Title: Mobile genetic elements that shape microbial diversity and functions in thawing permafrost soils

Authors

Jiarong Guo[1,2], Samuel Aroney[3], Guillermo Dominguez-Huerta[1,2], Derek Smith[4], Dean Vik[1,2], Carlos Owusu Ansah[1], Akbar Adjie Pratama[1], Sergei Solonenko[1], Funing Tian[1], Cristina Howard-Varona[1], ZhiPing Zhong[1,2,5], Aminata Fofana[1], Garrett Smith[1,2], Suzanne B. Hodgkins[1], Dylan Cronin[1], EMERGE Field Teams 2010-2019, EMERGE Coordinators, Ben J. Woodcroft[3], Gene W. Tyson[3], Virginia I. Rich[1], Matthew B. Sullivan[1,2]*, Simon Roux[6]*, Sarah C. Bagby[4]*

Affiliations

[1]Department of Microbiology, The Ohio State University; Columbus, OH, USA.
[2]Center of Microbiome Science, The Ohio State University; Columbus, OH, USA.
[3]Centre for Microbiome Research, School of Biomedical Sciences, Queensland University of Technology, Translational Research Institute; Woolloongabba, QLD, Australia.
[4]Department of Biology, Case Western Reserve University; Cleveland, OH, USA.
[5]Byrd Polar and Climate Research Center, Ohio State University; Columbus, OH, USA.
[6]DOE Joint Genome Institute, Lawrence Berkeley National Laboratory; Berkeley, CA, USA.
*Co-corresponding authors. MBS: sullivan.948@osu.edu, SR: sroux@lbl.gov, SCB: scb126@case.edu

One Sentence Summary

By developing microbiome-ready analytics and applying them to thawing permafrost microbiomes, we identified ~2.1 million diverse mobile genetic element (MGE) recombinases (integrons, insertion sequences, transposons, conjugative elements, phage-like elements) and revealed their function and activity to establish baseline ecological data and a framework for MGE study in complex natural systems.

Abstract

Microbiomes shape ecosystems through functional profiles influenced by gene gains and losses. While culture-based experiments demonstrate that mobile genetic elements (MGEs) can mediate gene flux, quantitative field data remains scarce. Here we leverage large-scale soil meta-omic data to develop and apply analytics for studying MGEs. In our model permafrost-thaw ecosystem, Stordalen Mire, we identify ~2.1 million MGE recombinases across 89 microbial phyla to assess ecological distributions, functions impacted, and past mobility and current activity. This revealed MGEs shape natural genetic diversity via lineage-specific engagement, and functions including canonical defense elements alongside diverse regulatory and metabolic genes affecting carbon flux and nutrient cycling. These findings and this systematic meta-omic framework open new avenues to better investigate MGE diversity, activity, mobility, and impacts in nature.

Main text

Microbiomes are increasingly recognized for their roles in biogeochemical and nutrient cycling across diverse systems (*1–4*). Mobile genetic elements (MGEs) likely alter these microbial impacts since culture-based studies show impacts such as horizontal gene transfer and/or genomic rearrangement (*5*, *6*), or by encoding key niche-defining genes involved in antimicrobial resistance

(AMR) (*7–9*), production of public goods (*10*), or defense against microbes, phages or other mobile elements (*11–13*). Further, MGEs are highly abundant and diverse with recent estimates across ~84K cultured bacterial genomes being that MGEs were broadly distributed across phyla and averaged 25 MGEs per genome (*8*). However, outside of cultivated isolates, the ecological and evolutionary effects of MGEs in nature remain poorly understood. This is because the same diversity and activity that drive MGE impacts also hamper metagenomic investigation – MGEs lacked well-developed marker genes until recently (*8*), microbial populations must be deeply and regularly sequenced to capture elements that move between subpopulations over time, and metagenomic assemblies tend to break when the same MGE is sampled in multiple contexts (*14*, *15*). As a result, MGEs in nature have largely been studied only one type at a time (*16–18*), which limits context for understanding their collective prevalence, distribution, and ecological impacts.

The potential scope for MGE impacts in natural systems is vast. MGEs can influence host phenotypes and niches through genes they carry (cargo) and/or the host genomic context of MGE insertion (**Fig. 1A**). MGE movements cause gains or losses of function and/or regulatory shifts (*19*, *20*) that should scale and diversify with the number of such movements (**Fig. 1B**). The wider an MGE's host range is, the greater is its range of partners—both hosts and other MGEs—for exchanging genetic modules (*21*). But which genes are affected, and what are the consequences? Do MGEs substantially affect just social functions intensively investigated in cultivated isolates (reviewed above), or might recently observed plasmid-borne genes affecting host "domestic" activities (e.g., stress response, nitrogen fixation) (*22*) be the tip of the iceberg for MGE-influenced processes in nature? Because such domestic activities directly affect microbial ecosystem outputs and adaptability, *in natura* mobilome activity and impacts constitute a knowledge gap that critically impairs our ability to understand and model microbial ecosystem processes and response to change.

