Figures and figure supplements

Freshwater monitoring by nanopore sequencing

Lara Urban *et al*

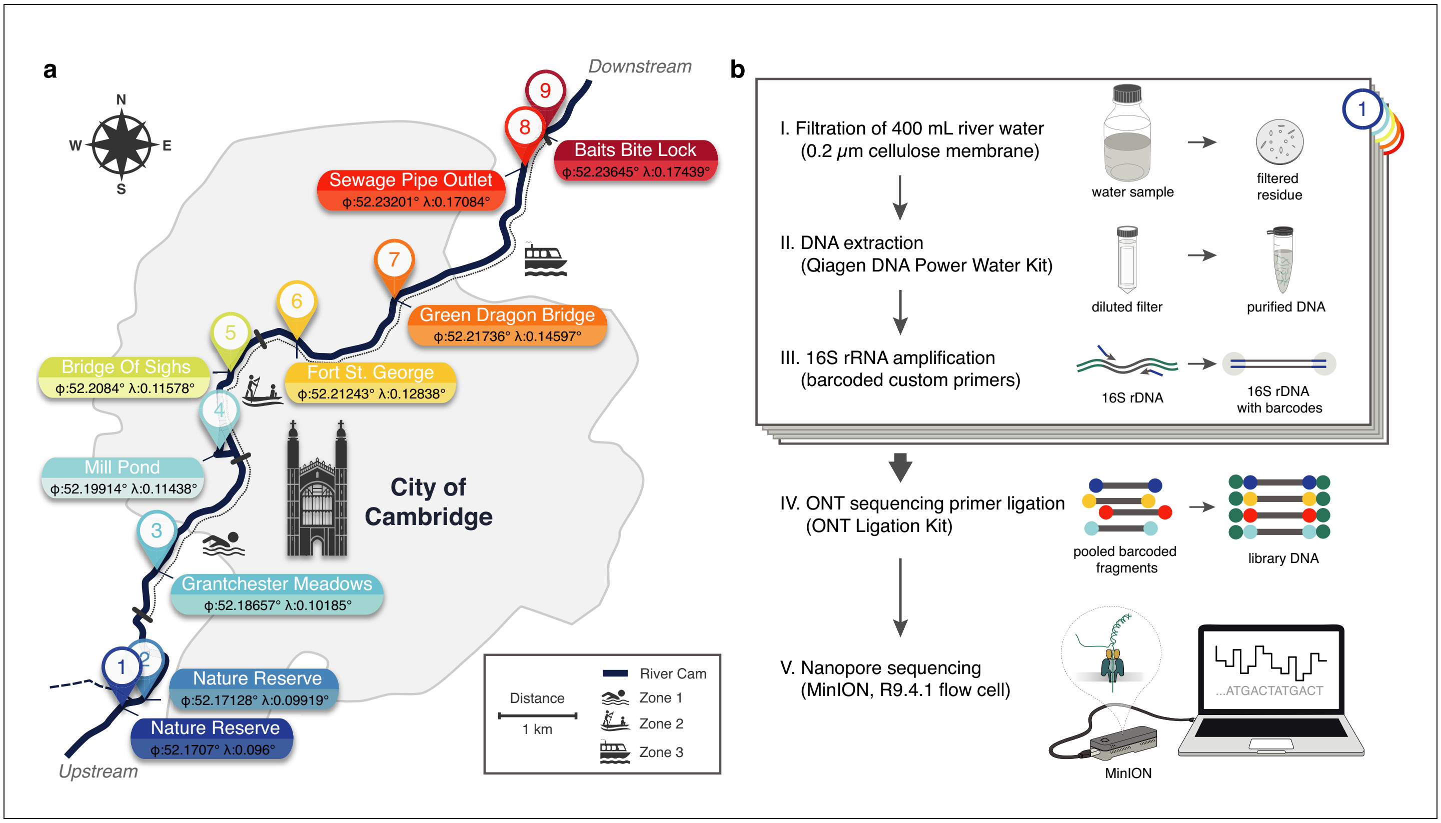


Figure 1. Freshwater microbiome study design and experimental setup. (**a**) Schematic map of Cambridge (UK), illustrating sampling locations (colour-coded) along the River Cam. Geographic coordinates of latitude and longitude are expressed as decimal fractions according to the global positioning system.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

