bioRxiv preprint doi: https://doi.org/10.64898/2026.10.06.756941; this version posted October 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Conserved synteny reveals a saccharolytic lifestyle across *Kiritimatiellia*

Aaron Oliver[a], Sheila Podell[a], Eric E. Allen[a,b#]

[a] Center for Marine Biotechnology and Biomedicine, Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA, USA

[b] Department of Molecular Biology, School of Biological Sciences, University of California San Diego, La Jolla, CA, USA

Address correspondence to Eric E. Allen, eallen@ucsd.edu

Running title: Kiritimatiellia similarity-colocalization networks

Keywords: Kiritimatiellia, Verrucomicrobiota, CAZyme gene clusters, nitrogen fixation, similarity-colocalization network

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.06.756941; this version posted October 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

Kiritimatiellia are a class of *Verrucomicrobiota* whose few cultured representatives are known for hosting large repertoires of carbohydrate-active enzymes and sulfatases. Many families of this class are restricted to metagenome-assembled genomes not yet discussed in literature, leaving their genome organization undescribed. Here, we survey 898 *Kiritimatiellia* genomes using similarity-colocalization networks (SCoNes), a data structure that captures conserved synteny among homologous genes across genomes. Using this approach, we investigate four metabolic adaptations to a glycan-rich environment: complex polysaccharide degradation, nitrogen cycling, hydrogenase activity, and natural product biosynthesis. Rather than unrelated functions, these capacities envelop a single anaerobic and saccharolytic strategy where sulfated glycan hydrolysis fuels fermentation, hydrogenases balance redox, and excess carbon supports nitrogen fixation.

Importance

Most microbes cannot be grown in the lab and are only identified by DNA fragments recovered from the environment, which makes it difficult to study the organization of their genomes. *Kiritimatiellia* are a class of sediment- and gut-associated bacteria that degrade and ferment complex sugars for energy, but our understanding of this lifestyle is based on a small number of cultured representatives. To understand how genes in core metabolic pathways are organized across hundreds of *Kiritimatiellia* genomes, we developed a new network analysis and visualization tool built to handle fragmented assemblies and flexible gene cluster boundaries. These network analyses reveal variable organization of catabolic enzymes and biosynthetic pathways, but a conserved core of metabolism that supports a lifestyle centered around polysaccharide utilization. These findings are of particular interest for understanding marine carbon cycling and discovering novel enzymes for the conversion of complex environmental sugars.

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.06.756941; this version posted October 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Introduction

Annotation of well-characterized marker genes can reliably predict metabolic potential from metagenomes, but the organization of those genes remains difficult to study[1,2]. Detecting conserved genome organization requires methods that scale to hundreds of genomes, tolerate fragmented assemblies, and function independently of curated cluster boundaries or cultured relatives[3–5]. Here, we address these needs with similarity-colocalization networks (SCoNes), a new data structure for comparing gene organization across large sets of related genomes. We apply SCoNes to *Kiritimatiellia*, a class of mostly uncultivated *Verrucomicrobiota*, to understand how the organization of their genomes relates to biosynthetic and catabolic capacity.

This style of inference has a long history: genomic context has served as an early predictor of protein function for decades[6–8], proving useful across problems as diverse as natural product discovery[9], antibiotic resistance[10], CRISPR-Cas systems[11], methane consumption[12], marine carbon cycling[13], and the annotation of metagenomic dark matter[14]. Meanwhile, homology search tools have matured beyond basic pairwise BLAST searches into structure-aware[15] and attention-based[16] similarity metrics, each partitioning sequence space differently. The sequence similarity network (SSN) harnesses these distance metrics to visualize the sequence space of individual protein families[17] or of entire genomes[18]. SSNs, through accessible webtools like EFI[19] and EFI-GNT[20], have been used to partition enzyme sets into functional subfamilies[21–23] for small numbers of genomes, but this strategy struggles to parse sets of genes that work together in complex pathways or similar enzymes whose substrate specificity depends on neighboring auxiliary genes.

Network approaches already underlie homology and localization resources such as the STRING database[24], but STRING encodes genomic context only for manually curated isolate genomes and relies heavily on prior experimental literature that most environmental lineages lack. Open-source tools that researchers can run on their own data fill more narrow niches: clinker[25,26], antiSMASH[27], and deepBGC[28] are tailored to biosynthetic gene cluster (BGC) annotation and

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.06.756941; this version posted October 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

depend on predefined marker genes and gene-order constraints, making them poorly suited for whole-genome comparisons of syntenic regions. Other tools like FlaGs[29] are more flexible in their metabolic scope but are rigid in their comparisons, interrogating a single query gene by considering only homologs from NCBI and immediate neighbors along a single contig. Despite this expanding toolkit, no single method visualizes conserved homologous gene neighborhoods across hundreds of genomes, a gap that becomes more consequential with the introduction of scalable whole-genome gene order comparison tools such as Spacedust[30].

The SCoNe data structure developed in this paper is designed to avoid these limitations. A SCoNe begins as a sequence similarity network constructed using any pairwise homology metric, collapses gene nodes into communities of homologous genes, and links those communities based on genomic position. Because the gene neighborhoods are supplied as input rather than internally calculated, a SCoNe can be built using cluster boundaries from a flexible variety of sources; an annotation tool, synteny blocks inferred *de novo*, or the immediate neighbors of a single gene of interest. Comparing homologous sets rather than individual genes and evaluating colocalization across many genomes, is intended to keep the network interpretable at scale and tolerant of fragmented metagenomic assemblies. We illustrate each of these input types in the Results section, with case studies spanning different metabolic pathways associated with an anaerobic lifestyle.

In marine and gut-associated microbes, anaerobes are often linked to the breakdown of complex polysaccharides using colocalized collections of carbohydrate-active enzymes. Polysaccharide utilization loci are gene clusters marked by the presence of CAZymes[31] and, in some environments[32], sulfatase genes. Unlike BGCs, CAZyme gene clusters (CGCs) lack conserved gene order and consistent marker genes across phyla[33], and they are disproportionately recovered from environmental samples whose fragmented, assembled contigs erode the performance of standard annotation tools[34]. Computational identification of CGCs is

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.06.756941; this version posted October 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

crucial for marine polysaccharides in particular, as fewer than 10% of experimentally validated CGCs act on marine substrates[35] and many marine polysaccharide-degrading microbes remain uncultured. Leveraging this metagenomic dark matter requires scalable methods that can link uncultured lineages to their degradative potential.

One understudied yet significant degrader of marine polysaccharides are the *Kiritimatiellia*, a class of Verrucomicrobiota (or a potential new phylum[36]) with few cultured representatives. Although there are 23 taxonomic orders of sequenced *Kiritimatiellia* metagenome-assembled genomes (MAGs), all isolates come from the single order *Kiritimatiellales* found in marine and coastal sediments[37,38]. Publications about *Kiritimatiellales* center on their sulfatase-rich genomes, where sulfatases organized in CGCs[39] can approach 10% of all coding sequences[40]. These CGCs are predicted to degrade sulfated algal polysaccharides such as agar, carrageenan, and fucoidan[41]. Some species have further been reported to fix nitrogen[38,42,43] and detoxify antimicrobial compounds[40] from marine polysaccharide producers. Whether these capacities represent independent adaptations or facets of one combined anaerobic strategy in *Kiritimatiellia* has not been explored. Here, we address this question by building SCoNes across nearly 900 public *Kiritimatiellia* genomes and benchmarking the effectiveness of SCoNes in visualizing and dividing enzyme families. Then, we investigate the anaerobic and saccharolytic lifestyle of this class through case studies in polysaccharide degradation, natural product biosynthesis, nitrogen fixation, and fermentative hydrogen metabolism.

Methods

Genome selection and annotation

All genomes of the class *Kiritimatiellia* were collected from the Genome Taxonomy Database version 226[44], yielding 898 genomes and 2,203,257 predicted genes. Each genome

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.06.756941; this version posted October 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

was categorized into an environmental source (host-associated, aquatic & sediment, wastewater & built environment, or terrestrial soil) based on manual classification of NCBI BioSample metadata (**Supplemental Table 1**). Genomes were dereplicated to obtain a set of 480 species representatives[45], but all predicted proteins from the *Kiritimatiellia* were leveraged for downstream analyses. A phylogenetic tree for all species representatives was constructed using GToTree v. 1.8.16, built upon MUSCLE v. 5.1 alignment and FastTree v. 2.1.11 phylogeny[46–48], and visualized using the R v. 4.5.2 packages ggtree v. 4.0.5[49] and ggtreeExtra v. 1.20.1[50].

