Genome-resolved metaproteomics decodes the microbial and viral contributions to coupled carbon and nitrogen cycling in river sediments

Authors: Josué A. Rodríguez-Ramos[1], Mikayla A. Borton[1,2], Bridget B. McGivern[1], Garrett J. Smith[3], Lindsey M. Solden[1], Michael Shaffer[1], Rebecca A. Daly[1], Samuel O. Purvine[4], Carrie D. Nicora[2], Elizabeth K. Eder[4], Mary Lipton[4], David W. Hoyt[4], James C. Stegen[2], and Kelly C. Wrighton[1*]

[1] Department of Soil and Crop Sciences, Colorado State University, Fort Collins, CO, USA

[2] Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA, USA

[3] Department of Microbiology, Radboud University, Nijmegen, Netherlands

[4] Environmental and Biological Sciences Directorate, Pacific Northwest National Laboratory, Richland, WA, USA

*** Corresponding author**

Kelly.Wrighton@colostate.edu

Keywords & Research Areas: hyporheic zone, river sediments, microbiome, metagenomics, mineralization, greenhouse gas, viruses, auxiliary metabolic genes

Abstract:

Rivers have a significant role in global carbon and nitrogen cycles, serving as a nexus for nutrient transport between terrestrial and marine ecosystems. Although rivers have a small global surface

area, they contribute substantially to global greenhouse gas emissions through microbially mediated processes within the river hyporheic zone. Despite this importance, microbial roles in these climatically relevant systems are mostly inferred from 16S rRNA amplicon surveys, which are not sufficiently resolved to inform biogeochemical models. To survey the metabolic potential and gene expression underpinning carbon and nitrogen biogeochemical cycling in river sediments, we collected an integrated dataset of over 30 metagenomes, metaproteomes, and paired metabolomes. We reconstructed over 500 microbial metagenome assembled genomes (MAGs), which we dereplicated into 55 unique genomes spanning 12 bacterial and archaeal phyla. We also reconstructed 2482 viral genomic contigs, which were dereplicated into 111 viral MAGs >10kb in size. As a result of integrating gene expression data with geochemical and metabolite data, we created a conceptual model that uncovers new roles for microorganisms in organic matter decomposition, carbon sequestration, nitrogen mineralization, nitrification, and denitrification. Integrated through shared resource pools of ammonium, carbon dioxide, and inorganic nitrogen we show how these metabolic pathways could ultimately contribute to carbon dioxide and nitrous oxide fluxes from hyporheic sediments. Further, by linking viral genomes to these active microbial hosts, we provide some of the first insights into viral modulation of river sediment carbon and nitrogen cycling.

Importance:

Here we created HUM-V (Hyporheic Uncultured Microbial and Viral), an annotated microbial and viral genome catalog that captures the strain and functional diversity encoded in river sediments. Demonstrating its utility, this genomic inventory encompasses multiple representatives of the most dominant microbial and archaeal phyla reported in river sediments and provides novel viral

genomes that can putatively infect these. Furthermore, we used HUM-V to recruit gene expression data to decipher the functional activities of these genomes and reconstruct their active roles in river sediment biogeochemical cycling. We show the power of genome resolved, multi-omics to uncover the organismal interactions and chemical handoffs shaping an intertwined carbon and nitrogen metabolic network and create a framework that can be extended to other river sediments. The accessible microbial and viral genomes in HUM-V will serve as a community resource to further advance more untargeted, activity-based measurements in these and related freshwater terrestrial-aquatic ecosystems.

Introduction

The hyporheic zone (HZ) is a transitional space between river compartments, where the mixing of nutrients and organic carbon from river and groundwater stimulate microbial activity (1–3). Characterized as the permanently saturated interface between the river surface channel and underlying sediments, the HZ is considered a biogeochemical hotspot for microbial biogeochemistry (1–3), ultimately contributing to the majority of river greenhouse gas (GHG) fluxes. For instance, it is estimated rivers contribute up to 85% of inland water carbon dioxide and 30% of nitrous oxide emissions (4–6). Microorganisms in the HZ also catalyze the transformation of pollutants and natural solutes, all while microbial biomass itself supports benthic food webs (7). Together these findings highlight that microbial metabolism in HZ sediments has substantial influence on overall river biogeochemistry and health.

Despite the importance of HZ microorganisms, research linking microbial identity to specific biogeochemical reactions in the carbon and nitrogen cycles is still nascent in these sediments. In conjunction with geochemistry, microbial functional genes or gene products (e.g.

70 *nirS* and *nrfA*) have been quantified to denote microbial contributions to specific biogeochemical
71 pathways (e.g. nitrate reduction) (8). However, these studies often do not identify the
72 microorganisms catalyzing the process and only focus on a few enzymatic reactions. Thus, a
73 comprehensive assessment of the interconnected microbial metabolisms fueling carbon and
74 nitrogen cycling in river sediments is underexplored.

75 More recently, 16S rRNA amplicon sequencing has shed new light on the identity of
76 bacteria and archaeal members in river sediments. These studies revealed that cosmopolitan and
77 dominant members in river sediments belong to six main phyla: *Acidobacteria, Actinobacteriota,*
78 *Firmicutes, Nitrospirota, Proteobacteria*, and *Thaumarchaeota* (9, 10). Furthermore, in some
79 instances cultivation paired to amplicon sequencing has assigned some of these microorganisms
80 (e.g. *Proteobacteria*) to specific biogeochemical process (e.g., denitrification) (11). Yet, most
81 functional inferences from taxonomic data alone are unreliable due to the dissociation between
82 microbial taxonomy and metabolic function (12, 13). Thus, many key biogeochemical pathways
83 in rivers (e.g., plant biomass deconstruction, denitrification, nitrogen mineralization) are not
84 holistically interrogated alongside microbial communities (14). Furthermore, amplicon sequencing
85 fails to sample viral communities. While it is likely viruses are key drivers of HZ microbial
86 mortality and biogeochemical cycling by dynamics of predation and auxiliary metabolic genes, the
87 evidence is even more sparse than for its microbial counterparts (15–18).

88 Cultivation-independent, community-wide, and genome resolved approaches are key to
89 addressing the knowledge gap of how microbial and viral communities influence river
90 biogeochemical cycling. However, metagenomic studies in river sediments are relatively sparse,
91 and have focused primarily on gene content as opposed to whole genomes (19, 20). To our
92 knowledge, only two river sediment studies have generated microbial genomes to allow linkages

of identity to functional processes, and these have focused on the impacts of nitrate oxidizing and comammox microorganisms to nitrification (21, 22). As such, despite these recent advances the chemical exchange points that interconnect the carbon and nitrogen cycles, and the metabolic handoffs between microorganisms that sustain them cannot be discerned from existing HZ microbiome studies.

With the overarching goal of providing enhanced resolution to microbial and viral contributions to carbon and nitrogen cycling in the HZ, we created the first of its kind Hyporheic Uncultured Microbial and Viral (HUM-V) genomic catalog. We then used HUM-V to recruit metaproteomic data collected from over 30 laterally and depth distributed HZ sediment samples. We further supported this gene expression data using chemical data from paired metabolomics and geochemical measurements. Our results (i) profiled the expressed microbial metabolic exchanges that support organic and inorganic carbon and nitrogen cycling in the HZ, (ii) uncovered roles for viruses that could modulate microbial activity in the HZ, and (iii) created a roadmap of the microbial metabolic circuitry that potentially contributes to greenhouse gas fluxes from rivers. We anticipate that this publicly available community resource will advance future microbial activity-based studies in HZ sediments and is a step towards the development of biologically aware, hydro-biogeochemical predictive models.

RESULTS and DISCUSSION

HUM-V greatly expands the genomic sampling of HZ microbial members

We collected samples from six HZ sediment cores, with each core subsampled into six 10cm depth increments (0-60 cm). This resulted in 33 samples that were analyzed for metagenomic sequencing, geochemistry, and metaproteomics (**Fig. 1abc**). For our metagenomics data, we

obtained 379Gbp of sequencing for all 33 samples which included i) the original shallow sequencing of all samples (1.7-4.9 Gbp/sample) (23) and ii) an additional deeper sequencing of 10 samples (15.3-49.2 Gbp/sample), which are reported here for the first time. We then used assembly, co-assembly, and sub-assembly approaches to reconstruct 655 bacterial and archaeal metagenome assembled genomes (MAGs). Of these genomes, 102 were denoted as medium or high-quality per the Genome Consortium Standards (24), and were dereplicated at 99% identity into the 55 unique genomic representatives that constitute the microbial component of HUM-V (see Data Availability) (**Fig. 1d**). Of the genomes retained in HUM-V, 36% were obtained from deeply sequenced, assembled, and binned samples; while 27% were from co-assemblies performed across samples. The ability to recover additional genomes relative to our prior effort which only used shallow sequencing (23) demonstrates how sequencing depth and integration of multiple assembly methods provided complementary information to generate these microbial HZ communities.

Given sparse metagenomic sampling of HZ sediments, it was not surprising that HUM-V contained the first genomic representatives of highly prevalent microorganisms (**Fig. 1d, Fig. 2ab**). Phylogenetic analyses of the 55 unique HUM-V genomes revealed they spanned 2 Archaeal and 9 Bacterial phyla, and that most genomes belonged to a subset of 3 bacterial phyla (*Desulfobacterota*, *Nitrospirota*, and *Proteobacteria*). To our knowledge, the 8 *Desulfobacterota* and 7 *Proteobacteria* genomes identified here represent the first HZ MAGs sampled from these commonly reported lineages. For the *Nitrospirota*, a prior study reported 21 MAGs that we dereplicated into 12 unique genomes (99% ANI) (21), a sampling we further expanded by an additional 20 genomes. We note the *Nitrospirota* genomes sampled here spanned 3 genera that we did not identify as being previously sampled from rivers, and that included new species within the

Nitrospiraceae 2-02-FULL-62-14, *40CM-3-62-11*, and *NS7*. Moreover, HUM-V contains one genome of the *Actinobacteriota* that may represent a new order, as well as 6 new genera from *Acidobacteriota*, *Actinobacteriota*, *CSP1-3*, *Desulfobacterota, Proteobacteria*, and *Thermoplasmatota* (**Fig 2ab**). Further highlighting the genomic novelty of this ecosystem, HUM-V contains genomes from entirely uncultivated members of different phyla (9 genomes from *CSP1-3* and *Eisenbacteria*) and classes (10 genomes from *Binatia* and *MOR-1*). Ultimately, HUM-V is a genome resource that will enable taxonomic analyses and metabolic reconstruction of microbial metabolisms in HZ sediments.

HUM-V recruits metaproteomes offering new insights into HZ microbiomes

Leveraging paired metaproteomes collected with the metagenomes allowed us to assign gene expression to each genome in HUM-V. These microbial genomes recruited 13,102 total peptides to 1,313 proteins. Because our genome analysis revealed that closely related strains shared overlapping metabolic potential (**Fig. 1d**), we analyzed the proteomic data using two approaches. First, we considered the 'unique' peptides that could only be assigned to proteins from a single microbial genome. These represented 67% of genes expressed in our proteome. Next, we considered proteins that recruited 'non-unique but conserved' peptides which we define as those assigned to proteins that (i) have identical functional annotation and (ii) are from more than one genome within the same genus. These proteins are shown in grey on **Fig. 2b**, and although they accounted for a smaller fraction of the genes expressed (14%), this prevented us excluding data due to strain overlap in our database.

In microbiome studies, dominance is often used as a proxy for microbial activity. Here, we evaluated this assumption using our paired metagenome and metaproteome data. When comparing

the genome relative abundances to protein expression patterns, we observed that the most abundant genomes were not necessarily those that were most actively expressing proteins at the time of sampling. The most abundant genomes included members of the *Binatia, Nitrospiraceae NS7*, and *Nitrososphaeraceae TA-21* (formerly *Thaumarchaeota*) (**Fig. 2b**). However, only the dominant *Nitrososphaeraceae* genomes had high recruitment of the uniquely assigned proteome. On the other hand, some low abundance members (e.g., *Actinobacteriota*) accounted for a sizeable fraction (30%) of the unique proteome (**Fig. 2b**).

Leveraging these metagenomic and metaproteomic datasets, we first examined metabolic traits that were conserved across nearly all HUM-V microbial genomes. Notably, all but one of the microbial genomes recovered from this site (CSP1_3_1) encoded the genomic capacity for aerobic respiration. We defined this capability by the recovery of genes indicating a complete electron transport chain and some form of terminal oxidase (**Fig. S1**). Consistent with this genomic data, resazurin reduction assays indicated these sediments were oxygenated, and could likely support aerobic microbial respiration (**Fig. S2**). However, while proteomic evidence for aerobic respiration (cytochrome c oxidase *aa3*) was detected in nearly 40% of samples, it could only be confidently assigned non-uniquely to nitrifying *Nitrososphaeraceae.* This is likely due to the highly conserved nature of this gene, as well as the limitations of detecting membrane, heme-containing cytochromes with metaproteomic data (25). As such, we consider it likely this metabolism was more active than was captured in the metaproteomic data.