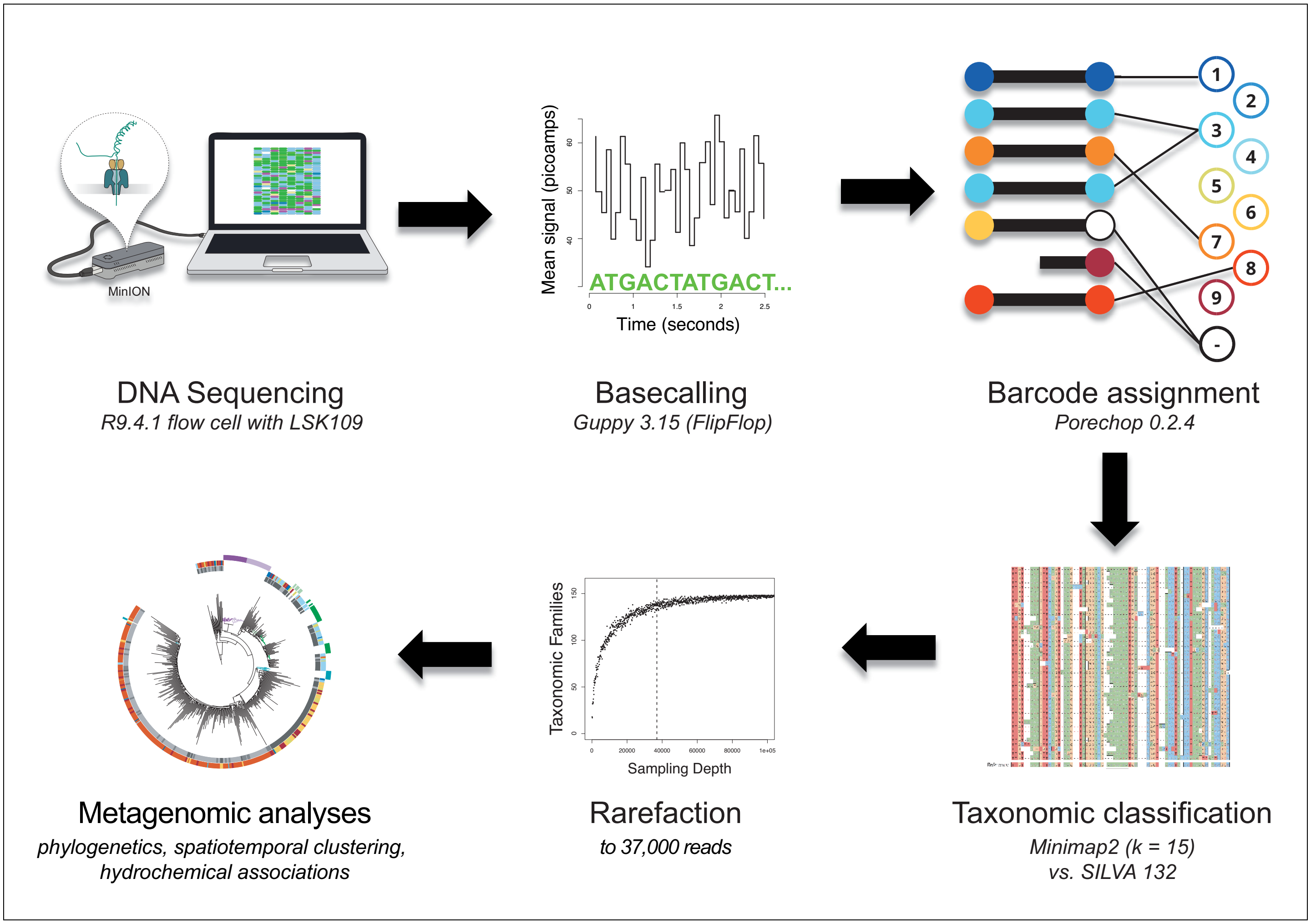
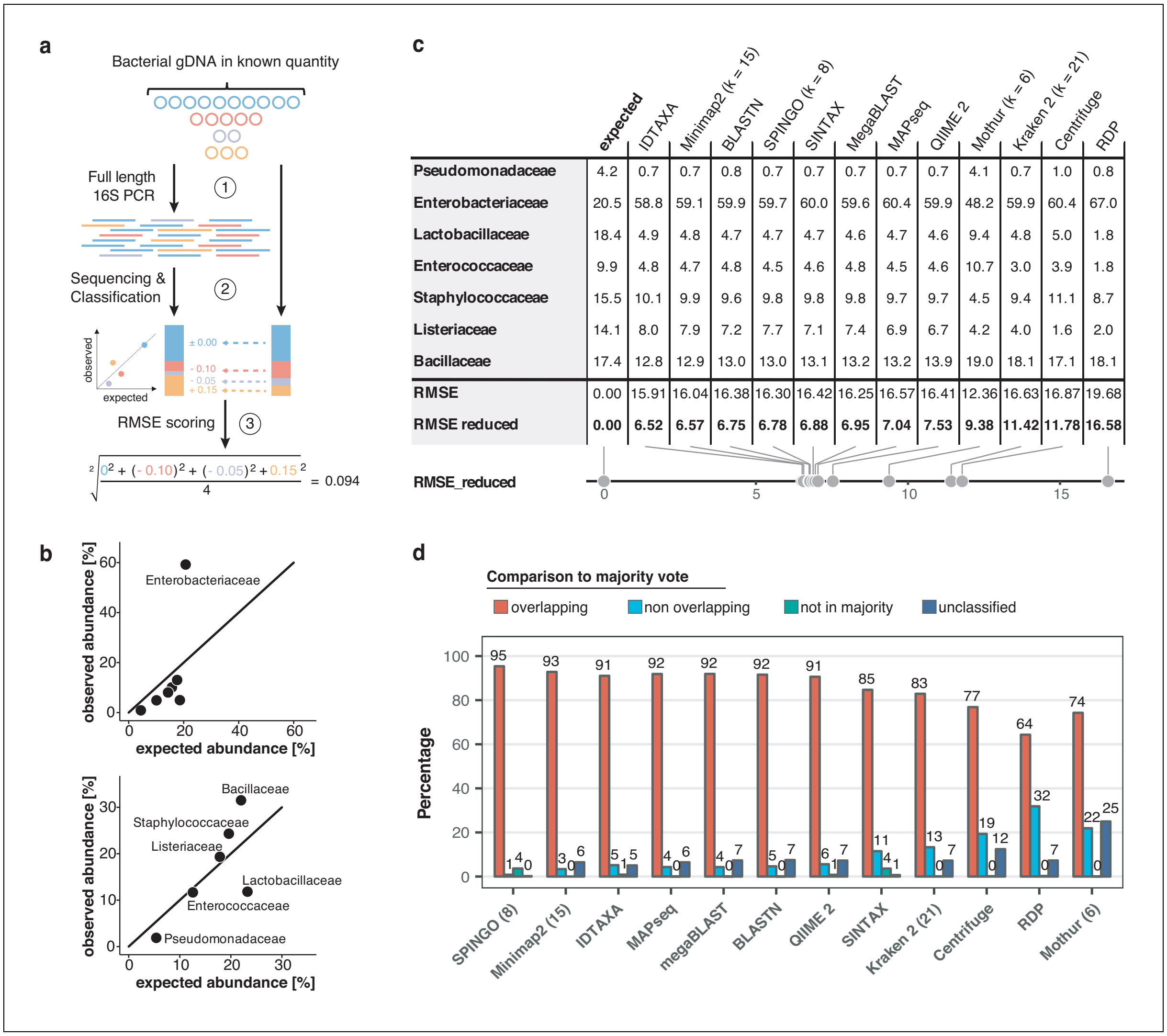


Figure 1—figure supplement 1. Bioinformatics consensus workflow. Essential data processing steps, from nanopore sequencing to spatiotemporal bacterial composition analysis (Materials and methods). After full-length 16S rDNA sequencing with the MinION (R9.4.1 flow cell), local basecalling of the raw fast5 files was performed using Guppy (***Wick et al., 2019***). Output fastq files were filtered for length and quality (Materials 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 analyses.



$$\sqrt[2]{\frac{0^2 + (-0.10)^2 + (-0.05)^2 + 0.15^2}{4}} = 0.094$$

	expected	IDTAXA	Minimap2 (k = 15)	BLASTN	SPINGO (k = 8)	SINTAX	MegaBLAST	MAPseq	QIIME 2	Mothur (k = 6)	Kraken 2 (k = 21)	Centrifuge	RDP
Pseudomonadaceae	4.2	0.7	0.7	0.8	0.7	0.7	0.7	0.7	0.7	4.1	0.7	1.0	0.8
Enterobacteriaceae	20.5	58.8	59.1	59.9	59.7	60.0	59.6	60.4	59.9	48.2	59.9	60.4	67.0
Lactobacillaceae	18.4	4.9	4.8	4.7	4.7	4.7	4.6	4.7	4.6	9.4	4.8	5.0	1.8
Enterococcaceae	9.9	4.8	4.7	4.8	4.5	4.6	4.8	4.5	4.6	10.7	3.0	3.9	1.8
Staphylococcaceae	15.5	10.1	9.9	9.6	9.8	9.8	9.8	9.7	9.7	4.5	9.4	11.1	8.7
Listeriaceae	14.1	8.0	7.9	7.2	7.7	7.1	7.4	6.9	6.7	4.2	4.0	1.6	2.0
Bacillaceae	17.4	12.8	12.9	13.0	13.0	13.1	13.2	13.2	13.9	19.0	18.1	17.1	18.1
RMSE	0.00	15.91	16.04	16.38	16.30	16.42	16.25	16.57	16.41	12.36	16.63	16.87	19.68
RMSE reduced	**0.00**	**6.52**	**6.57**	**6.75**	**6.78**	**6.88**	**6.95**	**7.04**	**7.53**	**9.38**	**11.42**	**11.78**	**16.58**

	overlapping	non overlapping	not in majority	unclassified
SPINGO (8)	95	1	4	0
Minimap2 (15)	93	3	0	6
IDTAXA	91	5	1	5
MAPseq	92	4	0	6
megaBLAST	92	4	0	7
BLASTN	92	5	0	7
QIIME 2	91	6	1	7
SINTAX	85	11	4	1
Kraken 2 (21)	83	13	0	7
Centrifuge	77	19	0	12
RDP	64	32	0	7
Mothur (6)	74	22	0	25

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* (Materials 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 (Materials 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.