Supplemental Table 1. Metadata and taxonomy of selected *Kiritimatiellia* genomes.

Genomes were annotated for the presence of carbohydrate-active enzymes using the dbCAN database v. 11 with additional eCAMI annotations[51,52]. A subset of CAZyme annotations were classified into substrate-specific activity using the classification system proposed by López-Mondéjar et al.[53]. Sulfatase genes were annotated using a DIAMOND BLASTp[54] screening using a coverage threshold of 0.35 and a minimum e-value of 1e-5 parameters, followed by the SulfAtlas v. 2.3.1 HMM webservice[39] on initial hits.

Synteny blocks across all Kiritimatiellia were identified using Spacedust v. 2[30]. Nitrogen-fixation genes were annotated using pyHMMER[55,56] from the Kofam[57] and TIGRFAM[58] HMM subset proposed by Pi et al.[59]. Hydrogenase genes were annotated using HydDB[22]. Sulfatase-aware CAZyme gene clusters, including relevant sugar transporters and regulatory genes, were identified using CGC-Finder[60], while biosynthetic gene clusters were annotated using antiSMASH v. 8.0.4[61] and BiG-SCAPE v. 2.0[9] with c = 0.3.

Network analyses

Sequence similarity networks were built in the visualization style of the EFI webtool[20] using global DIAMOND BLASTp all-vs-all alignment of all 2,203,257 predicted proteins from the 898 Kiritimatiellia genomes. Graph manipulation and visualization was performed in Cytoscape.js[62].

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.06.756941; this version posted October 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

In the sequence similarity network, nodes represent individual proteins and two proteins are connected by an unweighted edge if they are within a pairwise similarity score e-value cutoff as determined by maximum closeness centrality[63].

SSN communities of connected subgraphs ("homology communities") were automatically detected using the Leiden algorithm[64], and subgraphs of size 3 or less were dropped from the visualization. Homology communities were annotated using InterProScan v. 5.77[65] on the most central member of each community as calculated by betweenness centrality using Scikit-network v. 0.33.5[66] for interpretation of protein function. This includes the annotation of transporters and regulatory enzymes discussed in Case Studies 3 and 4.

SCoNe construction

The process of generating a simplified similarity-colocalization network from the underlying SSN and gene neighborhood definitions is illustrated schematically in **Figure 1**. In this example, the initial sequence similarity network (G_{SSN}) (**Figure 1b**) is generated using a pairwise BLASTp search, with e-value as the homology metric and alignment locations as a source of gene coordinates defining relative gene position (**Figure 1a**). A community detection algorithm divides the sequence space into communities based on pairwise distances ($G_{community}$) (**Figure 1c**). Then, the colocalization step generates an additional edge type between genes that colocalize in user-defined neighborhoods ($G_{neighborhood}$) (**Figure 1d**). Finally, homology communities are collapsed into nodes and colocalization edges between communities combined to form the edges of G_{ScoNe} (**Figure 1e**). A SCoNe construction script that accepts a list of pairwise homology scores and gene colocalization edges and outputs a Cytoscape[67] visualization file is provided at

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.06.756941; this version posted October 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

https://github.com/AaronAOliver/KiritimatiallesSCoNe.

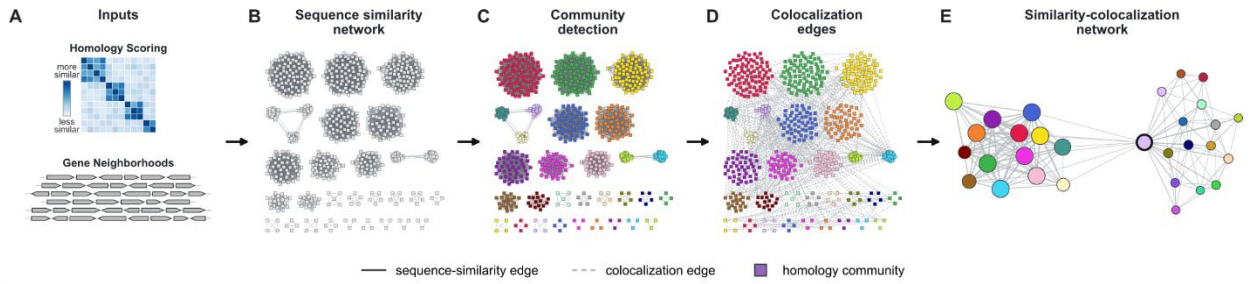

Figure 1. The multistep construction of a similarity-colocalization network using mock data. (A) Pairwise homology scores and gene coordinate data are sufficient to construct a SCoNe. (B) A sequence similarity network built from pairwise homology scores. (C) Clustering protein space through community detection. (D) Appending colocalization edges from gene cluster boundaries or anchored genome neighborhoods. (E) Collapsing homology communities into nodes to generate a similarity-colocalization network. The central pink node with two distinct subgraphs showcases unique gene contexts and potentially promiscuous biochemical function.

Results

Polysaccharide degradation repertoires in metagenomic *Kiritimatiellia*

To demonstrate the utility of SCoNes on metagenomic datasets, where contigs are often short and gene clusters fragmented, we collected all *Kiritimatiellia* genomes from the Genome Taxonomy Database and surveyed their capacity to degrade polysaccharides. Metagenome-assembled genomes of the *Kiritimatiellia* derive primarily from environmental and gut microbiomes, and only a handful of strains have been cultured to date[38]. The phylogenetic distribution of species representatives and their carbohydrate-degradation capacity is shown in **Figure 2**, with counts for a subset of CAZyme families of validated substrate specificity[53] plotted beneath the tree. Two families stand out for their exceptional sulfatase and algal-polysaccharide-degrading capacity and are highlighted on the tree: marine *Pontiellaceae* and freshwater

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.06.756941; this version posted October 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

CAIQIC01.

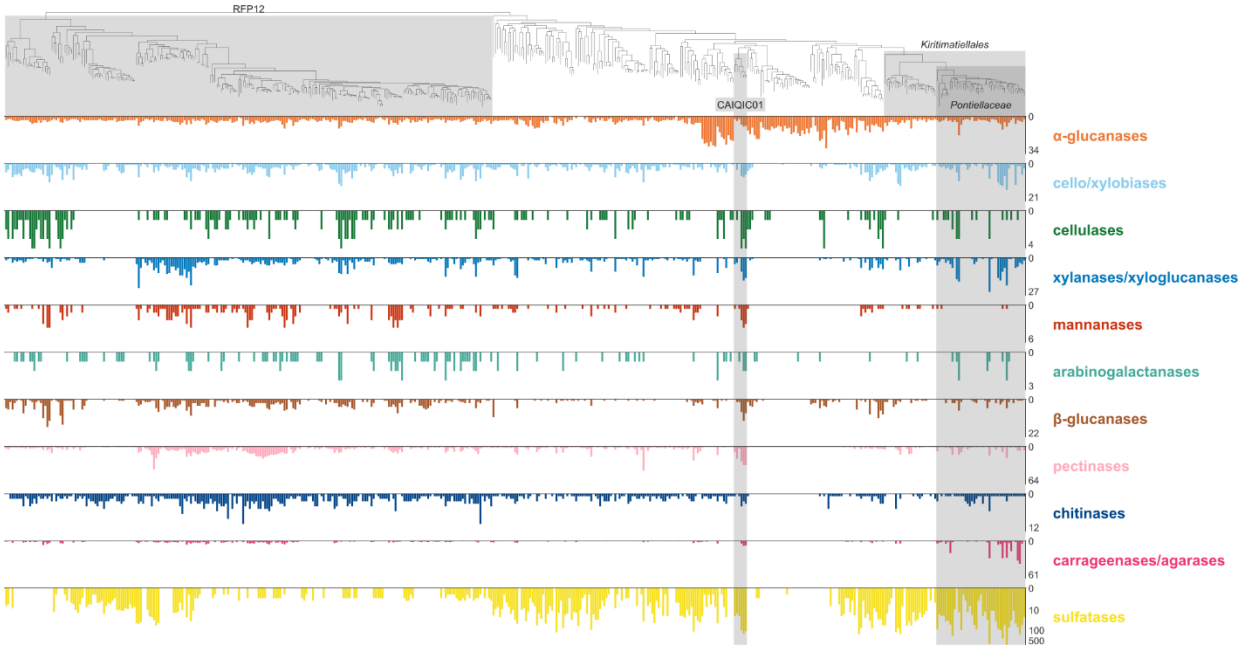

Figure 2. Polysaccharide degradation potential of the *Kiritimatiellia*. On top, a multigene marker phylogenetic tree of 480 species representative *Kiritimatiellia* genomes from the Genome Taxonomy Database. Two selected families with diverse enzyme repertoires, *Pontiellaceae* and CAIQIC01, highlighted in gray. Subsequent rings denote counts of catalytic carbohydrate enzymes and sulfatases per genome. Due to expanded sulfatase counts for some genomes, the bottommost ring uses a log scale.