Ordination analyses of our genome-resolved metaproteomes revealed that microbial gene expression did not cluster significantly by sediment depth or transect position (**Fig. S3ab**). In fact, over 90% of measured gene expression was shared across both transects (**Fig. S3c**). When considering explanations for this lack of spatial structuring, it is possible that the microbial

185 heterogeneity in these samples occurred over a finer spatial resolution (pore or biofilm scale, <10
186 cm) or larger (>60 cm) than the bulk 10 cm-depths sampled here. It is also possible that gene
187 expression is constitutive to sustain metabolic function during highly dynamic conditions that
188 occur in the HZ.
189
190 *An inventory of processes contributing to microbial carbon dioxide production and consumption*
191 To uncover the microbial food web contributing to organic carbon decomposition in HZ
192 sediments, we reconstructed the carbon degradation network using coordinated genome potential,
193 expression, and carbon metabolite data. Based on linkages to specific substrate classes, genomes
194 were assigned to the following different trophic levels in the carbon food chain: (i) plant polymer
195 degradation; (ii) sugar fermentation; (iii) smaller organic compounds (e.g., alcohols and fatty
196 acids), and (iv) single carbon compounds (carbon monoxide, carbon dioxide, methane) (**Fig. 3**).
197 It is well recognized that heterotrophic oxidation of organic carbon derived in HZ
198 sediments largely contributes to river respiration (2). Despite generally low organic concentrations
199 in our sediments (<10 mg/g), FTICR-MS analysis showed that lignin-like compounds were the
200 most abundant biochemical class detected in our samples suggesting that plant litter was likely
201 source of organic carbon (**Fig. S4**). In support of this, 38% of the HUM-V genomes encoded genes
202 for degradation of phenolic/aromatic monomers, while 11% could degrade the larger, more
203 recalcitrant polyphenolic polymers. In fact, our genome analyses revealed that seven unique
204 genomes constituting a new genus within the uncultivated *Binatia* encoded novel pathways for the
205 decomposition of aromatic compounds from plant biomass (phenylpropionic acid, phenylacetic
206 acid, salicylic acid), and xenobiotics (phthalic acid) (**Fig. S5**).

Gene expression of carbohydrate-active enzymes (CAZymes) also supported the degradation of plant polymers like starch and cellulose. We detected the expression of putative extracellular glucoamylase (GH15) and endo-glucanase (GH5) from an *Actinobacteriota* genome (Microm_1) and the *Nitrososphaeraceae* (Nitroso_2), respectively. The integration of our chemical and biological data revealed that heterotrophic metabolism in these sediments could be maintained, likely by inputs of plant biomass. In support of carbon depolymerization, sugars like glucose, and sucrose were detected by nuclear magnetic resonance (NMR).

We next sought to identify microorganisms that could utilize these sugars and found that members expressed transporters for fructose (Rhizo-Anders_1), glucose (Microm_1), and general sugar uptake (Actino_1, Nitroso_2, Nitroso_3). In support of further decomposition, we detected organic acids (acetate, butyrate, lactate, pyruvate, propionate) and alcohols (ethanol, methanol, isopropanol) by NMR. Similarly, proteomics supported interconversions of these smaller carbon molecules, with the *Myxococcota* (Anaerom_1) expressing genes for aerobic acetate respiration and the archaeal *Woeseia* (Woese_1) respiring methanol. In summary, the chemical scaffolding and overlayed gene expression patterns support an active heterotrophic metabolic network in these HZ, likely driven by plant biomass decomposition.

In addition to heterotrophy, our proteomics data revealed autotrophy might also generate carbon dioxide (CO_2). Dehydrogenase genes for the aerobic oxidation of carbon monoxide (CO) were among the most prevalent across these sediments. This metabolism was expressed by phylogenetically distinct lineages, including members of uncultivated lineages *Binatia* (Binatia_2) and *CSP1-3* (CSP1_3_1), as well as members of *Actinobacteriota* (Actino_1, Microm_1), *Methylomirabilota* (Roku_AR37_2), and *Proteobacteria* (Burk_1, Thioh_1). The wide range of bacteria and archaea that encoded dehydrogenase genes, combined with gene

expression data, suggests carbon monoxide oxidation may be an important metabolism for persistence in HZ sediments.

Given these sediments have relatively low total carbon concentrations (**Fig. S6**), we consider it possible that carbon monoxide may act as a supplemental microbial energy and/or carbon source. Based on genome content, we cautiously infer members of *Actinobacteriota* (Microm_1), *Binatia*, and *CSP1-3* may be capable of carboxydotrophy (i.e., using carbon monoxide as sole energy and carbon source), while the *Actinobacteriota* (Actino_1) is a likely carboxydovore (i.e., oxidize carbon monoxide, while requiring organic carbon). While this metabolism is poorly resolved environmentally, recent efforts have shown it is induced by organic carbon starvation to mediate aerobic respiration, thereby enhancing survival in oligotrophic conditions (26). Here we add river sediments to list of aerated environments (e.g. ocean and soils) where this metabolism may acts as sink or regulate the emission of this indirect GHG (27, 28).

Since proteomics indicated heterotrophy and carbon monoxide oxidation could generate carbon dioxide, we next tracked microorganisms that could fix this compound, sequestering its release. Genome analyses revealed four pathways for carbon fixation were encoded by 75% of HUM-V microbial genomes including (i) Calvin-Benson-Bassham cycle, (ii) reductive TCA cycle, (iii) 3-HydroxyPropionate /4-HydroxyButyrate cycle, and (iv) 3-Hydroxypropionate bi-cycle. The two nitrifying lineages were inferred chemolithoautotrophs, with *Nitrososphaeraceae* encoding HydroxyPropionate/4-HydroxyButyrate (3HP/4HB) and the *Nitrospiraceae* encoding the reductive tricarboxylic acid (TCA) cycle. Other phylogenetically diverse lineages, *Acidobacteriota*, *Binatia, CSP1_3, Proteobacteria*, and *Woeseiaceae* encoded redundant fixation pathways.

Our genome and proteome data revealed the prevalence and activity of single carbon metabolism in these sediments. Carbon monoxide and dioxide are likely the primary substrates, as HUM-V only had minimal evidence for methanol oxidation (*Woeseia*), no methanotrophs, and no methanogens. Together our proteogenomic findings hint at the importance of carbon monoxide and carbon dioxide in sustaining microbial metabolism in these aerated, but low, or fluctuating carbon environments. Further work is needed to understand physiochemical factors controlling carbon monoxide oxidation and carbon dioxide fixation activity, and the balance between production (via heterotrophy and carbon monoxide oxidation) and consumption (fixation) on overall river sediment carbon dioxide emissions.

Ammonium exchange can support coordinated nitrogen mineralization and nitrification pathways

The ratio of total carbon (C) and total nitrogen (N) (e.g., C/N) is a geochemical proxy used to denote the possible microbial metabolisms that can be supported in a habitat (29, 30). Our HZ sediments had C/N ratios with a mean of 6.5 ± 1.1 (maximum 8.4) (**Fig. S6**). Geochemical theory posits that sediments with low C/N ratios (<15) support organic mineralization that yields sufficient ammonium such that heterotrophic bacteria are not N-limited and nitrifying bacteria are able to compete successfully for ammonium enabling nitrification (30–32). Based on our sediment C/N ratios, we hypothesized organic nitrogen mineralization and nitrification co-occurred in these sediments. Here we profiled the microbial substrates (organic nitrogen metabolites, ammonium) and expressed pathways (mineralization and nitrification) to provide biological validation of this established geochemical theory.

To examine the microbial contributions to organic nitrogen mineralization, we examined metaproteomic data for peptidases, genes that mineralize organic nitrogen into amino acids and

free ammonium. Peptidase (n=31) gene expression was three times more abundant and prevalent than glycoside hydrolase genes catalyzing organic carbon transformations, hinting at the relative importance of the former. In support of active microbial N mineralization, FT ICR MS revealed that protein-like and amino sugar like organic nitrogen compounds were correlated to high microbial activity (23), while here we show hydrophobic, polar, and hydrophilic amino acids were abundant (more so than sugars) in the H^1-NMR characterized metabolites (**Fig. S7**).

The expression of peptidases *in situ*, combined with our genome resolution of their hosts, provided a new opportunity to interrogate the resource sharing and competition underpinning nitrogen mineralization. We first noted which microorganisms were expressing extracellular peptidases (inferred from PSORTb (33)), which would be enzymes that shape external organic nitrogen pools in the sediment. We categorized these expressed peptidases as either releasing free amino acids (end terminus cleaving families, e.g., M28) or releasing peptides (endocleaving families, e.g., S08A, M43B, M36, MO4) (**Fig. 4a**). Members of the *Actinobacteriota*, *Binatia, Methylomirabilota*, and *Thermoproteota* were found to express extracellular peptidases and as such, are likely candidates that contribute to sediment N mineralization.

We then profiled the expressed amino acid transporters, i.e., genes for the cellular uptake of these smaller organic nitrogen compounds (e.g., branched chain amino acids, glutamate, amines, and peptides were examined) (**Fig. 4b**). Some members functioned exclusively as 'producers', expressing peptidases for cleavage of organic N to liberating smaller peptides, yet we could not detect genes for the transport of these produced compounds. Other taxa were 'producers and consumers', as genomes in the *Actinobacteriota* and *Binatia* expressed genes for external peptidases and for transporting the organic N products into the cell. Alternatively, we entertain the possibility that members of the *CSP1-3, Proteobacteria*, and *Thermoplasmatota* could be

functioning as 'exploiters' given that these members expressed genes for assimilating peptidase products but did not contribute to the cost of their production. Expanding on previous carbon degradation frameworks of natural communities (34, 35), we provide a closer examination of the cooperative and competitive interactions that underpin nitrogen mineralization, and further offer new molecular insights into this important terrestrial process.

Finally, we examined the proteomes for evidence that nitrification co-occurred with organic nitrogen mineralization. Supporting this possibility, the substrate ammonium (NH_4^+) was detected in all 33 sediment samples (**Fig. S7**). We did not detect genomic evidence for comammox or anammox metabolisms in HUM-V genomes, suggesting aerobic nitrification by metabolic exchanges between organisms drives nitrification in these sediments. Proteomics confirmed aerobic ammonium oxidation to nitrite was performed by Archaeal *Nitrososphaeraceae*. In fact, ammonia monooxygenase (*amo*) subunits were within the top 5% most highly expressed functional proteins in this dataset. The next step in nitrification, nitrite oxidation to nitrate, was inferred from nitrite oxidoreductase (*nxr*) expressed by members of *Nitrospiraceae*. Demonstrating that new lineages first discovered in HUM-V could shape in situ biogeochemistry, we confirmed that 5 genomes from two new species of *Nitrospiraceae* expressed nitrification genes (Nitro_40CM-3_1, Nitro_NS7_3, Nitro_NS7_4, Nitro_NS7_5, and Nitro_NS7_14).

The proteome supported archaeal-bacterial nitrifying mutualism outlined here appear well adapted to the low nutrient conditions present in many HZ sediments, warranting future research on the universal variables that constrain nitrification rates (i.e., ammonium availability, dissolved oxygen, pH) and their role in driving nitrogen fluxes from these systems (36). In conclusion, our microbial data supports that nitrification is concomitant with mineralization in these samples,

320 providing biological evidence to substantiate inferences made from the C/N ratio of these
321 sediments.
322
323 *Metabolic handoffs between phylogenetically distinct microorganisms sustain denitrification*
324 Our proteomics suggests that aerobic nitrification could complement allochthonous nitrate
325 from groundwater discharges, contributing to measured nitrate concentrations in excess of 20 mg/L
326 (2, 37). HUM-V genomes with the capacity for nitrate reduction were phylogenetically diverse,
327 with *NarG* or *NapX* encoded in 11 genomes from the *Actinobacteriota*, *Binatia*, *Myxococcota*, and
328 *Proteobacteria,* yet unique peptides were only detected from *Binatia NarG*. Based on gene
329 expression data, nitrite was likely reduced to nitric oxide both by denitrifying *Burkholderia* and
330 nitrifying *Nitrososphaeraceae,* with the metabolic rationale in the archaeal ammonia oxidizers
331 including a detoxification mechanism or generating electron source for ammonia oxidation (38–
332 40).
333 We did not detect proteomic evidence for nitrous oxide production but did detect *nos* gene
334 expression for reducing nitrous oxide to nitrogen gas. Here the non-denitrifying
335 *Desulfobacterota_D* (Desulf_UBA2774_1, formerly *Dadabacteria* (41)) expressed the *nos* gene
336 for reducing nitrous oxide to nitrogen gas. Phylogenetic analysis revealed this sequence was a
337 "Clade II" *nosZ* sequence, an atypical variant adapted for lower, or atmospheric nitrous oxide
338 concentrations (42). Our activity data adds to emerging interest on these non-denitrifying clade II
339 *nosZ* microorganisms in terrestrial systems, as these microorganisms increase nitrous oxide sink
340 capacity (without contributing to its production) (43, 44). Notably, the activity of this enzyme
341 would have been missed using traditional *nosZ* primers, denoting the value of our untargeted,
342 expression-based approach (45).