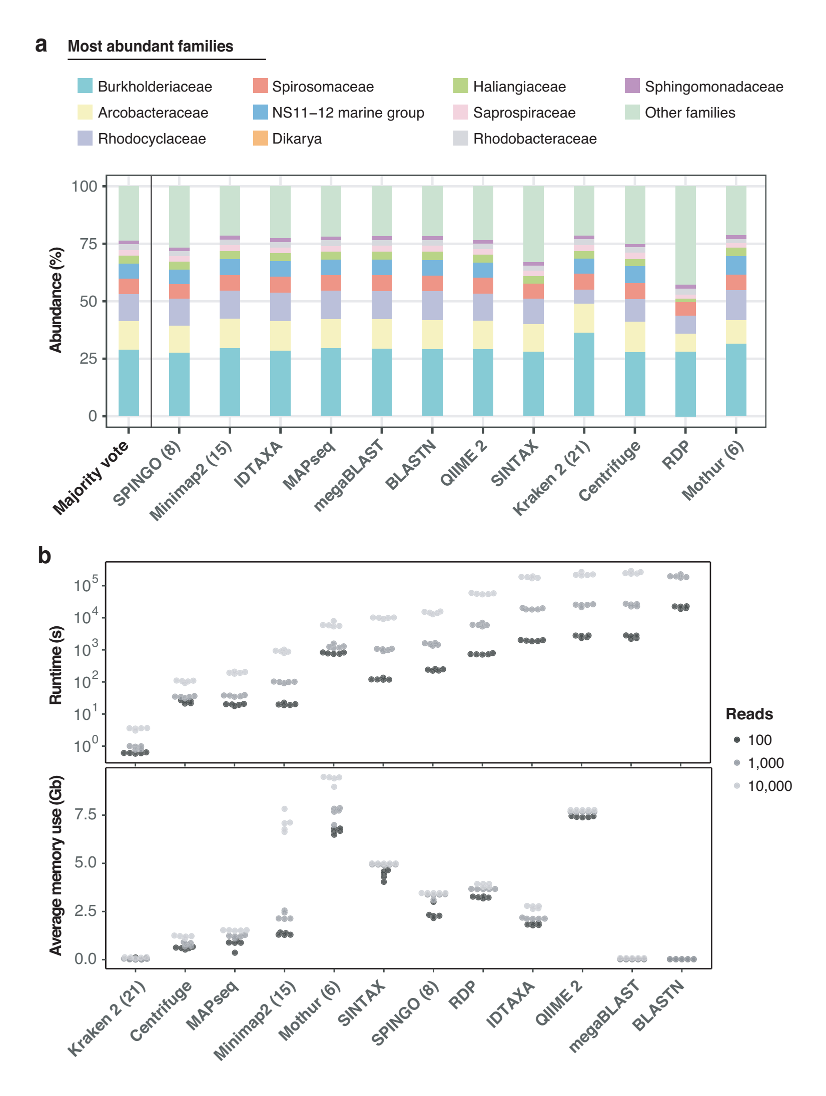


Figure 2—figure supplement 1. Benchmarking of twelve taxonomic classifiers with nanopore full-length 16S sequences. (a) Top 10 represented bacterial taxon families across all methods, based on the 10,000 aquatic reads used in ***Figure 2d***. (b) Comparison of computational performances with respect to (upper) runtime and average memory (lower) usage for the classification of 5 × 100, 5 × 1000, and 5 × 10,000 random read draws of the same sample. BLASTN-based classifications of 10,000 read sets are omitted, as their runtimes exceeded 14 days (> 10^6 s). Numbers in brackets for SPINGO, Minimap2, Kraken 2 and Mothur indicate kmer sizes.

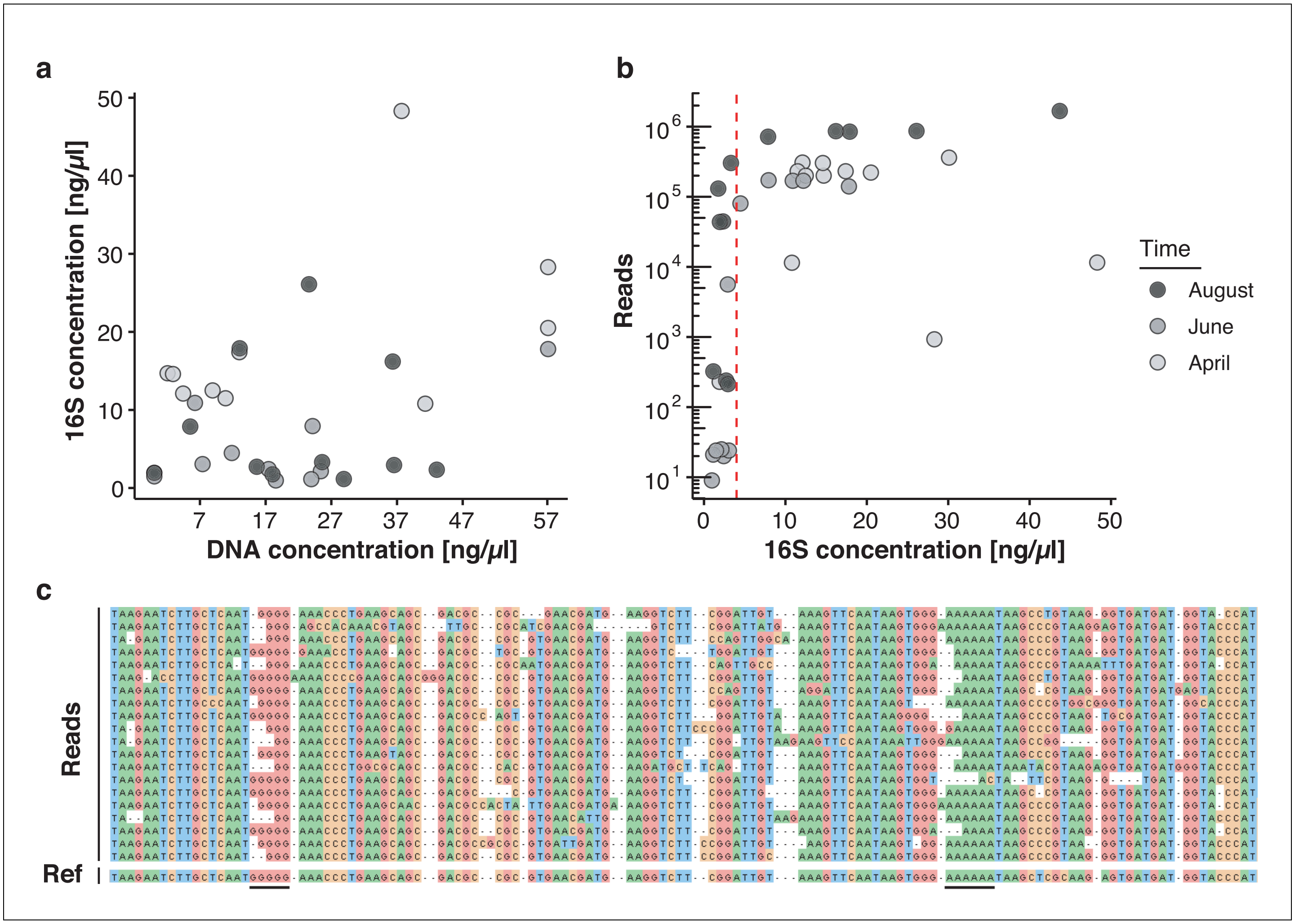


Figure 2—figure supplement 2. Key challenges of freshwater monitoring with nanopore sequencing. (**a–b**) Correlation analysis between DNA extraction yield, 16S amplification yield and raw sequencing output (***Supplementary file 2***). (**a**) DNA concentrations (x-axis) obtained from 30 freshwater samples after extraction with the DNeasy PowerWater Kit (Materials 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. (**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. 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%.

