

# National-scale biogeography and function of river and stream bacterial biofilm communities

Received: 10 March 2025

Accepted: 19 October 2025

Published online: 26 November 2025

Amy C. Thorpe<sup>1</sup>✉, Susheel Bhanu Busi<sup>1</sup>✉, Jonathan Warren<sup>2</sup>,  
Laura H. Hunt<sup>2</sup>, Kerry Walsh<sup>2</sup> & Daniel S. Read<sup>1</sup>✉

Biofilm-dwelling microorganisms coat the surfaces of stones in rivers and streams, forming diverse communities that are fundamental to biogeochemical processes and ecosystem functioning. Flowing water (lotic) ecosystems face mounting pressures from changes in land use, chemical pollution, and climate change. Despite their ecological importance, the taxonomic and functional diversity of river biofilms and their responses to environmental change are poorly understood at large spatial scales. We conducted a national-scale assessment of bacterial diversity and function using metagenomic sequencing from rivers and streams across England. We recovered 1,014 metagenome-assembled genomes (MAGs) from 450 biofilms collected across England's extensive river network. Substantial taxonomic novelty was identified, with ~20% of the MAGs representing novel genera. Here we show that biofilm communities, dominated by generalist bacteria, exhibit remarkable functional diversity and metabolic versatility, and likely play a significant role in nutrient cycling with the potential for contaminant transformation. Measured environmental drivers collectively explained an average of 71% of variation in the relative abundance of bacterial MAGs, with geology and land cover contributing most strongly. These findings highlight the importance of river biofilms and establish a foundation for future research on the roles of biofilms in ecosystem health and resilience to environmental change.

It is estimated that up to 80% of all bacterial and archaeal cells on Earth exist within biofilms<sup>1</sup>. These biofilms dominate microbial life in freshwater streams and rivers<sup>2</sup>, forming complex, interacting assemblages of taxonomically and functionally diverse heterotrophic and phototrophic microbes. These include bacteria, algae, fungi, protozoa, and viruses<sup>2</sup>, existing within an extracellular polymeric matrix that adheres to the surfaces of stones, aquatic plants, and sediments<sup>1</sup>. Biofilms are hotspots of metabolic activity, driving essential biogeochemical processes such as nutrient cycling, primary production, and respiration<sup>3</sup>. They play key roles in degrading pollutants, regulating water quality<sup>4</sup>, and supporting energy flow to higher trophic levels<sup>5</sup>. Microbes within

biofilms are therefore critical to the functioning of river ecosystems<sup>6</sup>, which are under increasing pressure from pollution and climate change<sup>7</sup>. Monitoring biofilm microbial communities offers valuable opportunities to assess ecosystem health, detect early signs of pollution, and inform restoration strategies, such as water quality monitoring, bioremediation, and improved wastewater treatment<sup>8</sup>. Consequently, a detailed understanding of the diversity and function of biofilm microbial communities in rivers is essential for the effective management and conservation of these vital ecosystems<sup>4</sup>.

Diverse microbial communities have been detected in river biofilms<sup>2,6</sup>. Their composition and functionality are likely influenced by

<sup>1</sup>UK Centre for Ecology & Hydrology, Benson Lane, Crowmarsh Gifford, Wallingford, UK. <sup>2</sup>Environment Agency, Horizon House, Bristol, UK.

✉ e-mail: [amytho@ceh.ac.uk](mailto:amytho@ceh.ac.uk); [susbus@ceh.ac.uk](mailto:susbus@ceh.ac.uk); [daniel.read@ceh.ac.uk](mailto:daniel.read@ceh.ac.uk)

a complex interplay of factors, including physiochemical parameters such as temperature, pH, and dissolved oxygen<sup>9–11</sup>; the concentrations of nutrients, organic matter, and pollutants<sup>12–14</sup>; flow dynamics<sup>15</sup>; land use<sup>16,17</sup>; ecological processes such as selection and dispersal<sup>18</sup>; and biotic interactions including cooperation and competition between microbes, and grazing by macroinvertebrates<sup>19,20</sup>. By shaping microbial community composition and function, environmental changes can significantly affect biogeochemical fluxes and overall ecosystem health.

Despite their ecological importance, the diversity of microbial communities and their functional roles in river biofilms remains poorly understood. Furthermore, identifying the environmental drivers that influence biofilm community dynamics and assessing their resilience to pressures, such as nutrient pollution, presents ongoing challenges. These challenges include the complex and heterogeneous nature of biofilm communities<sup>6</sup>, which are shaped not only by a multitude of external environmental factors, but also by the unique microenvironment and the intricate network of microbe-microbe interactions occurring within the biofilm<sup>3,21</sup>. Few studies capture a sufficiently comprehensive suite of environmental covariates to fully resolve these relationships. Moreover, much of the existing research on river biofilm communities has focused on individual rivers or catchments, providing valuable insights into local environmental drivers of microbial diversity and community dynamics<sup>22–24</sup>. However, these studies typically comprise a relatively small number of samples and cover a limited geographic area.

In recent years, national-scale studies have been undertaken to capture the wider diversity of freshwater microbes and explore their biogeographic patterns across a broad range of environments<sup>25–28</sup>. These national-scale datasets demonstrate the value of molecular approaches, including metagenomics, for uncovering both the taxonomic and functional diversity of microbial communities, and highlight community dynamics and adaptations in response to environmental drivers across large spatial scales. Collectively, national-scale datasets can provide a more comprehensive understanding of the ecological dynamics of freshwater microbes and their drivers at a global scale. However, there are few examples of such national-scale studies of river biofilm communities, and, to our knowledge, no national-scale metagenomic assessment of river microbes currently exists in England.

The Environment Agency (EA) established the River Surveillance Network (RSN) as a national initiative to monitor and assess long-term changes in the health of English rivers. The RSN comprises 1600 sites, representing the diversity of England's extensive river network. As part of this initiative, 450 river biofilms were systematically collected over a three-year period from 146 RSN sites spatially distributed across England for metagenomic sequencing (Fig. 1A). This comprehensive dataset spanned a latitudinal gradient of 645 km. It encompassed all major land cover types present in the country, ranging from woodlands and grasslands to arable land and urban areas, and represented a variety of catchment geologies (Supplementary Table 1, Supplementary Data 1). In addition to the diverse catchments, the RSN sites spanned considerable variability in their physicochemical conditions, including water temperature, alkalinity, pH, dissolved oxygen, dissolved organic carbon, orthophosphate, nitrate-N, and nitrite-N (Supplementary Table 2, Supplementary Data 1).

This systematic investigation of the river network, coupled with detailed environmental data, enables a robust assessment of microbial biofilms, comparable to existing large-scale efforts<sup>25,26,28,29</sup>, capturing their biogeographic distribution, taxonomic diversity, metabolic and functional potential, and environmental drivers. Specifically, this study addresses three fundamental research questions:

- (i) How are river biofilm bacterial communities structured across England's river network?

- (ii) What metabolic capabilities and functional traits underpin their roles in ecological processes such as biogeochemical cycling?
- (iii) How do environmental drivers shape river biofilm bacterial communities at a national-scale?

By addressing these questions, this study offers novel insights into the biogeography of river biofilm bacteria and their critical roles in ecosystem functioning. These findings help provide a foundation for effective monitoring, management, and conservation of river ecosystems.

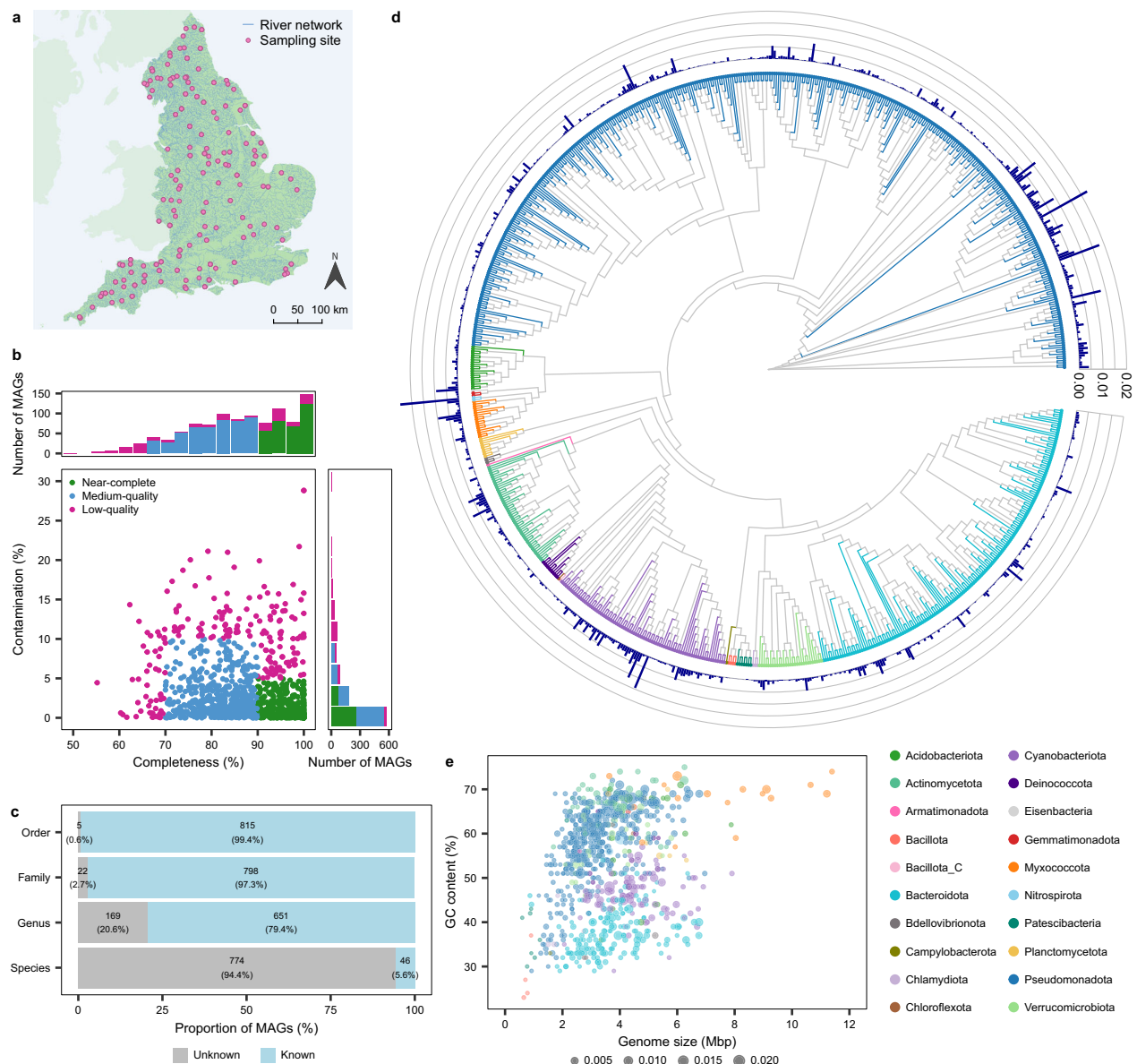
## Results and discussion

### Diversity and composition of river biofilm microbial communities

River biofilms are cosmopolitan in their distribution and known to host high microbial diversity, supported by continuous nutrient and community inputs from upstream<sup>6</sup>. However, our understanding of freshwater biofilm microbial communities is significantly less advanced than that of microbial communities in the water column<sup>25</sup> or sediments<sup>9</sup>. Although several studies have explored the taxonomic composition of river biofilms at the local scale<sup>10,11,23</sup>, the overall abundance, composition, and genomic traits of these communities across large spatial scales remain poorly understood.