Carbohydrate utilization potential aligns closely with environmental source in our dataset (**Supplemental Figure 1a**). Aquatic genomes are enriched in xylanases and sulfatases, while host-associated genomes have more cellulases, chitinases, and glucanases, consistent with environmental polysaccharide distributions. Marine *Pontiella* are outliers among aquatic *Kiritimatiellia*, with specialization toward sulfated algal polysaccharides such as carrageenan and agar (**Supplemental Figure 1b**). Individual genomes from *Pontiellaceae* have previously been found on macroalgal surfaces[43,68]. The cultured *P. desulfatans* and *P. sulfatireligans* encode 500 and 461 sulfatase genes respectively, more than any metagenome-assembled genome in the

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.06.756941; this version posted October 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

dataset. The freshwater CAIQIC01 are generalists, with a median of 217 CAZymes and 101 sulfatases per genome, the highest median sulfatase density of any family. Primarily found in lake and aquifer metagenomes[69], uncultured CAIQIC01 combine xylan- and sulfated-polysaccharide degradation repertoires with larger amounts of cellulases, glucanases, and other enzymes acting on labile sugars.

Supplemental Figure 1. **Polysaccharide degradation potential varies by environment and taxonomic lineage.** (A) A principal component analysis of glycoside hydrolase, polysaccharide lyase, carbohydrate esterase, and sulfatase abundance in species representative *Kiritimatiellia* normalized by genes per genome. Points are individual genomes colored by environmental source. (B) The same principal components analysis plot, with *Pontiella* and CAIQIC01 colored as outliers.

Case study 1: polysaccharide utilization gene clusters

To explore the genomic organization of polysaccharide degradation strategies of the *Kiritimatiellia*, CAZyme gene clusters (CGCs) were annotated in our genome set using dbCAN. In total, 2,298 gene clusters were identified, containing 4,825 CAZyme genes, 3,443 sulfatases, 4,522 transporters, 3,383 auxiliary genes, and 6,811 hypothetical genes. Multiple CAZymes were found in 44.0% of CGCs, with at least one sulfatase in 63.3%, and multiple sulfatases in 26.8%. We used the gene cluster boundaries as the basis of colocalization edges, and DIAMOND e-values for pairwise homology scores for all genes within a CGC. We then asked whether integrating genomic neighborhood information could identify conserved gene families whose organization is underexplored relative to well-annotated carbohydrate-active enzymes.

First, an e-value cutoff threshold of 1e-80 was chosen to construct a sequence similarity network of several thousand proteins (**Supplemental Figure 2**). Collapsing this sequence-similarity network into homology communities yielded the SCoNe in **Figure 3a**, in which each

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.06.756941; this version posted October 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

CAZyme, sulfatase, transporter, and regulatory family collapses to a small number of communities whose genomic neighborhoods can be examined individually. In total, the full set of CGC genes could be condensed into 188 homology communities of size 5 GCGs or larger. Sulfatases and auxiliary genes showed wider variance in sequence diversity and genomic context than CAZyme families.

Supplemental Figure 2. Sequence similarity network of CAZyme gene cluster proteins in *Kiritimatiellia*. Individual points are gene sequences from dbCAN-annotated CAZyme gene clusters, colored by putative function.

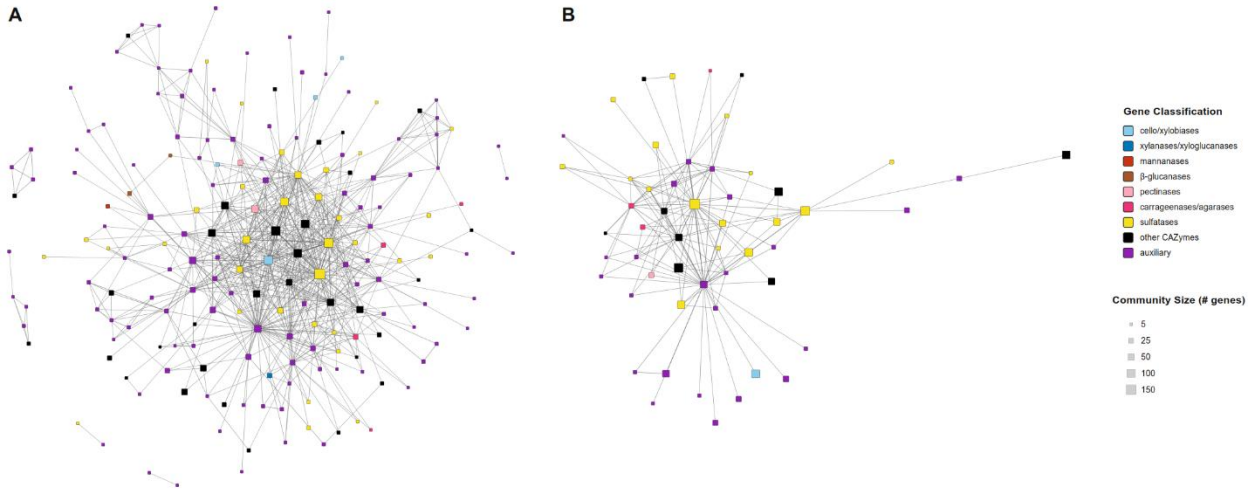

Figure 3. Similarity-colocalization network of conserved genes in CAZyme gene clusters of *Kiritimatiellia*. (A) The entire SCoNe built from dbCAN-annotated CGCs. (B) The SCoNe pruned to focus on genes associated with predicted agarases and carrageenases. Edges that colocalize in at least 5 CGCs shown. Singleton nodes and unannotated subgraphs are not shown.

Although a full SCoNe can span thousands of genes, the interpretability of the entire network at once is often quite difficult. One benefit of SCoNes is that network subgraphs or a selection of just a few nodes can be extracted, highlighting more detailed relationships for later ad-hoc exploration. Pictured in **Figure 3b** is a subset of the SCoNe in **Figure 3a**, looking only at

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.06.756941; this version posted October 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

genes colocalized with a carrageenase or agarase. Nodes in the center of each network colocalize with the highest diversity of genes, denoting core or multifunctional enzymes or auxiliary genes. This pruned SCoNe, containing only 42 homology communities, resolved a catalytic core of glycoside hydrolase subfamilies (including GH86, GH167, GH50, GH82, and GH117) frequently colocalized with more than a dozen sulfatase families (including S1_7, S1_15, S1_16, S1_17, S1_19, and S1_20) and 23 auxiliary families, whose conserved colocalization to the catalytic CGC core had not previously been appreciated. These conserved auxiliary families include primary (ATP-binding cassette) and secondary (major facilitator superfamily, solute:sodium symporter) sugar transporter families, two-component regulator systems previously shown to regulate polysaccharide utilization loci in marine bacteria[31,70], and formylglycine-generating enzymes required to post-translationally activate S1 sulfatase enzymes[71]. Both SCoNes pictured in **Figure 3** are available as Cytoscape visualization files at Zenodo 10.5281/zenodo.22103458.

