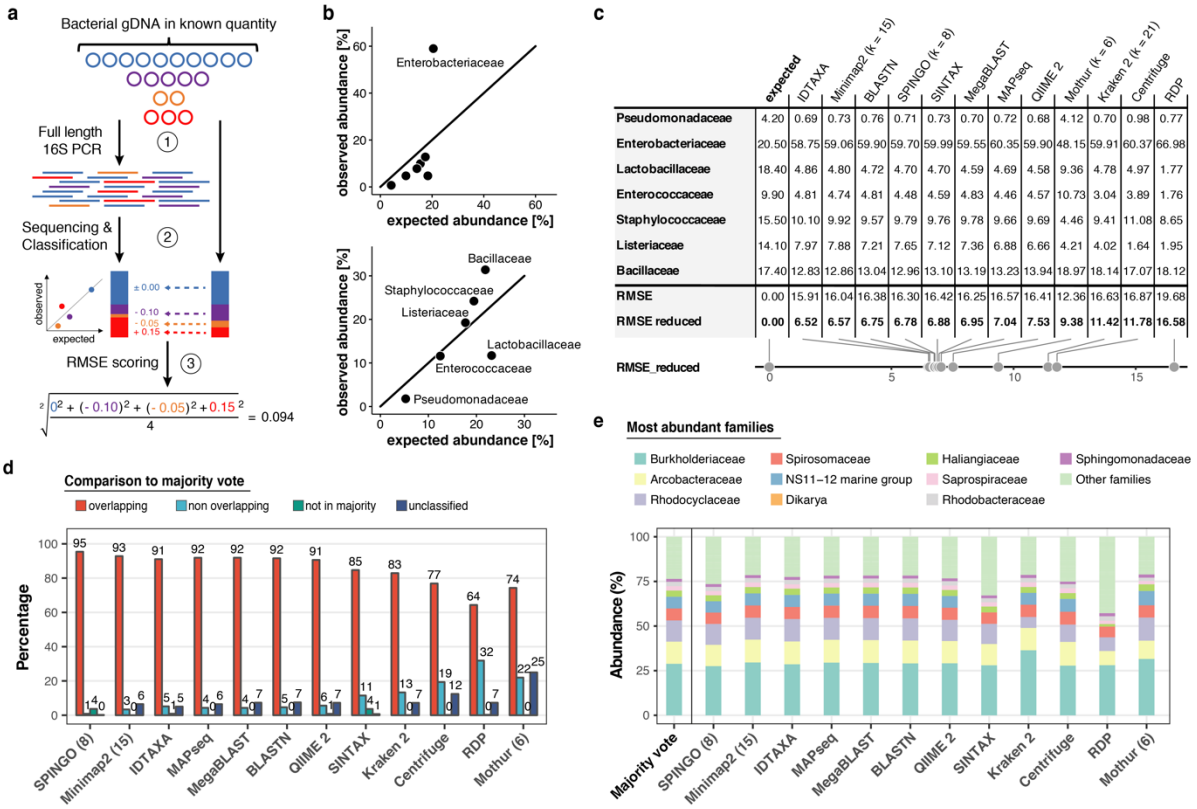
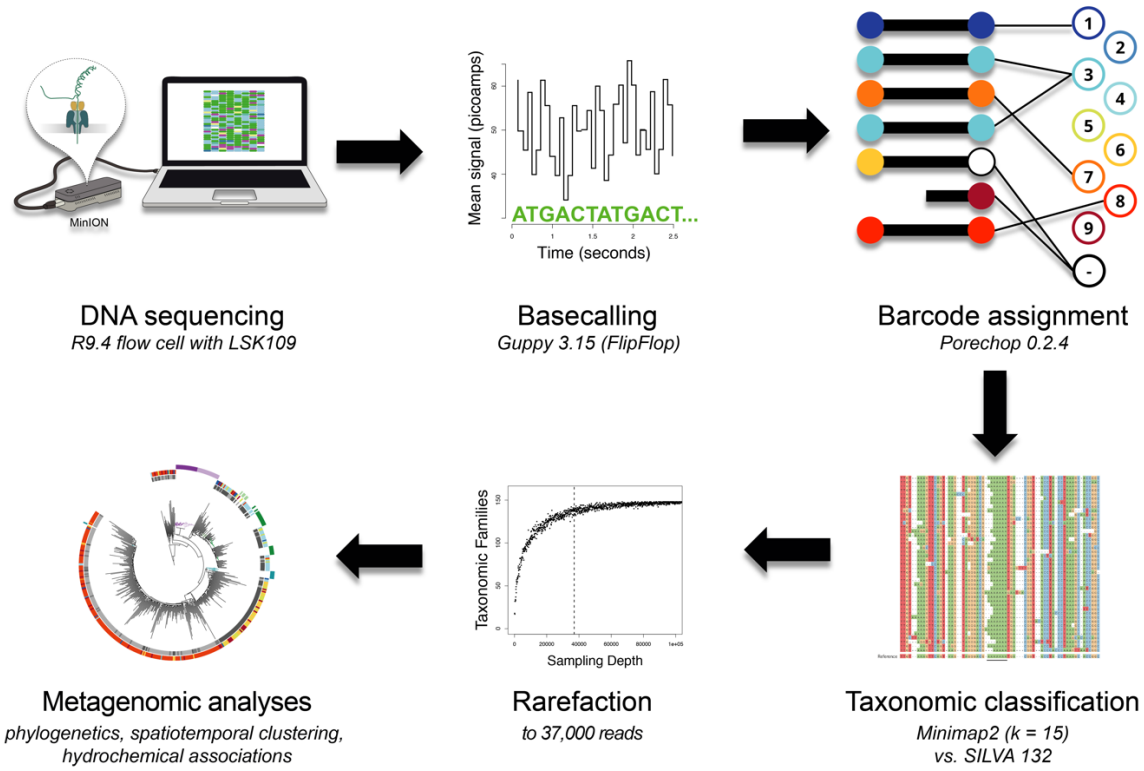