We do note that HUM-V capacity for nitrogen cycling exists beyond the proteomics detected instances highlighted above. For example, *Binatia* encoded dissimilatory nitrite reduction to ammonium (DNRA) and the potential for nitrous oxide production via *nor* was encoded by two *Gammaproteobacteria* (Steroid-FEN-1191_1, Steroid_1) and a member of the *Myxococcota* (Anaerom_1). HZ are characterized by dynamic changes in flow direction that lead to cyclic changes in aqueous compositions of dissolved oxygen, nitrate, and organic carbon, thus it is possible that this broader metabolic potential is manifested under other temporal conditions (46).

A key finding of our denitrification metabolic inventory is that complete denitrification by a single microorganism is likely the exception rather than the rule in natural systems (47, 48). In support of this, none of the genomes reconstructed here encoded a complete denitrification pathway for reducing nitrate to nitrous oxide or dinitrogen. Similarly, the metaproteomics data indicated that separate microbial members catalyzed each step of denitrification. Since cross-organism nitrogen chemical exchanges are necessary to support active denitrification, one must consider that physical processes (e.g., advection, diffusion) or the spatial colocalization of microorganisms, as well as organic carbon availability, may have disproportionate impacts on flux of nitrous oxide and dinitrogen in HZ sediments.

Viral influence on sediment carbon and nitrogen cycling

We reconstructed 2,482 vMAGs that dereplicated into 111 dereplicated viral populations (>10kb) in HUM-V (**Fig. 5a**), making this one of only a handful of genome-resolved studies that include viruses derived from rivers (16, 49, 50). To our knowledge, this is the first study to provide a coordinated analyses of microbial and viral community genomes, and given their sparse sampling, only 5 of the HUM-V viral genomes had taxonomic assignments using viral taxonomies

366 established from standard reference databases. To better understand if the remaining viral genomes
367 had been previously detected in other virally sampled freshwater systems, we compared the protein
368 content of vMAGs in our system to an additional 1,861 viral genomes we reconstructed *de novo*
369 or obtained from public metagenomes from North and South America (**Fig. 5b**). Of the 105
370 remaining viral genomes, 15% (n=17) clustered with these freshwater derived viral genomic
371 sequences indicating possible cosmopolitan viruses. Of the remaining viral genomes, 23% (n=26)
372 clustered only with genomes recovered in this data set, indicating multiple samplings of the same
373 virus across different sites and depths. The remaining 57% (n=63) of the viral genomes were
374 singletons, meaning they were sampled from these sediments once. Combined, these results hint
375 at the possible biogeographically diverse, as well as endemic viral lineages warranting further
376 exploration in river sediments.

377 We then assessed peptide recruitment to the viral portion of HUM-V (**Fig. 5a**). For viruses
378 and microbes alike, the most abundant genomes did not have the highest gene expression. While
379 the viral gene expression was not structured by edaphic or spatial factors (**Fig. S8ab**), it was
380 strongly coordinated to the microbial patterns (**Fig. S8c**). Like our microbial genome peptide
381 recruitment, 66% of the viral genomes uniquely recruited peptides. This exceeded prior viral
382 metaproteome recruitment from other environmental systems (e.g., wastewater, saliva, rumen,
383 with ranges from 0.4-15% (35, 51, 52). From this we infer a relatively large portion of the viral
384 community was active at the time of sampling.

385 The proteomic recruitment of viruses sampled in HUM-V hinted at the possibility that
386 viruses could structure the microbial biogeochemistry through predation. *In silico* analysis
387 assigned a putative host to 29% of the 111 viral genomes. Viruses were linked to 18 microbial
388 genomes that belong to bacterial members in *Acidobacteriota*, *Actinobacteriota*, *CSP1-3*,

389 *Eisenbacteria, Methylomirabilota*, *Myxococcota*, *Nitrospirota*, and *Proteobacteria* (**Fig. S9ab**).
390 Analysis of the metaproteomes for these putative phage-impacted genomes revealed these hosts
391 expressed genes for carbon monoxide oxidation (*Actinobacteriota*), carbon fixation (*CSP1-3*),
392 nitrogen mineralization (*Acidobacteriota, and Methylomirabilota*), methanol respiration
393 (*Myxococcota*), nitrification (*Nitrospirota*), and ammonia oxidation (*Proteobacteria*) (**Fig. 5cd**).
394 Thus, viral predation in HZs could govern carbon and nitrogen biogeochemistry and may explain
395 some of the strain and functional redundancy we observe encoded and expressed by microbes in
396 these sediments.

397 We next inventoried HUM-V viral genomes for auxiliary metabolic genes (AMGs) with
398 the potential to augment biogeochemistry. We detected 14 auxiliary metabolic genes (AMGs)
399 which we confirmed were not bacterial in origin and had viral-like genes on both flanks (53). These
400 highly ranked AMGs (see Methods) had the potential to augment carbon (CAZymes), sulfur
401 (sulfate adenylyltransferase), and nitrogen (amidase to cleave ammonium) metabolism (**Fig. S9c**).
402 One of our viral genomes that was putatively linked to a *Steroidobacteraceae* (Steroid_1) host
403 genome encoded a pectin lyase gene (PL1). This viral PL1 could enable its host to cleave pectin,
404 generating pectin oligosaccharides that could be used via two host-encoded glycoside hydrolases
405 (GH4 and GH2), ultimately freeing galactose for energy metabolism (**Fig. S9de**). While
406 theoretical, we include this as an example to illustrate how virally encoded genes could expand the
407 substrate ranges for their hosts and alter biogeochemical cycling in river sediments.

408 In support of their importance to modulating microbial activity and sediment
409 biogeochemistry, we noted that adding the viral genome abundance patterns to microbial genome
410 abundance patterns improved our predictions of river sediment carbon and nitrogen concentrations
411 (**Fig. S10**). In summary, these results indicate that viral predation and AMGs may contribute to

412 river sediment biogeochemistry, either through top down or bottom-up controls on the microbial
413 community. Further interrogation of river sediment viral communities through genomics will yield
414 a more comprehensive understanding of the ecological networks underpinning river
415 biogeochemistry.
416
417 Conclusions
418 *A multi-omics informed roadmap of carbon and nitrogen cycling in hyporheic zone sediments*
419 Despite the importance of the HZ and its relative accessibility in terms of sampling
420 locations, HZ microbial and viral communities are surprisingly under sampled in a genomic
421 context. Previous studies pertaining to this ecosystem are not genome resolved and used 16S rRNA
422 amplicons or unbinned metagenomes (8–11, 19–21), thus limiting the predictive and explanative
423 power of the study and often times not discriminating between metabolically active and inactive
424 organisms. Further, the few studies which are genome-resolved (21, 22) often focus on specific
425 lineages or single processes and not the entire microbial community, missing the complex
426 interplays between aspects of carbon and nitrogen cycles.
427 Here we created HUM-V, a genome resolved database to expand on prior non-genome
428 resolved analyses done in a previous publication by our team (23). In concordance to the findings
429 by Graham et al, we failed to detect any significant differences across spatial gradients in
430 community composition or proteomics data, supporting the notion that microbial phenotypic
431 plasticity is important in these sediments. However, our new resolution provided insights into this
432 metabolic plasticity and its mechanistic underpinnings, as we recovered strains with overlapping
433 metabolic potentials from different organisms. Beyond metabolic potentials, our genome-resolved
434 proteomic data further highlighted processes with overlapping expressed genes and provided some

435 of the first activity measurements for members of hyporheic microbiomes. As an example, we
436 focus on new insights gleaned from the 7 recovered *Binatia* genomes (one which included a
437 complete 16S rRNA gene), which are geographically widespread, uncultured, and have only been
438 described in terms of metabolic potential by one previous publication (54). Proteomics
439 demonstrated these bacteria (i) aerobically oxidized carbon monoxide, (ii) mineralized organic
440 nitrogen, and (iii) denitrified, vastly contributing to carbon and nitrogen cycling in these sediments
441 and being biogeographically dispersed (**Fig. S11a**). We also highlighted the metabolic capacity of
442 this lineage in polyphenolic metabolism, perhaps explaining the prevalence of this lineage in
443 terrestrial, plant impacted systems (**Fig. S11b**). These findings illustrate the power of HUM-V in
444 illuminating new roles for members of uncultivated, previously enigmatic microbial lineages in
445 hyporheic zone biogeochemistry.

446 Empowered by our genome-resolved, process-based metaproteomic analyses (**Fig. 2-Fig.**
447 **5**), we present a conceptual model outlining pathways of co-occurring microbial and viral
448 contributions to carbon and nitrogen biogeochemistry in these sediments (**Fig. 6**). Denoted by
449 black arrows, we highlight different microbial transformations that we discovered in our
450 proteomics data. Building on those, colored arrows show how these contributions were
451 complemented by inputs from aquatic sources such as surface or ground water (blue arrows),
452 terrestrial / biotic sources (purple arrows) or atmospheric sources (green arrows). Together, we
453 demonstrated how these could result in the formation and depletion of nutrients in shared resources
454 pools. We further revealed that modes of competition and cooperation form a network of metabolic
455 cross feeding that affected organic and inorganic carbon cycling, and is intertwined with nitrogen
456 mineralization, nitrification, and denitrification pathways.

In conclusion, while river carbon and nitrogen budgets are often quantified by direct measurements of inputs, and the concentration of inorganic and organic compounds exported from rivers, our findings put forth an integrated framework that advances microbial roles in hyporheic carbon and nitrogen transformations. It yields insights that could inform research strategies to reduce existing predictive uncertainties in river corridor models and resolves microbial contributions and ecological handoffs that were thought to occur but were poorly defined in river sediments. We also highlight previously enigmatic processes that could directly impact river GHG fluxes in unappreciated ways (e.g., carbon dioxide fixation, type II nitrous oxide reduction). Ultimately, we show that a genome-resolved database allows us to track the fates of resource pools of carbon dioxide, ammonium, and inorganic nitrogen to show how consumption and production of these compounds contributes to overall GHG fluxes.

Materials and Methods

Sample collection, DNA isolation, and chemical characterization

Samples were collected from the hyporheic zone of the Columbia River (46°22'15.80″N, 119°16'31.52″W) in March 2015 as previously described (23). Briefly, liquid nitrogen frozen sediment profiles (0-60cm) were collected along two transects separated by approximately 170 meters (**Fig. 1abc**). At each transect, three sediment cores up to 60 cm in depth were collected at 5-meter intervals perpendicular to the river flow. Each core was sectioned into 10 cm segments from 0-60-centimeter depths for downstream analyses.

DNA isolation was carried out as previously described (23), with sequencing sent to the Joint Genome Institute (JGI, n=33) under proposal 1781 (23). New deep sequencing described here was performed at the Genomics Shared Resource facility at The Ohio State University (OSU,

n=10) using a Nextera XT library System. Libraries at both facilities were sequenced using an Illumina HiSeq 2500 platform. **Table S1** details all sequencing information, including accession numbers.

Chemical analyses included geochemical and metabolite data, where geochemistry and Fourier-transform ion cyclotron resonance mass spectrometry (FTICR-MS) methods were performed as previously described (23). Additional metabolite data was obtained through ^{1}H Nuclear Magnetic Resonance (NMR) spectroscopy on sediment pore water. Sediment samples were mixed with 200, 300, or 600 µL of MilliQ water depending on the sediment mass (**Table S2**) and centrifuged to remove the sediment. Supernatant (180 µL) was then diluted by 10% (vol/vol) with 5 mM 2,2-dimethyl-2-silapentane-5-sulfonate-d_6 as an internal standard. All NMR spectra were collected using a Varian Direct Drive 600-MHz NMR spectrometer equipped with a 5-mm triple resonance salt-tolerant cold probe. Chemical shifts were referenced to the 1H or 13C methyl signal in DSS-d6 at 0 ppm. The 1D ^{1}H NMR spectra of all samples were processed, assigned, and analyzed using Chenomx NMR Suite 8.3 with quantification based on spectral intensities relative to the internal standard as described previously (55, 56). All geochemical data and methods can be found in **Text S1, Table S2,** and **Table S3**.