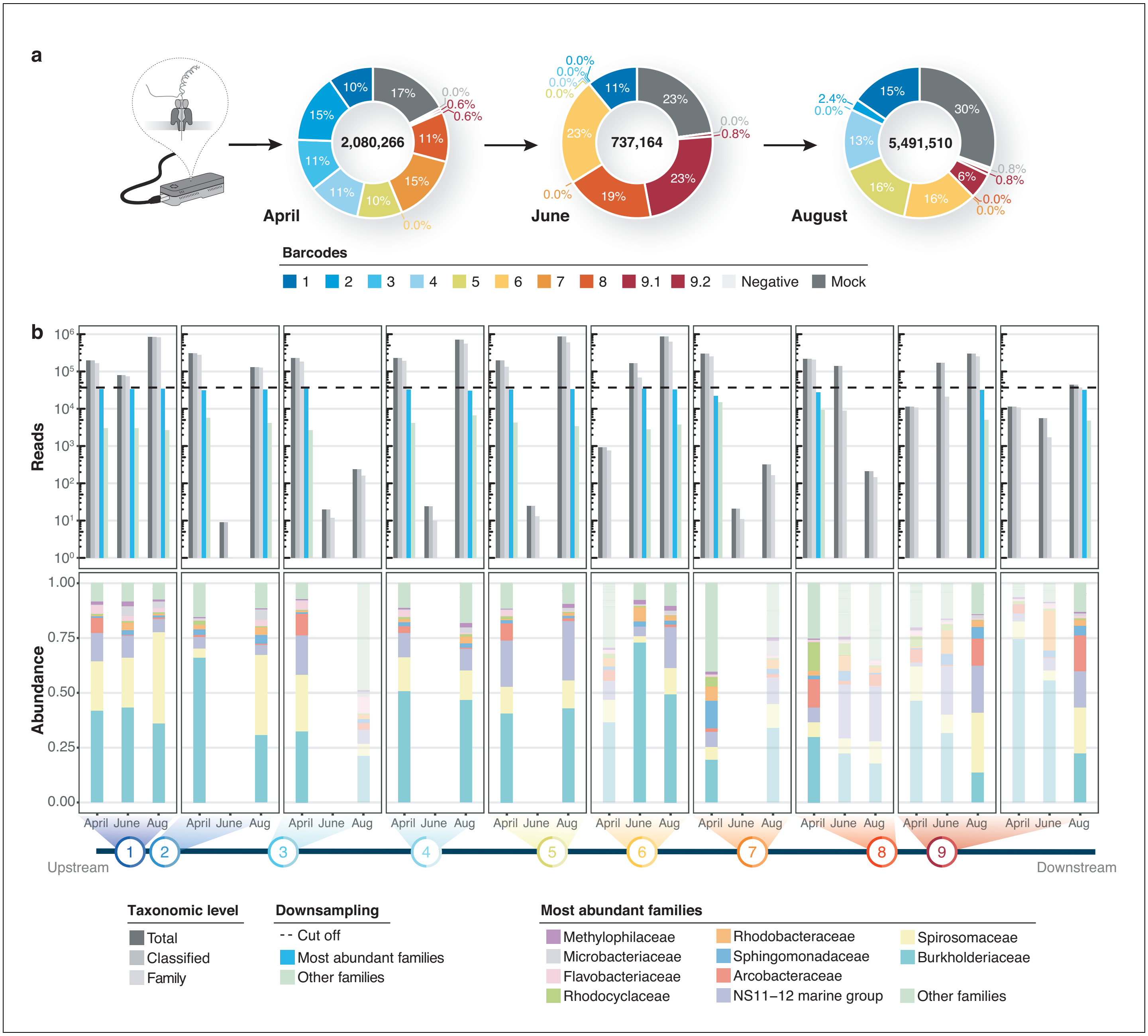


Figure 3. Bacterial diversity of the River Cam. (**a**) Nanopore sequencing output summary. Values in the centre of the pie charts depict total numbers of classified nanopore sequences per time point. Percentages illustrate representational fractions of locations and control barcodes (negative control 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.

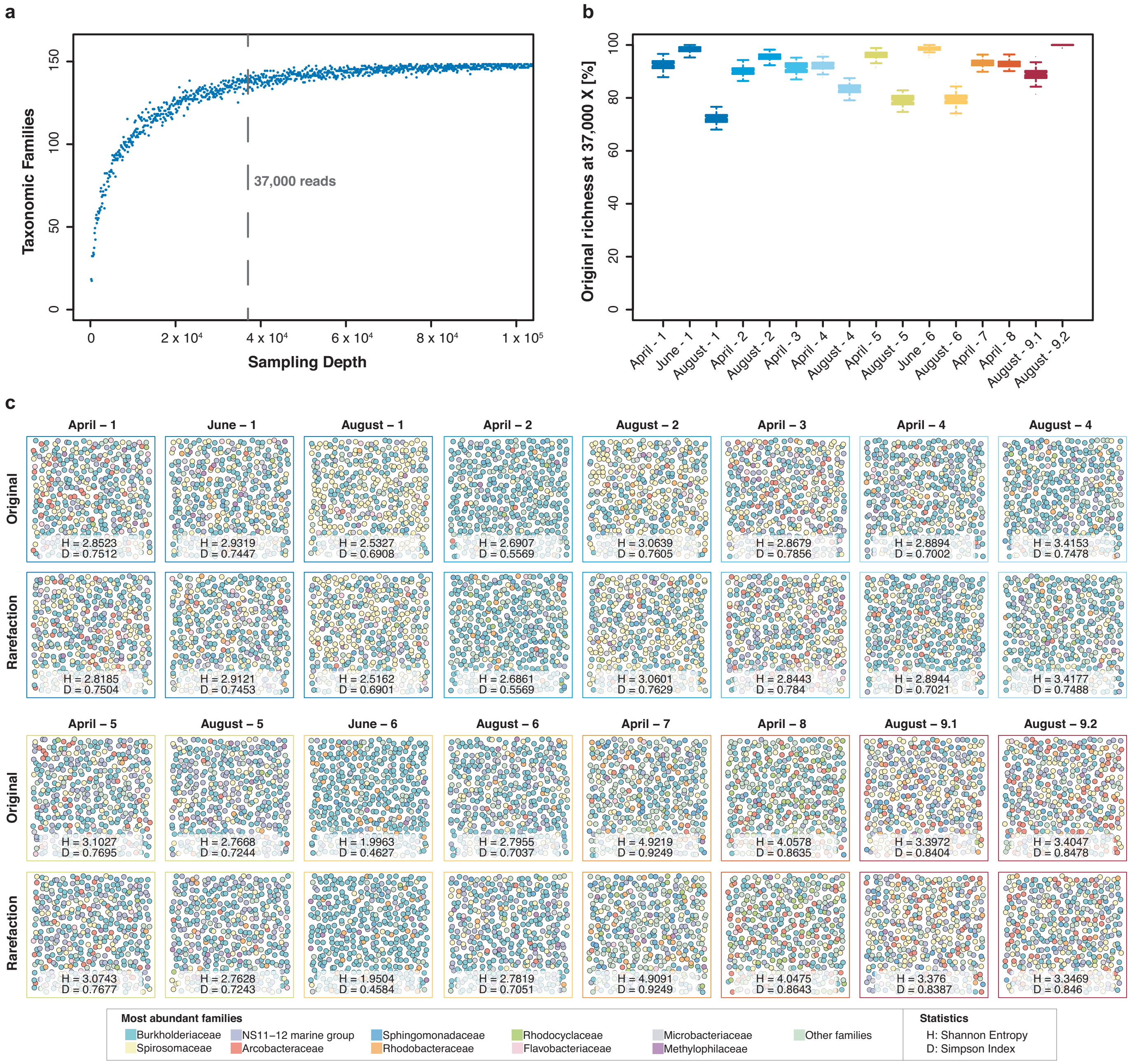


Figure 3—figure supplement 1. Impact of rarefaction on diversity estimation. (a) Example rarefaction curve for bacterial family classifications of the 'April-1' sample. The chosen cut-off preserves most (~90%) of the original family taxon richness (vertical line). (b) Difference between original and 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 3b***. Shannon (H) and Simpson (D) indices for all samples indicate marginal differences between pairs of original and rarefied sets.