Using metagenomic data averaging 66.80 million reads per sample ( $\pm 28.5$  million reads SD), we found that bacterial sequences comprised the majority (85.17%) of all metagenomic reads in the river biofilms. Eukaryotes and Archaea represented 11.56% and 2.64% of the reads, respectively (Supplementary Data 2). Because of their dominance in biofilms, we focused our analysis on bacterial communities. The assembled reads were used to reconstruct a total of 4027 bacterial bins (archaeal bins were not recovered), which were dereplicated into 1014 metagenome-assembled genomes (MAGs) (also known as species-level genome bins). Of these, 329 MAGs (32%) were identified as near-complete, 491 (48%) as medium-quality, and 194 (19%) as low-quality (Fig. 1B) with a mean completeness of 86.70% ( $\pm 9.3\%$  SD) and a mean contamination of 3.57% ( $\pm 4.1\%$  SD). Mean genome size was estimated to be 3.66 Mbp ( $\pm 1.4$  Mbp SD) and mean GC content was 52.66% ( $\pm 11.7\%$  SD) (Supplementary Fig. 1, Supplementary Data 3).

The 820 medium or high-quality bacterial MAGs retained for downstream analysis encompassed a diverse range of taxa, including 20 known phyla, 35 classes, 91 orders, 160 families, 311 genera and 46 species. The 46 MAGs assigned to known species according to GTDB-Tk had an average nucleotide identity (ANI) ranging from 95.0 to 95.4% (mean  $95.01 \pm 0.06$ ). There was also substantial taxonomic novelty detected, with 169 (20.6%) of the recovered MAGs representing previously uncharacterised genera, and 774 MAGs (94.4%) representing previously uncharacterised species with no suitable GTDB-Tk reference ( $\leq 65\%$  ANI across all species within the database; Supplementary Data 3) (Fig. 1C). Stringent assessment using only near-complete MAGs revealed that 306 (93%) were novel species-level genome bins, with no previously identified representatives. Across all river biofilms, Pseudomonadota was the most dominant phylum and comprised almost half of the total community, with a mean relative abundance of 48.49% ( $\pm 14.8\%$  SD) (Fig. 1D). Other abundant phyla included Cyanobacteriota with a mean relative abundance of 15.68% ( $\pm 16.9\%$ ), Bacteroidota with a mean relative abundance of 14.77% ( $\pm 9.5\%$ ), and Actinomycetota with a mean relative abundance of 6.27% ( $\pm 5.1\%$ ). Other less abundant phyla each comprised less than 5% of the total community on average. This community composition aligns with previous studies on benthic biofilms collected from a variety of river types globally, including the groundwater and rain-fed River Thames (United Kingdom)<sup>30</sup> and urban and rural rivers in China<sup>31</sup>, and glacier-fed streams in alpine regions such as the Southern Alps (New Zealand) and the Caucasus (Russia)<sup>32</sup>. These studies also reported the dominance of Pseudomonadota and Bacteroidota, which have been reported to play important roles in the



**Fig. 1 | Overview of the diversity, novelty, and genomic characteristics of river biofilm bacterial MAGs.** **a** Biofilm sampling sites in rivers across England (contains OS data © Crown copyright and database rights 2024). **b** Distribution of MAG completeness and contamination. **c** Taxonomic novelty of MAGs at the order, family, genus, and species level. **d** Phylogenetic tree of MAGs where colours

represent different bacterial phyla, and outer bars represent the mean relative abundance of MAGs. **e** Estimated GC content and genome size of MAGs where point size is scaled to mean relative abundance and colour represents bacterial phylum. Source data are provided as a Source Data file.

degradation of organic matter, and Cyanobacteriota, which contributes to primary productivity<sup>33</sup>. In contrast, river surface waters are typically dominated by Pseudomonadota and Bacteroidota, but also by Actinomycetota and Verrucomicrobiota<sup>25,30</sup>, highlighting differences in community assembly between benthic biofilms and pelagic habitats.

### Genomic traits of river biofilm bacteria

We found that genome size and GC content were correlated with phylogenetic background, delineating these traits across the bacterial tree of life (Fig. 1E). It was observed that many Myxococcota MAGs are large with a high GC content, Pseudomonadota MAGs are small to medium with a high GC content, and Cyanobacteriota and Bacteroidota MAGs are small to medium with a lower GC content. A similar phylogenetic relationship between GC content and genome size has

been observed in pelagic and sediment bacteria from lakes and rivers<sup>34–36</sup>, suggesting that the patterns observed for many phyla such as Myxococcota, Pseudomonadota, Cyanobacteriota, and Bacteroidota are consistent with broader trends across freshwater environments. These traits may reflect ecological niche adaptation. For example, bacteria with larger genomes are typically copiotrophs, with numerous genes associated with complex metabolic processes, allowing them to thrive in fluctuating or nutrient-rich environments<sup>37</sup>. In contrast, those with smaller genomes often have specific nutrient requirements and occupy more stable or oligotrophic environments<sup>37</sup>, consistent with genome streamlining theory<sup>38</sup>. Furthermore, lower GC content reduces nitrogen requirements, which may be advantageous for bacteria in nitrogen-limited environments<sup>36</sup>. Genome size is likely affected by a complex interplay of environmental factors, in addition to trophic strategies<sup>39</sup>, and biotic interactions<sup>40,41</sup>.

Functional annotation of these MAGs revealed multiple adaptations for biofilm matrix colonisation and maintenance, including genes associated with biofilm formation, quorum sensing, chemotaxis, flagellar assembly, the two-component system, exopolysaccharide (EPS) biosynthesis and ATP synthase-binding cassette (ABC) transporters (Supplementary Fig. 2). Notably, many of these genes were enriched in larger MAGs, which is consistent with the functional and metabolic flexibility of copiotrophs and may reflect the greater genomic complexity associated with larger bacterial genomes<sup>36</sup>. These traits are broadly conserved across bacterial biofilms in other freshwater environments, including glacier-fed streams<sup>32</sup> and other cryospheric systems<sup>42</sup>, where they have also been found across diverse taxonomic groups, highlighting the phylogenetically widespread contribution to the biofilm matrix through roles in biofilm formation, motility, sensing the environment, and nutrient acquisition.

### Biogeography of river biofilm bacteria

River biofilms host complex communities composed of both generalist and specialist taxa, each fulfilling distinct ecological roles<sup>2</sup>. Generalist taxa are highly adaptable, thriving across a diversity of habitats, and tolerating a broad range of environmental conditions, whereas specialist taxa are adapted to specific habitats, thriving only under specific environmental conditions<sup>43</sup>. Despite the ecological significance of these communities, the biogeographic patterns of river biofilm bacteria are poorly understood and have not yet been addressed in the context of ecological niche space. Using biogeographic mapping and niche breadth analysis coupled with measures of abundance and occupancy, we investigated the spatial distribution and ecological niches of river biofilm bacteria.

The distribution of MAGs across the river network revealed national-scale biogeographic patterns (Fig. 2). Pseudomonadota and Bacteroidota exhibited high relative abundances throughout the country (>5% in more than 90% of sites), dominating biofilm communities. In contrast, Cyanobacteriota displayed regional preferences, with higher relative abundance in the north and southwest of England, and a lower occupancy overall (Fig. 3A), indicating that they occasionally dominate biofilm communities (>5% in only 63% of sites). Other taxa, such as Acidobacteriota, Myxococcota, and Nitrospirata, had a widespread presence but exhibited localised hotspots where their relative abundance was notably higher. Campylobacter and Chloroflexota had more restricted biogeographic distributions and were present at low relative abundances in only 13 and 79 samples, respectively. However, most MAGs were widespread rather than limited to specific locations. Each MAG occupied an average of 413 samples (92%) spatially distributed across the river network, with 176 MAGs (21%) present in all samples. Many of these 176 high-occupancy MAGs were the most abundant members of the community, with a mean relative abundance of up to 2.49% (Fig. 3A). A similar relationship between occupancy and mean relative abundance has been observed in surface water bacterial communities in rivers across North America<sup>25</sup>. However, while river networks comprise mostly high-occupancy taxa, studies of bacterial communities in spatially distributed lakes have found a greater proportion of low-occupancy taxa<sup>26,44</sup>. This difference likely reflects the underlying ecological factors that differ between rivers and lakes, such as habitat connectivity, dispersal dynamics, and environmental stability<sup>45</sup>.

Using niche breadth analysis, we further categorised the MAGs as generalists (niche breadth,  $B_N > \text{median } B_N$ ) or specialists ( $B_N < \text{median } B_N$ ). This metric leverages the relative abundance of MAGs to quantify how evenly they are distributed across the river network, providing a measure of their ecological breadth. Among the 176 high-occupancy MAGs, 88% were identified as generalists. Generalists comprised 58.58% of Pseudomonadota MAGs, and 58.62% of Verrucomicrobiota MAGs, while 51.19% of Cyanobacteriota MAGs, 60.00% of Acidobacteriota MAGs, 64.67% of Bacteroidota MAGs, and 64.71% of

Myxococcota MAGs were categorised as specialists (Fig. 3B). Previous studies have also detected a high abundance of generalist bacteria in river and stream biofilms<sup>2,46,47</sup>. Furthermore, the distance-decay relationship of bacterial community dissimilarity and geographic distance was positive, but weak ( $R^2 = 0.04$ ,  $p < 0.001$ ) (Supplementary Fig. 3A, B), suggesting dispersal rates in bacterial biofilms are high, although other processes could also contribute, such as slow ecological or evolutionary turnover. Nevertheless, some degree of species sorting due to local environmental conditions likely occurs, consistent with the Baas Becking hypothesis, stating that everything is everywhere, but the environment selects<sup>48</sup>. Taken together, these results reveal a dominance of high-occupancy, generalist MAGs across the river network, emphasising the role of rivers as highly interconnected ecosystems with potential for widespread microbial dispersal. The ability of generalist bacteria to tolerate a diverse range of environmental conditions is likely an advantage in lotic ecosystems, where biofilm communities may become dislodged and recolonise further downstream<sup>46</sup>.