Case study 2: aryl polyene biosynthetic gene clusters

Biosynthetic gene clusters also serve as a relevant use case for visualizing hundreds of divergent gene neighborhoods at scale. Aryl polyene clusters were the second most abundant biosynthetic class across Kiritimatiellia species representatives, comprising 14.2% of total BGC annotations, exceeded only by terpene and terpene-precursor clusters (**Supplemental Figure 3**). The SCoNe pipeline condensed these genomic neighborhoods into 43 conserved SSN communities present in at least 10 clusters (**Figure 4a**). The network's central members consisted of two β-ketoacyl synthases and a ketoreductase family, with separate clusters within the SCoNe containing additional biosynthetic domains unique to separate gene cluster families. Two of the larger aryl polyene gene cluster families illustrate how auxiliary enzymes organize around a central backbone: GCF-3 (**Figure 4b**) and GCF-2 (**Figure 4c**). These two GCFs share a core set of five genes (two ketosynthases, a ketoreductase, an acyltransferase, and a dehydratase), where

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.06.756941; this version posted October 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

GCF-2 adds a hydrolase gene and GCF-3 includes a more complex cassette of two acyl carrier proteins, a lipoprotein carrier, and an MmpL-like transporter. Both GCFs were primarily found in UBA1067, an uncharacterized genus linked to human[72] and rumen[73] gut microbiomes.

Supplemental Figure 3. Biosynthetic potential of *Kiritimatiellia*. Each bar graph shows the number of biosynthetic gene clusters in species representative *Kiritimatiellia* genomes as annotated by antiSMASH. Percentages display the ratio of each biosynthetic gene cluster classification to the total BGC set. The bottom axis uses a log scale.

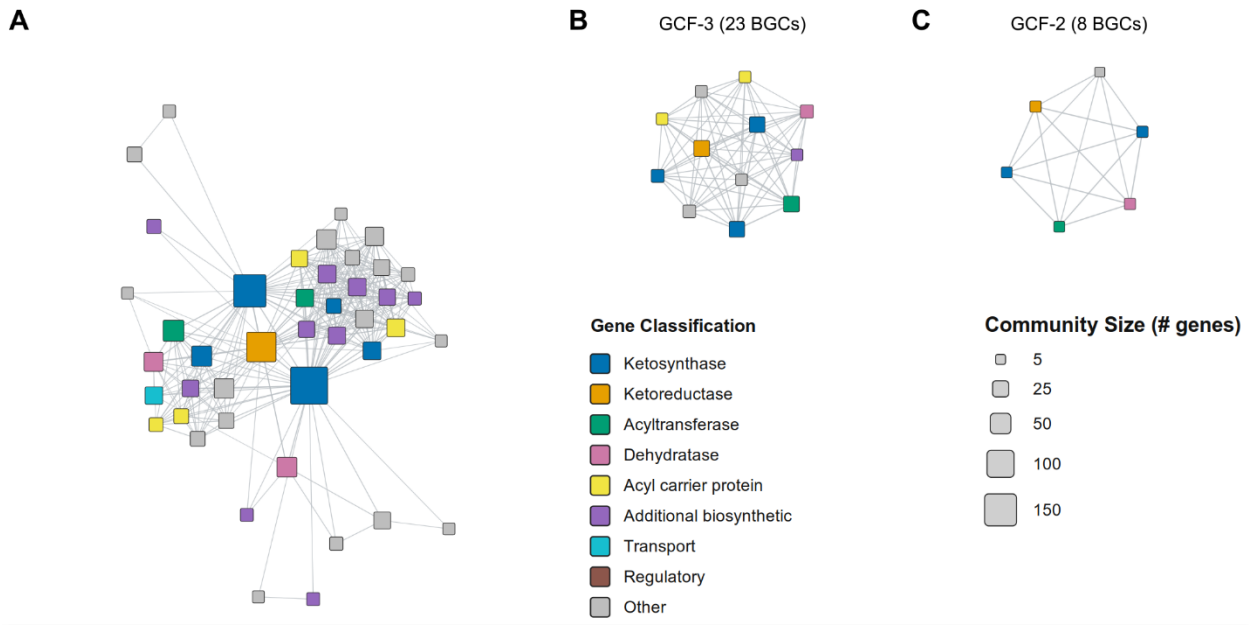

Figure 4. Biosynthetic potential of *Kiritimatiellia*. (A) A SCoNe of conserved aryl polyene biosynthesis genes shows a consistent backbone. Nodes are colored by antiSMASH annotation class and sized based on the number of homologous genes. (B, C) Pruning the main SCoNe produces SCoNes of individual gene cluster families.

Discovery and analysis of conserved synteny blocks

While visualization strategies that scale to hundreds or thousands of genomes are particularly useful for gene clusters with established annotation tools, such as biosynthetic or CAZyme gene clusters, the boundaries of other gene neighborhoods are less standardized.

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.06.756941; this version posted October 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Recent tools involving gene orthology and neighborhood now allow the discovery of conserved gene synteny strategies without requiring specialized databases or a priori gene-cluster boundaries. One example of these tools is Spacedust[30], which uses Foldseek and MMseqs2 to identify conserved blocks of orthologous, colocalized genes across many genomes. **Supplemental Figure 3** applies this to the 6,505 genes of the cultured *Pontiella desulfatans* genome[37] scored for conserved synteny across 479 species-representative *Kiritimatiellia* genomes, 41 of which are within the family Pontiellaceae (**Supplemental Figure 4a**) and 438 across the rest distributed across the order (**Supplemental Figure 4b**). Few gene neighborhoods are conserved across the entire order (**Supplemental Figure 4c**). Instead, the conserved blocks partition those retained within *Pontiella* and those shared more broadly, revealing lineage-specific genome organization. Because these blocks require neither pre-defined neighborhood boundaries nor core marker genes, the approach scales to the organization strategies of hundreds of related genomes at once.

Supplemental Figure 4. Conserved gene homology and synteny among *Pontiellaceae* and the wider *Kiritimatiellia*. Homology map showing shared synteny blocks between *Pontiella desulfatans* and (A) other *Pontiellaceae* and (B) other *Kiritimatiellia*. Each line row represents one genome, and each column represents one gene in *P. desulfatans*. (C) The proportion of genomes that share a homolog with *P. desulfatans* at each protein position. Gray areas show synteny blocks explored in other figures.

Case study 3: nitrogen fixation operons

One region of homologous genes highlighted by synteny block analysis is the nitrogen fixation (nif) operon. *Kiritimatiellia* have been shown to fix nitrogen in culture[38], and reference-free synteny analysis identified a nif gene cluster in 31 genomes across the class, 20 of which carry the complete nitrogen fixation catalytic core (**Figure 5**). Four marine Pontiella genomes, *P. sulfatireligans*[37] and three uncultivated representatives, encode a complete molybdate ABC

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.06.756941; this version posted October 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

transporter colocalized with the larger operon of Mo-dependent nitrogenases (**Figure 5c**). Despite the contig fragmentation common in metagenomic sequencing, the SCoNe approach tolerates unconserved gene order and still captures the operons with variable gene content and missing marker genes (**Figure 5b**). For example, several *Pontiella* genomes (e.g., *P. agarivorans*) carry *nifENB* upstream rather than downstream of *nifHDK*, and metagenome-assembled genomes frequently have molybdate transport in distinct operons from the primary nif cluster.