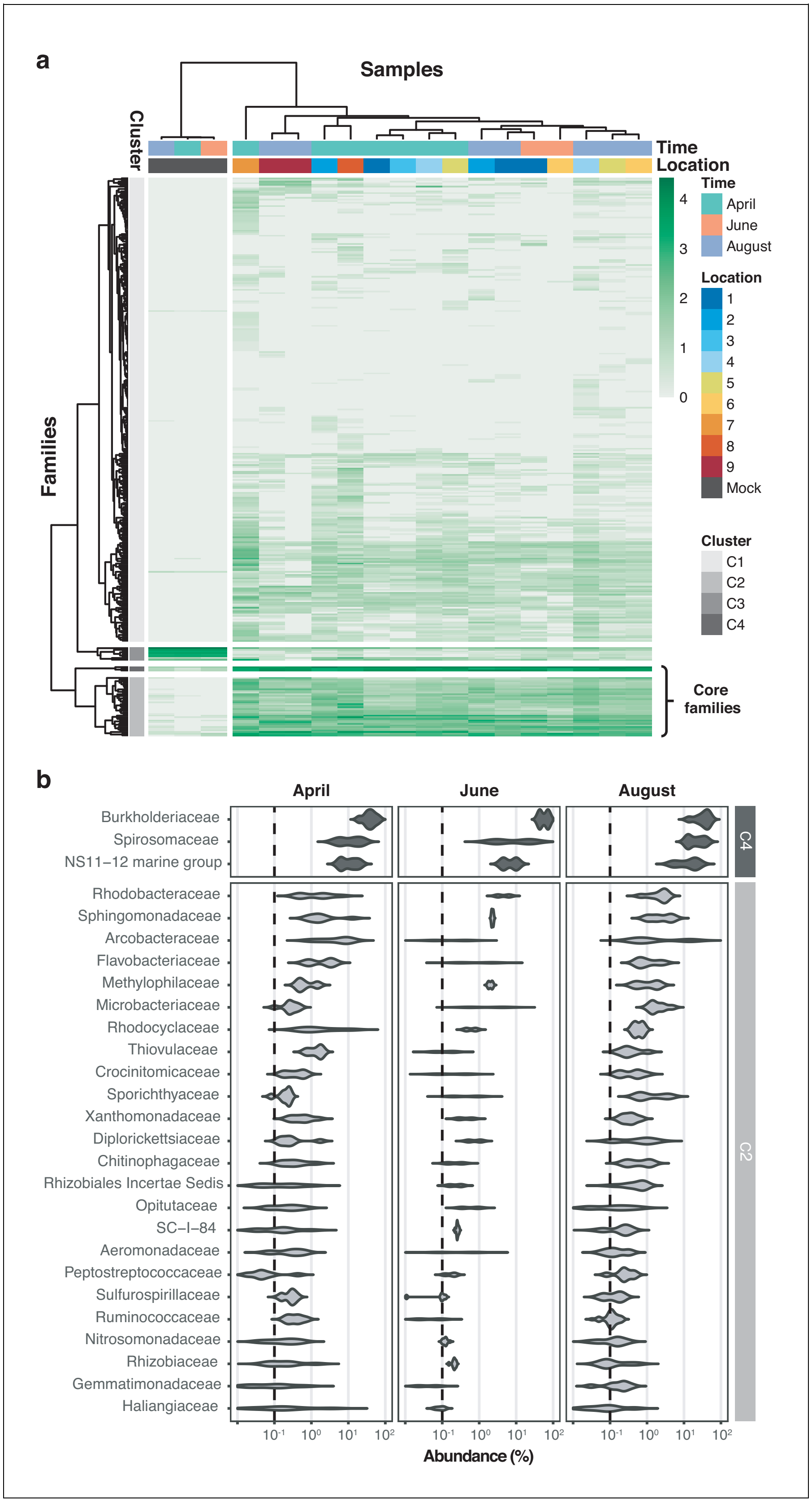


Figure 4. Core microbiome of the River Cam. (a) Hierarchical clustering of bacterial family abundances across freshwater samples after rarefaction, together with the mock community control. Four major clusters of bacterial

Figure 4 continued on next page

Figure 4 continued

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. (**b**) Detailed river core microbiome. Violin plots summarise fractional representation of bacterial families from clusters C2 and C4 (log_{10} scale of relative abundance [%] across all samples, n_{April} = 7, n_{June} = 2, n_{August} = 7), sorted by median total abundance. Vertical dashed lines depict 0.1% proportion.

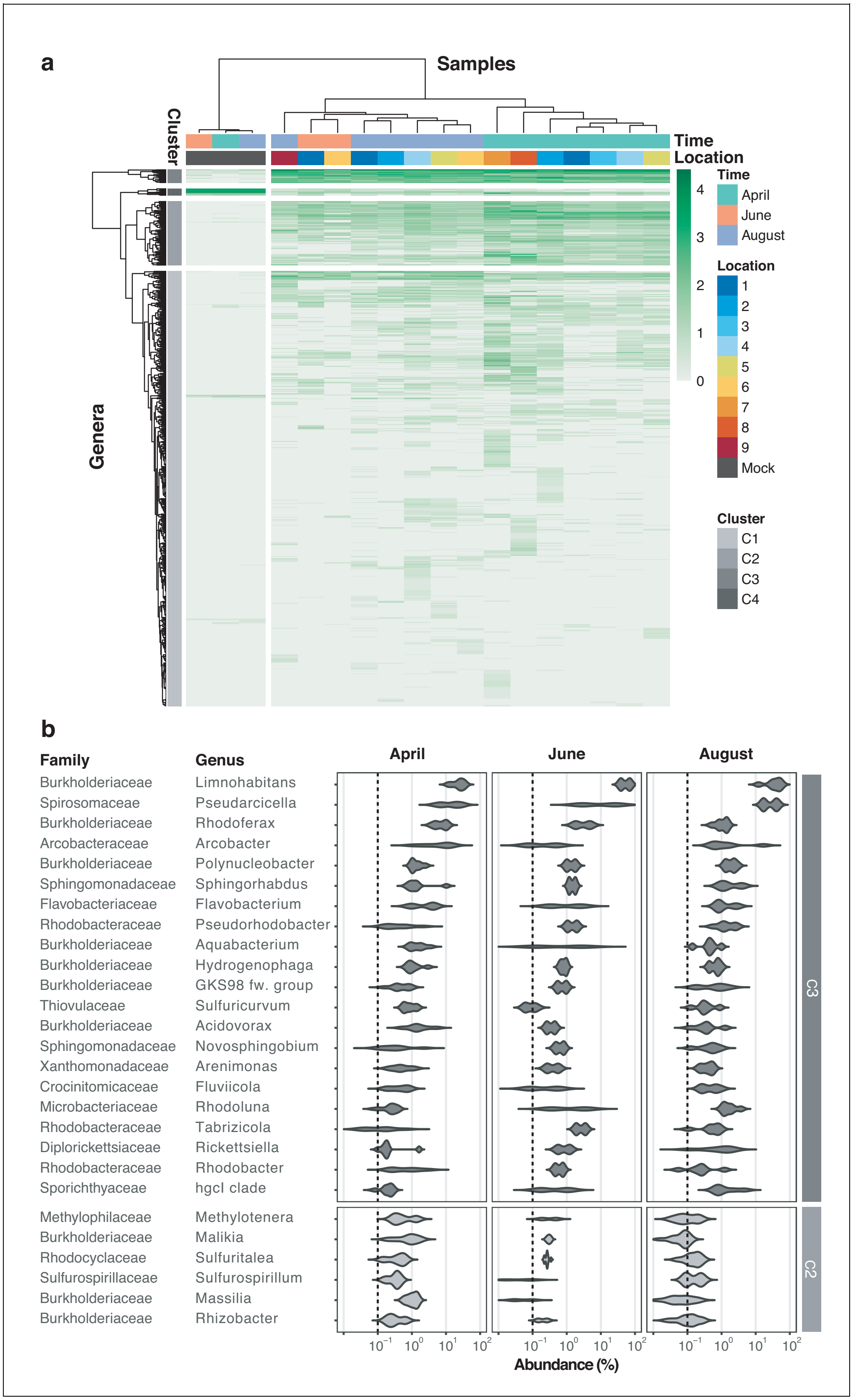


Figure 4—figure supplement 1. River Cam core microbiome analysis on the bacterial genus level. (a) Hierarchical clustering of bacterial genera abundances across freshwater samples after rarefaction, together with the mock community control. In similarity to the family analysis displayed in

Figure 4—figure supplement 1 continued on next page

Figure 4—figure supplement 1 continued

Figure 4, bacterial genera are clustered into four groups. Two of these (C3 and partially C2) correspond to the core microbiome of ubiquitously abundant genera, one (**C4**) corresponding to the main mock community genera and one (**C1**) corresponding to the majority of rare accessory taxa. (**b**) Dominant river core microbiome on the genus level. Violin plots ($\log_{10}$ scale of relative abundance [%] across all samples, n_{April} = 7, n_{June} = 2, n_{August} = 6) summarise fractional representation of the top 27 bacterial genera and corresponding families from clusters C2 and C3, sorted by median total abundance. Vertical dashed line depicts 0.1% proportion. Out of the top 16 core families (***Figure 4b***), only the NS11-12 marine group family was found not to be represented on the genus level; NS11-12 marine group genera are mainly composed of uncultured bacteria, which here could not be classified at higher resolution.