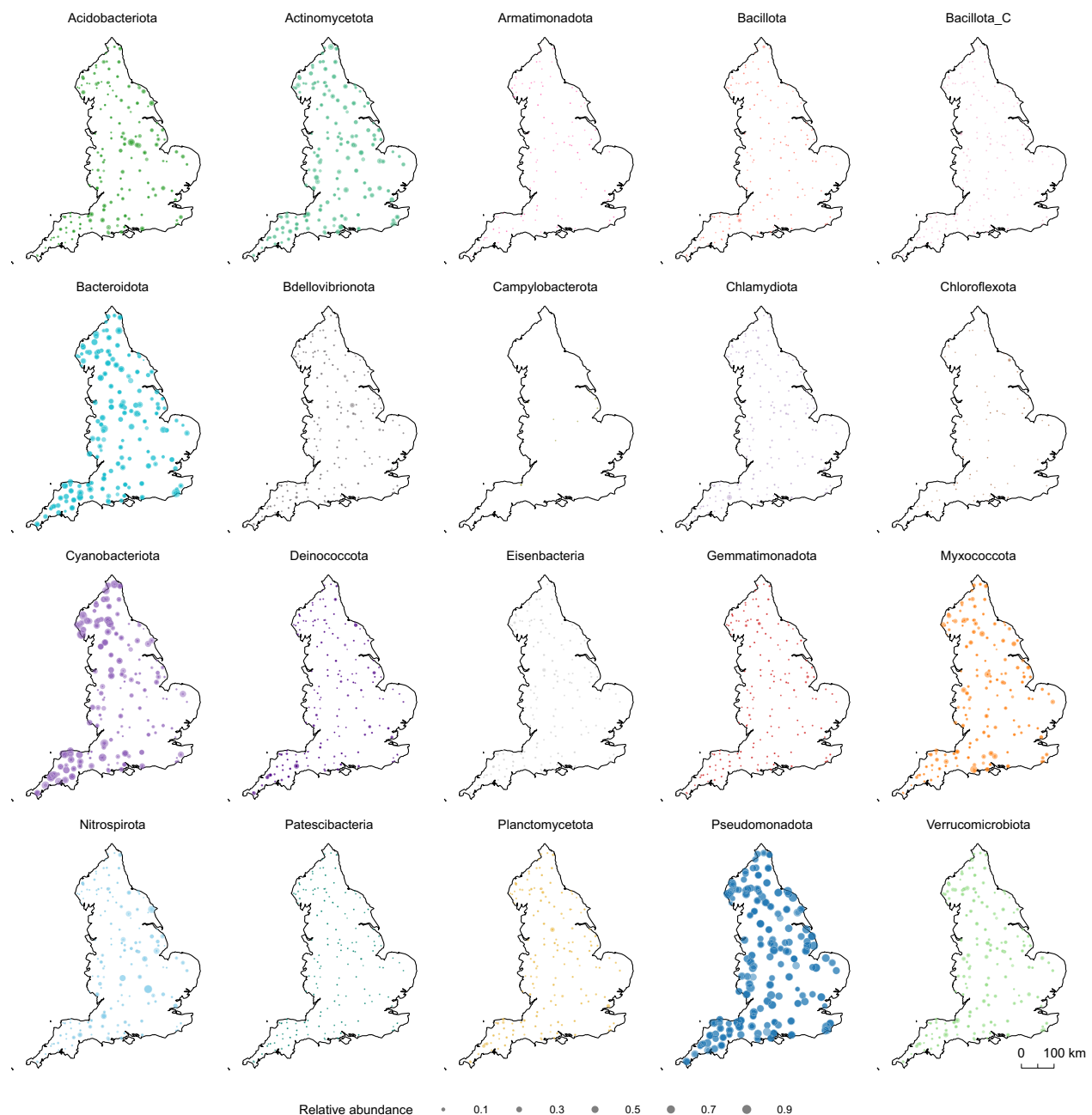
### Metabolic potential of river biofilm communities

Biofilm communities play critical roles in ecological processes such as primary production, nutrient cycling, and the transformation and breakdown of organic material and pollutants<sup>3,49</sup>. Their contributions to biogeochemical cycling can strongly influence the health and productivity of river ecosystems by mediating the availability of nutrients and other compounds to higher trophic levels in the food web<sup>5</sup>. Previous studies have emphasised the importance of assessing the metabolic capabilities of biofilms to better understand their functional roles in river ecosystems<sup>6</sup>. Despite this, there are few examples of studies that have conducted genome-resolved assessments of microbial metabolic potential within river biofilms<sup>32</sup>, particularly across broad spatial scales. As a result, detailed insights into the distribution and prevalence of key metabolic functions, which are necessary to understand the role of biofilm microbes in ecosystem functioning, remain limited.

The majority of bacterial MAGs were found to have the capacity to contribute to carbon, nitrogen, and sulfur cycling in rivers. Among these, Pseudomonadota, Cyanobacteriota, and Bacteroidota MAGs accounted for more than 52.41%, 16.62% and 10.95%, respectively, of the nutrient cycling genes identified in the biofilms (Supplementary Fig. 4). Across the 820 MAGs, the majority of steps in carbon, nitrogen, and sulfur cycling pathways were represented (Fig. 4A–C). Nearly all MAGs encoded for genes involved in organic carbon oxidation (99.39%), and a substantial proportion of the MAGs included genes involved in fermentation (86.83%), acetate oxidation (73.66%), sulfur oxidation (47.93%), and nitrite ammonification (43.29%). Many of these pathways represented by the river biofilms, including organic carbon oxidation and sulfur oxidation, were also prevalent in mountain stream biofilms<sup>50</sup>.

Carbon cycling genes, such as *sgdh* (pentose phosphate pathway) and *RuBisCO IV* (carbon fixation), were abundant in members of Pseudomonadota and Cyanobacteriota (Fig. 4D, Supplementary Fig. 5A). Nitrogen cycling genes, including *nirD* and *nirB* (nitrite reductases) (Fig. 4E, Supplementary Fig. 5B), and sulfur cycling genes, including *sdo* (sulfur dioxygenase), *fcbB* (sulfur dehydrogenase) and *sat* (sulfate reduction) (Fig. 4F, Supplementary Fig. 5C), were particularly abundant in members of Pseudomonadota, Cyanobacteriota, and Bacteroidota. Additionally, Pseudomonadota and Acidobacteriota MAGs encoded the most diverse range of genes involved in these pathways, with Pseudomonadota encoding 17, 16, and 10 genes and Acidobacteriota encoding 13, 15, and 9 genes involved in the carbon, nitrogen, and sulfur cycles, respectively. The biofilm MAGs also encoded a range of genes involved in oxygen and hydrogen cycling (Supplementary Fig. 5D, E, Supplementary Fig. 6, Supplementary Data 4). Many of these metabolic functions are associated with the production and attenuation of greenhouse gases such as carbon





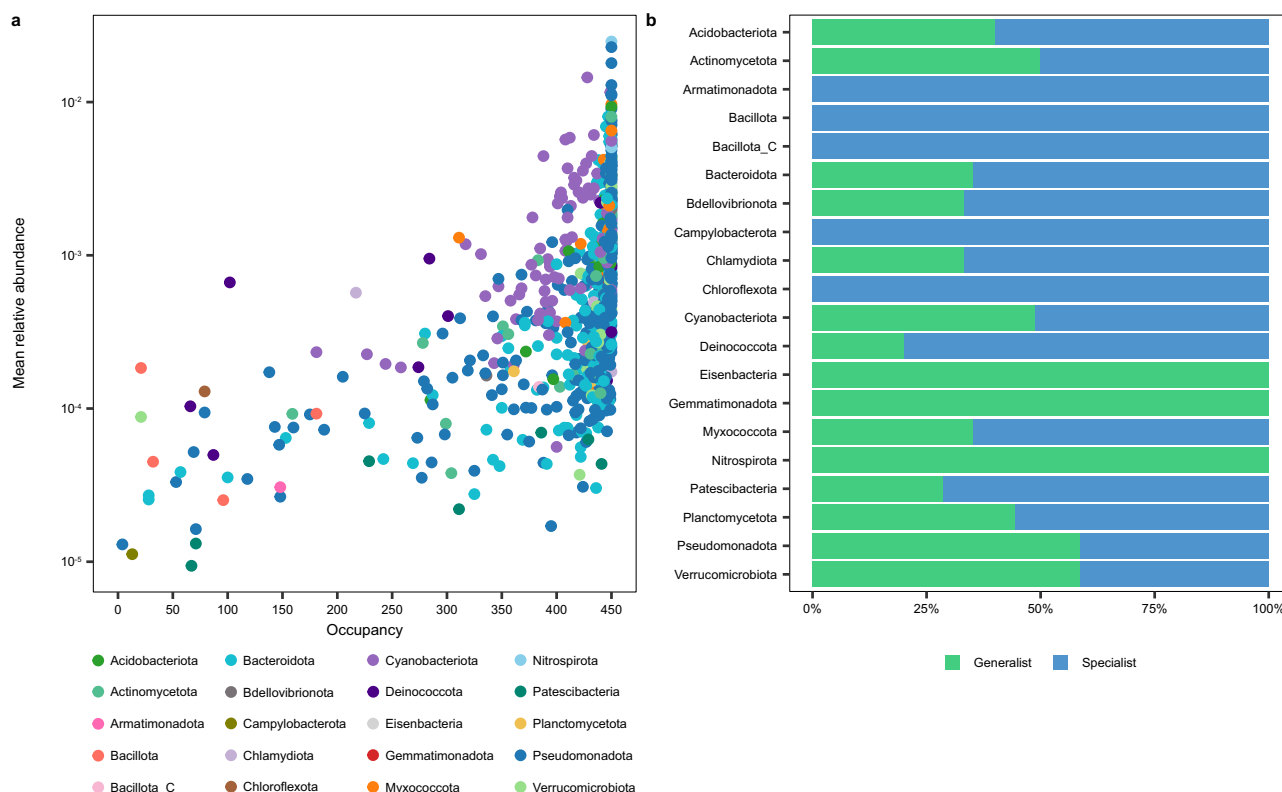
**Fig. 2 | Biogeographic distribution of bacterial MAGs in rivers across England.** Colours represent different bacterial phyla and points are scaled to relative abundance. Source data are provided as a Source Data file.

dioxide, methane, and nitrous oxide<sup>51</sup>. These results highlight the potential of biofilm bacteria to perform diverse biochemical roles, which may include contributing to the flux of greenhouse gases in river ecosystems.

### Functional potential of river biofilm communities

Biofilm bacteria exhibited an extensive range of functional capabilities, including resource acquisition, resource use, and stress tolerance (Fig. 5, Supplementary Fig. 7). Genes for the breakdown of compounds such as simple and complex carbohydrates, cellulose, and proteins were widely distributed across the community (present in more than 74.46% of MAGs) (Fig. 5A). This indicates that the biofilm community can utilise a broad range of organic compounds, highlighting their metabolic flexibility and resilience to fluctuating nutrient availability.

Up to 91.10% of biofilm MAGs contained genes for aerobic respiration, and 21.59% of MAGs, which included members of Pseudomonadota, Actinomycetota, Acidobacteriota, Deinococcota, and Planctomycetota, also possessed genes associated with anaerobic respiration. Additionally, some MAGs, notably those belonging to Pseudomonadota and Cyanobacteriota, had genes for both core and accessory photosynthetic pigments (Fig. 5B). The coexistence of the trophic capabilities identified among biofilm bacteria is consistent with studies showing that biofilms often exist as stratified micro-environments, in which aerobic and photosynthetic bacteria dominate the outer layers exposed to light and oxygen, while anaerobes occupy deeper oxygen-depleted layers<sup>3</sup>. Such trophic diversity may contribute to the versatility of biofilm bacteria, enabling them to occupy multiple ecological niches within the complex biofilm matrix.



**Fig. 3 | Ecological niche breadth of bacterial MAGs. a** Occupancy (number of samples each MAG is present in) and mean relative abundance (log scale) of MAGs where point colour represents bacterial phylum. **b** Proportion of MAGs of each

bacterial phylum classified as generalists (green) and specialists (blue) according to Levin's niche breadth index. Source data are provided as a Source Data file.