Here, we combine recent literature-provided MGE recombinase identification approaches (*8*), a sufficiently large longitudinal metagenomic dataset, and novel analytics development to yield broad new insights into their ecology and ecosystem impacts. We examined ~4.9 terabases of sequencing data collected over eight years along a permafrost peatland thaw gradient in Stordalen Mire, Sweden (**Fig. 2A**), a climate-critical, long-term ecological research site (*23–25*). We reasoned that a site undergoing rapid, landscape-scale change might be an observatory for mobilome-mediated microbiome adaptation to changes in a range of key environmental variables. From these data, we identified, characterized, and genomically and ecologically contextualized ~2.1M MGE recombinases to not only deepen our understanding of how the microbiome of thawing permafrost evolves, but also establish a conceptual and analytic foundation for studying all known, major MGE types in any complex system.

MGE recombinases from permafrost thaw gradient microbiomes

Stordalen Mire, as introduced in companion papers (REFs), is a climate-critical permafrost peatland, extensively studied for decades to assess the interplay of vegetation, microbiota, and climate along a three-stage thaw gradient of palsa, bog, and fen habitats that represent, respectively, the least to most thawed soils. Prior to our work, ~5.8 terabases of metagenomic and metatranscriptomic sequencing data had been generated from 8 years of field sampling (**Fig. 2A, table S1**). We first identified and quantified all major MGEs in these data via recently developed hidden Markov models (HMMs) and MGE identification protocols that leverage recombinases as MGE hallmark genes (*8*) (**table S2**). Specifically, we screened 131.5M genes predicted from 30.7M short-read metagenomic contigs, and this identified 2,107,204 MGE recombinases

(1,130,776 unique protein sequences) derived from 1,790,672 contigs (**Fig. 2B,C&D, fig. S1, fig. S2A&B, fig. S3A&B, table S3 item A**). On average, this represented 17.3 MGE recombinases per million base pairs (Mbp) of microbiome sequence data, dwarfing the representation found for even the most abundant KEGG gene family, the stress-response sigma factor *rpoE* (2.4 copies/Mbp; **Fig. 2C**).

We next classified these abundant recombinases into four previously-established types (*8*): (i) integrons, (ii) insertion sequences and transposons, herein "IS_Tn", (iii) phage / phage-like elements, herein "Phage", and (iv) conjugative elements and mobility islands, herein "CE". Classification facilitates evaluation of ecological impacts, since these types are known to differ in length, mode of propagation, and gene content. Prior work evaluating 187 metagenomes available in 2010 and >80K bacterial and archaeal isolate genomes (*8*, *26*) found IS_Tn dominant in highly abundant MGE pools. In Stordalen Mire, we also found thatIS_Tn was the most commonly recovered type (58.6%), distantly followed by Phage (16.1%), CE (2.8%), and integron (1.2%) (the remaining 21.3% were ambiguous, see Methods; **fig. S2A**). Ecologically, at this high level of recombinase types, this rank-order pattern remained remarkably constant across time (8 years of peak growing season sampling) and space (within and between the study site's three habitat types) (**Fig. 2D, fig. S2B, fig. S2C**). The same pattern largely persisted across phyla for both MAG- and contig-level taxonomic assignments (216,720 and 804,456 MGE recombinases, respectively, assigned to a total of 89 phyla with 93.7% agreement in assignments for the subset shared; **Fig. 3A,B**, **fig. S4A,B, table S3 item A**). MGE distribution broadly followed host lineage relative abundances, with most MGEs in ecosystem-dominant bacterial and archaeal phyla such as Acidobacteriota, Actinomycetota, Pseudomonadota, Verrucomicrobiota, and Halobacteriota (*24*) (**Fig. 3**, **fig. S4&S5**). Echoing isolate-based observations (*8*), the virtual absence of integrons in the abundant Actinomycetota stands out against a background of low cross-phylum variation in the frequency of each MGE type in our Stordalen microbiomes.

Next, we wondered how the MGE frequency of ~25 MGEs/genome robustly measured in isolates would compare with estimates from natural populations, given the potential for MGEs to be lost during laboratory cultivation due to reduced selective pressure in clonal isolates. To this end, we quantified MGEs per genome via two approaches applied to our Stordalen microbiomes. First, we used a contig-based estimate, comparing identified MGE recombinase counts to total genome counts estimated using a standard single-copy marker gene approach for each sample (*27*, *28*). This revealed a median of 36 MGE recombinases per genome in Stordalen microbiomes (**Fig. 2E**). Separately, we considered only the subset of MGEs that were binned into medium/high-quality metagenome-assembled genomes (MAGs) (≥70% completeness, ≤10% contamination) (**table S3 item B**). Scaling per-MAG MGE counts by MAG completeness revealed a median of 23 MGE recombinases per genome (**Fig. 2E**). Comparing the two approaches, we note that only 15.9% of MGE-bearing contigs were binned, even though Stordalen MAGs capture a remarkable 73% of sequence reads at genus level. Thus, we interpret the MAG-based estimate as a lower bound since it empirically captures only a subset of the contig-observed MGEs, and dynamic rearrangements in MGE-bearing genomic regions are a known challenge for short-read metagenome assembly and binning (*29*, *15*, *30*).