Metagenome assembly and binning

Raw reads were trimmed for length and quality using Sickle v1.33 (https://github.com/najoshi/sickle) and then subsequently assembled using IDBA-UD 1.1.0 (57) with an initial kmer of 40 or metagenomic metaSPAdes 3.13.0 (58) with default parameters. To further increase genomic recovery, for the ten samples that had shallow and deep sequencing, metagenomic reads were coassembled using IDBA-UD 1.1.0 with an initial kmer of 40. All

503 assemblies, including co-assemblies, were then individually binned using Metabat2 (59) with
504 default parameters to obtain MAGs.

505 For each bin, genome completion was estimated based on the presence of core gene sets
506 using Amphora2 (60). Bins were discarded for further analysis if completion was <70% or
507 contamination was >10% to select for only medium to high quality bins (24). This resulted in 102
508 MAGs that were then dereplicated using dRep (61) with default parameters and resulted in a final
509 set of 55 MAGs (>99% ANI). To further assess bin quality, we used the Distilled and Refined
510 Annotation of MAGs (DRAM) (53) to identify ribosomal ribonucleic acids (rRNAs) and transfer
511 ribonucleic acids (tRNAs). Genome quality information reported in **Table S4**.

512

513 *Metabolic analysis of MAGs*

514 Medium and high-quality MAGs were taxonomically classified using the Genome
515 Taxonomy Database (GTDB) Toolkit v1.3.0 on September 2020 (62). Novel taxonomy was
516 identified as the first taxonomic level with no designation using GTDB taxonomy. MAG scaffolds
517 were annotated using the DRAM pipeline (53). Phylogenetic analysis was performed on genes
518 annotated as respiratory nitrate reductase (*nar*) and nitrite oxidoreductase (*nxr*) to resolve novel
519 *Binatia* role in nitrogen cycling. Specifically, sequences from (63) were downloaded and combined
520 with *nar* and *nxr* amino acid sequences from dereplicated bins, aligned using MUSCLE, version
521 3.8.31, and run through ProtPipeliner, a Python script developed in-house for generation of
522 phylogenetic trees (https://github.com/TheWrightonLab). Phylogenetic trees are provided in
523 Zenodo here: https://doi.org/10.5281/zenodo.6339808. For polyphenol and organic polymer
524 degradation, we used functional annotation in addition to predicted secretion to assess functional
525 potential. To determine if the predicted genes encoded a secreted protein, we used pSortb (33) and

526 SignalP (64) to predict location; if those methods did not detect a signal peptide, the amino acid
527 sequence was queried to SecretomeP and a SecP score > 0.5 (65) was used as a threshold to report
528 non-canonical secretion signals. Metabolic information for each MAG discussed in this manuscript
529 are available in **Table S4.**
530
531 *Viral Analyses*
532 Metagenomic assemblies (n=43) were screened for DNA viral sequences using VirSorter
533 v1.0.3 with the ViromeDB database option (66), retaining viral contigs ranked 1, 2, 4 or 5 where
534 category 1-2 indicate high confidence predicted lytic viruses and 4-5 indicate high-confidence
535 prophage sequences from VirSorter output (66). Viral sequences were filtered based on size to
536 retain those greater than or equal to 10kb based on current standards (67). Viral scaffolds were
537 then clustered into vMAGs at 95% ANI across 85% of the shortest contig using ClusterGenomes
538 5.1 (https://github.com/simroux/ClusterGenome) (67). After clustering, vMAGs were manually
539 confirmed to be viral by looking at DRAM-v annotations and further assessing the total viral-like
540 genes relative to non-viral genes, where vMAGs containing more than 18% of non-viral genes
541 were deemed suspicious, manually confirmed non-viral, and subsequently discarded (J flag,
542 DRAM-v) (53).
543 To determine taxonomic affiliation, vMAGs were clustered to viruses belonging to viral
544 reference taxonomy databases NCBI Bacterial and Archaeal Viral RefSeq V85 with the
545 International Committee on Taxonomy of Viruses (ICTV) and NCBI Taxonomy using the
546 network-based protein classification software vContact2 v0.9.8 using default methods (68, 69). To
547 determine geographic distribution of viruses in freshwater ecosystems, we included viruses mined
548 from publicly available freshwater metagenomes in vContact2 analyses: 1) East River, CO

549 (PRJNA579838) 2) A previous study from the Columbia River, WA (PRJNA375338) 3) Prairie
550 Potholes, ND (PRJNA365086) and 4) the Amazon River (PRJNA237344). The viral sequences
551 that were identified from these systems and the genes used for vContact2 are deposited on Zenodo
552 with doi 10.5281/zenodo.6310084.

553 Viral contigs were annotated with DRAM-v (53). Genes that were identified by DRAM-v
554 as being high-confidence possible auxiliary metabolic genes (auxiliary scores 1-3) (53) were
555 subjected to protein modeling using Protein Homology / AnalogY Recognition Engine (PHYRE2)
556 (70). Auxiliary scores were assigned by DRAM (53), based on the following ranking system: A
557 gene is given an auxiliary score of 1 if there is at least one hallmark gene on both the left and right
558 flanks, indicating the gene is likely viral. An auxiliary score of 2 is assigned when the gene has a
559 viral hallmark gene on one flank and a viral-like gene on the other flank. An auxiliary score of 3
560 is assigned to genes that have a viral-like gene on both flanks. To identify likely vMAG hosts,
561 oligonucleotide frequencies between virus (n=111) and non-dereplicated hosts (n=102) were
562 analyzed using VirHostMatcher using a threshold of d2* measurements of <0.25 (71). The lowest
563 d2* value for each viral contig <0.25 was used. All vMAG data is reported in **Table S5**.

564

565 *Genome relative abundance calculations and their use in predictions*

566 To estimate the relative abundance of each MAG and vMAG, all metagenomic reads for
567 each sample were rarified to 3Gbp and mapped to 55 unique MAGs via Bowtie2 (72) (55, 73). For
568 MAGs, a minimum scaffold coverage of 75% and depth of 3x required for read recruitment at 7
569 mismatches. For vMAGs, reads were mapped using Bowtie2 (72) at a maximum mismatch of 15,
570 a minimum contig coverage of 75% and a minimum depth coverage of 2x. Relative abundances
571 for each MAG and vMAG were calculated as their coverage proportion from the sum of the whole

572 coverage of all bins for each set of metagenomic reads. Genome relative abundances per sample
573 for MAGs and vMAGs are reported in **Table S4** and **Table S5**. Correlations and sparse Partial
574 Least Squares Regression (sPLS) predictions (PLS R package (74)) were done using mapping data
575 pertaining to only the 10 deeply sequenced metagenomes rarified to 4.8Gbp (**Table S6**).
576
577 *Metaproteome generation and peptide mapping*
578 Sediment samples were prepared for metaproteome analysis as previously reported in
579 Graham et al. 2018 (23) and the protocol outlined by Nicora et al (75). As previously described
580 (55, 76), tandem mass spectrometry (MS/MS) spectra from all liquid chromatography tandem
581 mass spectrometry (LC-MS/MS) datasets were converted to ASCII text (.dta format) using
582 MSConvert (http://proteowizard.sourceforge.net/tools/msconvert.html) and the data files were
583 then interrogated via target-decoy approach (77) using MSGF+ (78). For protein identification,
584 spectra were searched against two files that included (i) 55 dereplicated MAG and (ii) 111
585 clustered vMAGs amino acid sequences. See **Text S1** for more details on metaproteome analysis.
586 Metaproteomic mapping results for MAGs and vMAGs can be found on **Table S7**.
587
588 *Data availability:*
589 The datasets supporting the conclusions of this article are publicly available. Sequencing data are
590 available in NCBI under BioProject PRJNA576070, with MAGs deposited under Biosamples
591 SAMN18867633-SAMN18867734 and 16S rRNA amplicon sequences under accession numbers
592 SRX9312157-SRX9312180. The raw annotations for each genome are deposited on Zenodo with
593 the following DOI: https://doi.org/10.5281/zenodo.5128772. 111 vMAGs have been deposited
594 NCBI under the BioProject ID PRJNA576070 and in Zenodo within the following DOI:

https://doi.org/10.5281/zenodo.5124937. Additionally, the dataset of freshwater viruses we used to cluster to the HUM-V viruses is also hosted on Zenodo with DOI https://doi.org/10.5281/zenodo.6310084. The raw annotations for each viral genome are on **Table S5**. Metaproteomics data are deposited in the MassIVE database under accession MSV000087330. Metabolomics data are publicly available and deposited in Zenodo doi https://doi.org/10.5281/zenodo.5076253. All scripts used in this manuscript are available at https://github.com/WrightonLabCSU/columbia_river.

Acknowledgements:

This work was supported by the Subsurface Biogeochemical Research (SBR) program (DE-SC0018170); the National Sciences Foundation Division of Biological Infrastructure [#1759874]; and the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Environmental System Science (ESS) program through subcontract from the River Corridor Scientific Focus Area project at Pacific Northwest National Laboratory. J.R.R. was partially supported by the National Science Foundation (NRT-DESE) [1450032], A Trans-Disciplinary Graduate Training Program in Biosensing and Computational Biology at Colorado State University. The NMR data, FTICR-MS data and MS-proteomics data in this work was collected using instrumentation in the Environmental Molecular Science Laboratory (grid.436923.9), a DOE Office of Science User Facility sponsored by the Office of Biological and Environmental Research and located at Pacific Northwest National Laboratory. Pacific Northwest National Lab is operated by Battelle for the DOE under Contract DE-AC05-76RL01830. Metagenomic sequencing for this research was performed by the Joint Genome Institute via a

large-scale sequencing award (Award 1781) and at the Genomics Shared Resource Core at The Ohio State University Comprehensive Cancer Center supported by P30 CA016058.

The authors would like to thank Tyson Claffey and Richard Wolfe for Colorado State University server management; Sandy Shew for management of computing resources retained from The Ohio State University Unity cluster; Dr. Pearlly Yan at the Genomics Shared Resource Core at The Ohio State University Comprehensive Cancer Center for management of metagenomic sequencing; and Dr. J John for the continuous support. The authors declare they have no competing interests. J.R.R. and M.A.B. contributed equally to this work.

REFERENCES

1. Boulton AJ, Findlay S, Marmonier P, Stanley EH, Valett HM. 1998. THE FUNCTIONAL SIGNIFICANCE OF THE HYPORHEIC ZONE IN STREAMS AND RIVERS. Annu Rev Ecol Syst 29:59–81.

2. Stegen JC, Johnson T, Fredrickson JK, Wilkins MJ, Konopka AE, Nelson WC, Arntzen EV, Chrisler WB, Chu RK, Fansler SJ, Graham EB, Kennedy DW, Resch CT, Tfaily M, Zachara J. 2018. Influences of organic carbon speciation on hyporheic corridor biogeochemistry and microbial ecology. Nat Commun 9:585.

3. Newcomer ME, Hubbard SS, Fleckenstein JH, Maier U, Schmidt C, Thullner M, Ulrich C, Flipo N, Rubin Y. 2018. Influence of hydrological perturbations and riverbed sediment characteristics on hyporheic zone respiration of CO2 and N2. Journal of Geophysical Research: Biogeosciences 123:902–922.

638 4. Hu M, Chen D, Dahlgren RA. 2016. Modeling nitrous oxide emission from rivers: a global
639 assessment. Glob Chang Biol 22:3566–3582.

640 5. Raymond PA, Hartmann J, Lauerwald R, Sobek S, McDonald C, Hoover M, Butman D,
641 Striegl R, Mayorga E, Humborg C, Kortelainen P, Dürr H, Meybeck M, Ciais P, Guth P.
642 2013. Global carbon dioxide emissions from inland waters. Nature 503:355–359.

643 6. Gómez-Gener L, Rocher-Ros G, Battin T, Cohen MJ, Dalmagro HJ, Dinsmore KJ, Drake
644 TW, Duvert C, Enrich-Prast A, Horgby Å, Johnson MS, Kirk L, Machado-Silva F, Marzolf
645 NS, McDowell MJ, McDowell WH, Miettinen H, Ojala AK, Peter H, Pumpanen J, Ran L,
646 Riveros-Iregui DA, Santos IR, Six J, Stanley EH, Wallin MB, White SA, Sponseller RA.
647 2021. Global carbon dioxide efflux from rivers enhanced by high nocturnal emissions. Nat
648 Geosci 14:289–294.

649 7. Lewandowski J, Arnon S, Banks E, Batelaan O, Betterle A, Broecker T, Coll C, Drummond
650 JD, Gaona Garcia J, Galloway J, Gomez-Velez J, Grabowski RC, Herzog SP, Hinkelmann R,
651 Höhne A, Hollender J, Horn MA, Jaeger A, Krause S, Löchner Prats A, Magliozzi C,
652 Meinikmann K, Mojarrad BB, Mueller BM, Peralta-Maraver I, Popp AL, Posselt M,
653 Putschew A, Radke M, Raza M, Riml J, Robertson A, Rutere C, Schaper JL, Schirmer M,
654 Schulz H, Shanafield M, Singh T, Ward AS, Wolke P, Wörman A, Wu L. 2019. Is the
655 Hyporheic Zone Relevant beyond the Scientific Community? Water 11:2230.