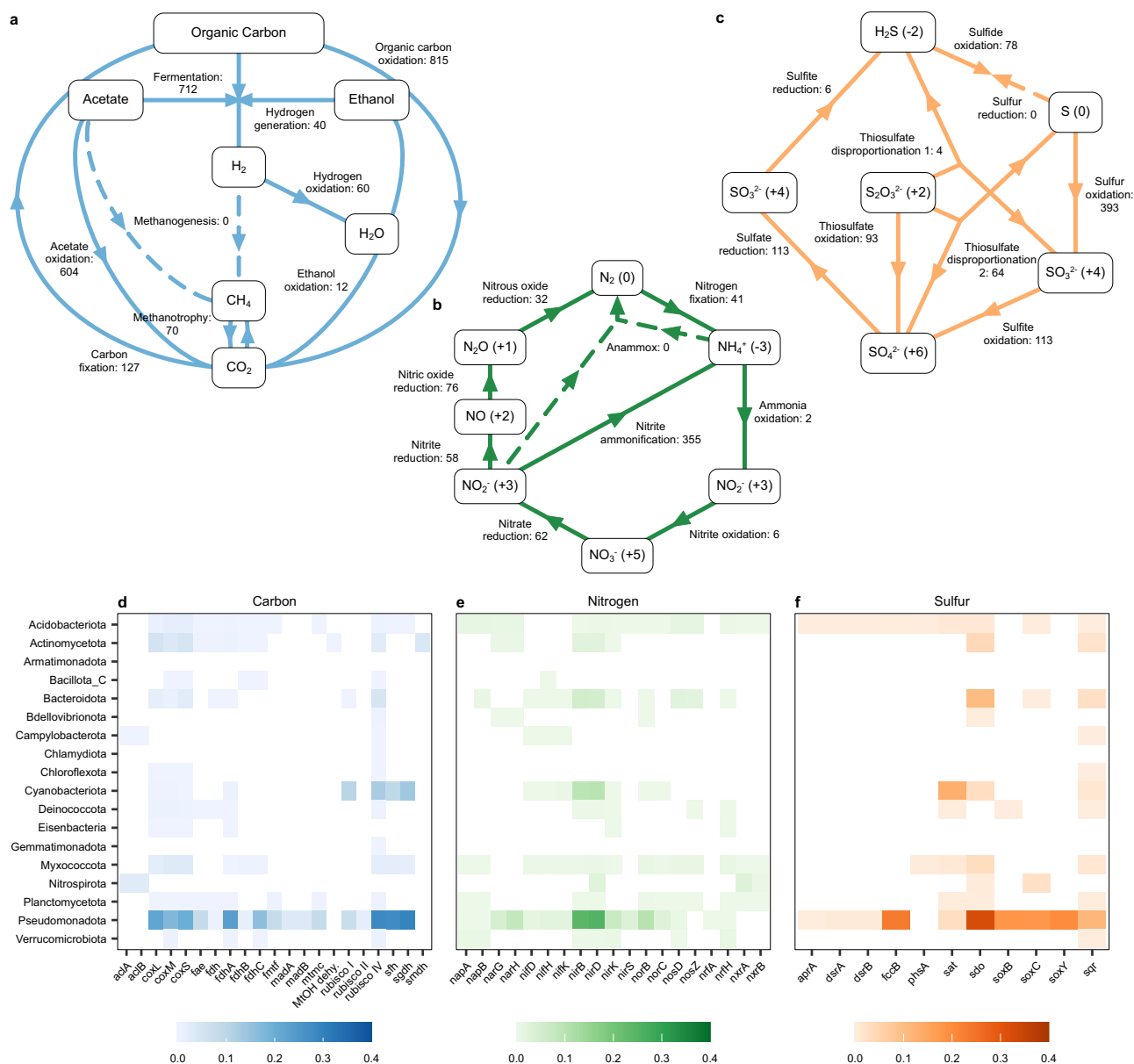
Biofilm formation is a key survival strategy for many freshwater microbes, allowing them to gain physical protection from the extracellular polymeric matrix, accumulate and readily access nutrients which may include biopolymers derived from the matrix itself, and interact with other taxa<sup>3</sup>. Further to the observed prevalence of biofilm-associated genes such as those involved in quorum sensing, chemotaxis, flagellar assembly, and EPS biosynthesis (Supplementary Fig. 2), genes associated with biofilm formation were also identified in 96.95% of MAGs encompassing nearly all bacterial phyla detected (Fig. 5C). Exceptions to this were Patescibacteria, which are known for their reduced genomes and symbiotic lifestyle<sup>52</sup> and likely depend on interactions with biofilm-forming bacteria rather than contributing to biofilm formation directly, and Bacillota, which may play a less prominent role in biofilm formation, possibly by relying on an alternative survival strategy such as sporulation<sup>53</sup>. These findings highlight the widespread importance of biofilm formation as a strategy for microbial survival in dynamic river ecosystems, facilitating interactions among taxa, promoting community stability, and acting as a nutrient resource that may buffer potential effects from nutrient fluctuations and other environmental disturbances<sup>3,54</sup>. Additionally, genes involved in a variety of other stress tolerance strategies were identified in the biofilm MAGs, such as compatible solutes accumulation, membrane stability, scavenging of reactive oxygen species, and repair and degradation of damaged proteins (Fig. 5C). Many of these stress tolerance strategies have also been identified in biofilm bacteria in glacier-fed streams<sup>32</sup>, highlighting that the widespread presence of stress tolerance genes may support the persistence of biofilm bacteria under fluctuating or unfavourable condition in freshwater environments.

Biofilm communities play an essential role in the uptake, transformation, and degradation of organic compounds, including those that are harmful pollutants in river ecosystems<sup>49</sup>. The vast majority of

MAGs (92.20%) contained genes involved in organophosphorus transport (Fig. 5A), which is a common pollutant in rivers derived from agricultural pesticides<sup>55</sup>. A total of 20.49% of MAGs, which included members of Pseudomonadota, Actinomycetota, Deinococcota, Myxococcota, Cyanobacteriota, and Verrucomicrobiota, had genes for the transport of aromatic acids. These compounds, while naturally present in ecosystems, can also include polycyclic aromatic hydrocarbons (PAHs) introduced into rivers from industrial activities<sup>56</sup>. Furthermore, 99.39% of MAGs, which encompassed all bacterial phyla detected, had genes involved in the transport of metals. These transportation genes are widespread in bacteria for acquiring essential metals, but they may also participate in the cycling of metals in rivers, which alongside natural sources, may enter rivers through industrial activities<sup>57</sup>. The widespread presence of these genes indicates that biofilm bacteria may be capable of transporting and therefore potentially transforming harmful organic pollutants that pose a significant risk to aquatic life in rivers. Biofilm bacteria may contribute to metal detoxification by immobilising metals and reducing their bioavailability in the water. This demonstrates the importance of biofilm bacteria in maintaining water quality<sup>4</sup> and their valuable application as bioindicators for monitoring ecosystem health<sup>58</sup>, particularly as rivers are under increasing threat from pollution and contaminants worldwide<sup>7</sup>.

### Environmental drivers of river biofilm communities

Our findings demonstrate the taxonomic diversity of biofilm bacteria and their extensive metabolic and functional roles which are essential for riverine ecosystem functioning. However, while previous studies have explored the environmental drivers of river benthic biofilm communities at the individual catchment scale<sup>10,13,59</sup>, large-scale patterns and drivers remain poorly understood<sup>27</sup>, particularly in comparison to surface water communities<sup>25</sup>. This knowledge gap limits our



**Fig. 4 | Nutrient cycling and metabolic potential of bacterial MAGs.** **a** Carbon, **b** nitrogen, and **c** sulfur cycling pathways, identified using METABOLIC, where numbers denote the number of MAGs positive for genes involved in each step and dashed lines show steps with no positive MAGs. Heatmaps of **d** carbon, **e** nitrogen,

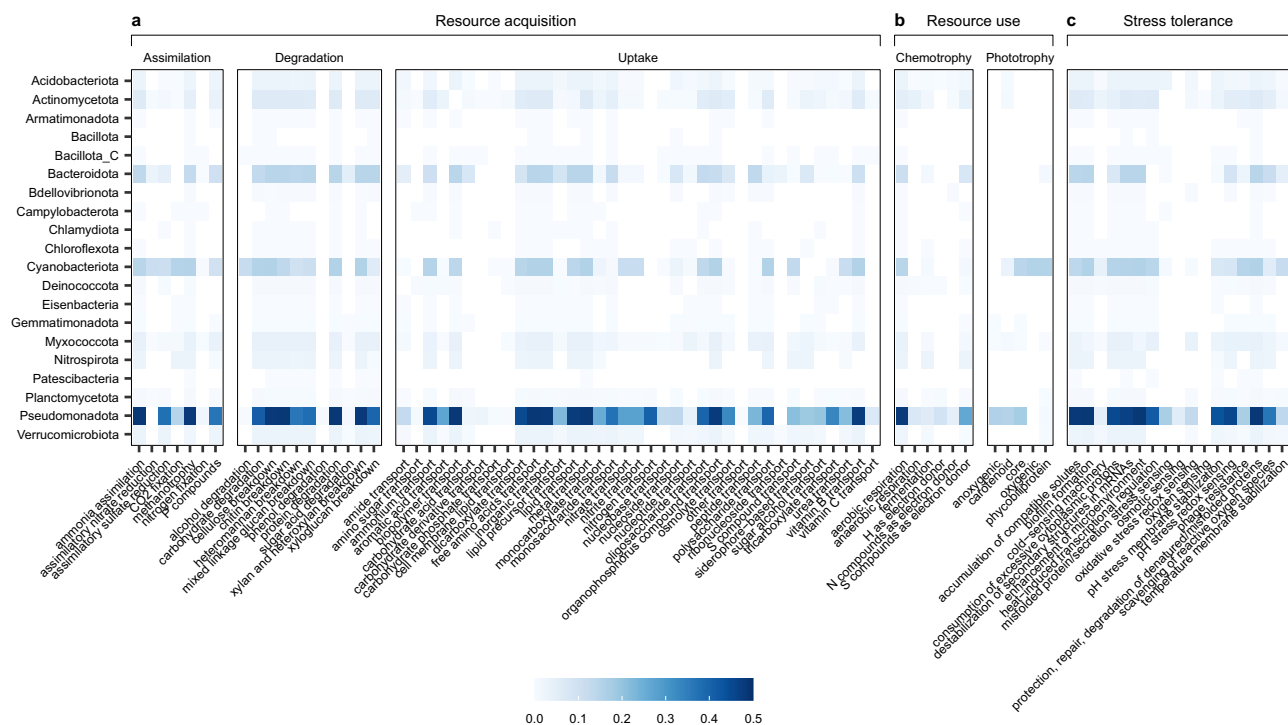
and **f** sulfur cycling genes, based on annotations made using metabolisHMM, where colour scales represent gene counts weighted by mean relative abundance and summed by bacterial phylum. Source data are provided as a Source Data file.

ability to understand and predict how biofilms and wider river eco-systems respond to environmental change across diverse landscapes. By incorporating detailed spatial information of the upstream catchment alongside high-resolution monitoring of water chemistry, this study represents a novel and comprehensive effort to assess the environmental drivers of river biofilm bacteria across large spatial scales.

Variance partitioning analysis was used to identify the environmental factors shaping river biofilm community composition across England. Environmental variables were categorised as catchment land cover, catchment geology, water chemistry and watershed characteristics. The total variance in the relative abundance of each MAG explained by the full set of measured environmental variables averaged 70.63% ( $\pm 19.5\%$ ). When grouped by phylum, the total variance explained by all environmental variables ranged from 46.55%

(Bacillota) to 97.36% (Gemmatimonadota) (Fig. 6A). Catchment land cover has previously been shown to significantly influence bacterial community composition in New Zealand stream biofilms<sup>16,27</sup>, and in North American river surface water<sup>25</sup>. Consistent with these findings, upstream catchment land cover was also identified as an important driver of river biofilm communities in England, accounting for up to 24.81% of observed variance at the phylum level, with a mean of 12.78% ( $\pm 7.3\%$ ), although variance explained by land cover was particularly low for some phyla such as Gemmatimonadota (0.66%). The most dominant land cover types included acid grassland, improved grassland, and arable and horticulture, although the proportion of variance explained by these factors varied among bacterial phyla (Fig. 6C).