To emperically quantify the impact of assembly and binning challenges on short-read counts, we performed long-read sequencing on six samples and evaluated MGE recombinase recovery across the assembly and binning process. In total, these data suggested that short-read assembly and binning recovered 35–60% of MGE recombinases, with IS_Tn recombinases disproportionately under-recovered (**Fig. 2F**, **fig. S6&S7, Supplementary text**). Using this

benchmarking to scale our MAG-based counts, we obtain a revised estimate of 38–66 MGE recombinases per genome. We interpret this number to be comparable with the contig-based estimate and not ruling out the hypothesis that isolates might lose MGEs during lab cultivation.

MGE impacts range far beyond biotic social interactions

We next wondered whether MGEs affected genes of ecological importance, and how these fit into microbiome activity in Stordalen Mire. Prior genome-resolved metagenomic analyses (*24*) show that intercellular interactions and public goods matter in this system: cross-feeding relationships between carbon generalists and methanogens are persistent and strongly associated with methane flux, and extracellular degradation of polysaccharides is a bottleneck in carbon flow. Beyond microbes, there are also abundant viral community members that harbor auxiliary metabolic genes that likely mediate exopolysaccharide breakdown (*25*). Such outward-directed social functions are expected to be affected by MGEs, as are some stress responses (*22*), but these are only a small slice of microbial activities. Each cell must also manage its own routine "domestic" processes (e.g., genome maintenance, gene expression, nutrient homeostasis; (*31–34*)), functions that shape physiology and ecosystem outputs but that have been little studied in the mobilome to date.

Thus, we asked which Stordalen Mire functions could be confidently detected as mobilome-affected (i.e., encoded in MGEs or interrupted by them) and how far beyond biotic interactions the mobilome's direct influence extends. To ensure that host gene activities were not misattributed to MGEs, we adopted a highly conservative approach for curating MGE boundaries. While this permitted assessment of only a small subset of the total 2.1M MGE recombinases identified, it let us identify conservative boundaries for 6,467 cargo-carrying MGEs (1,365 CEs, 681 integrons, and 4,421 phage and phage-like element [PhLs]) and interrupted host genes for 12,332 IS_Tns (**fig. S1, Methods**). Among the 5,188 KEGG gene annotations carried or interrupted by this subset of MGEs, our manually curated annotation identified functional predictions with high confidence for 23.5% (1,220 from 3,577 MGEs; **table S3 item C-F**, **table S4, Supplementary text**).

As expected, this reduced set of curated MGE-affected genes included many related to host biotic interactions (n=1321; **Fig. 4, table S4**). This included diverse phage defense systems (379 carried, 168 interrupted) that can abort infection, degrade phage DNA by restriction-modification, or inhibit phage DNA synthesis by nucleotide depletion (dCTP deaminase) (*11*, *35*). MGEs also impact genes for YD-repeat-containing proteins (10 carried, 29 interrupted), a contact-dependent growth inhibitor thought to shape responses to high population densities (*36*). As well, public goods appear MGE-impacted as MGEs impact genes encoding carbohydrate-active enzymes (CAZymes) involved in synthesis or degradation of extracellular polysaccharides (154 carried, 125 interrupted). The former produce public goods and molecules involved in cell-phage and cell-cell interactions, while the latter can become public goods through direct secretion (extracellular enzymes) or outer membrane vesicle budding (periplasmic enzymes) (*37*). Separately, MGEs appear to impact synthesis and transport of other public goods like siderophores (39 carried, 58 interrupted). Finally, MGEs can impactgenes that encode a public service (22 carried, 13 interrupted), the detoxification of diffusible peroxides (*38*).

Beyond biotic interactions, we identified diverse MGEs seemingly affecting domestic processes. For example, we uncovered MGEs affecting central carbon metabolism genes (6 carried, 45 interrupted; **table S4**). Intriguingly, even in the small subset of conservatively-cuated MGEs, there were two direct connections to methane metabolism: the key hydrogenotrophic metabolism gene formate dehydrogenase (*39*) was carried by a Phage from Halobacteriota

(*Methanoregula* sp003136075), a group of methanogens active at Stordalen Mire (*23*, *24*), and an IS_Tn interrupted a methane monooxygenase (*pmoC*) subunit in a Pseudomonadota (unbinned Alphaproteobacterial contig). Thus, even within this highly conservative subset of MGE-affected genes, we find genomic evidence that MGEs may alter Stordalen methane fluxes both directly (by changing genomic potential for methanogenesis and methanotrophy) and indirectly (by changing the distribution of carbon-processing pathways that control the flow of substrates to methanogens).