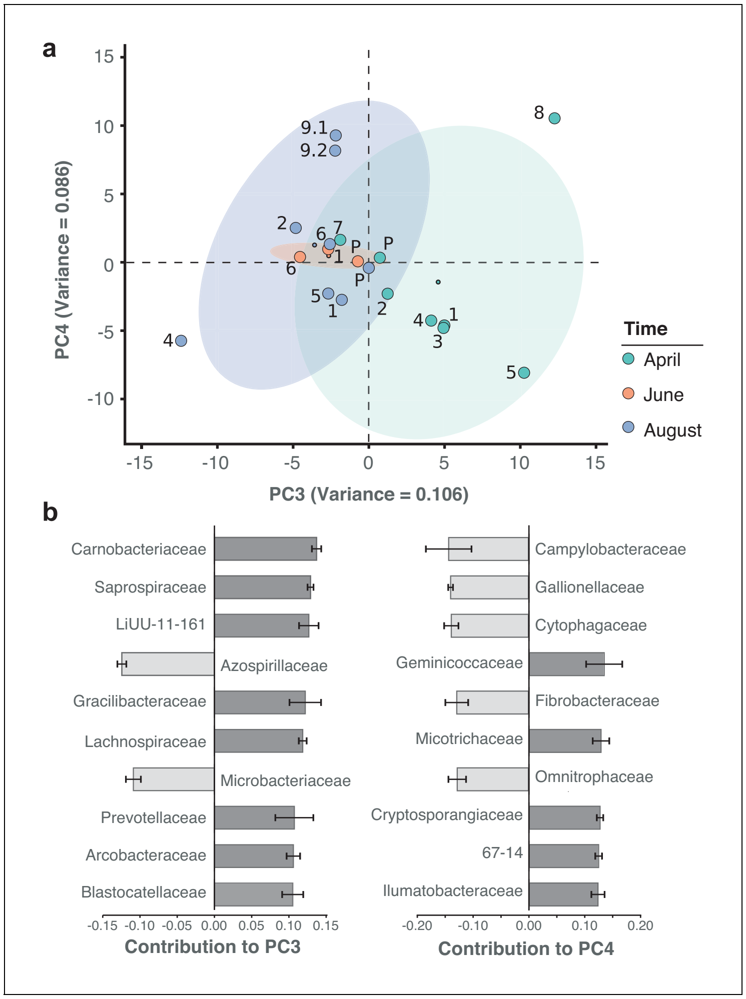


Figure 5. Spatiotemporal axes of taxonomic diversity in the River Cam. (**a**) PCA of bacterial composition across locations, indicating community dissimilarities along the main time (PC3) and spatial (PC4) axes of variation; dots coloured according to time points. Kruskal-Wallis test on PC3 component contributions, with post-hoc Mann-Whitney U rank test (April vs. August): p = 2.2*10^{-3}. (**b**) Contribution of individual bacterial families to the PCs in (**a**). Error bars represent the standard deviation of these families across four independent rarefactions.

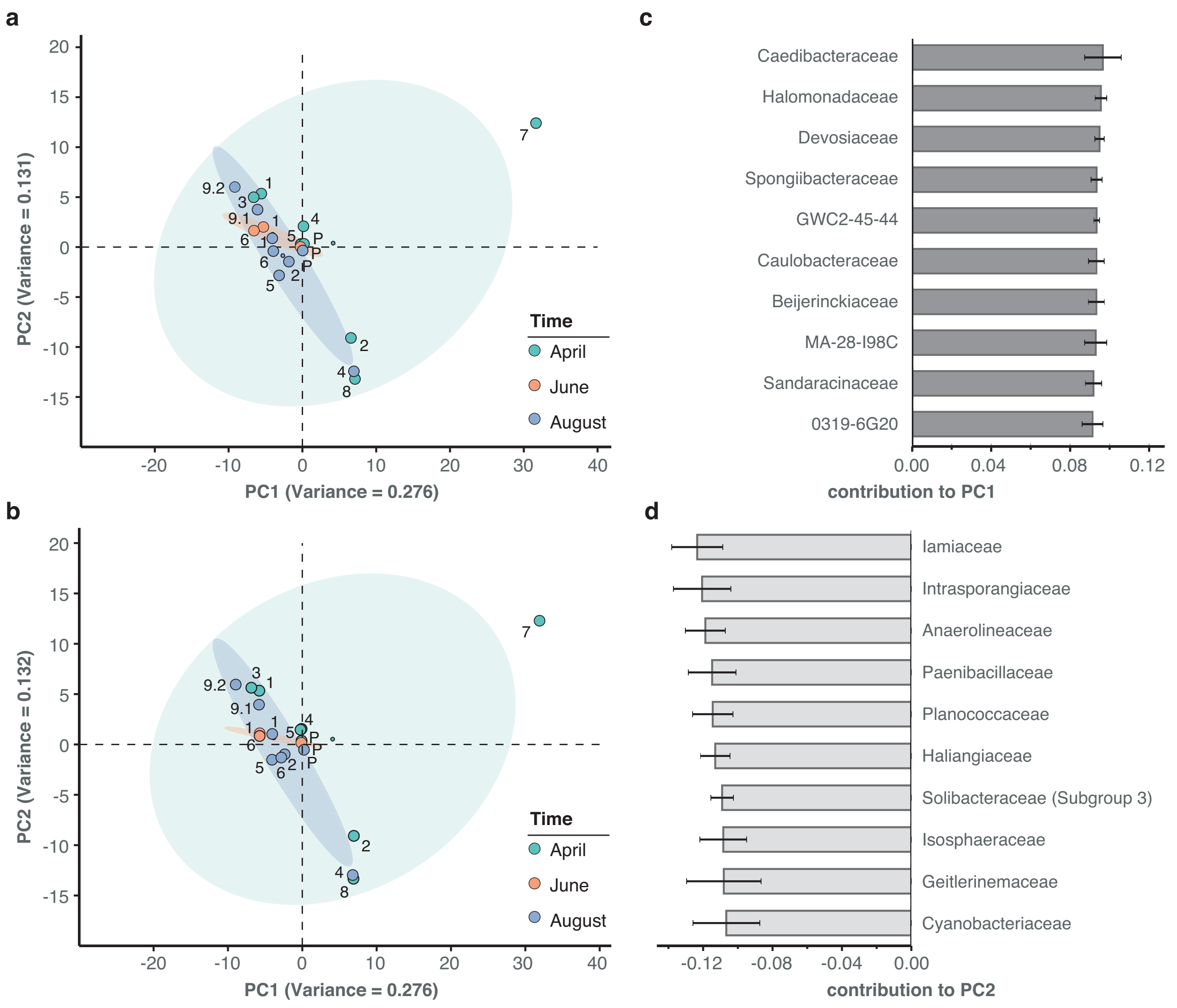


Figure 5—figure supplement 1. Principal component analysis of river bacterial family compositions. (**a–b**) PCA with two independent rarefaction sets to 37,000 reads in all freshwater sequencing samples. Numbers and coloured dots indicate locations for each time point. The first and second principal 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.