Upstream catchment geology was found to account for the largest proportion of variance in biofilm bacterial community composition, explaining between 10.94 to 93.52% across bacterial phyla, with a mean



**Fig. 5 | Functional genomic traits of bacterial MAGs.** Traits associated with **a** resource acquisition (substrate assimilation, degradation, and uptake), **b** resource use (chemotrophy and phototrophy), and **c** stress tolerance, based on

annotations made using microTrait. The heatmap colour scale represents gene counts per category weighted by mean relative abundance and summed by bacterial phylum. Source data are provided as a Source Data file.

of 45.68% ( $\pm 18.7\%$ ). The proportion of the upstream catchment represented by calcareous and siliceous geologies were identified as the dominant geological drivers, explaining up to 37.46% (mean  $18.38\% \pm 8.4\%$ ) and 27.76% (mean  $13.55\% \pm 5.3\%$ ) of variation, respectively (Fig. 6D). To further evaluate the influence of these geological types, non-metric multidimensional scaling (NMDS) was performed, which revealed clear clustering of river biofilms along a gradient of calcareous to siliceous geology, suggesting that calcareous and siliceous geology structure bacterial communities in river biofilms differently (Supplementary Fig. 8, Supplementary Data 5). Furthermore, calcareous geology correlated positively with surface water pH ( $r = 0.36$ ,  $p < 0.001$ ), alkalinity ( $r = 0.33$ ,  $p < 0.001$ ), and conductivity ( $r = 0.29$ ,  $p < 0.001$ ), while siliceous geology correlated negatively with these variables ( $r = -0.32$ ,  $r = -0.53$ , and  $r = -0.31$ , respectively,  $p < 0.001$ ) (Supplementary Fig. 9, Supplementary Data 5). These findings are consistent with established relationships between geology and river water chemistry, particularly regarding  $\text{CaCO}_3$ -based buffering capacity (i.e., alkalinity), which is higher in catchments underlain by calcareous geology<sup>60,61</sup>. The influence of geology on biofilm communities therefore likely operates through its effects on pH, alkalinity, and conductivity<sup>47,62</sup>.

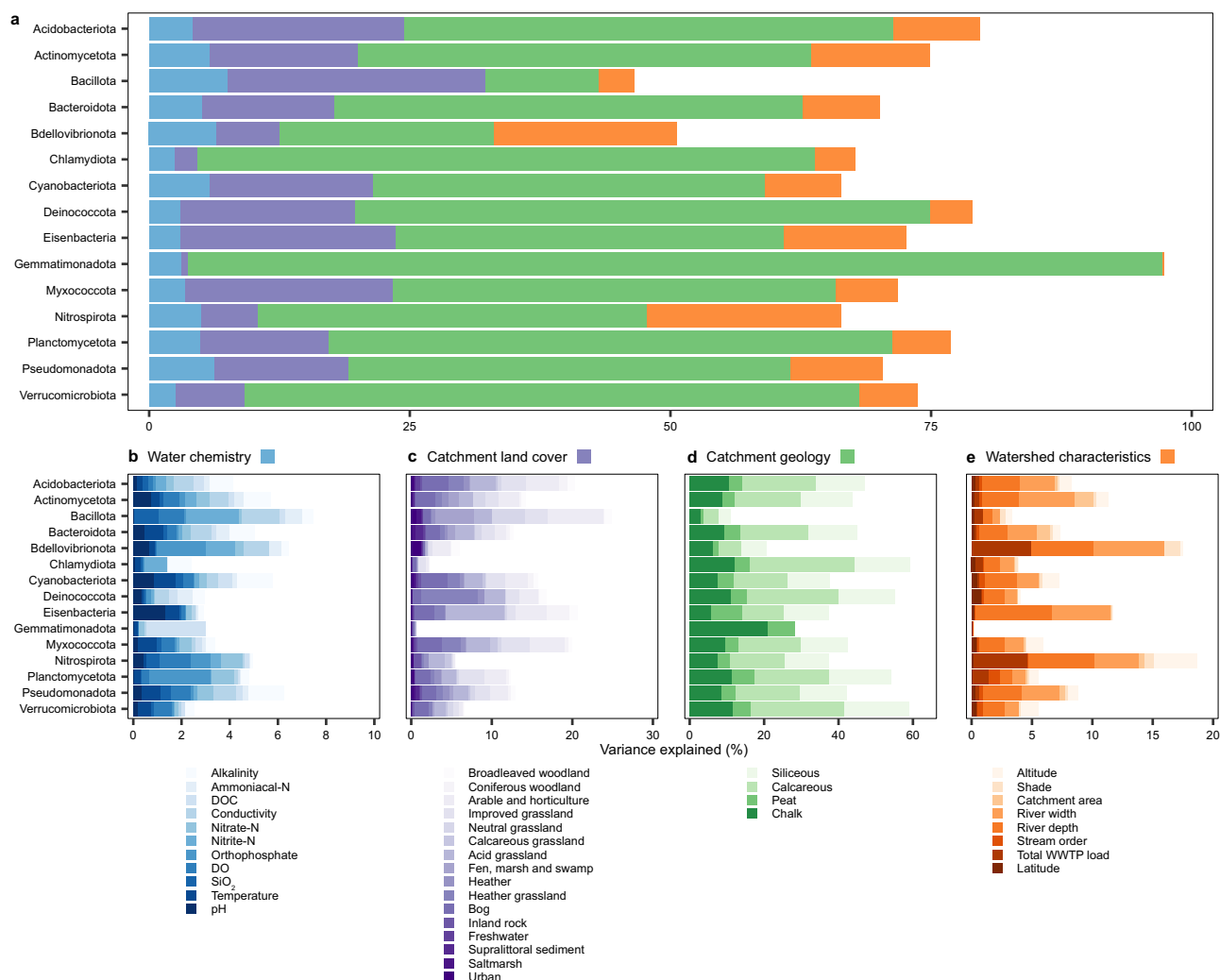
Despite these associations, water chemistry variables explained a relatively small proportion of variance in biofilm community composition (2.41 to 7.45%, with a mean of  $4.51\% \pm 1.6\%$  per phylum). However, individual variables such as alkalinity, conductivity, pH, dissolved oxygen, and water temperature were still influential, each accounting for up to 1.6% of the observed variation (Fig. 6B). While water chemistry was measured and averaged over a three-month period to capture temporal variability, these measurements may still underrepresent the influence of catchment geology, which acts as a persistent landscape-level driver influencing river systems and therefore biofilm development over much longer timescales. Additionally, the biofilm matrix may act as a buffer against physiochemical fluctuations in the water

column<sup>3</sup>, reducing the influence of water chemistry on biofilm communities.

In addition to geological and chemical drivers, orthophosphate, nitrate-N, and nitrite-N emerged as key nutrients shaping biofilm community composition, with orthophosphate most notably influencing the relative abundance of Planctomycetota (2.55%) and Bdellovibrionota (2.03%), nitrate-N influencing the relative abundance of Planctomycetota (0.95%) and Nitrospirota (0.90%), and nitrite-N influencing the relative abundance of Bacillota (2.21%) and Bdellovibrionota (1.24%). Wastewater treatment plant (WWTP) load, which correlated positively with orthophosphate ( $r = 0.66$ ,  $p < 0.001$ ), nitrate-N ( $r = 0.48$ ,  $p < 0.001$ ) and nitrite-N ( $r = 0.56$ ,  $p < 0.001$ ) (Supplementary Fig. 9), was also identified as a key driver of Bdellovibrionota, Nitrospirota, and Planctomycetota relative abundance, accounting for 4.89, 4.51, and 1.33% of variance, respectively (Fig. 6E). These results suggest that WWTP effluent, which introduces organic nutrients into the watercourse, may impact the dynamics of biofilm bacterial communities, with possible implications for microbial nutrient cycling and water quality<sup>14</sup>.

The River Continuum Concept (RCC)<sup>63</sup> describes how continuous physical gradients along a river's course, from headwaters to mouth, shape its biological and chemical properties. Low-order streams are shallow, narrow and strongly influenced by shading and allochthonous input from vegetation. As stream order, and therefore river width and depth, increase towards the mouth, autochthonous inputs play a progressively larger role<sup>63</sup>. Successional shifts in surface water bacterial communities along the river continuum have been well-documented within individual catchments<sup>64–66</sup>. This trend extends to the national-scale, where stream order was identified as the most significant environmental driver shaping bacterial communities in the surface water of rivers across North America<sup>25</sup>, ranging from stream order 1 to 12. In contrast, stream order within the RSN, which ranged from 1 to 8, was not a dominant driver of bacterial community





**Fig. 6 | Environmental drivers of river biofilm bacterial communities.** **a** Total variance explained by water chemistry, upstream catchment land cover, upstream catchment geology, and watershed characteristics, and variance explained by **b** water chemistry variables, **c** upstream catchment land cover types, **d** upstream

catchment geology types, and **e** watershed characteristics. Variance explained for each bacterial phylum is calculated as a mean of variance explained across individual MAGs. Source data are provided as a Source Data file.

composition in river biofilms across England, explaining less than 1% of variance for each bacterial phylum (Fig. 6e). Associated physical factors such as river depth and width explained up to 6.31 and 5.81%, respectively, but shading accounted for less than 1.32% of bacterial community composition. Bacterial communities within stream biofilms from a pre-alpine catchment in Austria (stream order 1–5) were also found to show patterns contradictory to the RCC<sup>67</sup>. These contrasting findings between river surface water and biofilm bacterial communities suggest that successional changes along the river continuum may affect these distinct microbial communities differently. While variation in surface water bacterial communities can be primarily explained by the RCC, local factors such as geology, land cover, nutrients and WWTP load appear to be more influential in shaping river biofilm communities. However, the smaller range of stream orders in the English RSN, compared with those investigated in North America, may partially contribute to the weaker relationship observed between stream order and biofilm community composition. Further studies of biofilm bacterial communities in larger rivers and comparisons between co-located water column and biofilm samples are required to fully understand the influence of the RCC on freshwater bacteria. Seasonal fluctuations could also influence river biofilm community composition, functional potential, and the relative importance of

environmental drivers, though seasonal effects have yet to be comprehensively explored at a large-scale. Furthermore, cross-domain biotic interactions may also be an important driver of biofilm bacterial community dynamics and require further study<sup>54,67</sup>.

## Conclusion

Rivers are valuable, biodiverse habitats and their microbial communities, particularly those within biofilms, are critical to ecosystem functioning, biogeochemical cycling, pollutant degradation, and maintaining water quality<sup>2,6</sup>. By leveraging comprehensive national-scale sampling coupled with high-resolution environmental data, this study provides novel insights into the biogeography, taxonomic diversity, metabolic potential, and functional roles of river biofilm microbial assemblages.

River biofilms are revealed to be dominated by high-occupancy generalist bacteria, most notably members of Pseudomonadota, Cyanobacteriota and Bacteroidota, reflecting the interconnectedness of river ecosystems that facilitate microbial dispersal across the river network and colonisation by generalist taxa. These results further highlight the significant metabolic and functional versatility of river biofilms, which harbour a diverse array of genes supporting carbon, nitrogen and sulfur cycling, the utilisation of a broad range of organic

compounds, the transportation of organic pollutants, and various trophic strategies. Their capacity to perform diverse biochemical roles reflects niche adaptation to the biofilm microenvironment and provides resilience in dynamic river ecosystems, allowing them to sustain key biogeochemical processes under fluctuating or unfavourable conditions.

The RCC is a strong driver of surface water bacterial communities<sup>25,64–66</sup>. However, this framework is less applicable to river biofilm communities. Instead, upstream catchment geology and land cover emerge as the primary determinants of biofilm communities, which likely shape community composition through their long-term effects on the physicochemical conditions of the water column and importantly the biofilm microenvironment. Furthermore, anthropogenic activities, including changes in land use and the release of WWTP effluents and pollutants into rivers, may significantly impact biofilm community dynamics. River biofilms are valuable bioindicators of ecosystem health<sup>58</sup>, and our findings provide baseline data that may support water quality monitoring and management, and, importantly, explain resilience to environmental change observed in river biofilm bacterial communities.