MGEs also might influence other major nutrient cycles, trace metal cycles, and cellular housekeeping functions. In nitrogen cycling, we found MGEs influencing nitrogen processing (18 carried, 57 interrupted; **Fig. 4, table S4**) across the thaw gradient, predominantly in amino acid metabolism but also touching other nitrogen processes, including fixation. Because fixed nitrogen is thought to be a partially "leaky" public good (*40*), IS_Tn interruption of an Acidobacterial nitrogen fixation regulator (*41*) illustrates the potential for MGEs influencing growth beyond the host cell by controlling nitrogen. The CAZyme genes highlighted above include 10 chitinase genes carried and 1 interrupted by MGEs; chitin is hypothesized to serve as a key alternative nitrogen source at our N-limited site. In sulfur cycling, we found MGEs that affect genes putatively related to inorganic sulfur limitation (19 carried, 15 interrupted). Curiously, their distribution, though with low statistical power, suggests possible niche-driven selection as IS_Tn elements interrupted 11 sulfatases, all in the palsa, that might otherwise degrade sulfated polysaccharides under sulfur limitation (*32*), whereas 4 bog MGEs carried cysteine desulfurases and selenocysteine lyases that might scavenge thiol groups and reduce sulfur demand when sulfur is very limiting (*42*, *43*). Beyond macronutrient cycling, the siderophore transporters (highlighted above as public goods) might also connect MGEs to iron cycling (45 carried, 37 interrupted). Finally, MGEs also encode core housekeeping genes involved in genome maintenance (DNA repair, *radC*; genome replication, *dinB*), regulation of host gene transcription (RNA polymerase subunits, transcriptional regulators), and translation (elongation factor, ribosomal proteins), as well as proteome maintenance (chaperonin, ubiquitin activation, proteases).

Critically, even though our stringent analysis criteria reduced the set of MGE-affected genes analyzed to 0.02% of the total MGEs identified, we still find abundant examples of mobilome-affected social and domestic functions beyond previous observations (*44*), and add many further functional MGE-impacted examples including those influence carbon flow through this climate-critical ecosystem, both directly (exopolysaccharide degradation, central carbon metabolism) and indirectly (via nitrogen, sulfur, and/or iron availability to support carbon flux). Although research to date has focused on MGE impacts on biotic interactions, MGE impacts on domestic functions do not appear to be rare, at the level of either the functions (~55% domestic vs. ~38% social) or the MGEs themselves (~52% affect domestic functions). If this conservative subset is representative of the Stordalen mobilome as a whole, then as sequencing approaches, functional annotation, and databases improve, we would expect ~1.1M of the full ~2.1M MGE recombinases to influence domestic functions, with significant implications for nutrient cycling, as well as microbiome activity, ecology, and evolution.

MGE recombinase mobility and activity

To assess ecological and evolutionary change of Stordalen Mire MGEs, we evaluated their mobility and activity from three different viewpoints: (i) snapshots of metatranscriptomic MGE recombinase expression as a proxy for mobilization underway, (ii) genomic neighborhood change as a marker of completed MGE movement between insertion sites, and (iii) partial carriage of an MGE by the host population as the result of evolution and selection over longer time scales.

Integrating results across these analyses can help highlight individual MGE dynamics. For example, high recombinase expression rates together with low rates of genomic neighborhood change for a given MGE would suggest a strong preference among insertion sites. Our overall findings are as follows.

First, we asked whether MGE recombinases were "transcriptionally active" as measured via read mapping from 50 available metatranscriptomes against contigs from paired metagenomes (**table S5**), taking ribosomal gene mapping rates as a baseline for constitutive transcriptional activity of actively growing cells (*45*) (**table S6**). This revealed transcriptional activity in ~3.4–56.4% (median: 27.8%) of ribosomal genes and ~0.04-9.1% (median: 1.2%) of MGE recombinase genes (**Fig. 5A, table S3 item G**) – with MGE recombinases the second-most-transcribed non-ribosomal genes in Stordalen Mire (**Fig. 5A**). Stratified by recombinase type, we found roughly equal levels of transcriptional activity for IS_Tn and Phage recombinases versus ~five- to ten-fold lower integron and CE recombinase transcription (**Fig. 5B**). Ecologically, IS_Tn recombinase transcription increased along the thaw gradient (significant from palsa to fen, $p \leq 0.01$, Wilcoxon rank-sum test) (**Fig. 5C, fig.S8B**), suggesting increased MGE activity in the more-disturbed fen. Taxonomically, transcriptionally active MGE recombinases were detected across hosts spanning 30 phyla with activity largely proportional to MGE- and phylum abundance (**Fig. 5D, fig. S8C**). However, one of these, the archaeal phylum *Halobacteriota,* was unique in that it is neither abundant nor a MGE recombinase-rich phylum, and yet was the most active phylum for MGE recombinase transcription. Given that Stordalen Mire *Halobacteriota* are known to be methanogens that are active and selected for in thawed fen and bog habitats (*23*, *46*), we speculate that their disproportionate MGE recombinase activity might be adaptive.