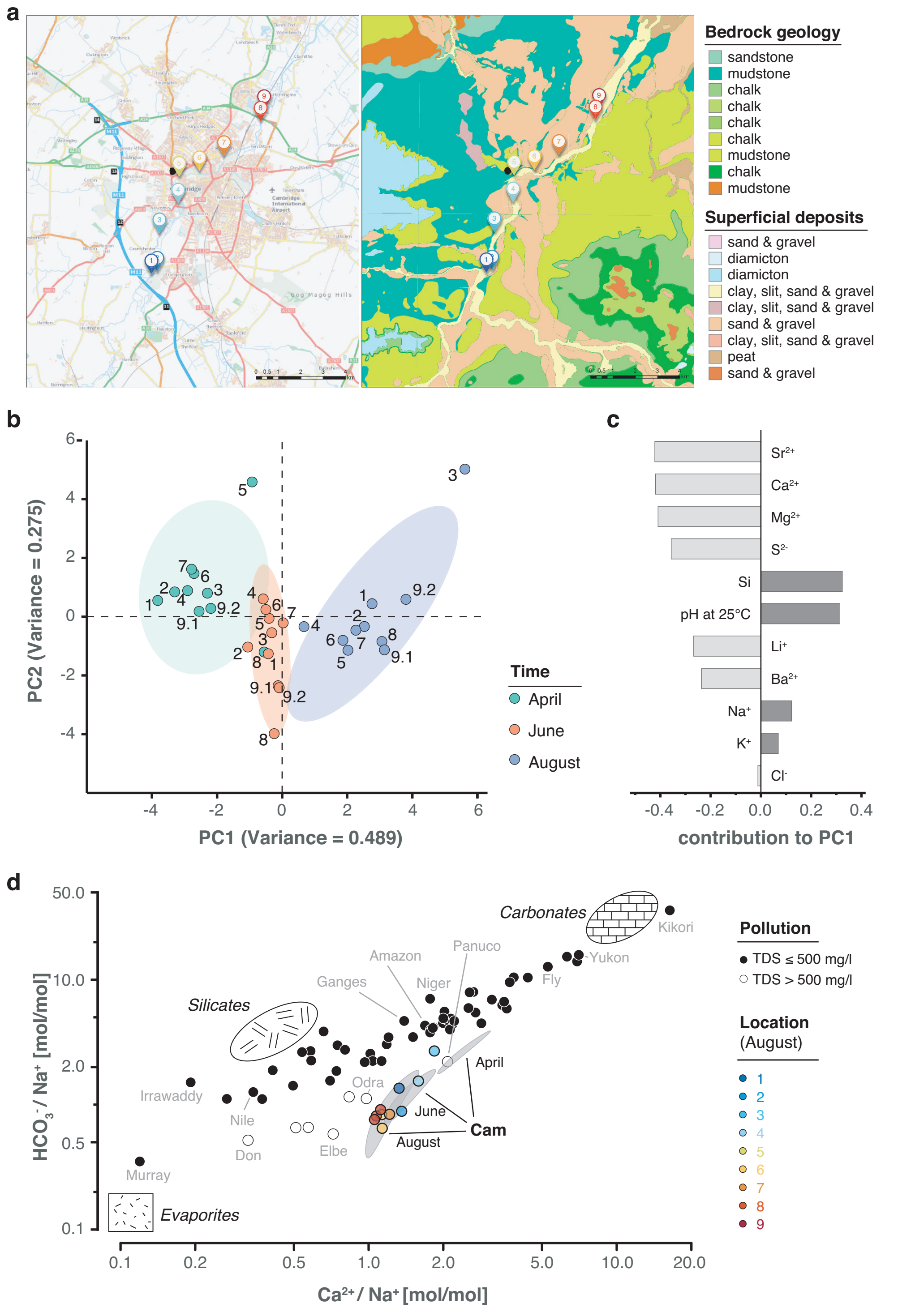


Figure 6. Geological and hydrochemical profile of the River Cam and its basin. (**a**) Outline of the Cam River catchment surrounding Cambridge (UK), and its corresponding lithology. Overlay of bedrock geology and superficial deposits (British Geological Survey data: DiGMapGB-50, 1:50,000 scale) is shown 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%

Figure 6 continued on next page

Figure 6 continued

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 the 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 upstream-to-downstream 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).

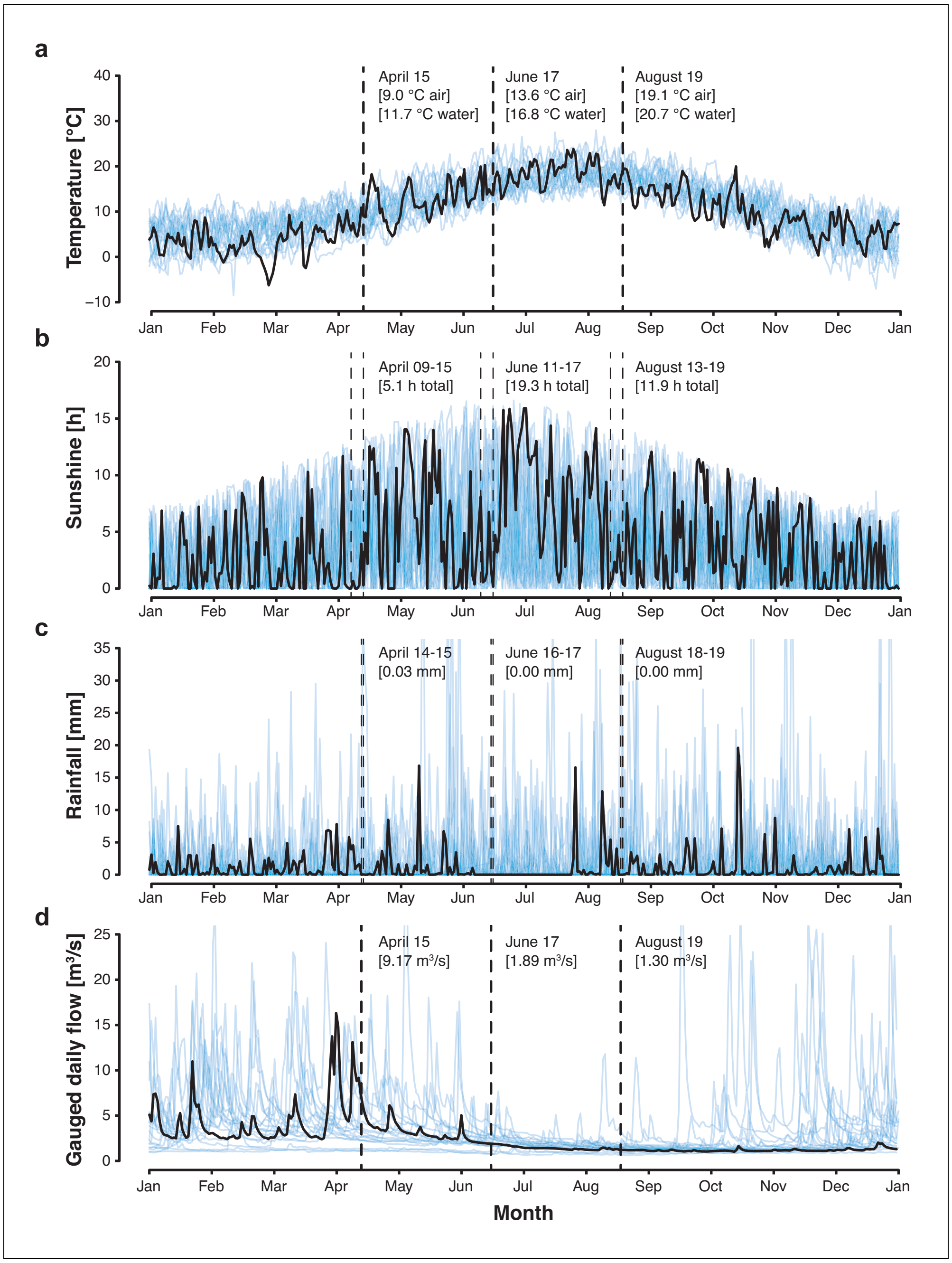


Figure 6—figure supplement 1. Cambridge weather and River Cam flow rate. (**a**) Daily air temperature [°C], (**b**) daily sunshine [hr], and (**c**) daily rainfall [mm] of Cambridge in 2018 (black trend line) vs. 1998–2017 (blue background trend lines). (**d**) Cam River gauged daily flow [m^3s^{-1}]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