## Methods

### Sample collection

A total of 450 river biofilm samples were collected over a three-year period (2021–2023) from 146 sites within the EA's RSN programme. The sampling sites were selected using a randomised, spatially balanced design to ensure representative and unbiased spatial coverage across England<sup>68</sup>. A total of 174 samples were collected in 2021, 178 in 2022, and 98 in 2023, with 62 sites being sampled in all three years. Biofilms were collected from the river benthos by scraping biofilm-covered stones from the riverbed at each site with a sterile toothbrush and deionised water<sup>69</sup>. At sites where stones were not accessible, biofilms were collected by macrophyte scrape. The suspended biofilm was transferred to a 15 ml tube and preserved in the field in 5 ml nucleic acid preservation buffer (3.5 M ammonium sulphate, 17 mM sodium citrate and 13 mM EDTA)<sup>70</sup>. Samples were immediately transported to the EA's National Laboratory at Starcross, Exeter where they were concentrated by centrifugation and frozen. Samples were then transported to the UK Centre for Ecology & Hydrology (UKCEH), Wallingford on dry ice and stored at  $-20^{\circ}\text{C}$ .

### Water chemistry

Water chemistry variables, including water temperature ( $^{\circ}\text{C}$ ), pH, alkalinity to pH 4.5 as  $\text{CaCO}_3$  ( $\text{mg L}^{-1}$ ), conductivity, and the concentration of chloride ion ( $\text{mg L}^{-1}$ ), dissolved oxygen ( $\text{DO}$ ,  $\text{mg L}^{-1}$ ), dissolved oxygen saturated ( $\text{mg L}^{-1}$ ), dissolved organic carbon ( $\text{DOC}$ ,  $\text{mg L}^{-1}$ ), total phosphorus ( $\text{TP}$ ,  $\text{mg L}^{-1}$ ), orthophosphate ( $\text{mg L}^{-1}$ ), total nitrogen ( $\text{TN}$ ,  $\text{mg L}^{-1}$ ), total oxidised nitrogen ( $\text{TON}$ ,  $\text{mg L}^{-1}$ ), nitrate-nitrogen ( $\text{nitrate-N}$ ,  $\text{mg L}^{-1}$ ), nitrite-nitrogen ( $\text{nitrite-N}$ ,  $\text{mg L}^{-1}$ ), ammoniacal nitrogen ( $\text{ammoniacal-N}$ ,  $\text{mg L}^{-1}$ ), ammonia-N unionised ( $\text{mg L}^{-1}$ ), and reactive  $\text{SiO}_2$  ( $\text{mg L}^{-1}$ ), were measured using surface water samples collected from each sampling site. All water chemistry data is available at: <https://environment.data.gov.uk/water-quality/view/landing>. For each variable, a mean was calculated from up to 5 independent measurements taken on separate occasions at each site across a 3-month period up to and including the day of biofilm sampling and is provided in Supplementary Data 1.

### Upstream catchments

The upstream catchment of each RSN site was determined using 10 m flow direction and accumulation grids and the sampling points which were snapped to the maximum flow accumulation cell within 50 m of its location. The digital elevation model has a 10 m resolution and was based on NEXTMap which was derived from airborne Interferometric Synthetic Aperture Radar (IFSAR) and is available on the UKCEH

Catchment Management Modelling Platform (<https://catalogue.ceh.ac.uk/cmp/documents>). The area of the derived shapefile was then used as the area of the upstream catchment ( $\text{km}^2$ ). The upstream catchment shapefiles were intersected with the 10 m scale 2020 and 2021 UKCEH land cover maps available at: <https://www.ceh.ac.uk/data/ukceh-land-cover-maps>, and the British Geological Survey (BGS) geology map available at: <https://www.bgs.ac.uk/datasets/bgs-geology-250k>. The percentage of the upstream catchment area covered by each land cover type or geology was calculated for each sampling site.

### Watershed characteristics

Strahler numbers for the branches of the rivers where sampling sites are located were determined by spatially joining the GRTS points with the 2021 OS Open Rivers Network layer (<https://www.ordnancesurvey.co.uk/products/os-open-rivers>). To provide an estimate of river depth and width at sampling sites, sampling points were intersected with the gridded  $1\text{ km}^2$  physical river characteristics for the UK v2 dataset (<https://doi.org/10.5285/8df65124-68e9-4c68-8659-1c6b82c735e9>).

To determine the number of wastewater treatment plants (WWTPs) in the upstream catchments of each site and the total population equivalent loads, a spatial intersect was performed between the upstream catchments and a dataset of WWTPs in England and Wales with their population equivalent loads mapped to discharge location (<https://www.data.gov.uk/dataset/Of76a1c3-1368-476b-a4df-7ef32bfd9a8b/urban-waste-water-treatment-directive-treatment-plants>). Count and sum functions were used to provide the number of WWTPs and their total population equivalent loads respectively.

Average shading from landforms and surface model objects was determined for each sampling by spatially joining the sampling points with a 25 m buffer zone to the Environment Agency Keeping Rivers Cool relative shading map available at: <https://data.catchmentbasedapproach.org/maps/therivertrust::riparian-shade-england>.

### DNA extraction

Each biofilm sample was defrosted and briefly vortexed to homogenise the sample. The Quick-DNA fecal/soil microbe kit (Zymo Research, California, U.S.) was used to extract DNA from  $100\text{ }\mu\text{L}$  of sample with the following amendments to the manufacturer's protocol to maximise DNA yield: Zymo DNA/RNA shield was used as the lysis buffer, samples were lysed at 20 Hz for 20 min using the TissueLyser II (Qia-gen, Germany), and  $20\text{ }\mu\text{L}$  of proteinase K was added to the lysate prior to incubation at  $65^{\circ}\text{C}$  for 20 min<sup>71</sup>. Purified DNA was eluted in  $100\text{ }\mu\text{L}$  of elution buffer. The purity of extracted DNA was checked using the NanoDrop 8000 spectrophotometer (Thermo Fisher Scientific, MA, U.S.). The concentration of DNA was measured using the QuantiFluor ONE dsDNA kit (Promega, Madison, WI, U.S.) with a BioTek Cytation 5 imaging reader (Agilent Technologies, California, U.S.) according to the manufacturer's protocol. The DNA was stored frozen at  $-70^{\circ}\text{C}$  for long-term archiving at UKCEH, Wallingford.

### Metagenomic sequencing

Extracted DNA was sent to Novogene UK for library preparation and  $2 \times 150\text{ bp}$  shotgun metagenomic sequencing on an Illumina NovaSeq 6000 with an S4 flow cell to achieve a sequencing depth of at least 10 Gb raw data per sample. Between 2.85 million and 188.26 million raw reads were generated per sample, with a median of 72.16 million raw reads per sample.

### Metagenomic data processing

Illumina adaptor sequences were trimmed, and metagenomic reads were filtered to a minimum quality score of 25 using Trim Galore v0.6.5 (<https://github.com/FelixKrueger/TrimGalore>). Reads mapping to the human reference genome (GRCh38; downloaded on 2023-11-16 from

[https://ftp.ncbi.nlm.nih.gov/refseq/H\\_sapiens/annotation/GRCh38\\_latest/refseq\\_identifiers/GRCh38\\_latest\\_genomic.fna.gz](https://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/GRCh38_latest/refseq_identifiers/GRCh38_latest_genomic.fna.gz) were removed. After quality filtering and trimming, between 4.28 million and 90.17 million reads were obtained per sample, with a median of 36.37 million reads per sample. To determine the overall profile and percentage of archaea, bacteria, and eukaryotes in the dataset, singleM v0.16.0<sup>72</sup> was run on the pre-processed reads. MultiQC v1.17<sup>73</sup> was used to process all filtered reads to obtain basic statistics per sample. Megahit v1.2.9<sup>74</sup> was used to assemble the reads into contigs, which were subsequently used for binning, functional annotation, and downstream analyses. For functional analysis, open-reading frames (ORFs) were predicted using Prodigal v2.6.3<sup>75</sup>, which were fed into EggNOG-mapper v2.1.9<sup>76</sup> for functional annotations via the built-in v6.0 database. The ORFs were used in featureCounts<sup>77</sup> to generate per-gene coverage, while Kraken2 v2.1.2<sup>78</sup> and Bracken v2.6.0<sup>79</sup> were used to taxonomically annotate the contigs, using the PlusPFP (<https://benlangmead.github.io/aws-indexes/k2>) database, which includes RefSeq protozoa, fungi, plant, archaea, bacteria, virus, plasmid, and human sequences. For stringency, a 0.7 confidence threshold was used in Kraken2. The contigs were also used to obtain metagenome-assembled genomes (MAGs), by dereplicating bins obtained via MetaBAT v2.15<sup>80</sup>, MetaBinner v1.4.3<sup>81</sup> and CONCOCT v1.1.0<sup>82</sup> using dRep v3.2.2<sup>83</sup>. The completion and contamination of the dereplicated bins were estimated with CheckM2 v1.2.2<sup>84</sup>, alongside taxonomic annotation using the comprehensive and phylogenetically consistent prokaryotic genome taxonomy database, GTDB-Tk v2.3.2<sup>85</sup>. Species were delineated using standard GTDB-Tk thresholds of  $\geq 95\%$  average nucleotide identity (ANI) across  $\geq 65\%$  of the genome alignment fraction.

MicroTrait v1.0.0<sup>86</sup> was used to extract fitness traits from the MAGs, based on protein family sequence similarities and Kyoto Encyclopedia of Genes and Genomes (KEGG) orthologs (KOs). The metabolic and biogeochemical traits and pathways of the MAGs were also analysed using METABOLIC v4.0, which integrates numerous annotation databases, including KEGG and curated hidden Markov model (HMM) databases<sup>87</sup>. The metabolic pathways were further assessed via metabolisHMM v2.22, which uses curated HMMs for functional annotations of metabolic characteristics spanning carbon, nitrogen, sulfur, and hydrogen cycling pathways<sup>88</sup>. MicrobeAnnotator v2.0.5 was used to annotate the MAGs using the -light option against the Kofam and Swissprot databases<sup>89</sup>. Biofilm associated KOs were then identified according to the BBSdb of bacterial biofilm-associated proteins<sup>90</sup> supplemented with additional KOs found in KEGG pathways such as quorum sensing (map02024), flagellar assembly (map02040), chemotaxis (map02030), and biofilm formation (map02025, map02026, map05111). The workflow was implemented using the Snakemake workflow management system v7.8.2<sup>91</sup>, and is available at: [https://github.com/amycthorpe/metag\\_analysis\\_EA](https://github.com/amycthorpe/metag_analysis_EA) for preprocessing the metagenomic reads to MAG assembly. The analysis workflow, including dereplication, taxonomy assignment and the implementation of microTrait, METABOLIC, and metabolisHMM, is available at: [https://github.com/amycthorpe/EA\\_metag\\_post\\_analysis](https://github.com/amycthorpe/EA_metag_post_analysis).

## Data analysis

Using contamination and completeness statistics estimated with CheckM2, the MAGs were categorised as near-complete ( $>90\%$  complete and  $<5\%$  contamination) or medium-quality ( $>70\%$  complete and  $<10\%$  contamination)<sup>92</sup>, with remaining MAGs that did not meet the thresholds categorised as low-quality. Medium and high-quality MAGs were retained for downstream analysis. A phylogenetic tree was constructed using the maximum likelihood method with the ggtree R package v3.12.0<sup>93</sup> using the multiple sequence alignment from the genome taxonomy database toolkit analysis (GTDB-Tk). For biogeographic mapping, relative abundance was determined for each MAG, summed at the phylum level and plotted according to site latitude and longitude using rnatuarearth R package v1.10.9000<sup>94</sup>.