Figure 3: Rare taxa and potential pathogens of the River Cam. (a) PCA of bacterial composition across locations, indicating community dissimilarities along the main time (PC3) and spatial (PC4) axes of variation; dots coloured according to time points. (b) Contribution of individual bacterial families to the PCs in (a). Error bars represent the standard deviation of these families across four independent rarefactions. (c) Abundance and distribution of potentially pathogenic bacteria and wastewater treatment related bacteria, at genus level resolution. The boxplots on the left show the abundance distribution across locations per bacterial genus. Error bars represent

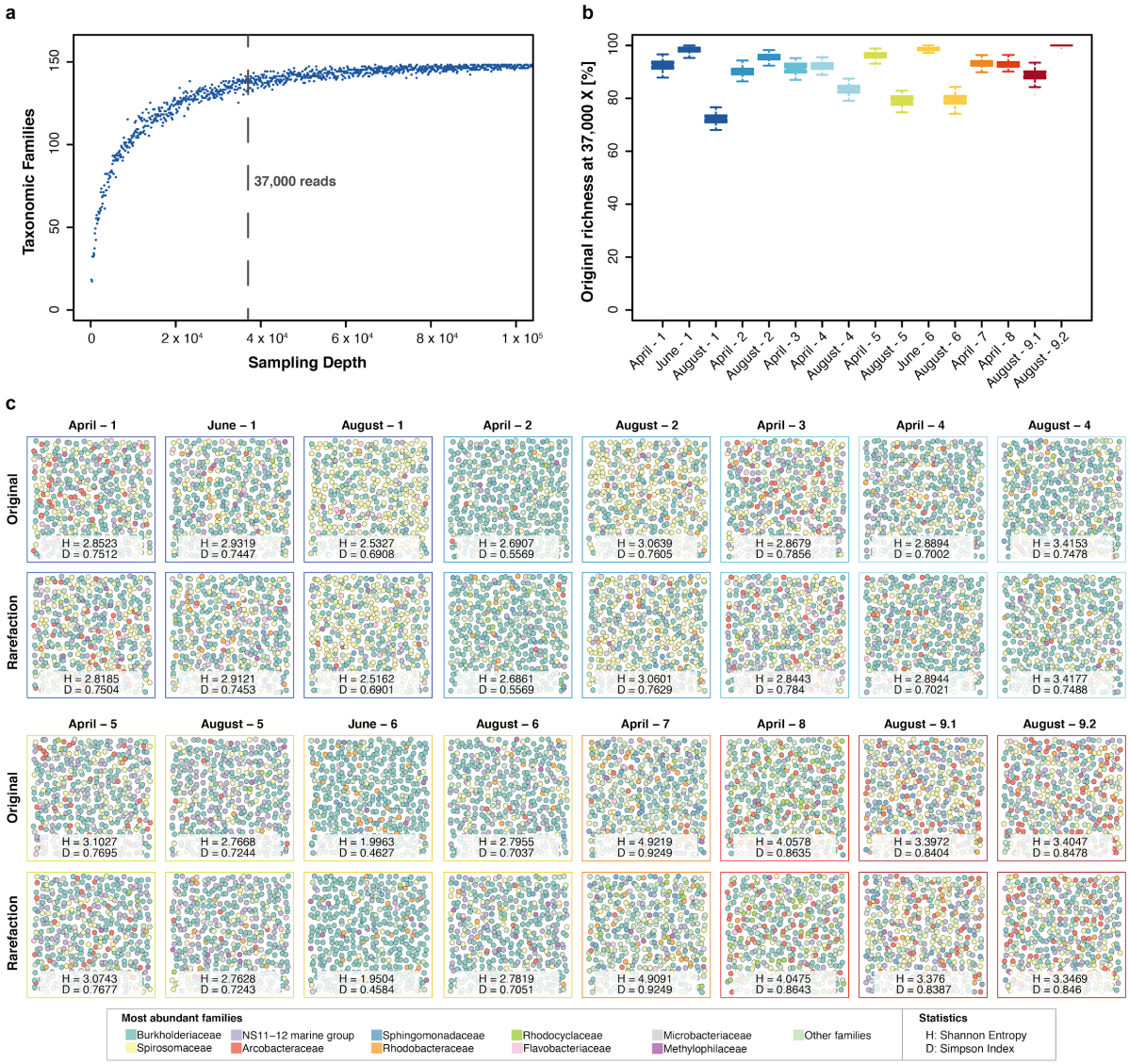
Q1 – 1.5*IQR (lower), and Q3 + 1.5*IQR (upper), respectively; Q1: first quartile, Q3: third quartile, IQR: interquartile range. The centre colour-scale table depicts the categorisation of subsets of genera as waterborne bacterial pathogens (WB), drinking water pathogens (DWP), potential drinking water 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. (d) Phylogenetic tree illustrating the multiple sequence alignment of all nanopore reads classified as *Leptospira*, together with known *Leptospira* reference sequences ranging from pathogenic to saprophytic species[29] (Supplementary Table 4a).