656 8. Raes EJ, Karsh K, Kessler AJ, Cook PLM, Holmes BH, van de Kamp J, Bodrossy L, Bissett
657 A. 2020. Can We Use Functional Genetics to Predict the Fate of Nitrogen in Estuaries? Front
658 Microbiol 11:1261.

659 9. Hou Z, Nelson WC, Stegen JC, Murray CJ, Arntzen E, Crump AR, Kennedy DW, Perkins
660 MC, Scheibe TD, Fredrickson JK, Zachara JM. 2017. Geochemical and Microbial
661 Community Attributes in Relation to Hyporheic Zone Geological Facies. Sci Rep 7:12006.

662 10. Nelson AR, Sawyer AH, Gabor RS, Saup CM, Bryant SR, Harris KD, Briggs MA, Williams
663 KH, Wilkins MJ. 2019. Heterogeneity in Hyporheic Flow, Pore Water Chemistry, and
664 Microbial Community Composition in an Alpine Streambed. J Geophys Res Biogeosci
665 124:3465–3478.

666 11. Sackett JD, Shope CL, Bruckner JC, Wallace J, Cooper CA, Moser DP. 2019. Microbial
667 Community Structure and Metabolic Potential of the Hyporheic Zone of a Large Mid-Stream
668 Channel Bar. Geomicrobiol J 36:765–776.

669 12. Sun S, Jones RB, Fodor AA. 2020. Inference-based accuracy of metagenome prediction tools
670 varies across sample types and functional categories. Microbiome 8:46.

671 13. Djemiel C, Maron P-A, Terrat S, Dequiedt S, Cottin A, Ranjard L. 2022. Inferring microbiota
672 functions from taxonomic genes: a review. Gigascience 11.

673 14. Thornton PE, Lamarque J-F, Rosenbloom NA, Mahowald NM. 2007. Influence of carbon-
674 nitrogen cycle coupling on land model response to CO2fertilization and climate variability.
675 Global Biogeochem Cycles 21.

676 15. Weitz JS, Wilhelm SW. 2012. Ocean viruses and their effects on microbial communities and
677 biogeochemical cycles. F1000 Biol Rep 4:17.

678 16. Peduzzi P. 2016. Virus ecology of fluvial systems: a blank spot on the map? Biol Rev Camb
679 Philos Soc 91:937–949.

680 17. Pollard PC, Ducklow H. 2011. Ultrahigh bacterial production in a eutrophic subtropical
681 Australian river: Does viral lysis short-circuit the microbial loop? Limnol Oceanogr 56:1115–
682 1129.

683 18. Ma L, Sun R, Mao G, Yu H, Wang Y. 2013. Seasonal and spatial variability of virioplanktonic
684 abundance in Haihe River, China. Biomed Res Int 2013:526362.

685 19. Samson R, Shah M, Yadav R, Sarode P, Rajput V, Dastager SG, Dharne MS, Khairnar K.
686 2019. Metagenomic insights to understand transient influence of Yamuna River on taxonomic
687 and functional aspects of bacterial and archaeal communities of River Ganges. Sci Total
688 Environ 674:288–299.

689 20. Huber DH, Ugwuanyi IR, Malkaram SA, Montenegro-Garcia NA, Lhilhi Noundou V,
690 Chavarria-Palma JE. 2018. Metagenome Sequences of Sediment from a Recovering
691 Industrialized Appalachian River in West Virginia. Genome Announc 6.

692 21. Liu S, Wang H, Chen L, Wang J, Zheng M, Liu S, Chen Q, Ni J. 2020. Comammox Nitrospira
693 within the Yangtze River continuum: community, biogeography, and ecological drivers.
694 ISME J https://doi.org/10.1038/s41396-020-0701-8.

695 22. Black EM, Just CL. 2018. The Genomic Potentials of NOB and Comammox Nitrospira in
696 River Sediment Are Impacted by Native Freshwater Mussels. Front Microbiol 9:2061.

697 23. Graham EB, Crump AR, Kennedy DW, Arntzen E, Fansler S, Purvine SO, Nicora CD, Nelson
698 W, Tfaily MM, Stegen JC. 2018. Multi'omics comparison reveals metabolome biochemistry,
699 not microbiome composition or gene expression, corresponds to elevated biogeochemical
700 function in the hyporheic zone. Science of the total environment 642:742–753.

701 24. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, Schulz
702 F, Jarett J, Rivers AR, Eloe-Fadrosh EA, Tringe SG, Ivanova NN, Copeland A, Clum A,
703 Becraft ED, Malmstrom RR, Birren B, Podar M, Bork P, Weinstock GM, Garrity GM,
704 Dodsworth JA, Yooseph S, Sutton G, Glöckner FO, Gilbert JA, Nelson WC, Hallam SJ,
705 Jungbluth SP, Ettema TJG, Tighe S, Konstantinidis KT, Liu W-T, Baker BJ, Rattei T, Eisen
706 JA, Hedlund B, McMahon KD, Fierer N, Knight R, Finn R, Cochrane G, Karsch-Mizrachi I,
707 Tyson GW, Rinke C, Genome Standards Consortium, Lapidus A, Meyer F, Yilmaz P, Parks
708 DH, Eren AM, Schriml L, Banfield JF, Hugenholtz P, Woyke T. 2017. Minimum information
709 about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG)
710 of bacteria and archaea. Nat Biotechnol 35:725–731.

711 25. Yang F, Bogdanov B, Strittmatter EF, Vilkov AN, Gritsenko M, Shi L, Elias DA, Ni S,
712 Romine M, Paša-Tolić L. 2005. Characterization of Purified c-Type Heme-Containing
713 Peptides and Identification of c-Type Heme-Attachment Sites in Shewanella o neidenis
714 Cytochromes Using Mass Spectrometry. Journal of proteome research 4:846–854.

715 26. Cordero PRF, Bayly K, Man Leung P, Huang C, Islam ZF, Schittenhelm RB, King GM,
716 Greening C. 2019. Atmospheric carbon monoxide oxidation is a widespread mechanism
717 supporting microbial survival. ISME J 13:2868–2881.

718 27. Zafiriou OC, Andrews SS, Wang W. 2003. Concordant estimates of oceanic carbon monoxide
719 source and sink processes in the Pacific yield a balanced global "blue-water" CO budget.
720 Global Biogeochem Cycles 17.

721 28. Haszpra L, Ferenczi Z, Barcza Z. 2019. Estimation of greenhouse gas emission factors based
722 on observed covariance of CO2, CH4, N2O and CO mole fractions. Environmental Sciences
723 Europe 31:1–12.

724 29. Xia X, Zhang S, Li S, Zhang L, Wang G, Zhang L, Wang J, Li Z. 2018. The cycle of nitrogen
725 in river systems: sources, transformation, and flux. Environ Sci Process Impacts 20:863–891.

726 30. Strauss EA, Lamberti GA. 2000. Regulation of nitrification in aquatic sediments by organic
727 carbon. Limnol Oceanogr 45:1854–1859.

728 31. Brust GE. 2019. Chapter 9 - Management Strategies for Organic Vegetable Fertility, p. 193–
729 212. *In* Biswas, D, Micallef, SA (eds.), Safety and Practice for Organic Food. Academic
730 Press.

731 32. Verhagen FJ, Laanbroek HJ. 1991. Competition for Ammonium between Nitrifying and
732 Heterotrophic Bacteria in Dual Energy-Limited Chemostats. Appl Environ Microbiol
733 57:3255–3263.

734 33. Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, Dao P, Sahinalp SC, Ester M, Foster
735 LJ, Brinkman FSL. 2010. PSORTb 3.0: improved protein subcellular localization prediction
736 with refined localization subcategories and predictive capabilities for all prokaryotes.
737 Bioinformatics 26:1608–1615.

738 34. Emerson JB, Roux S, Brum JR, Bolduc B, Woodcroft BJ, Jang HB, Singleton CM, Solden
739 LM, Naas AE, Boyd JA, Hodgkins SB, Wilson RM, Trubl G, Li C, Frolking S, Pope PB,
740 Wrighton KC, Crill PM, Chanton JP, Saleska SR, Tyson GW, Rich VI, Sullivan MB. 2018.
741 Host-linked soil viral ecology along a permafrost thaw gradient. Nat Microbiol 3:870–880.

742 35. Solden LM, Naas AE, Roux S, Daly RA, Collins WB, Nicora CD, Purvine SO, Hoyt DW,
743 Schückel J, Jørgensen B, Willats W, Spalinger DE, Firkins JL, Lipton MS, Sullivan MB,
744 Pope PB, Wrighton KC. 2018. Interspecies cross-feeding orchestrates carbon degradation in
745 the rumen ecosystem. Nature Microbiology 3:1274–1284.

746 36. Strauss EA, Richardson WB, Bartsch LA, Cavanaugh JC, Bruesewitz DA, Imker H, Heinz
747 JA, Soballe DM. 2004. Nitrification in the Upper Mississippi River: patterns, controls, and
748 contribution to the NO3- budget. Journal of the North American Benthological Society 23:1–
749 14.

750 37. Stegen JC, Fredrickson JK, Wilkins MJ, Konopka AE, Nelson WC, Arntzen EV, Chrisler
751 WB, Chu RK, Danczak RE, Fansler SJ, Kennedy DW, Resch CT, Tfaily M. 2016.
752 Groundwater-surface water mixing shifts ecological assembly processes and stimulates
753 organic carbon turnover. Nat Commun 7:11237.

754 38. Bartossek R, Nicol GW, Lanzen A, Klenk H-P, Schleper C. 2010. Homologues of nitrite
755 reductases in ammonia-oxidizing archaea: diversity and genomic context. Environmental
756 microbiology 12:1075–1088.

757 39. Kozlowski JA, Stieglmeier M, Schleper C, Klotz MG, Stein LY. 2016. Pathways and key
758 intermediates required for obligate aerobic ammonia-dependent chemolithotrophy in bacteria
759 and Thaumarchaeota. ISME J 10:1836–1845.

760 40. Martens-Habbena W, Qin W, Horak REA, Urakawa H, Schauer AJ, Moffett JW, Armbrust
761 EV, Ingalls AE, Devol AH, Stahl DA. 2015. The production of nitric oxide by marine
762 ammonia-oxidizing archaea and inhibition of archaeal ammonia oxidation by a nitric oxide
763 scavenger. Environ Microbiol 17:2261–2274.

764 41. Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, Probst AJ, Thomas BC, Singh
765 A, Wilkins MJ, Karaoz U, Brodie EL, Williams KH, Hubbard SS, Banfield JF. 2016.
766 Thousands of microbial genomes shed light on interconnected biogeochemical processes in
767 an aquifer system. Nat Commun 7:13219.

768 42. Jones CM, Spor A, Brennan FP, Breuil M-C, Bru D, Lemanceau P, Griffiths B, Hallin S,
769 Philippot L. 2014. Recently identified microbial guild mediates soil N2O sink capacity. Nat
770 Clim Chang 4:801–805.

771 43. Conthe M, Wittorf L, Kuenen JG, Kleerebezem R, van Loosdrecht MCM, Hallin S. 2018.
772 Life on N2O: deciphering the ecophysiology of N2O respiring bacterial communities in a
773 continuous culture. ISME J 12:1142–1153.

774 44. Hallin S, Philippot L, Löffler FE, Sanford RA, Jones CM. 2018. Genomics and Ecology of
775 Novel N2O-Reducing Microorganisms. Trends Microbiol 26:43–55.

776 45. Orellana LH, Rodriguez-R LM, Higgins S, Chee-Sanford JC, Sanford RA, Ritalahti KM,
777 Löffler FE, Konstantinidis KT. 2014. Detecting nitrous oxide reductase (NosZ) genes in soil

778 metagenomes: method development and implications for the nitrogen cycle. MBio 5:e01193-
779 14.

780 46. Li M, Li R, Gao Y, Resch CT, Qian W-J, Shi T, Shi L, Liu H, Liu C. 2020. Nitrate
781 bioreduction dynamics in hyporheic zone sediments under cyclic changes of chemical
782 compositions. J Hydrol 585:124836.

783 47. Nelson WC, Graham EB, Crump AR, Fansler SJ, Arntzen EV, Kennedy DW, Stegen JC.
784 2020. Distinct temporal diversity profiles for nitrogen cycling genes in a hyporheic
785 microbiome. PLoS One 15:e0228165.

786 48. Kuypers MMM, Marchant HK, Kartal B. 2018. The microbial nitrogen-cycling network. Nat
787 Rev Microbiol 16:263–276.

788 49. Moon K, Jeon JH, Kang I, Park KS, Lee K, Cha C-J, Lee SH, Cho J-C. 2020. Freshwater
789 viral metagenome reveals novel and functional phage-borne antibiotic resistance genes.
790 Microbiome 8:75.