A Bray–Curtis beta diversity dissimilarity matrix was generated based on MAG relative abundance using the vegan R package v2.6.10<sup>95</sup>. The relationship between the dissimilarity matrix and geographic distance between samples was investigated with a linear regression. Non-metric multidimensional scaling was performed based on the dissimilarity matrix. Environmental variables were fitted to the ordination space using the vegan envfit() function with 999 permutations. To investigate sample clustering within the ordination space by dominant geological drivers, the log ratio of the proportion of the upstream catchment represented by calcareous to siliceous geologies was calculated.

For occupancy and relative abundance analyses, mean relative abundance was calculated for each MAG across all samples, and the number of samples each MAG was present in was used as a measure of occupancy. Levins' niche breadth ( $B_N$ ) was calculated for each MAG using the MicroNiche R package v1.0.0<sup>96</sup>. This metric quantifies the proportional similarity of resource use, where a  $B_N$  closer to one is indicative of a more uniform distribution and therefore a broad niche and generalist lifestyle, while a  $B_N$  closer to zero indicates a narrower niche with a preference for specific conditions<sup>96</sup>. MAGs with a  $B_N$  greater than the median  $B_N$  (0.03) were categorised as generalists, and those with a  $B_N$  less than the median were categorised as specialists. All MAGs were above the recommended limit of quantification threshold of 1.65. The percentage of generalists and specialist MAGs was calculated for each bacterial phylum. The presence of metabolic and functional traits identified for each MAG using METABOLIC, metabolisHMM, and microTrait were weighted by mean relative abundance and summed at the phylum level and genus level. Biofilm-associated KOs identified with MicrobeAnnotator were normalised by coding density per MAG.

Variance partitioning was used to identify the environmental drivers of bacterial community composition at the MAG level. A linear mixed model using maximum likelihood estimation was fitted to the matrix of MAG relative abundance to estimate the contribution of each environmental variable to community variation using the variancePartition R package v1.34.0<sup>97</sup>. Due to low variance for some MAGs, variance partitioning was performed based on a subset of 340 MAGs. Samples ( $n = 450$ ) were filtered to complete observations according to environmental data availability ( $n = 401$ ). Pearson correlations (two-sided) were computed using the cor() function in R to investigate correlation among the environmental variables and with bacterial phyla on the subset of 401 samples with complete observations. To optimise variance partitioning model performance, some measured water chemistry parameters, including total phosphorus, were not included in the analysis due to significant co-correlation ( $r > 0.80$ ,  $p < 0.001$ ) with other, more biologically available measures (e.g., orthophosphate, Supplementary Data 5). Variance explained per phylum was calculated as a mean of variance explained per MAG. Data analysis was performed in R v4.4.0<sup>98</sup>, and R scripts for data analysis and visualisation are available at: [https://github.com/amycthorpe/biofilm\\_MAG\\_analysis](https://github.com/amycthorpe/biofilm_MAG_analysis).

## Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

The metagenomic data generated in this study have been deposited in the European Nucleotide Archive (ENA) at EMBL-EBI under accession number [PRJEB85861](https://doi.org/10.1038/s41467-025-65620-3). Sample accession codes and all the data generated, including the environmental metadata associated with each sample, MAG coverage, taxonomy, and CheckM2 statistics, outputs from METABOLIC, metabolisHMM, and microTrait, and the niche breadth index, variance partitioning, and correlation analysis results are available on Zenodo at: <https://doi.org/10.5281/zenodo.14762144>. Source data are provided with this paper.

## Code availability

Snakemake workflows used to process and analyse the metagenomic data are available on GitHub ([https://github.com/amycthorpe/metag\\_analysis\\_EA](https://github.com/amycthorpe/metag_analysis_EA), [https://github.com/amycthorpe/EA\\_metag\\_post\\_analysis](https://github.com/amycthorpe/EA_metag_post_analysis)) and archived on Zenodo<sup>99,100</sup>. R scripts for data analysis and visualisation are available on GitHub ([https://github.com/amycthorpe/biofilm\\_MAG\\_analysis](https://github.com/amycthorpe/biofilm_MAG_analysis)) and archived on Zenodo<sup>101</sup>.

## References

- Flemming, H. C. & Wuertz, S. Bacteria and archaea on Earth and their abundance in biofilms. *Nat. Rev. Microbiol.* **17**, 247–260 (2019).
- Besemer, K. Biodiversity, community structure and function of biofilms in stream ecosystems. *Res. Microbiol.* **166**, 774–781 (2015).
- Penesyan, A., Paulsen, I. T., Kjelleberg, S. & Gillings, M. R. Three faces of biofilms: a microbial lifestyle, a nascent multicellular organism, and an incubator for diversity. *npj Biofilms Microbiomes* **7**, 80 (2021).
- Sentenac, H., Loyau, A., Leflaive, J. & Schmeller, D. S. The significance of biofilms to human, animal, plant and ecosystem health. *Funct. Ecol.* **36**, 294–313 (2022).
- Weitere, M. et al. The food web perspective on aquatic biofilms. *Ecol. Monogr.* **88**, 543–559 (2018).
- Battin, T. J., Besemer, K., Bengtsson, M. M., Romani, A. M. & Packmann, A. I. The ecology and biogeochemistry of stream biofilms. *Nat. Rev. Microbiol.* **14**, 251–263 (2016).
- Best, J. Anthropogenic stresses on the world's big rivers. *Nat. Geosci.* **12**, 7–21 (2019).
- Sagova-Mareckova, M. et al. Expanding ecological assessment by integrating microorganisms into routine freshwater biomonitoring. *Water Res.* **191**, 116767 (2021).
- Engloner, A. I., Vargha, M., Kós, P. & Borsodi, A. K. Planktonic and epilithic prokaryota community compositions in a large temperate river reflect climate change related seasonal shifts. *PLOS ONE* **18**, e0292057 (2023).
- Gautam, A., Lear, G. & Lewis, G. D. Time after time: detecting annual patterns in stream bacterial biofilm communities. *Environ. Microbiol.* **24**, 2502–2515 (2022).
- Pineda-Mora, D. et al. Diversity and functions of epilithic riverine biofilms. *Water Air Soil Pollut.* **231**, 1–19 (2020).
- Kobayashi, Y. et al. Longitudinal changes in bacterial community composition in river epilithic biofilms: influence of nutrients and organic matter. *Aquat. Microb. Ecol.* **54**, 135–152 (2009).
- Ponsati, L. et al. Multiple-stressor effects on river biofilms under different hydrological conditions. *Freshw. Biol.* **61**, 2102–2115 (2016).
- Tamminen, M. et al. Wastewater constituents impact biofilm microbial community in receiving streams. *Sci. Total Environ.* **807**, 151080 (2022).
- Besemer, K., Singer, G., Hödl, I. & Battin, T. J. Bacterial community composition of stream biofilms in spatially variable-flow environments. *Appl. Environ. Microbiol.* **75**, 7189–7195 (2009).
- Lear, G. et al. The biogeography of stream bacteria. *Glob. Ecol. Biogeogr.* **22**, 544–554 (2013).
- Qu, X. et al. Influences of anthropogenic land use on microbial community structure and functional potentials of stream benthic biofilms. *Sci. Rep.* **7**, 15117 (2017).
- Wang, J. et al. Phylogenetic beta diversity in bacterial assemblages across ecosystems: deterministic versus stochastic processes. *ISME J.* **7**, 1310–1321 (2013).
- Guo, K., Wu, N., Li, W., Baattrup-Pedersen, A. & Riis, T. Microbial biofilm community dynamics in five lowland streams. *Sci. Total Environ.* **798**, 149169 (2021).
- Lyautey, E., Jackson, C. R., Cayrou, J., Rols, J. L. & Garabétian, F. Bacterial community succession in natural river biofilm assemblages. *Microb. Ecol.* **50**, 589–601 (2005).
- Henderson, A., Del Panta, A., Schubert, O. T., Mitri, S. & van Vliet, S. Disentangling the feedback loops driving spatial patterning in microbial communities. *npj Biofilms Microbiomes* **11**, 32 (2025).
- Kamjunke, N., Mages, M., Büttner, O., Marcus, H. & Weitere, M. Relationship between the elemental composition of stream biofilms and water chemistry - a catchment approach. *Environ. Monit. Assess.* **187**, 1–13 (2015).
- Malazarte, J. et al. Bacterial communities in a subarctic stream network: Spatial and seasonal patterns of benthic biofilm and bacterioplankton. *Mol. Ecol.* **31**, 6649–6663 (2022).
- Zancarini, A. et al. Deciphering biodiversity and interactions between bacteria and microeukaryotes within epilithic biofilms from the Loue River, France. *Sci. Rep.* **7**, 4344 (2017).
- Borton, M. A. et al. A functional microbiome catalogue crowd-sourced from North American rivers. *Nature* **637**, 103–112 (2024).
- Garner, R. E. et al. A genome catalogue of lake bacterial diversity and its drivers at continental scale. *Nat. Microbiol.* **8**, 1920–1934 (2023).
- Hermans, S. et al. Exploring freshwater stream bacterial communities as indicators of land use intensity. *Environ. Microbiome* **19**, 45 (2024).
- Singleton, C. M. et al. Microflora Danica: the atlas of Danish environmental microbiomes. *bioRxiv*. **06**, 2024 (2024).
- Sunagawa, S. et al. Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).
- Gweon, H. S. et al. Contrasting community assembly processes structure lotic bacteria metacommunities along the river continuum. *Environ. Microbiol.* **23**, 484–498 (2021).
- Lin, Q., Sekar, R., Marrs, R. & Zhang, Y. Effect of river ecological restoration on biofilm microbial community composition. *Water* **11**, 1244 (2019).
- Busi, S. B. et al. Genomic and metabolic adaptations of biofilms to ecological windows of opportunity in glacier-fed streams. *Nat. Commun.* **13**, 2168 (2022).
- Newton, R. J., Jones, S. E., Eiler, A., McMahon, K. D. & Bertilsson, S. A guide to the natural history of freshwater lake bacteria. *Microbiol. Mol. Biol. Rev.* **75**, 14–49 (2011).
- Cabello-Yeves, P. J. et al. Microbiome of the deep Lake Baikal, a unique oxic bathypelagic habitat. *Limnol. Oceanogr.* **65**, 1471–1488 (2020).
- Chiriack, M. C., Haber, M. & Salcher, M. M. Adaptive genetic traits in pelagic freshwater microbes. *Environ. Microbiol.* **25**, 606–641 (2023).
- Shah, M. et al. Genome-resolved metagenomics reveals the effect of nutrient availability on bacterial genomic properties across 44 European freshwater lakes. *Environ. Microbiol.* **26**, e16634 (2024).
- Serra Moncadas, L., Hofer, C., Bulzu, P. A., Pernthaler, J. & Andrei, A. S. Freshwater genome-reduced bacteria exhibit pervasive episodes of adaptive stasis. *Nat. Commun.* **15**, 3421 (2024).
- Giovannoni, S. J., Cameron Thrash, J. & Temperton, B. Implications of streamlining theory for microbial ecology. *ISME J.* **8**, 1553–1565 (2014).
- Rodríguez-Gijón, A. et al. A genomic perspective across Earth's microbiomes reveals that genome size in Archaea and Bacteria is linked to ecosystem type and trophic strategy. *Front. Microbiol.* **12**, 761869 (2022).
- Ramonedá, J., Hoffert, M., Stallard-Olivera, E., Casamayor, E. O. & Fierer, N. Leveraging genomic information to predict environmental preferences of bacteria. *ISME J.* **18**, 195 (2024).