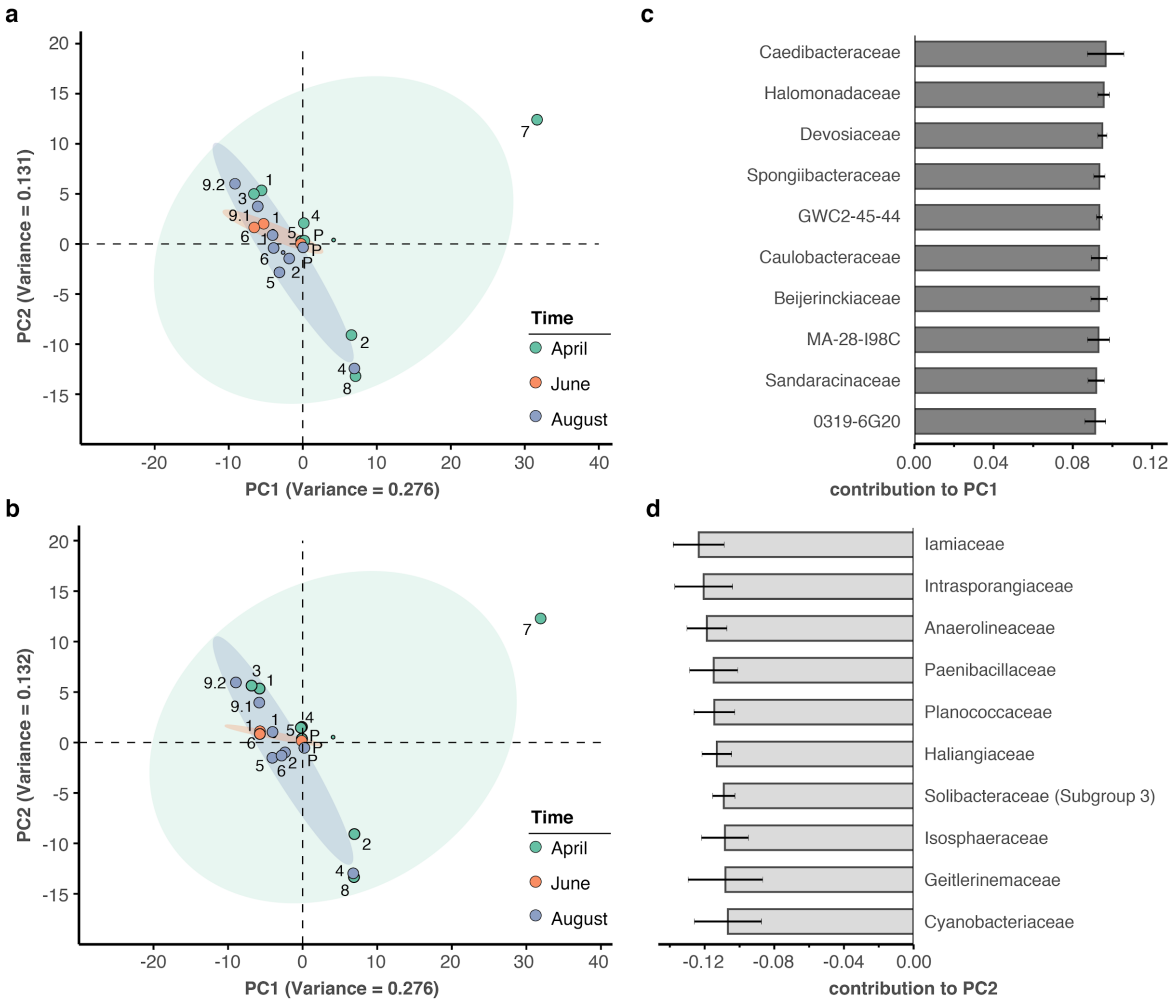
Extended Data Figure 1: Benchmarking of classification tools with nanopore full-length 16S reads. (a) Schematic of mock community quantification performance testing. (b) Observed vs. expected read fraction of bacterial families present in 10,000 nanopore reads randomly drawn from mock community sequencing data. Example representation of Minimap2 (kmer length 15) quantifications with (upper) and without (lower) *Enterobacteriaceae* (Methods). (c) Mock community classification output summary for twelve classification tools tested against the 10,000 nanopore 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 (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 twelve classifiers based on the 10,000 aquatic reads used in (d).