Second, where MGE recombinases were observed multiple times, we assessed the genomic context (300 bp upstream and downstream) surrounding. We then interpreted this genomic context as a proxy for intra/inter-genomic mobility, either of the recombinase itself or of other recombinase-proximal genes (see examples of changing neighborhoods in **fig. S9, Methods**). From the original ~2.1M MGE recombinases, we identified 74,257 MGE recombinases clusters (100% average amino acid identity[AAI])) observed more than once in our dataset. Among these, most had identical genomic neighborhoods across observations, but some (0.1-10%, depending on MGE type) did show neighborhood variation (**Fig. 5B**, **table S3 item H**). Neighborhood variability was most common among IS_Tn recombinases (~10%), and much less so for Phage (~1%), integron (0.2%) and CE (0.1%) recombinases. This was consistent among years and habitats, with larger variation detected across diverse host phyla (up to 37% in Desulfobacterota_G) (**fig. S10**). Previous analysis in human gut microbiomes has detected changing gene contexts for ~half of mobile genes identified (*47*), using an approach that captured changing linkages among mobile genes, including MGE cargo genes, rather than MGE mobility with respect to host genomes. Mindful that hypervariable islands in cellular genomes can often be impacted by multiple MGE types, we quantified this and found that less than 1 in 5 of our changed-genomic-neighborhood examples was due to co-occurring IS_tn insertions (**fig. S9**). Considering our approach only detects insertion at immediate neighborhoods MGE recombinase, this is a lower bound estimate of mobility at hypervariable multiple-MGE genomic islands among MGE context changes at Stordalen.

Third, we sought to assess what fraction of a population contains any given MGE recombinase by evaluating read-mapping-based coverage of the recombinase against that of its surrounding genomic context (see Methods). Where MGE recombinase coverage was significantly lower, we infer that only a subset of the host population's genotypes encoded the MGE recombinase,

presumably a function of MGE mobility and possible selection over longer evolutionary timescales than the genomic context analyses above. Overall, across MGE types, we found 8,033 recombinase clusters (100% AAI) that had ≥20 kbp of genomic context. Of these, ~5–23% showed variation in this genomic context consistent with at least part of the cellular population having the MGE and part not having the MGE (**Fig. 5B, table S3 item I**). These findings were again highly consistent across years and habitats (**fig. S11B&C**). Because of variation in MGE sizes and inherent noise in read mapping analysis for lesser-sampled taxa, we interpret these numbers as conservative lower bounds on the true MGE penetrance into cellular population genomes. In Stordalen Mire, is appears probable that IS_Tn and Phage are particularly prone to being variably uncovered in member genomes of any given host population. Functionally this can have impacts as, for example, about 1 in 5 IS_Tn elements variably interrupt a protein-coding gene or not (36 out of 162 amenable for this analysis). These results all suggest that MGE penetrance within member genomes of any given population can differentially provide individuals access to a diverse repertoire of genes while shielding the population from loss of function due to MGE insertion (*48*, *49*).

Conclusions

MGEs have long been known from cultivated isolate genomes to be important drivers of strain-level variation and key functional adaptations, particularly in biomedical contexts related to AMR (*7–9*). However, virtually no data exist for diverse MGE types in natural settings. Our analyses presented here for Stordalen Mire microbiomes illuminate the mechanisms and scope of MGE impact on a climate-sensitive ecosystem by showing that, well beyond biotic interactions, the mobilome affects functions intrinsic to a wide range of physiologically and ecologically significant processes. Strikingly, the IS_Tn elements that heavily dominate recovered MGE recombinases in our dataset are also the most active by two of our three measures (i.e., transcription and genomic neighborhood change) (**Fig. 5B**), highlighting the need for future studies to examine context-dependent MGE effects like modulation of host gene expression downstream of MGE insertion sites.

As improved methods for studying the genomic potential of microbes in nature come on-line, the field has progressed from community-level taxonomic and functional descriptions towards species-resolved links between taxa and their functions (*24*, *27*). Continued advances in sequencing and analytical approaches are increasing this resolution to reveal intra-population variation, early speciation, and selection (*50*) with recent MGE typing efforts (REF) and further analytical advances made here finally bringing a quantitative understanding of the mobilome within reach, even in complex natural systems. This work benchmarks sequencing artifacts and provides empirical baselines and inferred impacts for all major MGEs in a natural soil system. These findings suggest that ~1% of MGE recombinases are active, genomes contain 36–66 MGE recombinases on average, and that MGEs impact (via carriage or interruption) diverse ecosystem-relevant functions. As technologies advance, the mobilome will come more fully into focus, and, given this new toolkit and baseline findings, we predict it will help reveal a subtler layer of change that could serve as the canary in the coalmine for early detection of change.

Acknowledgements

Non-author contribution: we thank Ohio Supercomputer Center for providing computing resources.

References

(See the full list at the end of this doc)

Funding: this work is supported by following sources of funding:
National Science Foundation, Biology Integration Institutes Program, Award # 202207. The work conducted by the U.S. Department of Energy Joint Genome Institute (JGI; https://ror.org/04xm1d337), a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy operated under Contract No. DE-AC02-05CH11231. This work by JGI includes BER Support Science Proposal 503530 (used for the majority of samples sequenced by JGI), and Facilities Integrating Collaborations for User Science (FICUS) initiative proposal 503547 (used for sequencing samples from the SIP experiment).