791 50. Wolf YI, Silas S, Wang Y, Wu S, Bocek M, Kazlauskas D, Krupovic M, Fire A, Dolja VV,
792 Koonin EV. 2020. Doubling of the known set of RNA viruses by metagenomic analysis of an
793 aquatic virome. Nat Microbiol https://doi.org/10.1038/s41564-020-0755-4.

794 51. Püttker S, Kohrs F, Benndorf D, Heyer R, Rapp E, Reichl U. 2015. Metaproteomics of
795 activated sludge from a wastewater treatment plant--A pilot study. Proteomics 15:3596–3601.

796 52. Rudney JD, Xie H, Rhodus NL, Ondrey FG, Griffin TJ. 2010. A metaproteomic analysis of
797 the human salivary microbiota by three-dimensional peptide fractionation and tandem mass
798 spectrometry. Molecular oral microbiology 25:38–49.

799 53. Shaffer M, Borton MA, McGivern BB, Zayed AA, La Rosa SL, Solden LM, Liu P, Narrowe
800 AB, Rodríguez-Ramos J, Bolduc B, Gazitúa MC, Daly RA, Smith GJ, Vik DR, Pope PB,
801 Sullivan MB, Roux S, Wrighton KC. 2020. DRAM for distilling microbial metabolism to
802 automate the curation of microbiome function. Nucleic Acids Res 48:8883–8900.

803 54. Murphy CL, Sheremet A, Dunfield PF, Spear JR, Stepanauskas R, Woyke T, Elshahed MS,
804 Youssef NH. 2021. Genomic Analysis of the Yet-Uncultured Binatota Reveals Broad
805 Methylotrophic, Alkane-Degradation, and Pigment Production Capacities. Mbio 12.

806 55. Borton MA, Hoyt DW, Roux S, Daly RA, Welch SA, Nicora CD, Purvine S, Eder EK,
807 Hanson AJ, Sheets JM, Morgan DM, Wolfe RA, Sharma S, Carr TR, Cole DR, Mouser PJ,
808 Lipton MS, Wilkins MJ, Wrighton KC. 2018. Coupled laboratory and field investigations
809 resolve microbial interactions that underpin persistence in hydraulically fractured shales. Proc
810 Natl Acad Sci U S A 115:E6585–E6594.

811 56. Weljie AM, Newton J, Mercier P, Carlson E, Slupsky CM. 2006. Targeted profiling:
812 quantitative analysis of 1H NMR metabolomics data. Anal Chem 78:4430–4442.

813 57. Peng Y, Leung HCM, Yiu SM, Chin FYL. 2012. IDBA-UD: a de novo assembler for single-
814 cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28:1420–
815 1428.

816 58. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. 2017. metaSPAdes: a new versatile
817 metagenomic assembler. Genome Res 27:824–834.

818 59. Kang D, Li F, Kirton ES, Thomas A, Egan RS, An H, Wang Z. 2019. MetaBAT 2: an adaptive
819 binning algorithm for robust and efficient genome reconstruction from metagenome
820 assemblies. e27522v1. PeerJ Preprints.

821 60. Wu M, Scott AJ. 2012. Phylogenomic analysis of bacterial and archaeal sequences with
822 AMPHORA2. Bioinformatics 28:1033–1034.

823 61. Olm MR, Brown CT, Brooks B, Banfield JF. 2017. dRep: a tool for fast and accurate genomic
824 comparisons that enables improved genome recovery from metagenomes through de-
825 replication. ISME J 11:2864–2868.

826 62. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. 2018. GTDB-Tk: a toolkit to classify
827 genomes with the Genome Taxonomy Database. Bioinformatics 36:1925–1927.

828 63. Castelle CJ, Hug LA, Wrighton KC, Thomas BC, Williams KH, Wu D, Tringe SG, Singer
829 SW, Eisen JA, Banfield JF. 2013. Extraordinary phylogenetic diversity and metabolic
830 versatility in aquifer sediment. Nat Commun 4:1–10.

831 64. Armenteros JJA, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, von Heijne
832 G, Nielsen H. 2019. SignalP 5.0 improves signal peptide predictions using deep neural
833 networks. Nature biotechnology 37:420–423.

834 65. Bendtsen JD, Kiemer L, Fausbøll A, Brunak S. 2005. Non-classical protein secretion in
835 bacteria. BMC Microbiol 5:58.

836 66. Roux S, Enault F, Hurwitz BL, Sullivan MB. 2015. VirSorter: mining viral signal from
837 microbial genomic data. PeerJ 3:e985.

838 67. Roux S, Adriaenssens EM, Dutilh BE, Koonin EV, Kropinski AM, Krupovic M, Kuhn JH,
839 Lavigne R, Brister JR, Varsani A, Amid C, Aziz RK, Bordenstein SR, Bork P, Breitbart M,
840 Cochrane GR, Daly RA, Desnues C, Duhaime MB, Emerson JB, Enault F, Fuhrman JA,
841 Hingamp P, Hugenholtz P, Hurwitz BL, Ivanova NN, Labonté JM, Lee K-B, Malmstrom RR,
842 Martinez-Garcia M, Mizrachi IK, Ogata H, Páez-Espino D, Petit M-A, Putonti C, Rattei T,
843 Reyes A, Rodriguez-Valera F, Rosario K, Schriml L, Schulz F, Steward GF, Sullivan MB,
844 Sunagawa S, Suttle CA, Temperton B, Tringe SG, Thurber RV, Webster NS, Whiteson KL,
845 Wilhelm SW, Wommack KE, Woyke T, Wrighton KC, Yilmaz P, Yoshida T, Young MJ,
846 Yutin N, Allen LZ, Kyrpides NC, Eloe-Fadrosh EA. 2018. Minimum Information about an
847 Uncultivated Virus Genome (MIUViG). Nat Biotechnol https://doi.org/10.1038/nbt.4306.

848 68. Merchant N, Lyons E, Goff S, Vaughn M, Ware D, Micklos D, Antin P. 2016. The iPlant
849 Collaborative: Cyberinfrastructure for Enabling Data to Discovery for the Life Sciences.
850 PLoS Biol 14:e1002342.

851 69. Bin Jang H, Bolduc B, Zablocki O, Kuhn JH, Roux S, Adriaenssens EM, Brister JR,
852 Kropinski AM, Krupovic M, Lavigne R, Turner D, Sullivan MB. 2019. Taxonomic
853 assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks.
854 Nat Biotechnol 37:632–639.

855 70. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. 2015. The Phyre2 web portal
856 for protein modeling, prediction and analysis. Nat Protoc 10:845–858.

857 71. Ahlgren NA, Ren J, Lu YY, Fuhrman JA, Sun F. 2017. Alignment-free oligonucleotide
858 frequency dissimilarity measure improves prediction of hosts from metagenomically-derived
859 viral sequences. Nucleic acids research 45:39–53.

860 72. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods
861 9:357–359.

862 73. Daly RA, Borton MA, Wilkins MJ, Hoyt DW, Kountz DJ, Wolfe RA, Welch SA, Marcus
863 DN, Trexler RV, MacRae JD, Krzycki JA, Cole DR, Mouser PJ, Wrighton KC. 2016.
864 Microbial metabolisms in a 2.5-km-deep ecosystem created by hydraulic fracturing in shales.
865 Nature Microbiology 1:16146.

866 74. Lê Cao K-A, Rossouw D, Robert-Granié C, Besse P. 2008. A sparse PLS for variable
867 selection when integrating omics data. Statistical applications in genetics and molecular
868 biology 7.

869 75. Nicora CD, Burnum-Johnson KE, Nakayasu ES, Casey CP, White RA III, Chowdhury TR,
870 Kyle JE, Kim Y-M, Smith RD, Metz TO. 2018. The MPLEx protocol for multi-omic analyses
871 of soil samples. Journal of visualized experiments: JoVE.

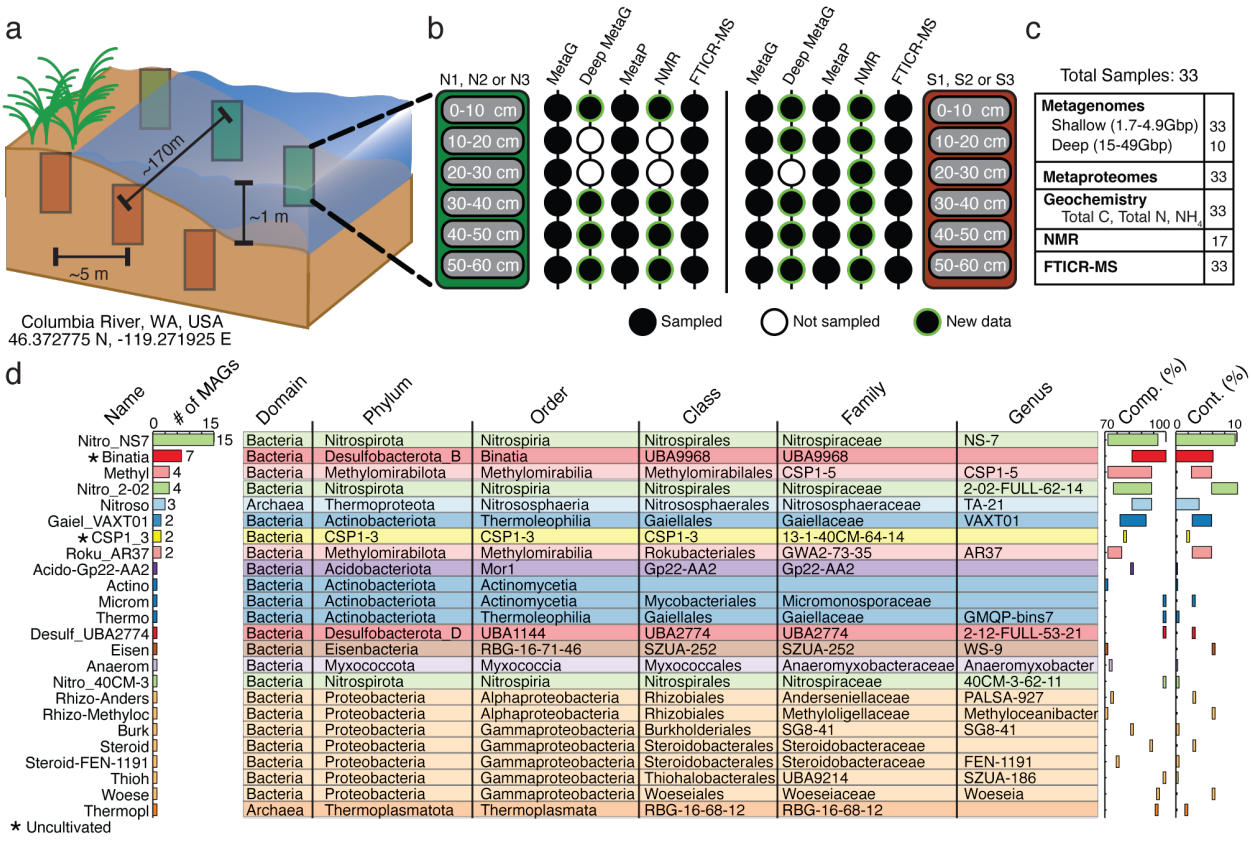
872 76. McGivern BB, Tfaily MM, Borton MA, Kosina SM, Daly RA, Nicora CD, Purvine SO, Wong
873 AR, Lipton MS, Hoyt DW, Northen TR, Hagerman AE, Wrighton KC. 2021. Decrypting
874 bacterial polyphenol metabolism in an anoxic wetland soil. Nat Commun 12:1–16.

875 77. Elias JE, Gygi SP. 2010. Target-decoy search strategy for mass spectrometry-based
876 proteomics, p. 55–71. *In* Proteome bioinformatics. Springer.

877 78. Kim S, Pevzner PA. 2014. MS-GF+ makes progress towards a universal database search tool
878 for proteomics. Nat Commun 5:5277.