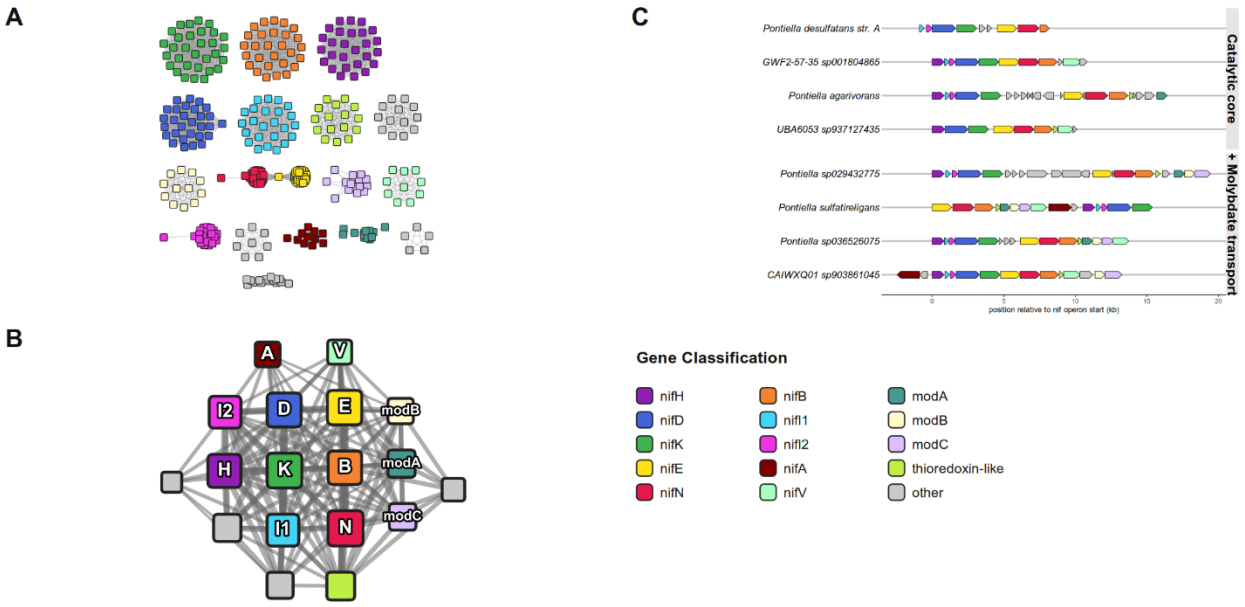

Figure 5. Paired SSN and SCoNe of nitrogen fixation gene clusters in the *Kiritimatiellia*. (A) A sequence similarity network of *nif* gene operons. (B) A SCoNe showing organization of *nif* and molybdate transport clusters, using the same color scheme as (A). (C) A gene neighborhood map showing two styles of *nif* operon synteny.

Case study 4: [FeFe]-hydrogenase loci and fermentative hydrogen metabolism

Synteny block analysis also identified a [FeFe]-hydrogenase locus conserved across Kiritimatiellia (**Figure 6**), typically a trimeric *hydABC* operon accompanied by a small [2Fe-2S] ferredoxin. 78% of the Kiritimatiellia genomes analyzed encode at least one [FeFe]-hydrogenase, and many encode more than one type: the electron-bifurcating Group A3 enzyme is abundant

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.06.756941; this version posted October 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(74% of all genomes), while the sensory *hydS*-containing Group C (32%), ancestral Group B (30%), and fermentative Group A1 (5%) types occur less frequently. A SCoNe built from all hydrogenase loci (**Figure 6b**) resolves this diversity into distinct operon architectures. In the *hydS* architecture, *hydB* is absent and the sensory *hydS* colocalizes with a PP2C-family phosphatase and an alcohol dehydrogenase, consistent with signal transduction and control of fermentative pathways. Elsewhere, the *hydB* subunit is instead extended: in 163 genomes it fuses a FAD/NAD(P)-binding oxidoreductase module and additional [4Fe-4S] clusters onto the ancestral core, and 132 genomes carry both the fused and the unfused form. The cultured *K. glycovorans* and *T. aerotolerans* genomes carry the fused *hydB* within a tandemly duplicated *hydABC* array, adjacent to but distinct from its canonical unfused operon (**Figure 6a**).

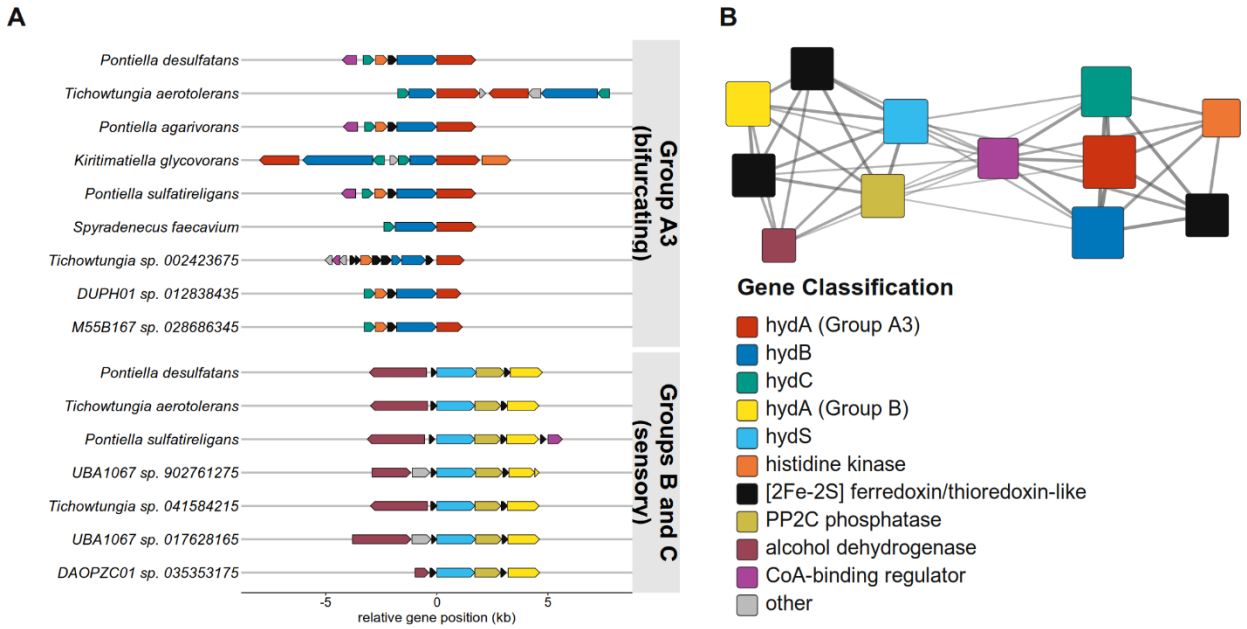

Figure 6. Genomic organization of [FeFe]-hydrogenase synteny blocks. (A) Gene neighborhood diagram of selected genomic regions containing the hydrogenase cassette *hydABC* or standalone hydrogenase *hydS* and conserved auxiliary genes. (B) A similarity-colocalization network of genes associated with [FeFe]-hydrogenase, following the same color scheme as (A).

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.06.756941; this version posted October 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Discussion

In this study, we present the first class-wide investigation of the potential of Kiritimatiellia to degrade terrestrial and marine polysaccharides. We have identified trends across the phylogenetic sub-groups of the Kiritimatiellia, highlighting families with unique metabolic potential to degrade a wide variety of glycan substrates. We also confirm that several biochemical processes previously limited to a few cultured isolates, such as nitrogen fixation, hydrogen production, and natural product biosynthesis, are widespread among uncultured *Kiritimatiellia* genomes.

For glycan degradation, we see clear differentiation in enzyme repertoire by environmental source. In particular, sulfatase enzymes are enriched in genomes belonging to bacteria that live in aquatic environments and marine sediment. This is consistent with the enrichment of sulfated monosaccharides in algal sugars commonly found in the ocean. Sinking marine debris also explains the enrichment of these enzymes in sediment microbes as it is well established that sediment microbes play an important role in cycling carbon of particulate algal biomass[42]. While most *Kiritimatiellia* in our dataset are rich in CAZymes, two groups stand out in their adaptation towards degrading algal polysaccharides, as marked by an abundance of sulfatase enzymes.

The first, the family *Pontiellaceae*, is the most well-described lineage across all of the *Kiritimatiellia* and includes several cultured isolates. *Pontiella* have been isolated from anoxic marine sediments and microbial mats[38,74], and our analysis of uncultured microbes shows that this algal polysaccharide targeting is consistent across the family regardless of environmental source. Still, the isolate *Pontiella* genomes do contain a very high number of sulfatase enzymes, higher than any *Pontiella* MAG (**Figure 2**). **Supplemental Figure 1b** further displays the unique repertoire of carbohydrate degradation genes of Pontiella relative to the rest of the class, driven by this expansion of sulfatases together with an abundance of carrageenases and agarases. This

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.06.756941; this version posted October 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

specialization is further apparent as the relative abundance of glucanases and chitinases in Pontiella is far lower than in Kiritimatiellia from wastewater and mammalian gut microbiomes.