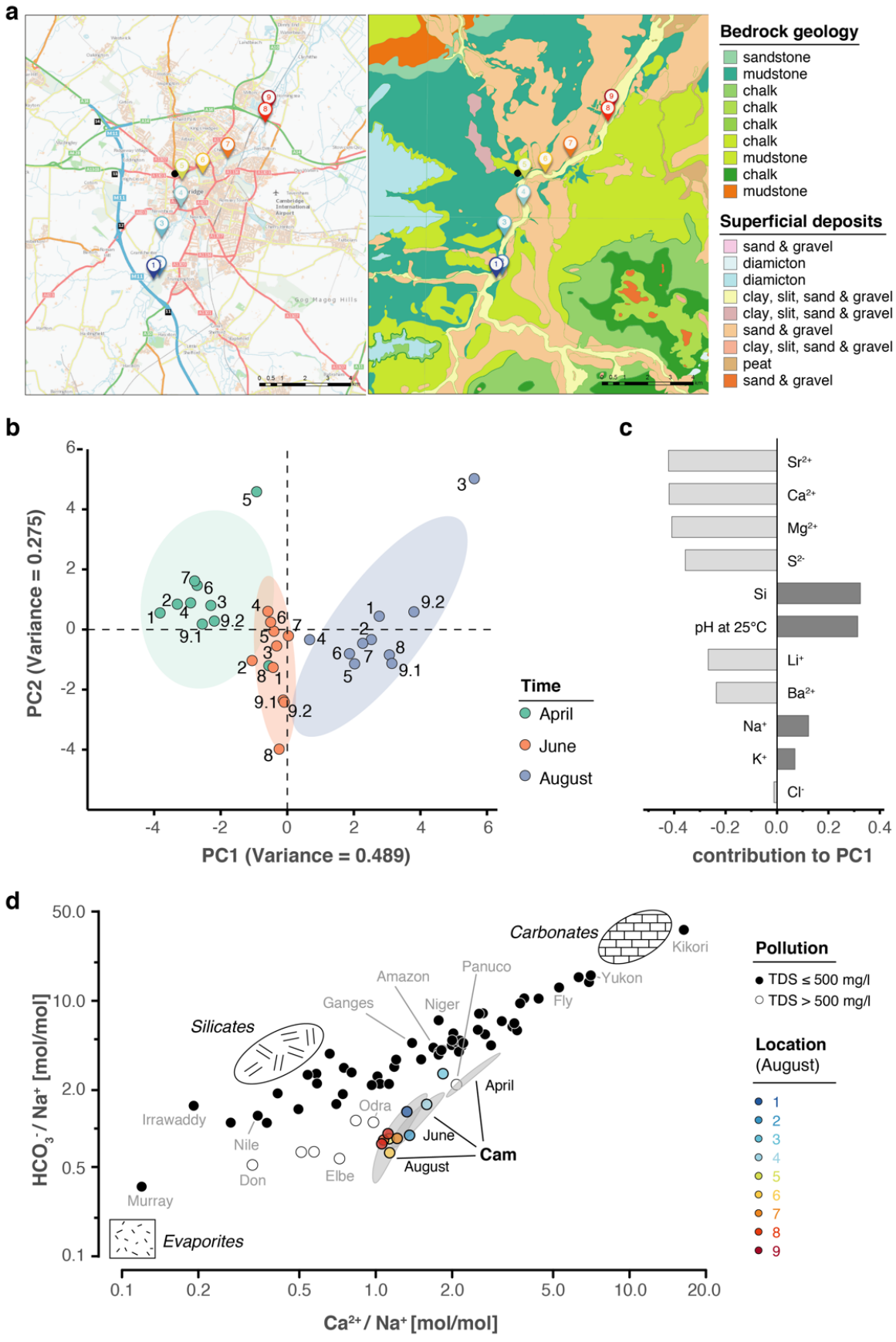
Extended Data Figure 2: Bioinformatics consensus workflow. Essential data processing steps, from nanopore sequencing to spatiotemporal bacterial composition analysis (Methods). After full-length 16S rDNA sequencing with the MinION (R9.4 flow cell), local basecalling of the raw fast5 files was performed using Guppy[69]. Output fastq files were filtered for length and quality (Methods), and reads assigned to their location barcode using Porechop. We then used Minimap2[15] (k = 15) and the SILVA v.132 database[16] 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.