41. Rodríguez-Gijón, A. et al. The ecological success of freshwater microorganisms is mediated by streamlining and biotic interactions. *bioRxiv*, <https://doi.org/10.1101/2025.03.24.644981> (2025).
42. Bourquin, M. et al. The microbiome of cryospheric ecosystems. *Nat. Commun.* **13**, 3087 (2022).
43. von Meijenfildt, F. B., Hogeweg, P. & Dutilh, B. E. A social niche breadth score reveals niche range strategies of generalists and specialists. *Nat. Ecol. Evol.* **7**, 768–781 (2023).
44. Pearman, J. K. et al. The role of environmental processes and geographic distance in regulating local and regionally abundant and rare bacterioplankton in lakes. *Front. Microbiol.* **12**, 793441 (2022).
45. Liu, X. et al. Microbiome analysis in Asia's largest watershed reveals inconsistent biogeographic pattern and microbial assembly mechanisms in river and lake systems. *iScience* **27**, 110053 (2024).
46. Anderson-Glenne, M. J., Bakkestuen, V. & Clipson, N. J. Spatial and temporal variability in epilithic biofilm bacterial communities along an upland river gradient. *FEMS Microbiol. Ecol.* **64**, 407–418 (2008).
47. Wilhelm, L., Singer, G. A., Fasching, C., Battin, T. J. & Besemer, K. Microbial biodiversity in glacier-fed streams. *ISME J.* **7**, 1651–1660 (2013).
48. Baas-Becking, L. G. M. *Geobiologie of inleiding tot de milieukunde* (The Hague: W. P. Van Stockum & Zoon, 1934).
49. Romani, A. M. et al. Biofilm structure and function and possible implications for riverine DOC dynamics. *Microb. Ecol.* **47**, 316–328 (2004).
50. Michoud, G. et al. Unexpected functional diversity of stream biofilms within and across proglacial floodplains despite close spatial proximity. *Limnol. Oceanogr.* **68**, 2183–2194 (2023).
51. Stein, L. Y. The long-term relationship between microbial metabolism and greenhouse gases. *Trends Microbiol.* **28**, 500–511 (2020).
52. Hu, H. et al. Global abundance patterns, diversity, and ecology of Patescibacteria in wastewater treatment plants. *Microbiome*. **12**, 55 (2024).
53. Galperin, M. Y., Yutin, N., Wolf, Y. I., Vera Alvarez, R. & Koonin, E. V. Conservation and evolution of the sporulation gene set in diverse members of the Firmicutes. *J. Bacteriol.* **204**, e00079–22 (2022).
54. Busi, S. B. et al. Cross-domain interactions confer stability to benthic biofilms in proglacial streams. *Front. Microbiomes*. **2**, 1280809 (2024).
55. Ore, O. T., Adeola, A. O., Bayode, A. A., Adedipe, D. T. & Nomngongo, P. N. Organophosphate pesticide residues in environmental and biological matrices: occurrence, distribution and potential remedial approaches. *Environ. Chem. Ecotoxicol.* **5**, 9–23 (2023).
56. Mojiri, A., Zhou, J. L., Ohashi, A., Ozaki, N. & Kindaichi, T. Comprehensive review of polycyclic aromatic hydrocarbons in water sources, their effects and treatments. *Sci. Total Environ.* **696**, 133971 (2019).
57. Zhou, Q. et al. Total concentrations and sources of heavy metal pollution in global river and lake water bodies from 1972 to 2017. *Glob. Ecol. Conserv.* **22**, e00925 (2020).
58. Pu, Y., Ngan, W. Y., Yao, Y. & Habimana, O. Could benthic biofilm analyses be used as a reliable proxy for freshwater environmental health?. *Environ. Pollut.* **252**, 440–449 (2019).
59. Zhou, Y. et al. Climate and environmental variables drive stream biofilm bacterial and fungal diversity on tropical mountainsides. *Microb. Ecol.* **87**, 28 (2024).
60. Hamid, A., Bhat, S. U. & Jehangir, A. Local determinants influencing stream water quality. *Appl. Water Sci.* **10**, 1–16 (2020).
61. Rothwell, J. J. et al. A spatial and seasonal assessment of river water chemistry across North West England. *Sci. Total Environ.* **408**, 841–855 (2010).
62. Fierer, N., Morse, J. L., Berthrong, S. T., Bernhardt, E. S. & Jackson, R. B. Environmental controls on the landscape-scale biogeography of stream bacterial communities. *Ecology* **88**, 2162–2173 (2007).
63. Vannote, R. L., Minshall, G. W., Cummins, K. W., Sedell, J. R. & Cushing, C. E. The river continuum concept. *Can. J. Fish. Aquat. Sci.* **37**, 130–137 (1980).
64. Henson, M. W. et al. Nutrient dynamics and stream order influence microbial community patterns along a 2914 kilometer transect of the Mississippi River. *Limnol. Oceanogr.* **63**, 1837–1855 (2018).
65. Read, D. S. et al. Catchment-scale biogeography of riverine bacterioplankton. *ISME J.* **9**, 516–526 (2015).
66. Savio, D. et al. Bacterial diversity along a 2600 km river continuum. *Environ. Microbiol.* **17**, 4994–5007 (2015).
67. Besemer, K. et al. Headwaters are critical reservoirs of microbial diversity for fluvial networks. *Proc. R. Soc. B Biol. Sci.* **280**, 20131760 (2013).
68. Brown, J. A., Robertson, B. L. & McDonald, T. Spatially balanced sampling: application to environmental surveys. *Procedia Environ. Sci.* **27**, 6–9 (2015).
69. Kelly, M. G. et al. Development of a novel metric for evaluating diatom assemblages in rivers using DNA metabarcoding. *Ecol. Indic.* **118**, 106725 (2020).
70. Warren, J. et al. Influence of storage time on the stability of diatom assemblages using DNA from riverine biofilm samples. *Metabarcoding Metagenomics* **8**, 129227 (2024).
71. Newbold, L. K. et al. DNA extraction methodology has a limited impact on multitaxa riverine benthic metabarcoding community profiles. *Environ. DNA* **7**, 70102 (2025).
72. Woodcroft, B. J. et al. Comprehensive taxonomic identification of microbial species in metagenomic data using SingleM and Sandpiper. *Nat. Biotechnol.* <https://doi.org/10.1038/s41587-025-02738-1> (2025).
73. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
74. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
75. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinforma.* **11**, 1–11 (2010).
76. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
77. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
78. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 1–13 (2019).
79. Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: estimating species abundance in metagenomics data. *PeerJ Comput. Sci.* **3**, e104 (2017).
80. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*. **3**, e1165 (2015).
81. Wang, Z., Huang, P., You, R., Sun, F. & Zhu, S. MetaBinner: a high-performance and stand-alone ensemble binning method to recover individual genomes from complex microbial communities. *Genome Biol.* **24**, 1 (2023).

82. Alneberg, J. et al. Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144–1146 (2014).
83. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).
84. Chklovskii, A., Parks, D. H., Woodcroft, B. J. & Tyson, G. W. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat. Methods* **20**, 1203–1212 (2023).
85. Chaumeil, P. A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* **38**, 5315–5316 (2022).
86. Karaoz, U. & Brodie, E. L. microTrait: A toolset for a trait-based representation of microbial genomes. *Front. Bioinforma.* **2**, 918853 (2022).
87. Zhou, Z. et al. METABOLIC: high-throughput profiling of microbial genomes for functional traits, metabolism, biogeochemistry, and community-scale functional networks. *Microbiome* **10**, 33 (2022).
88. McDaniel, E. A., Anantharaman, K. & McMahon, K. D. metabo-lishMM: Phylogenomic analysis for exploration of microbial phylogenies and metabolic pathways. *bioRxiv* <https://doi.org/10.1101/2019.12.20.884627> (2019).
89. Ruiz-Perez, C. A., Conrad, R. E. & Constantinidis, K. T. Microbe-Annotator: a user-friendly, comprehensive functional annotation pipeline for microbial genomes. *BMC Bioinforma.* **22**, 1–16 (2021).
90. Zhang, Z., Pan, Y., Hussain, W., Chen, G. & Li, E. BBSdb, an open resource for bacterial biofilm-associated proteins. *Front. Cell. Infect. Microbiol.* **14**, 1428784 (2024).
91. Mölder, F. et al. Sustainable data analysis with Snakemake. *F1000 Res* **10**, 33 (2021).
92. Bowers, R. M. et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).
93. Yu, G., Smith, D., Zhu, H., Guan, Y. & Lam, T. T. ggTree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).
94. Massicotte, P. & South, A. rnatuarearth: World map data from natural earth. <https://cran.r-project.org/web/packages/rnatuarearth/index.html> (2025).
95. Oksanen, J. et al. vegan: Community ecology packages. <https://cran.r-project.org/package=vegan> (2019).
96. Finn, D. R. et al. MicroNiche: an R package for assessing microbial niche breadth and overlap from amplicon sequencing data. *FEMS Microbiol. Ecol.* **96**, fiae131 (2020).
97. Hoffman, G. E. & Schadt, E. E. variancePartition: interpreting drivers of variation in complex gene expression studies. *BMC Bioinformatics*, **17**. <https://doi.org/10.1186/s12859-016-1323-z> (2016).
98. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from: <https://www.r-project.org> (2022).
99. Thorpe, A. C. and Busi, S. B. GitHub release metag\_analysis\_EA v1.0.0. <https://doi.org/10.5281/zenodo.17252617> (2025).
100. Thorpe, A. C. & Busi, S. B. GitHub release EA\_metag\_post\_analysis v1.0.0. <https://doi.org/10.5281/zenodo.17252627> (2025).
101. Thorpe, A. C. & Busi, S. B. GitHub release biofilm\_MAG\_analysis v1.0.0. <https://doi.org/10.5281/zenodo.17252633> (2025).

## Acknowledgements

Thank you to Dr Jonathan Porter, Sean Butler, and Alan Wan from the Environment Agencies National Monitoring Laboratories for their support collating and shipping the biofilm samples for further processing. This work was funded by the Environment Agency under research project SC220034. A.C.T., S.B.B. and D.S.R. were supported by Natural Environment Research Council (NERC) grant NE/X015947/1. K.W. and J.W. were supported by NERC grant NE/X015777/1. The authors acknowledge the support of the Biotechnology and Biological Sciences Research Council (BBSRC), part of UK Research and Innovation; Earlham Institute Strategic Programme Grant Decoding Biodiversity BB/X011089/1 and its constituent work package - BBS/E/ER/230002C (Decode WP3 Linking Fine-Scale Microbial Diversity to Ecosystem Functions).

## Author contributions

A.C.T.: conceptualisation, data curation, investigation, formal analysis, visualisation, writing – original draft, review and editing. S.B.B.: conceptualisation, data curation, formal analysis, writing – review and editing. J.W.: conceptualisation, data curation, writing – review and editing. L.H.H.: data curation, writing – review and editing. K.W.: conceptualisation, supervision, writing – review and editing. D.S.R.: conceptualisation, supervision, writing – review and editing.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-65620-3>.

**Correspondence** and requests for materials should be addressed to Amy C. Thorpe, Susheel Bhanu Busi or Daniel S. Read.

**Peer review information** *Nature Communications* thanks Sanjay Swarup, and the other, anonymous, reviewers for their contribution to the peer review of this work. A peer review file is available.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2025