On the other hand, the undescribed metagenomic CAIQIC01 family is composed solely of freshwater microbes from water column or freshwater sediment samples[69] that are simultaneously rich in sulfatases and in xylanases, pectinases, chitinases, and cellulases. This sulfatase abundance also extends to some host-associated lineages, which does serve an important biological purpose, as host mucins and glycosaminoglycans are often heavily sulfated[32]. For example, UBA1067, found in mammalian gut microbiomes[73,75], contains up to 11 sulfatases per genome, and two UBA6053 genomes recovered from coral mucus[76] contain 80 and 82 sulfatases. Members of these uncultured families carry sulfatase levels otherwise seen only in *Pontiella* and a few Verrucomicrobiota[13,77], yet they have never been described in prior literature as a potential source of novel polysaccharide degradation strategies.

Pontiella also display the ability to fix nitrogen using a molybdenum-dependent system. The relationship between sulfated polysaccharide degradation and nitrogen fixation may stem from large amounts of carbon brought on by an influx of sinking particulate organic matter, raising local C:N ratios[78]. The bottleneck of nitrogen supply can be bypassed using the excess carbon and energy provided by this set of recalcitrant polysaccharides. Other heterotrophic marine bacteria, including Verrucomicrobiota and Proteobacteria, likewise raise their rates of nitrogen fixation in the presence of bioavailable mono- and polysaccharides[79–81]. In the eastern Mediterranean, addition of glucose increased nitrogen fixation by up to tenfold[82], an effect attributed to the high C:N, low-oxygen microenvironments created by carbon-rich marine particles. Carbon is also readily available in the phycosphere due to glycan seepage by kelp[83], and one study found that *Pontiella* represented up to 45% of nifH gene activity of the kelp-surface microbiome, following a diurnal cycle[43]. Previous studies have shown that macroalgal blooms

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.06.756941; this version posted October 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

select for heterotrophic nitrogen fixers as macroalgae utilize nitrogen in the water column[84] but provide a niche for polysaccharide degraders.

Complex polysaccharide degradation is only one component of an anaerobic lifestyle, and our survey shows strong genomic evidence of fermentation across Kiritimatiellia. [FeFe]-hydrogenases appeared in most species representative genomes, with conserved auxiliary genes immediately surrounding the hydABC cluster. Previous analysis of bifurcating hydrogenases noted that *P. sulfatireligans* and *P. desulfatans* appeared to lose the hydD subunit present in related phyla such as *Planctomycetota* and *Hydrogenedentota*[85]. This analysis extends that observation, as hydD is absent across the class, supporting the hypothesis that hydD loss preceded diversification of the class rather than occurring independently in Pontiella. However, tandem duplication of the fused hydB subunit in *K. glycovorans* and *T. aerotolerans* suggests hydrogenase metabolism continues to diversify in this lineage[86].

Under the metabolic model described by our case studies, sulfatases and glycoside hydrolases release monosaccharides from algal and host glycans, fermentation extracts energy from those monosaccharides, and hydrogenases discharge those electrons as H_2. This coupling is explicit in *P. agarivorans,* where approximately 75% of fermentation electrons are recovered in organic products and the remainder are available for nitrogen fixation, with hydrogen gas appearing as a minor fermentation product[38]. Aryl polyenes, known to quench reactive oxygen species and support biofilm formation[87,88], shield these anoxic processes from the outside environment.

Enzymes acting on sulfated polysaccharides are of considerable interest for the valorization of algal biomass[89], and targeted culturing of novel *Kiritimatiellia* for enzyme discovery has been established by our group in bioreactor systems[40]. Uncultured marine microbiomes are similarly a continuing source of novel natural products[90]. Our survey represents the first comprehensive study of polysaccharide utilization and natural product potential across

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.06.756941; this version posted October 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Kiritimatiellia, which should serve as a useful reference when further mining these genomes for enzymes relevant to bioenergy and new pharmaceuticals. The gene neighborhood approach enabled by SCoNe in this manuscript is particularly effective in highlighting conserved hypothetical and novel genes in widespread gene clusters like CGCs and BGCs, which would otherwise require manual scanning to find putatively novel enzymes[91].

We have demonstrated that SCoNes are useful at three levels of gene neighborhood characterization. First (**Figure 3**), SCoNes apply to well-defined gene cluster boundaries produced by established tools. This requires only the gene sequences within each cluster and their relative arrangement, so transforming the output of tools such as dbCAN, antiSMASH, or deepBGC into a SCoNe demands little additional computation and can visualize the metabolic or catabolic potential of hundreds or thousands of genomes at once, which popular gene arrangement plots like clinker[25] cannot match at scale. Second (**Figure 5**), because SCoNes do not require predetermined cluster or neighborhood boundaries, they are especially suited to the conserved gene neighborhoods of biochemically important regions, regardless of the existence of specialized annotation tools. Basic gene organization and similarity metrics quantified with pairwise homology tools are sufficient to identify and analyze synteny blocks in a scalable way. Third**,** a SCoNe visualization can quickly predict multifaceted activity of a single gene based on multiple conserved gene neighborhoods**.** In our hydrogenase vignette (**Figure 6**), the annotation of a single *hydA* gene was sufficient hypothesis generation to subset the Kiritimatiellia SCoNe and identify two divergent gene cassettes for bifurcating and sensory hydrogenases.

Conclusion

These analyses confirm *Kiritimatiellia* as anaerobic and saccharolytic specialists with metabolism organized around breaking down sulfated glycans. These bacteria live in low-oxygen environments, such as marine sediments or animal guts, and interface with complex, often sulfated polysaccharides. In response, *Kiritimatiellia* hydrolyze large glycans into simple

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.06.756941; this version posted October 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

monomers for fermentation, using hydrogenases to balance redox. Some lineages, particularly *Pontiella*, divert energy from fermentation into molybdenum-dependent nitrogen fixation when bioavailable nitrogen is scarce. Describing the breadth of glycan cycling strategies in uncultured marine lineages is crucial for understanding the fate of carbon in the oceans.

The novel SCoNe data structure offers a scalable way to visualize and explore genome organization directly from environmental assemblies. This approach succeeds when cluster boundaries are defined by specialized tools, inferred from conserved synteny, or seeded from a single query gene. We find SCoNe particularly helpful for exploring metagenomic dark matter, where genes are distant from database representatives and gene ontologies are often undescriptive. We release our SCoNe generation software at https://github.com/AaronAOliver/KiritimatiallesSCoNe to support its further use in visualization and hypothesis generation for meta-analyses and user-generated data.

Acknowledgements

Computational analyses were performed using the San Diego Supercomputer Center's Triton Shared Computing Cluster. This work was funded by United States Department of Agriculture Agriculture and Food Research Initiative Predoctoral Fellowship 2026-67011-46317 to A.O. and National Science Foundation grant OCE-2414798 and National Institutes of Health NIEHS grant P01ES035541-01 to E.E.A.

Data Availability

Code used to generate SCoNes is available at https://github.com/AaronAOliver/KiritimatiallesSCoNe. SCoNes generated for Figures 3, 4, 5, and 6 are available as Cytoscape-compatible files on Zenodo 10.5281/zenodo.22103458. Metadata on genomes included in this study are included as **Supplemental Table 1**.