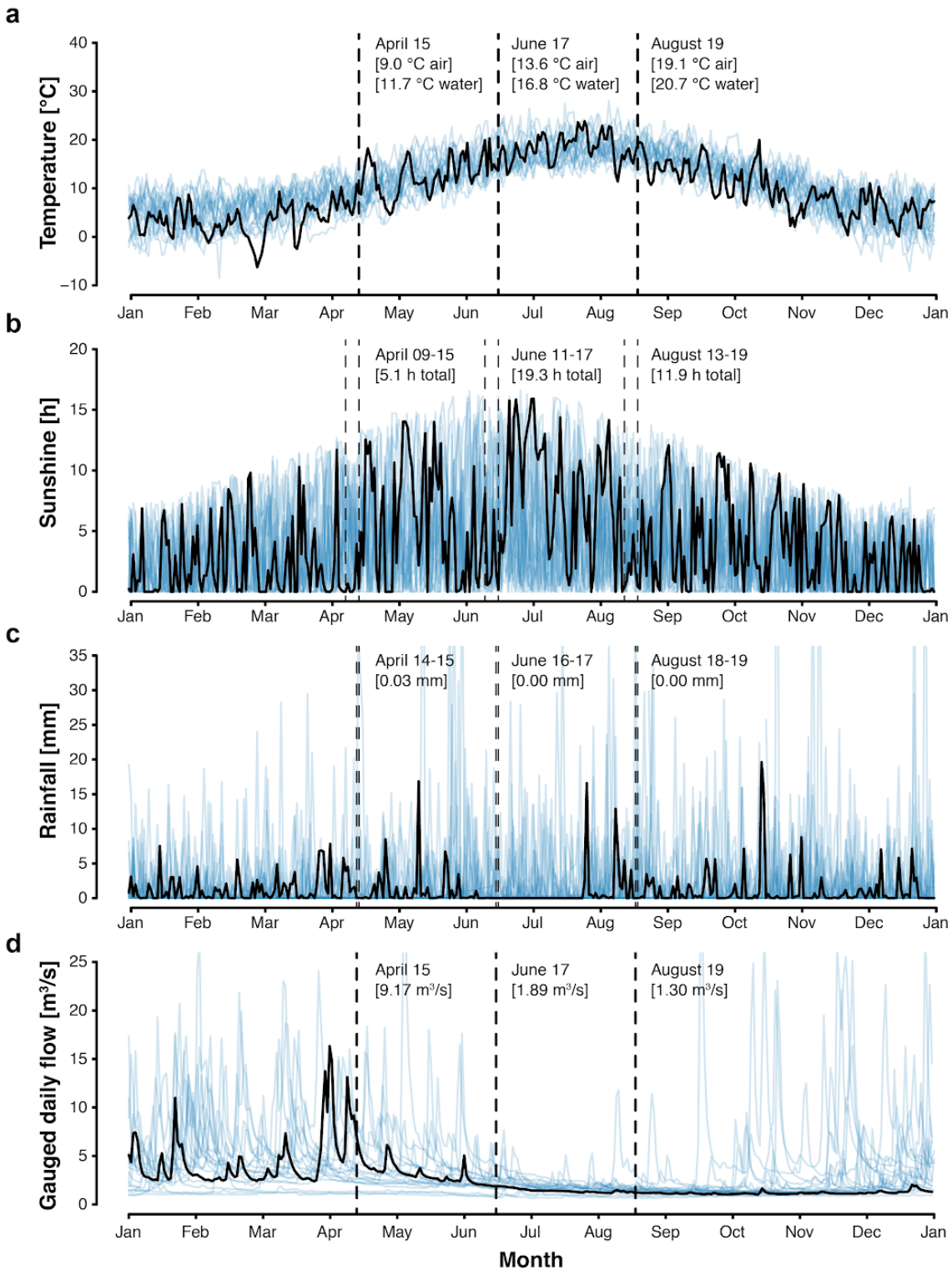
Extended Data Figure 3: Impact of rarefaction on diversity estimation. (a) Example rarefaction curve for bacterial family classifications of the 'April-1' sample. The chosen cut-off preserves most (~90 %) of the original family taxon richness (vertical line). (b) Difference between original and rarefied 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.



Extended Data Figure 4: 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.