879

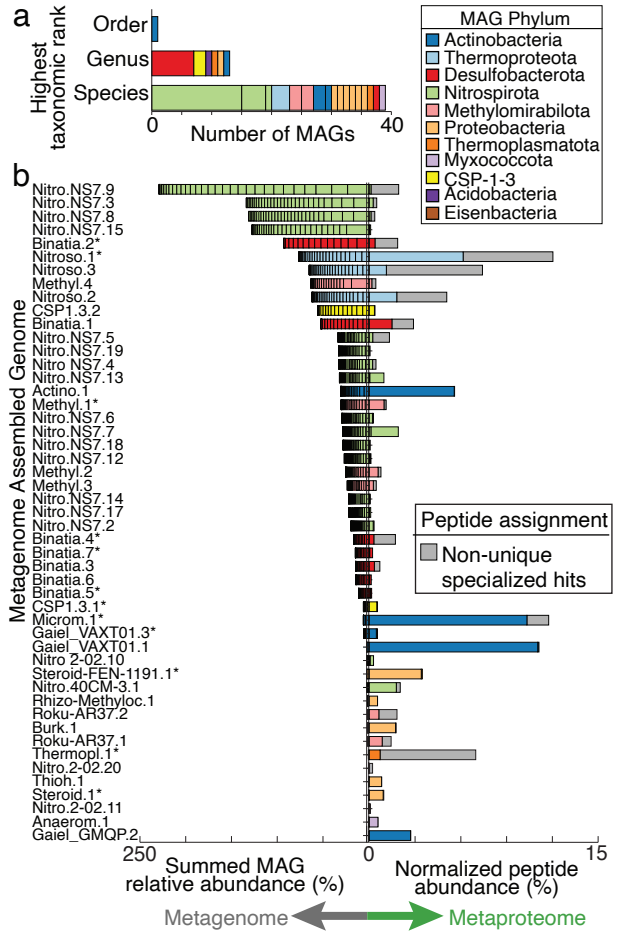
880 ***Main text figures and legends:***



881

882 **Fig. 1. Overview of hyporheic zone sampling and the microbial genomes included in the**
883 **HUM-V database a)** Samples were collected from two transects, each with 3 sediment cores,
884 with each core sectioned into 10-centimeter segments from 0-60 centimeters in depth and paired
885 with metaproteomics and geochemistry. **b)** Schematic of the data types available for each of the
886 depth samples within a core. Black-filled circles indicate depth samples for a particular data type,
887 open circles denote missing analyses (due to limited sample availability), and black-filled circles
888 with green outlines indicate new sequencing that was performed as part of this project and not

889 available in Graham et al. **c)** Summarized catalog of the total samples for each analysis. **d)** The
890 total recovered genomes (# of MAGs), taxonomic string, inferred genome completion (Comp.,
891 %), and contamination (Cont., %) for the dereplicated microbial genomes retained in HUM-V.
892 Asterisks on names indicate uncultivated lineages.
893



894

895 **Fig. 2. Bacterial and archaeal members of HUM-V database coupled to metaproteomics**
896 **reveals active members in the hyporheic zone microbiome. a)** Stacked bar graph indicates the
897 taxonomic novelty of the de-replicated MAGs colored by phylum and stacked according to the
898 first empty position within the taxonomic string provided by the Genome Taxonomy Database
899 GTDB-Tk. Each color represents a MAG phylum according to the MAG Phylum legend, a

coloring maintained across this manuscript. **b)** Butterfly plot reports the summed genomic relative abundance across all samples (left side) and the normalized mean proteomic relative abundance (right side) for dereplicated MAGs (55 total, 49 shown), with bars colored by phylum. MAGs that contain a partial or complete 16S rRNA sequence are denoted with and asterisk (*). Non-unique specialized peptide assignment is defined in the methods and is shown with grey bars.

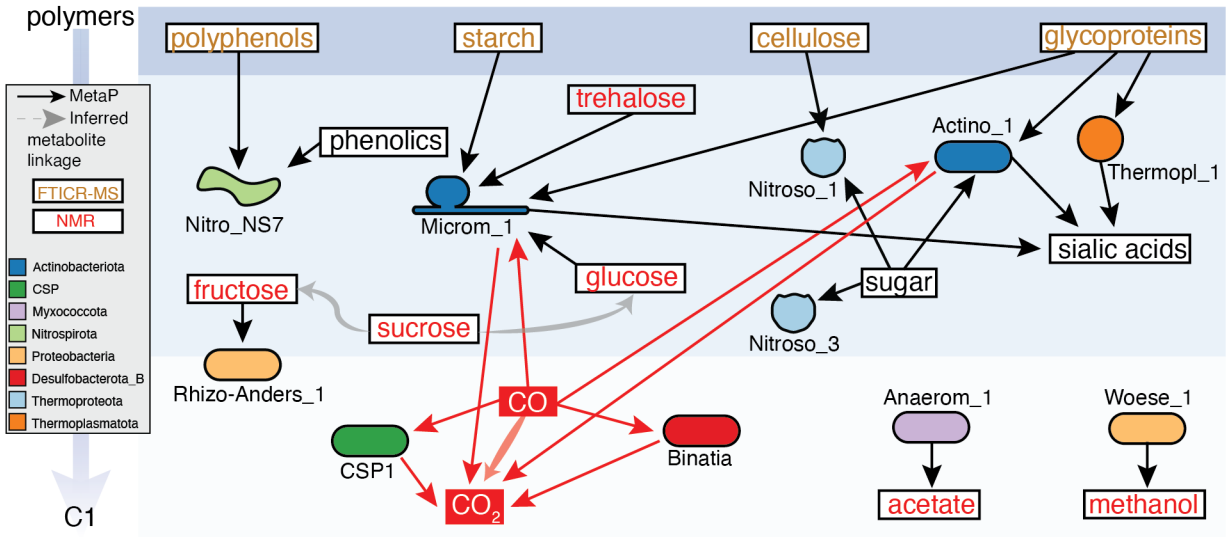


Fig. 3. Metaproteomics and metabolomics reveal microbial metabolic handoffs that support carbon cycling in river sediments. Detected metabolites are given in boxes, with NMR-detected compounds listed in red, polymers from FTICR-MS in orange, undetected metabolites in black. These polymers were inferred from FTICR-MS assigned biochemical classes and the specificity of CAZymes detected in metaproteome, where starch and cellulose were within the "polysaccharide-like" class and glycoproteins were in the "amino sugar-like" class. MAG-resolved metaproteome information is indicated by solid arrows, with MAG shape colored by phylum. Red arrows indicate processes leading to CO_2 production, while black arrows indicate other microbial carbon transforming genes expressed in the proteome. Shaded bold arrows

indicate chemical connections, where (1) grey indicates a metabolite was detected along with putative downstream products (e.g., sucrose conversion to glucose) but metaproteomic lacked evidence for the transformation or (2) red indicates a metabolite not measured but metaproteomic evidence supported transformation (e.g., CO conversion to CO_2).

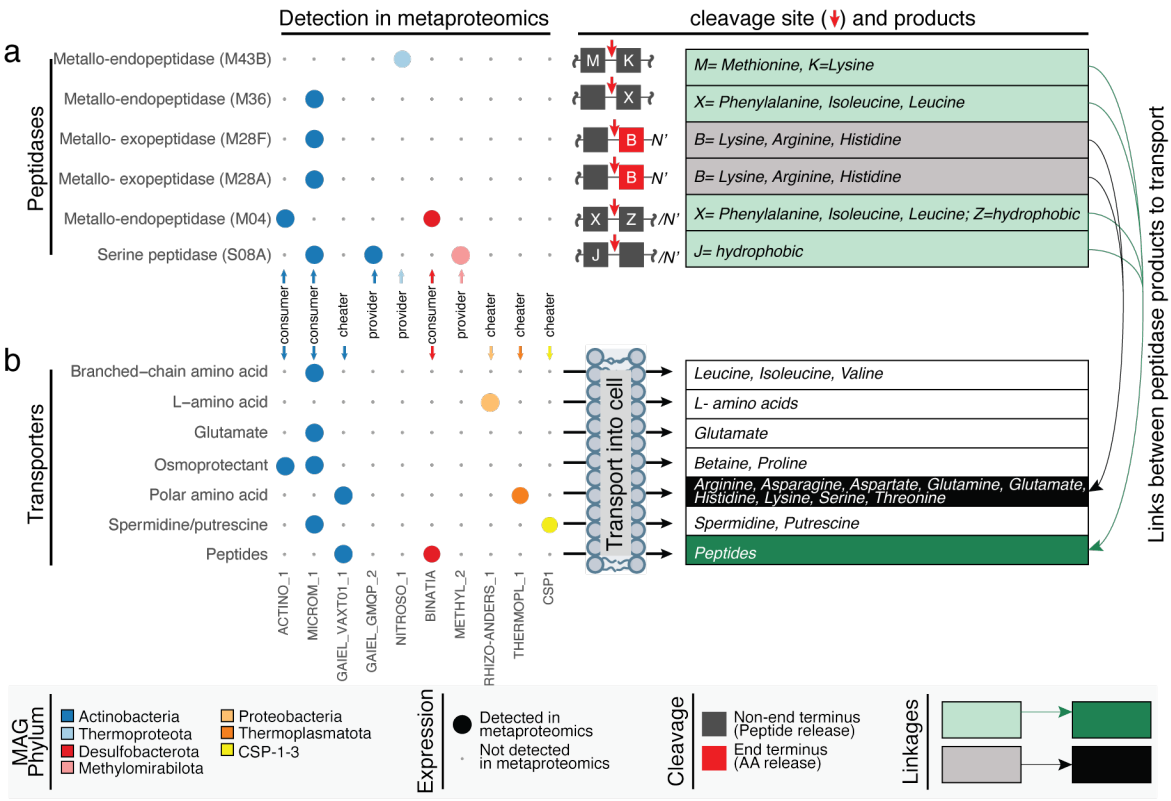


Fig. 4. Organic nitrogen mineralization and cellular transport are active microbial processes in river sediments. Bubble plots indicate the expressed genes that were uniquely assigned to specific genome including (**a**) extracellular peptidases and (**b**) cellular transporters for organic nitrogen. Unique peptides detected in at least 3 samples are reported as bubbles and colored by phylum. Table on the right shows putative amino acids cleaved or transported by respective peptidases or transporters, shades of color (green or grey) denote peptides that are cleaved into amino acids that could be transported, providing linkages between extracellular

organic nitrogen transformation and transport of nitrogen into the cell. White boxes indicate an organic nitrogen transporter that recruited peptides but could not be linked to outputs of specific peptidases.

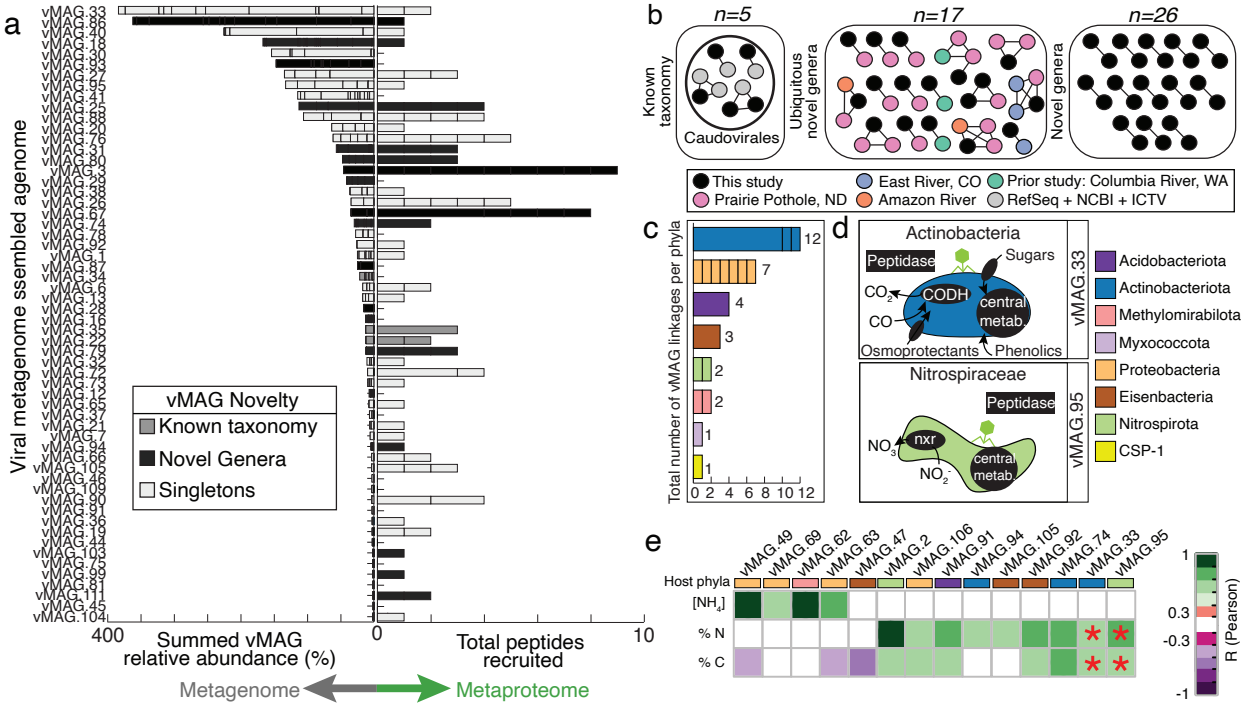


Fig. 5 Viruses in HUM-V are active, taxonomically novel, and can play key roles in microbial host metabolism and river geochemistry a) Butterfly plot showing summed genomic relative abundance (left side) and total peptides recruited for each vMAG population (total 111, 58 shown). Bars are colored by clustering of vMAGs from this study with (i) viruses of known taxonomy in RefSeq, ICTV and NCBI Taxonomy (dark grey), (ii) novel genera, both only from this study and ubiquitous (black), and no clustering from any database (light grey, singletons). **b)** Similarity network of the few vMAGs from our study (black) that clustered to viruses belonging to the default RefSeq, ICTV and NCBI Taxonomy databases (gray), as well as clustering of our vMAGs to other freshwater, publicly available dataset we mined (Pink, Purple, Orange, and Turquoise). The remaining clusters of viruses that were novel (e.g., did not cluster