References

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.06.756941; this version posted October 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

1. Pellegrinetti, T. A., Molligan, J., Mendes, L. W., Pedrinho, A. & Pérez-López, E. Rethinking metagenome-assembled genome completeness: are we truly recovering complete genomes? *Front. Microbiol.* **17**, (2026).
2. Trigodet, F., Sachdeva, R., Banfield, J. F. & Eren, A. M. Assemblies of long-read metagenomes suffer from diverse errors. 2025.04.22.649783 Preprint at https://doi.org/10.1101/2025.04.22.649783 (2025).
3. Lloyd, K. G., Steen, A. D., Ladau, J., Yin, J. & Crosby, L. Phylogenetically Novel Uncultured Microbial Cells Dominate Earth Microbiomes. *mSystems* **3**, 10.1128/msystems.00055-18 (2018).
4. Steen, A. D. *et al.* High proportions of bacteria and archaea across most biomes remain uncultured. *ISME J.* **13**, 3126–3130 (2019).
5. Lewis, W. H., Tahon, G., Geesink, P., Sousa, D. Z. & Ettema, T. J. G. Innovations to culturing the uncultured microbial majority. *Nat. Rev. Microbiol.* **19**, 225–240 (2021).
6. Dandekar, T., Snel, B., Huynen, M. & Bork, P. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.* **23**, 324–328 (1998).
7. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D. & Maltsev, N. The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci.* **96**, 2896–2901 (1999).
8. Marcotte, E. M. *et al.* Detecting Protein Function and Protein-Protein Interactions from Genome Sequences. *Science* **285**, 751–753 (1999).
9. Draisma, A. *et al.* BiG-SCAPE 2.0 and BiG-SLiCE 2.0: scalable, accurate and interactive sequence clustering of metabolic gene clusters. *Nat. Commun.* **17**, 2000 (2026).
10. Mungan, M. D., Blin, K. & Ziemert, N. ARTS-DB: a database for antibiotic resistant targets. *Nucleic Acids Res.* **50**, D736–D740 (2022).
11. Shmakov, S. A., Makarova, K. S., Wolf, Y. I., Severinov, K. V. & Koonin, E. V. Systematic prediction of genes functionally linked to CRISPR-Cas systems by gene neighborhood analysis. *Proc. Natl. Acad. Sci.* **115**, E5307–E5316 (2018).

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.06.756941; this version posted October 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

12. Delherbe, N. A. *et al.* Atmospheric methane consumption in arid ecosystems acts as a reverse chimney and is accelerated by plant-methanotroph biomes. *ISME J.* **19**, wraf026 (2025).
13. Sichert, A. *et al.* Verrucomicrobia use hundreds of enzymes to digest the algal polysaccharide fucoidan. *Nat. Microbiol.* **5**, 1026–1039 (2020).
14. Pavlopoulos, G. A. *et al.* Unraveling the functional dark matter through global metagenomics. *Nature* **622**, 594–602 (2023).
15. van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* **42**, 243–246 (2024).
16. Vig, J. *et al.* Bertology Meets Biology: Interpreting Attention in Protein Language Models. in (2021).
17. Song, N., Joseph, J. M., Davis, G. B. & Durand, D. Sequence Similarity Network Reveals Common Ancestry of Multidomain Proteins. *PLOS Comput. Biol.* **4**, e1000063 (2008).
18. Halary, S., McInerney, J. O., Lopez, P. & Bapteste, E. EGN: a wizard for construction of gene and genome similarity networks. *BMC Evol. Biol.* **13**, 146 (2013).
19. Zallot, R., Oberg, N. & Gerlt, J. A. The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways. *Biochemistry* **58**, 4169–4182 (2019).
20. Oberg, N., Zallot, R. & Gerlt, J. A. EFI-EST, EFI-GNT, and EFI-CGFP: Enzyme Function Initiative (EFI) Web Resource for Genomic Enzymology Tools. *J. Mol. Biol.* **435**, 168018 (2023).
21. Viborg, A. H. *et al.* A subfamily roadmap of the evolutionarily diverse glycoside hydrolase family 16 (GH16). *J. Biol. Chem.* **294**, 15973–15986 (2019).
22. Søndergaard, D., Pedersen, C. N. S. & Greening, C. HydDB: A web tool for hydrogenase classification and analysis. *Sci. Rep.* **6**, 34212 (2016).

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.06.756941; this version posted October 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

23. Copp, J. N., Akiva, E., Babbitt, P. C. & Tokuriki, N. Revealing Unexplored Sequence-Function Space Using Sequence Similarity Networks. *Biochemistry* **57**, 4651–4662 (2018).
24. Szklarczyk, D. *et al.* The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* **51**, D638–D646 (2023).
25. Gilchrist, C. L. M. & Chooi, Y.-H. clinker & clustermap.js: automatic generation of gene cluster comparison figures. *Bioinformatics* **37**, 2473–2475 (2021).
26. Gilchrist, C. L. M. *et al.* cblaster: a remote search tool for rapid identification and visualization of homologous gene clusters. *Bioinforma. Adv.* **1**, vbab016 (2021).
27. Blin, K. *et al.* antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Res.* **49**, W29–W35 (2021).
28. Hannigan, G. D. *et al.* A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Res.* **47**, e110 (2019).
29. Saha, C. K., Sanches Pires, R., Brolin, H., Delannoy, M. & Atkinson, G. C. FlaGs and webFlaGs: discovering novel biology through the analysis of gene neighbourhood conservation. *Bioinformatics* **37**, 1312–1314 (2021).
30. Zhang, R., Mirdita, M. & Söding, J. De novo discovery of conserved gene clusters in microbial genomes with Spacedust. *Nat. Methods* **22**, 2065–2073 (2025).
31. Grondin, J. M., Tamura, K., Déjean, G., Abbott, D. W. & Brumer, H. Polysaccharide Utilization Loci: Fueling Microbial Communities. *J. Bacteriol.* **199**, 10.1128/jb.00860-16 (2017).
32. Luis, A. S. *et al.* A single sulfatase is required to access colonic mucin by a gut bacterium. *Nature* **598**, 332–337 (2021).
33. Lapébie, P., Lombard, V., Drula, E., Terrapon, N. & Henrissat, B. Bacteroidetes use thousands of enzyme combinations to break down glycans. *Nat. Commun.* **10**, 2043 (2019).

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.06.756941; this version posted October 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

34. Oliver, A. *et al.* Enrichable consortia of microbial symbionts degrade macroalgal polysaccharides in Kyphosus fish. *mBio* **0**, e00496-24 (2024).
35. Ausland, C. *et al.* dbCAN-PUL: a database of experimentally characterized CAZyme gene clusters and their substrates. *Nucleic Acids Res.* **49**, D523–D528 (2021).
36. Spring, S. *et al.* Characterization of the first cultured representative of Verrucomicrobia subdivision 5 indicates the proposal of a novel phylum. *ISME J.* **10**, 2801–2816 (2016).
37. van Vliet, D. M. *et al.* Pontiella desulfatans gen. nov., sp. nov., and Pontiella sulfatireligans sp. nov., Two Marine Anaerobes of the Pontiellaceae fam. nov. Producing Sulfated Glycosaminoglycan-like Exopolymers. *Microorganisms* **8**, 920 (2020).
38. Liu, N. *et al.* Pontiella agarivorans sp. nov., a novel marine anaerobic bacterium capable of degrading macroalgal polysaccharides and fixing nitrogen. *Appl. Environ. Microbiol.* **90**, e00914-23 (2024).
39. Stam, M. *et al.* SulfAtlas, the sulfatase database: state of the art and new developments. *Nucleic Acids Res.* **51**, D647–D653 (2023).
40. Plominsky, A. M. *et al.* Detoxifying and depolymerizing microorganisms reveal intertwined guild collaborations in the gut microbiome of the generalist macro-algivorous fish Kyphosus cinerascens. *mBio* **0**, e03382-25 (2026).
41. Bäumgen, M., Dutschei, T. & Bornscheuer, U. T. Marine Polysaccharides: Occurrence, Enzymatic Degradation and Utilization. *ChemBioChem* **22**, 2247–2256 (2021).
42. Peoples, L. M. *et al.* A deep-sea isopod that consumes Sargassum sinking from the ocean's surface. *Proc. R. Soc. B Biol. Sci.* **291**, 20240823 (2024).
43. Wang, Z. *et al.* Diurnal fluctuations of bacterial communities and diazotrophs in the kelp phycosphere. *Algal Res.* **92**, 104363 (2025).
44. Parks, D. H. *et al.* GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkab776 (2021) doi:10.1093/nar/gkab776.

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.06.756941; this version posted October 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

45. Parks, D. H. *et al.* A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat. Biotechnol.* **38**, 1079–1086 (2020).

46. Lee, M. D. GToTree: a user-friendly workflow for phylogenomics. *Bioinformatics* **35**, 4162–4164 (2019).

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

48. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLOS ONE* **5**, e9490 (2010).

49. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. 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).