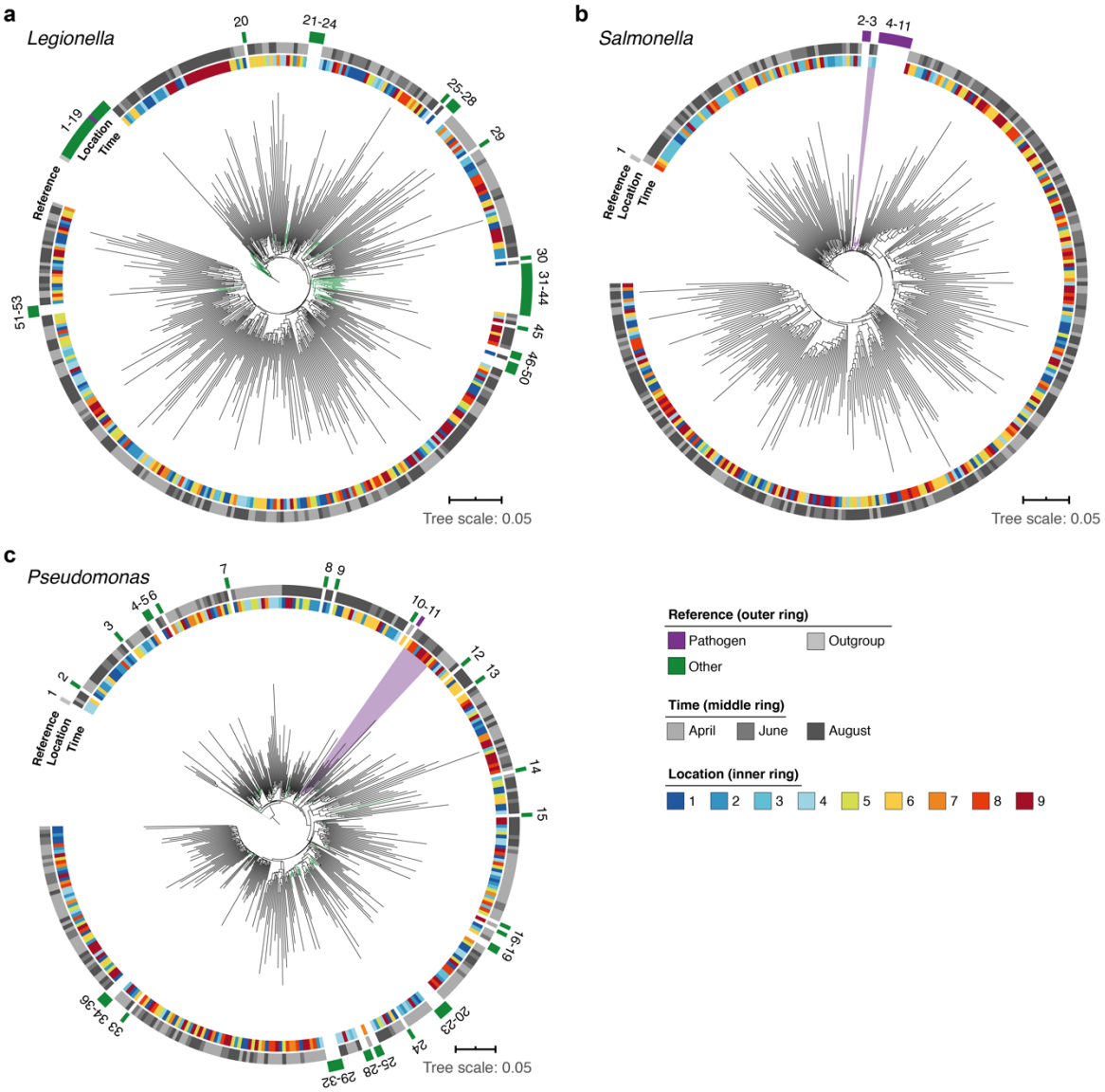


Extended Data Figure 5: Geological and hydrochemical profile of the River Cam and its basin. (a) Outline of the Cam River catchment surrounding Cambridge (UK), and its corresponding lithology. Overlay of bedrock geology and superficial deposits (British Geological Survey data: DiGMapGB-50, 1:50,000 scale) 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[21]; 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).

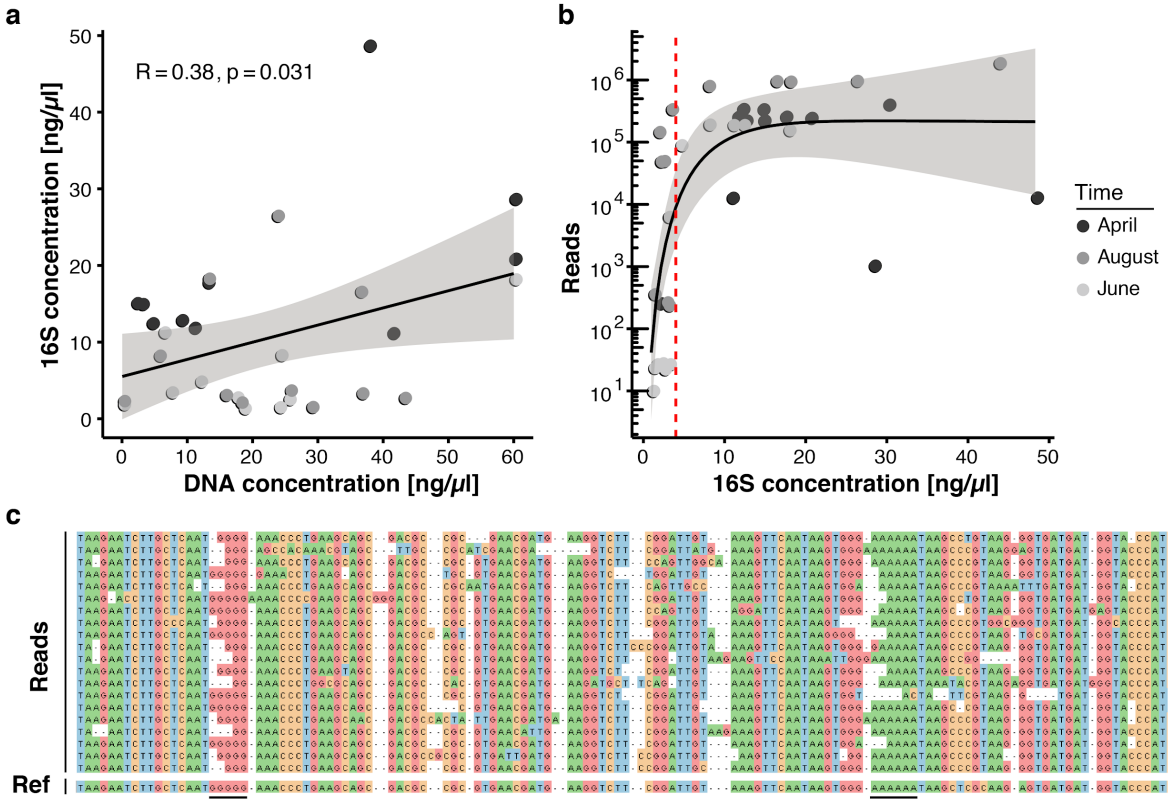


Extended Data Figure 6: Cambridge weather and Cam River flow rate. (a) Daily air temperature [°C], (b) daily sunshine [hours], and (c) daily rainfall [mm] of Cambridge in 2018 (black trend line) vs. 1998-2017 (blue background trend lines). (d) Cam River gauged daily flow [m^3s^{-1}] in 2018 (black trend 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).



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



Extended Data Figure 8: Key challenges of environmental 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 (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 %.

MATERIALS 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[51]. 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 (Extended Data Figures 5 and 6, Supplementary Tables 1b and 1c).

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 (Extended Data Figure 5). 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[52] (Extended Data Figure 5). 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[21], exhibiting enrichment in Na^+ (Extended Data Figure 5).

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 (Extended Data Figure 5b). We highlighted the ten most important features (i.e., features with the largest weights) and their contributions to PC1 (Extended Data Figure 5c).