944 with prior viral genomes) are shown, with the full network file including singletons shown in
945 **Table S5**. **c)** Stacked bar chart of the total number of vMAGs (n=32) that have putative host
946 linkages. Each bar represents a phylum and lines within bars indicate the linkages for specific
947 genomes within each phylum. For example, there are three genomes within the Actinobacteriota
948 phylum that collectively have 12 viral linkages and of the three genomes that have linkages, one
949 host has 10 viruses linked, while the other two hosts have 1 virus linked. **d)** Genome cartoons of
950 microbial metabolisms for two representative genomes that could be predated by vMAGs, with
951 the genes shown in black text boxes denoting processes detected in proteomics. These two
952 microorganisms were selected as examples because they were active members in shaping carbon
953 and nitrogen metabolism in these river sediments but could be impacted by viral predation; other
954 virus-host relationships are reported in **Supplementary Figure 9**. **e)** Heatmap reports
955 correlations between a subset of vMAGs with rectangle colors denoting the putative phyla for the
956 respective host. Correlations between these vMAGs and ecosystem geochemistry (NH_4 μg/gram,
957 %N, %C) are reported with significant correlation coefficients denoted by purple-green shading
958 according to the legend. Red asterisks (*) indicate the vMAG relative abundance predicted a key
959 environmental variable by sparse partial least squares (sPLS) regression. Note two of these
960 predicted vMAGs are shown in **(d)**.
961

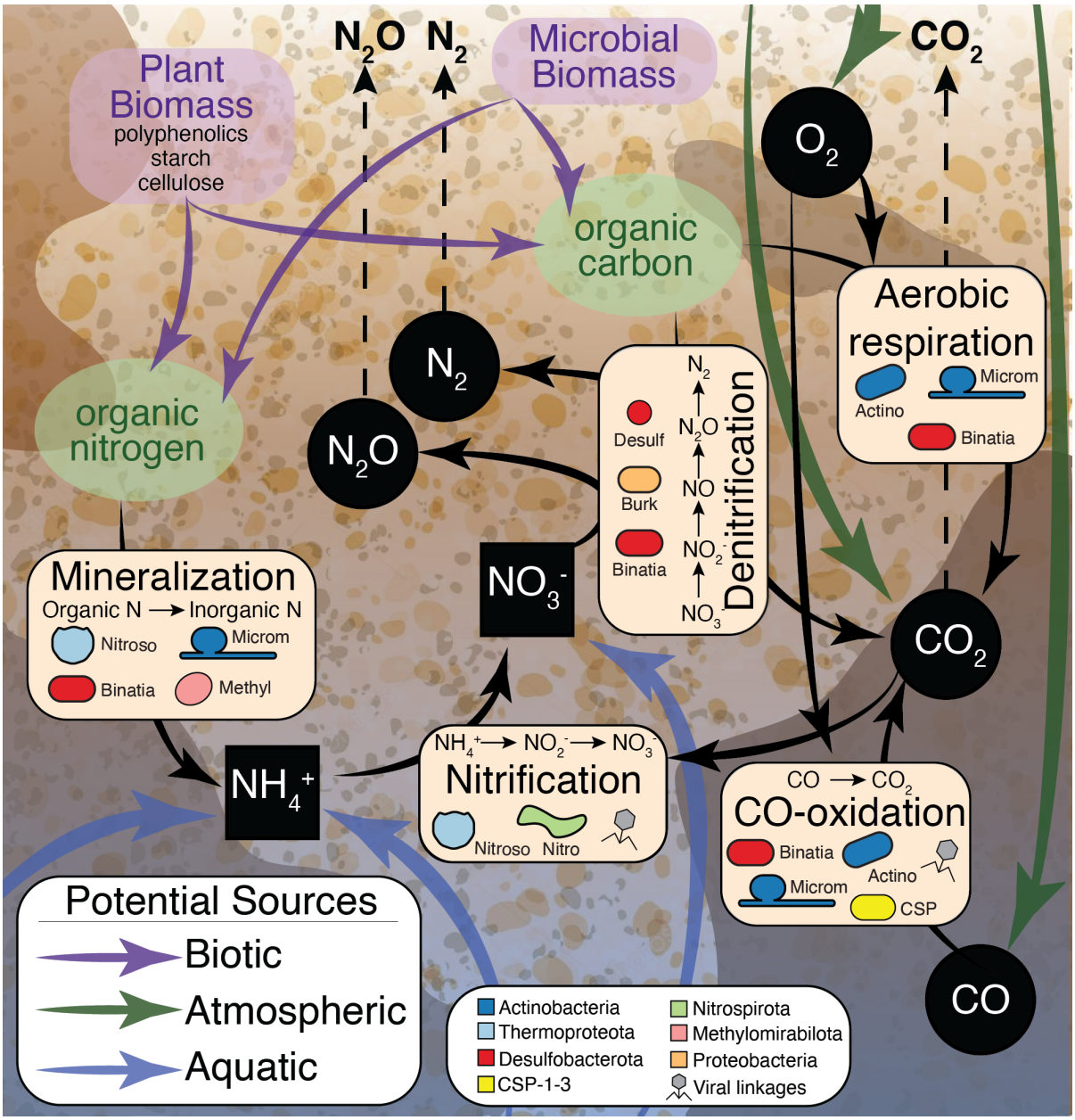


Fig. 6. Conceptual model uncovering microbes and processes contributing to carbon and nitrogen cycling in river sediments. Integration of multi-omic data uncovered the microbial and viral effects on carbon and nitrogen cycling in river sediments. Black arrows signify microbial transformations uncovered in our metaproteomic data. Specific processes (e.g., mineralization, nitrification, CO-oxidation, denitrification, and aerobic respiration) are highlighted in beige boxes, with microorganisms inferred to carry out the specific process

969 denoted by overlaid cell shapes colored by phylum. Prior to this research, little was known about
970 the specific enzymes and organisms responsible for river organic nitrogen mineralization and
971 CO-oxidation, thus this research adds new content to microbial roles in carbon and nitrogen
972 transformations in these systems. Possible (biotic, atmospheric, and aquatic) carbon and nitrogen
973 sources are shown by purple, green, and blue arrows respectively. Inorganic carbon and nitrogen
974 sources are shown by black squares (aqueous) and black circles (gaseous) with white text and
975 dashed arrows indicating possible gasses that could be released to the atmosphere. Processes that
976 could be impacted by viruses are marked with grey viral symbols.

977

978 **Supplementary Figure Legends**

979 **Figure S1: DRAM annotation of MAGs.** Heatmap showing the DRAM product output for
980 medium and high-quality genomes (n=102) from HUM-V. The interactive version of this
981 heatmap is available here: https://zenodo.org/record/5124964

982

983 **Figure S2: Resazurin reduction assay measurements.** Box and whisker plots showing the
984 values reported for a resazurin reduction assay. Different boxes represent the different depths (0-
985 60cm) that were sampled per sediment core.

986

987 **Figure S3: NMDS and Venn Diagram of differences in transect (N / S) or Depth (0-60cm)**
988 **for MAG proteome. a)** NMDS of north (green) and south (orange) transects of recruited
989 proteome peptides for MAGs. ANOSIM is reported. **b)** NMDS of sediment core depth (0-60cm)
990 of recruited proteome peptides for MAGs. ANOSIM is reported. **b)** Euler diagram showing the

number of total proteins recruiting peptides in each transect, where the overlap represents the proportion of these proteins recruiting peptides in both transects. Only proteins recruiting 2 or more total peptides were included (n=898) to allow for recruitment of at least 1 peptide for a given protein to both transects or both depths and reduce false positives.

Figure S4: Relative abundance of FTICR-MS classes across samples. The relative abundance of biochemical classes identified in FTICR-MS across samples is not statistically structured by depth. Bar plots show the average relative abundance for each class, with error bars representing one standard deviation (n=33). Individual data points are plotted for each sample with point size increasing with depth. Peaks were classified as described in **Methods**. Raw data is provided in **Additional File 4.**

Figure S5: Binatia MAGs possess the ability to metabolize compounds related to carbon and nitrogen cycling. The seven HUM-V genomes reconstructed here encode diverse pathways for transforming phenolic compounds. Pathways are shown with boxes corresponding to each enzyme in the pathway, colored by the number of Binatia MAGs that encode each step. Gene information shown here is reported in detail in **Additional File 6**.

Figure S6: C:N ratio and total carbon and nitrogen measurements across depths reveal no significant differences. Box and whisker plots report concentrations. Graphs show C:N ratio, percent total carbon, and percent total nitrogen measurements across different depths (0-60cm). No significant differences were observed.

Figure S7: NMR detected metabolites and ammonium indicate prevalence of nitrogen-containing compounds, saccharides, organic acids, and alcohols. Bar graphs showing specific NMR metabolites and ammonium with the percent of total samples that they were found in. Shading denotes the different categories of detected compounds. For detailed information on how these were collected see methods section. Only metabolites detected in more than 4 samples are shown.

Figure S8: vMAG expressed peptides are not structured by transect of depth, and relative abundance NMDS of vMAGs is coordinate with relative abundance NMDS of MAGs: a) NMDS of north (green) and south (orange) transects of recruited proteome peptides for vMAGs. ANOSIM is reported. **b)** NMDS of sediment core depth (0-60cm) of recruited proteome peptides for vMAGs. ANOSIM is reported. **c**) Procrustes ordination of MAG and vMAG NMDS ordinations using relative abundance genome data across 33 shallower sequenced samples.

Figure S9: Virus-Host associations show viruses could infect key players in river sediment carbon and nitrogen cycling. a) Genome cartoons colored by their phyla. Gray dotted circles represent vMAGs that putatively infect each genome. **b)** A breakdown of each vMAG name for each of the infected host genomes. Colors match genome host phyla assignment. **c)** Bar chart showing the different types of AMGs identified in viral genomes with putative hosts, with the predicted substrate for each AMG denoted by colored boxes on the y-axis. **d)** Viruses encode AMGs which can expand the metabolic repertoire of their microbial hosts. Conceptual models of a putative glycoside hydrolase (GH) and a pectin lyase (PL) are shown. **e)** For one putative host, Steroidobacteraceae, with a linkage to a viral genome that encodes a PL1, a conceptual model of

an integrated metabolism is shown. Specifically, integration of vMAG.63 genome into the host genome may provide the capacity to cleave the pectin backbone (PL1), followed by oligo cleavage of pectin, resulting in galactose monomers for host cell metabolism.

Figure S10: sPLS regressions show vMAG abundance is better predictor of carbon and nitrogen relative to MAGs. sPLS plots showing predicted vs measured values for percent carbon (c_per) and percent nitrogen (n_per) for our samples using n=10 deep sequences. Shown sPLS are **a)** vMAGs only, **b)** MAGs only, and **c)** vMAGs and MAGs combined. Values corresponding to cor.test function output in R for the predicted and measured values are shown in each box (t, degrees of freedom, p-value, confidence interval, and correlation).

Figure S11: **Uncultured Binatia are widely dispersed across ecosystems and express nitrogen and carbon genes *in situ*. a)** Using the 16S rRNA gene (from Binatia_7), we inventoried the distribution of closely related species to our HUM-V genomes (>97% similarity) in the Sequence Read Archive (SRA) samples, uncovering the ecological distribution of these organisms from soils, as well as a wide variety of terrestrial, terrestrial-aquatic, marine samples, indicating the processes uncovered by proteomics here are likely applicable to a wide range of ecosystems. **b)** Metabolic genome cartoon for the major functions encoded by the Binatia MAGs. Dotted arrows are functions encoded in the metagenome, with black arrows corresponding to metaproteome-detected enzymes. (BCAA, Branched chain amino acids).

Supplementary Files Index:

Table S1: Metagenome read counts, accession numbers, biosamples.

Table S2: Geochemistry data and NMR metabolite data.

Text S1: Supplementary materials and methods

Table S3: FTICR-MS Data.

Table S4: MAGs. Completion information, accession numbers, annotations, Read mapping information, metabolism summary from DRAM.

Table S5: vMAGs. MIUViG information, accession numbers, annotations, read mapping information, vContact2 results, downloaded MetaG information for vContact2 analyses, virhostmatcher2 results.

Table S6: SPLS data and correlations results

Table S7: Metaproteomes. MAG Unique peptides, MAG unique peptide NSAF relative abundance, MAG Per genome NSAF Relative abundance, vMAG unique peptides