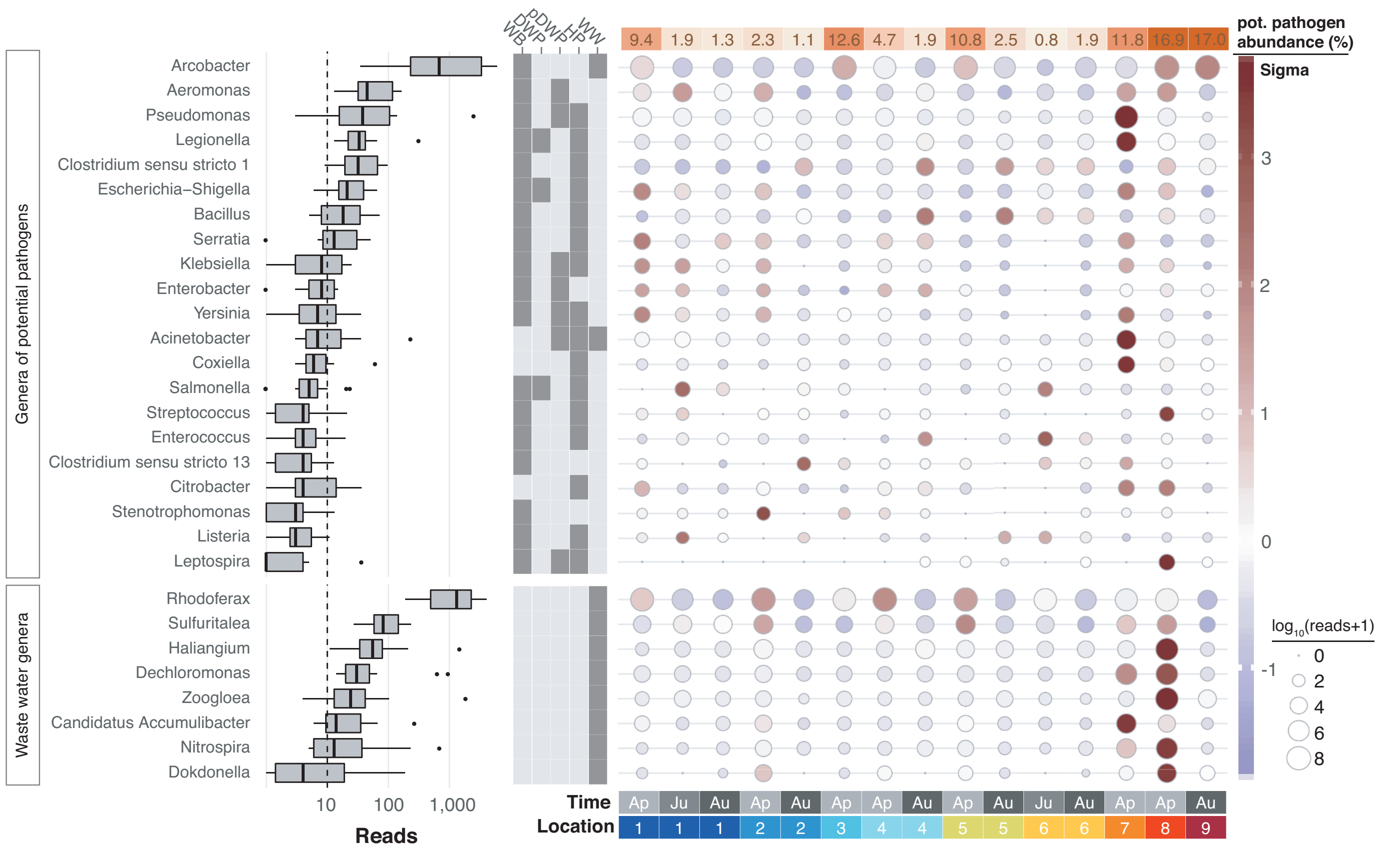


Figure 7. Potentially pathogenic and wastewater treatment related bacteria in the River Cam. Boxplots on the left show the abundance distribution across locations per bacterial genus. Error bars represent Q1 – 1.5*IQR (lower), and Q3 + 1.5*IQR (upper), respectively; Q1: first quartile, Q3: third quartile, IQR: interquartile range. The central 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 file 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.

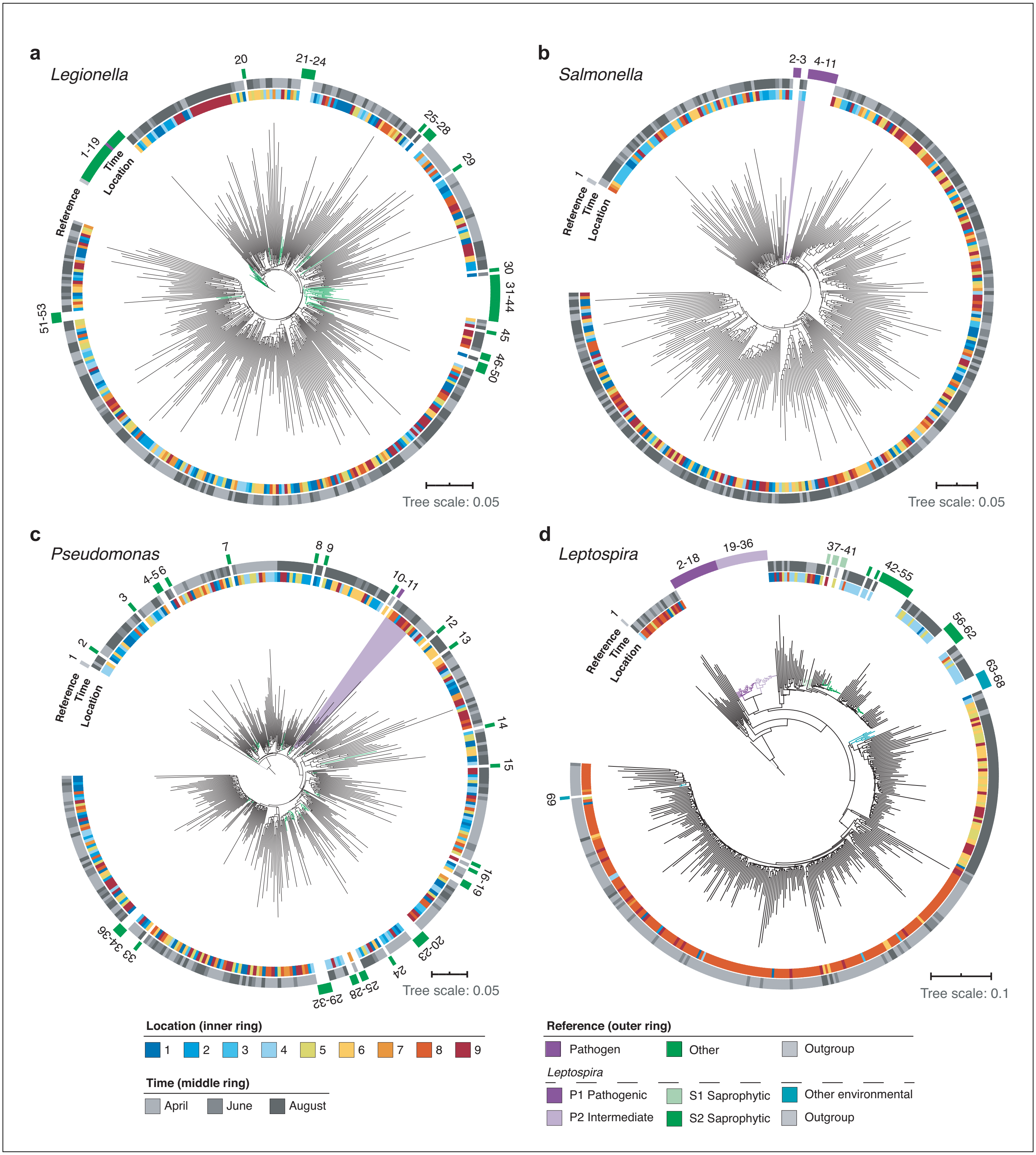


Figure 8. High-resolution phylogenetic clustering of candidate pathogenic genera in the River Cam. Phylogenetic trees illustrating multiple sequence alignments of exemplary River Cam nanopore reads (black branches) classified as (a) *Legionella*, (b) *Salmonella*, (c) *Pseudomonas*, or (d) *Leptospira*, together with known reference species sequences ranging from pathogenic to saprophytic taxa within the same genus (coloured branches). Reference species sequences are numbered in clockwise orientation around the tree (***Supplementary file 4***). Nanopore reads highlighted in light violet background display close clustering with pathogenic isolates of (b) *Salmonella spp.* and (c) *Pseudomonas aeruginosa*.