We integrated sensor data sets on mean daily air temperature, sunshine hours and total rainfall from a public, Cambridge-based weather station (Extended Data Figure 6a-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 (Extended Data Figure 6d)

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) (Extended Data Figure 1b-c), which was previously assessed using nanopore shotgun metagenomics[42]. 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 the following conditions:

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

0.2 µL Taq (Qiagen)

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[53] 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 gathered read statistics such as 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[54,55]	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[56]	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[19]	Implemented in R	load("SILVA_SSU_r132_March2018.RData")

	DECIPHER v.2.10.2	IdTaxa(Cam16S.fa, trainingSet, strand = "both", threshold = 0)
Kraken 2[57]	v.2.0.7	# build database kraken2 --db kraken2_16s_database --output Cam16S,out --report Cam16S.kreport --gzip-compressed --threads 1 Cam16S.fa
MAPseq[18]	v.1.2.3	mapseq Cam16S.fa silva_ref.fa > Cam16S.mseq
MegaBLAST[55,58]	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[15]	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[59]	v.1.43.0	align.seqs(candidate=Cam16S.fa, template=mothur.silva.nr_v132.align, processors=1, ksize=6, align=needleman)
QIIME 2[60]	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[61]	Implemented in R *DADA2* v.1.12.1[62]	assignTaxonomy(seqs = Cam16S.fa, refFasta = silva_nr_v132_train_set.fa.gz", tryRC = T, outputBootstraps=T,minBoot=0)
SINTAX[63]	Implemented in VSEARCH v.2.13.3[64]	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[17]	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[16] (Extended Data Figure 1).

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 (Extended Data Figure 1b-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 (Extended Data Figure 1d-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 2b, Extended Data Figure 3, 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[65] 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, 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[66] (v.1.3.1) stats function.

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 (Extended Data Figure 3; 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[22] and data on known waterborne pathogens[23] (Supplementary Table 3a). Additionally, we integrated known genera from a large wastewater reference collection[24] (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 4a). 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[67] (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[29]. Additional published river water *Leptospira* that did not fall within these clades were included as 'Other Environmental'[68]. Similarly, we constructed phylogenies for the *Legionella*, *Salmonella* and *Pseudomonas* genus, using established full-length 16S reference species sequences from NCBI (Supplementary Table 4b-d).

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.

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, Extended Data Figures 1, 3, 4, 5, 7, and 8. 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: Extended Data Figure 6.

The are no restrictions on data availability.

CODE AVAILABILITY

Our Github repository (https://github.com/d-j-k/puntseq/) provides 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 downstream analyses.

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 *Leptospira* (a), *Legionella* (b), *Salmonella* (c) and *Pseudomonas* (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

1 Bartram, J., Lewis, K., Lenton, R. & Wright, A. Focusing on improved water and sanitation for health. *The Lancet* **365**, 810-812 (2005).

2 Schewe, J. *et al.* Multimodel assessment of water scarcity under climate change. *Proc Natl Acad Sci U S A* **111**, 3245-3250 (2014).

3 Haddeland, I. *et al.* Global water resources affected by human interventions and climate change. *Proc Natl Acad Sci U S A* **111**, 3251-3256 (2014).

4 Gardy, J. L. & Loman, N. J. Towards a genomics-informed, real-time, global pathogen surveillance system. *Nat Rev Genet* **19**, 9-20 (2018).

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

6 Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol* **17**, 239 (2016).

7 Payne, A., Holmes, N., Rakyan, V. & Loose, M. BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. *Bioinformatics* **35**, 2193-2198 (2019).

8 Quick, J. *et al.* Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella. *Genome Biol* **16**, 114 (2015).

9 Quick, J. *et al.* Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530**, 228-232 (2016).

10 Faria, N. R. *et al.* Establishment and cryptic transmission of Zika virus in Brazil and the Americas. *Nature* **546**, 406-410 (2017).

11 Faria, N. R. *et al.* Genomic and epidemiological monitoring of yellow fever virus transmission potential. *Science* **361**, 894 (2018).

12 Kafetzopoulou, L. E. *et al.* Metagenomic sequencing at the epicenter of the Nigeria 2018 Lassa fever outbreak. *Science* **363**, 74 (2019).

13 Chan, J. F.-W. *et al.* A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *The Lancet* (2020).

14 Frank, J. A. *et al.* Critical evaluation of two primers commonly used for amplification of bacterial 16S rRNA genes. *Appl Environ Microbiol* **74**, 2461-2470 (2008).

15 Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100 (2018).

16 Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41**, D590-596 (2013).

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

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

19 Murali, A., Bhargava, A. & Wright, E. S. IDTAXA: a novel approach for accurate taxonomic classification of microbiome sequences. *Microbiome* **6**, 140 (2018).

20 Lawson, P. A. & Caldwell, M. E. in *The Prokaryotes: Firmicutes and Tenericutes* (eds Eugene Rosenberg *et al.*) 19-65 (Springer Berlin Heidelberg, 2014).

21 Gaillardet, J., Dupré, B., Louvat, P. & Allègre, C. J. Global silicate weathering and CO2 consumption rates deduced from the chemistry of large rivers. *Chemical Geology* **159**, 3-30 (1999).

22 Wattam, A. R. *et al.* Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. *Nucleic Acids Res* **45**, D535-D542 (2017).

23 Jin, D. *et al.* Bacterial communities and potential waterborne pathogens within the typical urban surface waters. *Sci Rep* **8**, 13368 (2018).