50. Xu, S. *et al.* ggtreeExtra: Compact visualization of richly annotated phylogenetic data. *Mol. Biol. Evol.* **38**, 4039–4042 (2021).

51. Zheng, J. *et al.* dbCAN3: automated carbohydrate-active enzyme and substrate annotation. *Nucleic Acids Res.* **51**, W115–W121 (2023).

52. Xu, J., Zhang, H., Zheng, J., Dovoedo, P. & Yin, Y. eCAMI: simultaneous classification and motif identification for enzyme annotation. *Bioinformatics* **36**, 2068–2075 (2020).

53. López-Mondéjar, R., Tláskal, V., da Rocha, U. N. & Baldrian, P. Global Distribution of Carbohydrate Utilization Potential in the Prokaryotic Tree of Life. *mSystems* **7**, e00829-22 (2022).

54. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).

55. Eddy, S. R. Accelerated Profile HMM Searches. *PLOS Comput. Biol.* **7**, e1002195 (2011).

56. Larralde, M. & Zeller, G. PyHMMER: a Python library binding to HMMER for efficient sequence analysis. *Bioinformatics* **39**, btad214 (2023).

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.06.756941; this version posted October 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

582 57. Aramaki, T. *et al.* KofamKOALA: KEGG Ortholog assignment based on profile HMM and
583 adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).
584 58. Haft, D. H. *et al.* TIGRFAMs and Genome Properties in 2013. *Nucleic Acids Res.* **41**, D387–
585 D395 (2013).
586 59. Pi, H.-W. *et al.* Nitrogen-fixing microbes gain genes in diverse types of living environments.
587 *Proc. Natl. Acad. Sci.* **122**, e2523106122 (2025).
588 60. Zhang, H. *et al.* dbCAN2: a meta server for automated carbohydrate-active enzyme
589 annotation. *Nucleic Acids Res.* **46**, W95–W101 (2018).
590 61. Blin, K. *et al.* antiSMASH 8.0: extended gene cluster detection capabilities and analyses of
591 chemistry, enzymology, and regulation. *Nucleic Acids Res.* **53**, W32–W38 (2025).
592 62. Franz, M. *et al.* Cytoscape.js 2023 update: a graph theory library for visualization and
593 analysis. *Bioinformatics* **39**, btad031 (2023).
594 63. Hornung, B. V. H. & Terrapon, N. An objective criterion to evaluate sequence-similarity
595 networks helps in dividing the protein family sequence space. *PLOS Comput. Biol.* **19**,
596 e1010881 (2023).
597 64. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-
598 connected communities. *Sci. Rep.* **9**, 5233 (2019).
599 65. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics*
600 **30**, 1236–1240 (2014).
601 66. Bonald, T., Lara, N. de, Lutz, Q. & Charpentier, B. Scikit-network: Graph Analysis in Python.
602 *J. Mach. Learn. Res.* **21**, 1–6 (2020).
603 67. Shannon, P. *et al.* Cytoscape: A Software Environment for Integrated Models of
604 Biomolecular Interaction Networks. *Genome Res.* **13**, 2498–2504 (2003).
605 68. Zhang, Y.-S. *et al.* Metagenomic insights into the dynamic degradation of brown algal
606 polysaccharides by kelp-associated microbiota. *Appl. Environ. Microbiol.* **90**, e02025-23
607 (2024).

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.06.756941; this version posted October 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

69. Anantharaman, K. *et al.* Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat. Commun.* **7**, 13219 (2016).
70. Ficko-Blean, E. *et al.* Carrageenan catabolism is encoded by a complex regulon in marine heterotrophic bacteria. *Nat. Commun.* **8**, 1685 (2017).
71. Helbert, W. Marine Polysaccharide Sulfatases. *Front. Mar. Sci.* **4**, (2017).
72. Youngblut, N. D., de la Cuesta-Zuluaga, J. & Ley, R. E. Incorporating genome-based phylogeny and functional similarity into diversity assessments helps to resolve a global collection of human gut metagenomes. *Environ. Microbiol.* **24**, 3966–3984 (2022).
73. Hinsu, A. T. *et al.* Characterizing rumen microbiota and CAZyme profile of Indian dromedary camel (Camelus dromedarius) in response to different roughages. *Sci. Rep.* **11**, 9400 (2021).
74. van Vliet, D. M. *et al.* Anaerobic Degradation of Sulfated Polysaccharides by Two Novel Kiritimatiellales Strains Isolated From Black Sea Sediment. *Front. Microbiol.* **10**, (2019).
75. Gilroy, R. *et al.* Metagenomic investigation of the equine faecal microbiome reveals extensive taxonomic diversity. *PeerJ* **10**, e13084 (2022).
76. Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2**, 1533–1542 (2017).
77. Orellana, L. H. *et al.* Verrucomicrobiota are specialist consumers of sulfated methyl pentoses during diatom blooms. *ISME J.* **16**, 630–641 (2022).
78. Zakem, E. J. & Levine, N. M. Systematic Variation in Marine Dissolved Organic Matter Stoichiometry and Remineralization Ratios as a Function of Lability. *Glob. Biogeochem. Cycles* **33**, 1389–1407 (2019).
79. Raut, Y. *et al.* Autochthonous carbon loading of macroalgae stimulates benthic biological nitrogen fixation rates in shallow coastal marine sediments. *Front. Microbiol.* **14**, (2024).

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.06.756941; this version posted October 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

632 80. Cohen, R. A., Walker, K. & Carpenter, E. J. Polysaccharide addition effects on rhizosphere
633 nitrogen fixation rates of the California cordgrass,Spartina foliosa. *Wetlands* **29**, 1063–1069
634 (2009).
635 81. Geisler, E., Bogler, A., Rahav, E. & Bar-Zeev, E. Direct Detection of Heterotrophic
636 Diazotrophs Associated with Planktonic Aggregates. *Sci. Rep.* **9**, 9288 (2019).
637 82. Rahav, E., Giannetto, M. J. & Bar-Zeev, E. Contribution of mono and polysaccharides to
638 heterotrophic N2 fixation at the eastern Mediterranean coastline. *Sci. Rep.* **6**, 27858 (2016).
639 83. Singh, R. P. & Reddy, C. R. K. Seaweed–microbial interactions: key functions of seaweed-
640 associated bacteria. *FEMS Microbiol. Ecol.* **88**, 213–230 (2014).
641 84. Zhang, X., Song, Y., Liu, D., Keesing, J. K. & Gong, J. Macroalgal blooms favor
642 heterotrophic diazotrophic bacteria in nitrogen-rich and phosphorus-limited coastal surface
643 waters in the Yellow Sea. *Estuar. Coast. Shelf Sci.* **163**, 75–81 (2015).
644 85. Zheng, X. & Huang, L. Diverse non-canonical electron bifurcating [FeFe]-hydrogenases of
645 separate evolutionary origins in Hydrogenedentota. *mSystems* **9**, e00999-24 (2024).
646 86. Peters, J. W. *et al.* [FeFe]- and [NiFe]-hydrogenase diversity, mechanism, and maturation.
647 *Biochim. Biophys. Acta BBA - Mol. Cell Res.* **1853**, 1350–1369 (2015).
648 87. Johnston, I. *et al.* Identification of essential genes for Escherichia coli aryl polyene
649 biosynthesis and function in biofilm formation. *Npj Biofilms Microbiomes* **7**, 56 (2021).
650 88. Schöner, T. A. *et al.* Aryl Polyenes, a Highly Abundant Class of Bacterial Natural Products,
651 Are Functionally Related to Antioxidative Carotenoids. *ChemBioChem* **17**, 247–253 (2016).
652 89. Adalsteinsson, B. T. *et al.* Targeted metagenomics – Enrichment for enzymes active on
653 sulfated polysaccharides from seaweeds. *Enzyme Microb. Technol.* **182**, 110528 (2025).
654 90. Chen, J. *et al.* Global marine microbial diversity and its potential in bioprospecting. *Nature*
655 **633**, 371–379 (2024).
656 91. Scherlach, K. & Hertweck, C. Mining and unearthing hidden biosynthetic potential. *Nat.*
657 *Commun.* **12**, 3864 (2021).

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.06.756941; this version posted October 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

658

659