Author contribution: J.G., S.R., M.B.S., and S.C.B. conceptualized, designed, coordinated the study. S.A., J.G. performed long read and short read comparison. S.A., A.F., J.G. performed MGE recombinase host lineage analyses. G.D.-H., D.S., D.V., S.C.B., A.A.P., G.S., F.T., C.H.-V., S.S., Z.Z., C.O.A., J.G. performed MGE impacted function analyses. D.C., S.A., S.B.H. performed metagenome assembly and binning. J.G., S.R., M.B.S., and S.C.B. drafted the manuscript with contribution from B.J.W., G.W.T., V.I.R. All authors read, edited, commented on, and approved the final manuscript.

Competing interests: The authors declare that they have no competing interests.

Data and material availability: The authors declare that all the data in this study are freely and publicly available without restriction. The metagenomic and metatranscriptomic raw reads are available in SRA / ENA under Umbrella BioProject PRJNA888099 with individual accessions cited in **table S1**. The data provided by the DOE Joint Genome Institute were generated under proposals "10.46936/10.25585/60001148" and "10.46936/fics.proj.2017.49950/60006215". MAGs used in this study are available in Zenodo with https://doi.org/10.5281/zenodo.13984630 (main set of 36,419 MAGs) and https://doi.org/10.5281/zenodo.14207415 (733 short-read, long-read, and hybrid MAGs from six paired samples). The paired long read and short read data from six samples are available in PRJNA888099 with accessions cited in **table S7**. Processed data including tables for MGE recombinase identification, activity by metatranscriptome, mobility by genomic neighborhood change, and partial carriage, and MGE impacted functions are available through supplementary tables and a large file data package shared through CERN's Zenodo repository (https://zenodo.org/records/14287562). The code produced in this study is also included in the above repository, which is an archive of the latest version of the project's GitHub repository (https://github.com/emerge-bii/mobilome-submit-202407).

Supplementary Materials

Methods and Materials
Supplementary Text
Figs. S1 - S11
Tables S1 - S7
References (*53 - 107*)