24 Wu, L. *et al.* Global diversity and biogeography of bacterial communities in wastewater treatment plants. *Nat Microbiol* **4**, 1183-1195 (2019).

25 Kayman, T. *et al.* Emerging pathogen Arcobacter spp. in acute gastroenteritis: molecular identification, antibiotic susceptibilities and genotyping of the isolated arcobacters. *J Med Microbiol* **61**, 1439-1444 (2012).

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

27 Numberger, D. *et al.* Characterization of bacterial communities in wastewater with enhanced taxonomic resolution by full-length 16S rRNA sequencing. *Sci Rep* **9**, 9673 (2019).

28 Wynwood, S. J. *et al.* Leptospirosis from water sources. *Pathogens and Global Health* **108**, 334-338 (2014).

29 Vincent, A. T. *et al.* Revisiting the taxonomy and evolution of pathogenicity of the genus Leptospira through the prism of genomics. *PLoS Negl Trop Dis* **13**, e0007270 (2019).

30 Rang, F. J., Kloosterman, W. P. & de Ridder, J. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biol* **19**, 90 (2018).

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

32 Karst, S. M. *et al.* Enabling high-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. *bioRxiv*, 645903 (2020).

33 Ramirez-Castillo, F. Y. *et al.* Waterborne pathogens: detection methods and challenges. *Pathogens* **4**, 307-334 (2015).

34 Tan, B. *et al.* Next-generation sequencing (NGS) for assessment of microbial water quality: current progress, challenges, and future opportunities. *Front Microbiol* **6**, 1027 (2015).

35 Rowe, W. *et al.* Comparative metagenomics reveals a diverse range of antimicrobial resistance genes in effluents entering a river catchment. *Water Sci Technol* **73**, 1541-1549 (2016).

36 Rowe, W. *et al.* Overexpression of antibiotic resistance genes in hospital effluents over time. *J Antimicrob Chemother* **72**, 1617-1623 (2017).

37 Darby, B. J., Todd, T. C. & Herman, M. A. 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**, 5456-5471 (2013).

38 Benitez-Paez, A., Portune, K. J. & Sanz, Y. Species-level resolution of 16S rRNA gene amplicons sequenced through the MinION portable nanopore sequencer. *Gigascience* **5**, 4 (2016).

39 Kerkhof, L. J., Dillon, K. P., Haggblom, M. M. & McGuinness, L. R. Profiling bacterial communities by MinION sequencing of ribosomal operons. *Microbiome* **5**, 116 (2017).

40 Cusco, A., Catozzi, C., Vines, J., Sanchez, A. & Francino, O. 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 (2018).

41 Krehenwinkel, H. *et al.* Nanopore sequencing of long ribosomal DNA amplicons enables portable and simple biodiversity assessments with high phylogenetic resolution across broad taxonomic scale. *Gigascience* **8** (2019).

42 Nicholls, S. M., Quick, J. C., Tang, S. & Loman, N. J. Ultra-deep, long-read nanopore sequencing of mock microbial community standards. *Gigascience* **8** (2019).

43 Gowers, G. F. *et al.* Entirely Off-Grid and Solar-Powered DNA Sequencing of Microbial Communities during an Ice Cap Traverse Expedition. *Genes (Basel)* **10** (2019).

44 Stewart, R. D. *et al.* Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. *Nat Biotechnol* **37**, 953-961 (2019).

45 Leggett, R. M. *et al.* Rapid MinION profiling of preterm microbiota and antimicrobial-resistant pathogens. *Nat Microbiol* (2019).

46 Hamner, S. *et al.* Metagenomic Profiling of Microbial Pathogens in the Little Bighorn River, Montana. *Int J Environ Res Public Health* **16** (2019).

47 Acharya, K. *et al.* A comparative assessment of conventional and molecular methods, including MinION nanopore sequencing, for surveying water quality. *Sci Rep* **9**, 15726 (2019).

48 Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science* **348** (2015).

49 Bahram, M. *et al.* Structure and function of the global topsoil microbiome. *Nature* **560**, 233-237 (2018).

50 Boykin, L. M. *et al.* Tree Lab: Portable genomics for Early Detection of Plant Viruses and Pests in Sub-Saharan Africa. *Genes (Basel)* **10** (2019).

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

52 Rose, S. The effects of urbanization on the hydrochemistry of base flow within the Chattahoochee River Basin (Georgia, USA). *Journal of Hydrology* **341**, 42-54 (2007).

53 Köster, J. & Rahmann, S. Snakemake--a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520-2522 (2012).

54 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403-410 (1990).

55 Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

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

57 Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol* **20**, 257 (2019).

58 Morgulis, A. *et al.* Database indexing for production MegaBLAST searches. *Bioinformatics* **24**, 1757-1764 (2008).

59 Schloss, P. D. *et al.* Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75**, 7537-7541 (2009).

60 Bolyen, E. *et al.* Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol* **37**, 852-857 (2019).

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

62 Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods* **13**, 581-583 (2016).

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

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

65 Salter, S. J. *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology* **12**, 87 (2014).

66 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

67 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-1797 (2004).

68 Ganoza, C. A. *et al.* Determining risk for severe leptospirosis by molecular analysis of environmental surface waters for pathogenic Leptospira. *PLoS Med* **3**, e308 (2006).

69 Wick, R. R., Judd, L. M. & Holt, K. E. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol* **20**, 129 (2019).