Figures

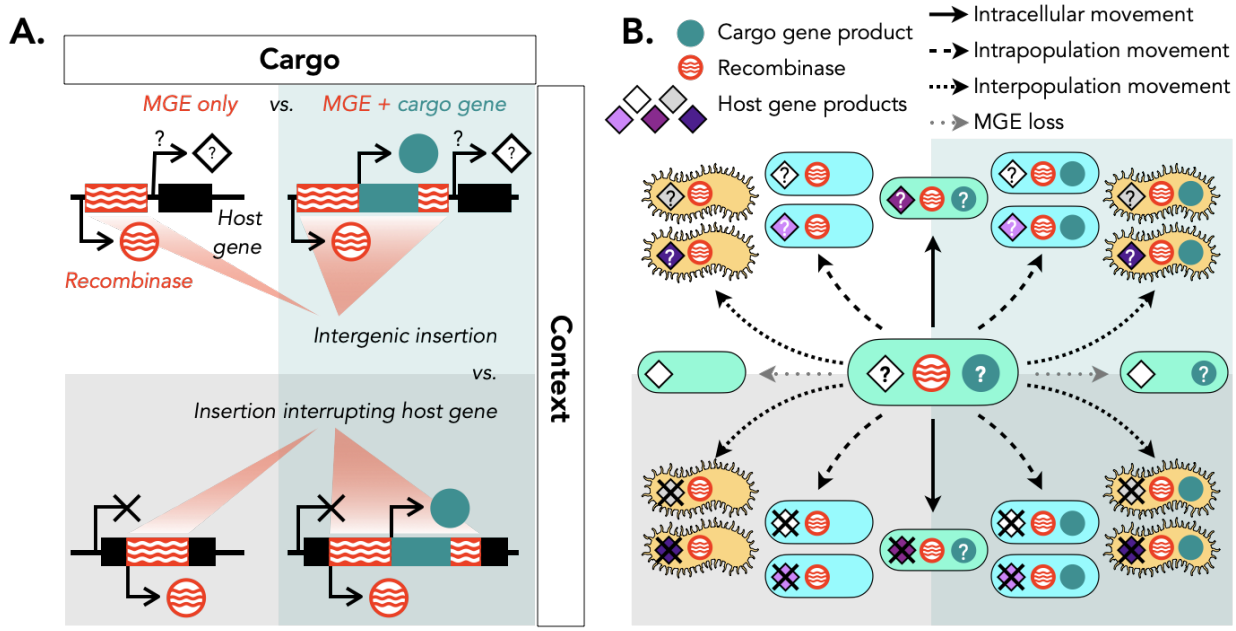

Figure 1. MGE cargo and context determine the impact of MGE activity on host coding capacity and population genetic diversity. A. An MGE may or may not carry cargo and may insert either intergenically, potentially altering transcriptional regulation downstream, or into a host gene, causing function loss. MGE and recombinase gene product, red and white waves; cargo gene and product, solid teal; host gene, black; host gene product, white diamond. Disrupted genes yield no product (X); genes downstream of MGE insertions may or may not be transcribed (?). **B.** Each cycle of excision and integration may shift the MGE and its insertion site and/or cargo between contexts, generating new genetic diversity. Intracellular movement (solid arrows, teal cells) may interrupt (X) or influence gene regulation of (?) a new host gene (diamonds); intrapopulation movement (dashed arrows, blue cells) lets the MGE interrupt or influence the same or different host genes in a new genetic background; interpopulation movement (dot-dashed arrows, yellow cells) lets the MGE interrupt or influence a new suite of genes; and MGE loss (gray dotted arrows) may remove or leave behind cargo genes and potentially restore function of previously disrupted host genes. Not depicted: capture of host genes as new cargo genes.

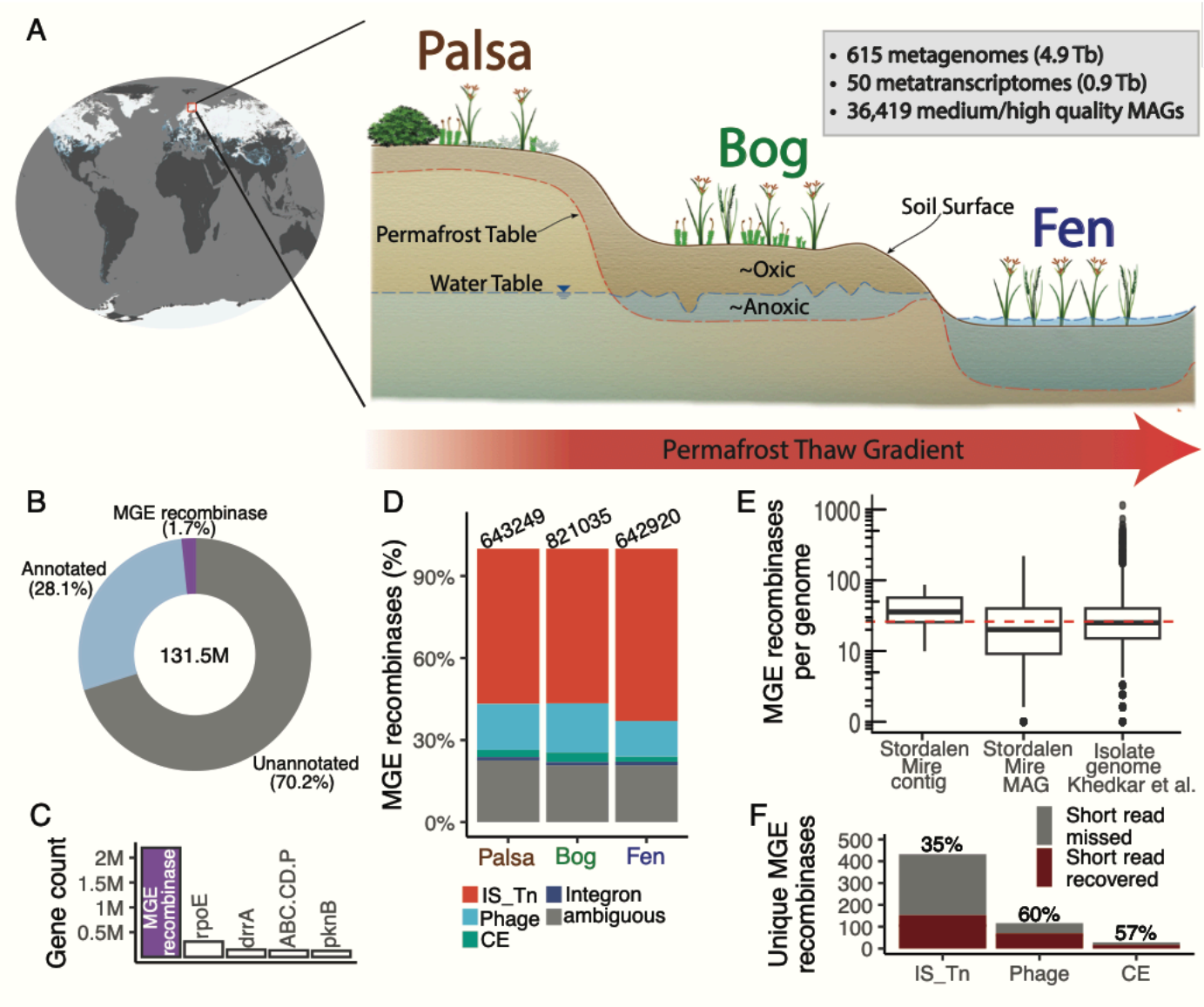

Figure 2. The Stordalen Mire model ecosystem and its mobile genetic element (MGE) recombinases. A. Long-term study site (image credit: NASA) with large datasets (gray inset) along a palsa-bog-fen thaw gradient. **B.** MGE recombinase percentage in all genes from Stordalen contigs (≥1,500 bp). **C**. Counts of most abundant genes: MGE recombinase; *rpoE*, RNA polymerase sigma 70 factor; *drrA*, antibiotic transport system ATP binding protein; *pknB*, serine/threonine protein kinase; *ABC.CD.P*, putative ABC transport system permease. **D.** MGE recombinase type distributions: "IS_Tn", insertion sequence or transposon; "Phage", phage and phage-like elements; "CE", conjugative elements or mobility islands; "Integron", integron; "ambiguous", recombinases that could be found in >1 MGE types. **E.** MGE estimates per genome for Stordalen Mire contigs (median=36) or MAGs (median=23) and from cultured isolates in Khedkar et al. (median=25; red dashed line). Filled circles, points >1.5 times the interquartile range below the first quartile or above the third quartile. **F.** Comparison of MGE recombinase recovery between short-read and hybrid assemblies (n = 17). "Short read missed": recombinase found only in hybrid MAGs; "short read recovered": recombinase found in both hybrid and short-read MAGs (also shown as percentages on top of bars). Recombinases were

matched across assemblies based on 100% identical protein sequence matches.

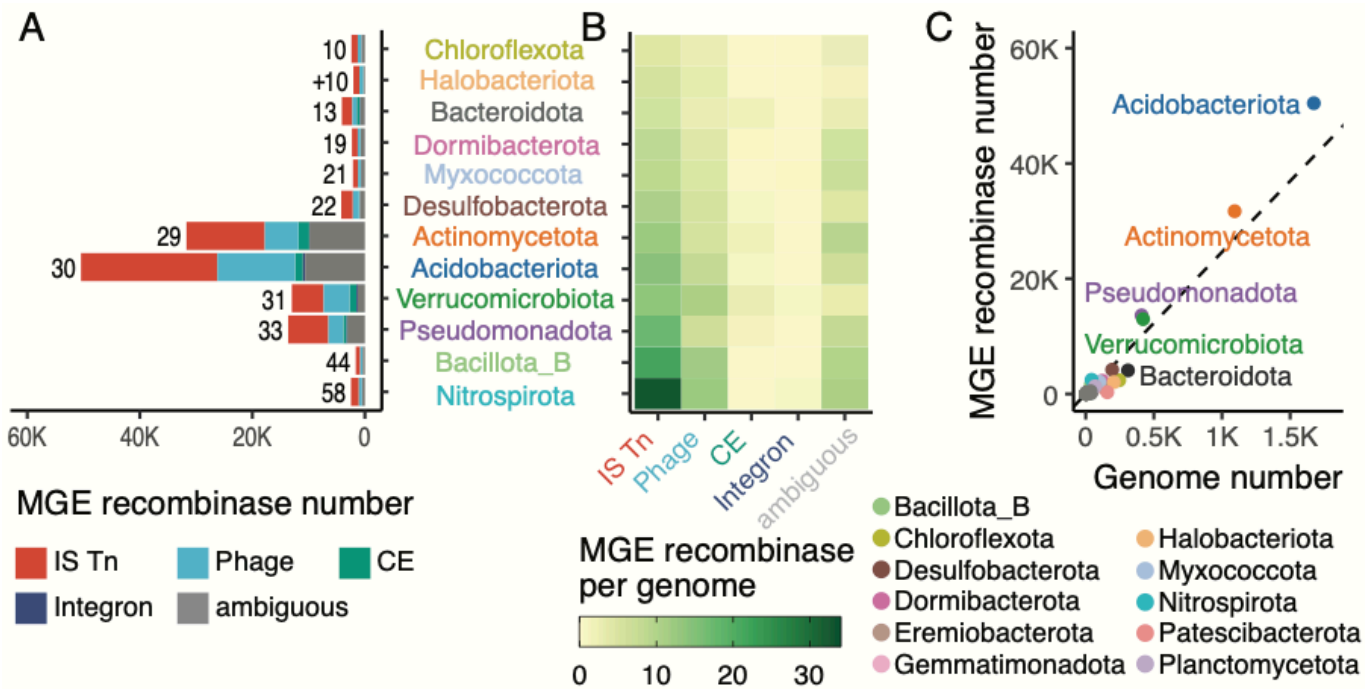

Figure 3. MGE recombinase type distribution across microbial host lineages (MAG-based) of the full Stordalen Mire short read dataset. A. Number of each MGE type found in each phylum, as identified from contigs binned into MAGs. MGE recombinase counts shown for the top 12 phyla (out of 46 total), ordered by average MGE recombinase number per genome (indicated to the left of each bar with "+" indicating Archaea). **B.** Number of MGE recombinases identified per genome grouped by MGE type and host phylum. **C.** Number of MGE recombinases (y-axis) against estimated number of genomes per phylum (x-axis). Dotted line represents the global phylum-averaged number of MGE recombinases per genome.

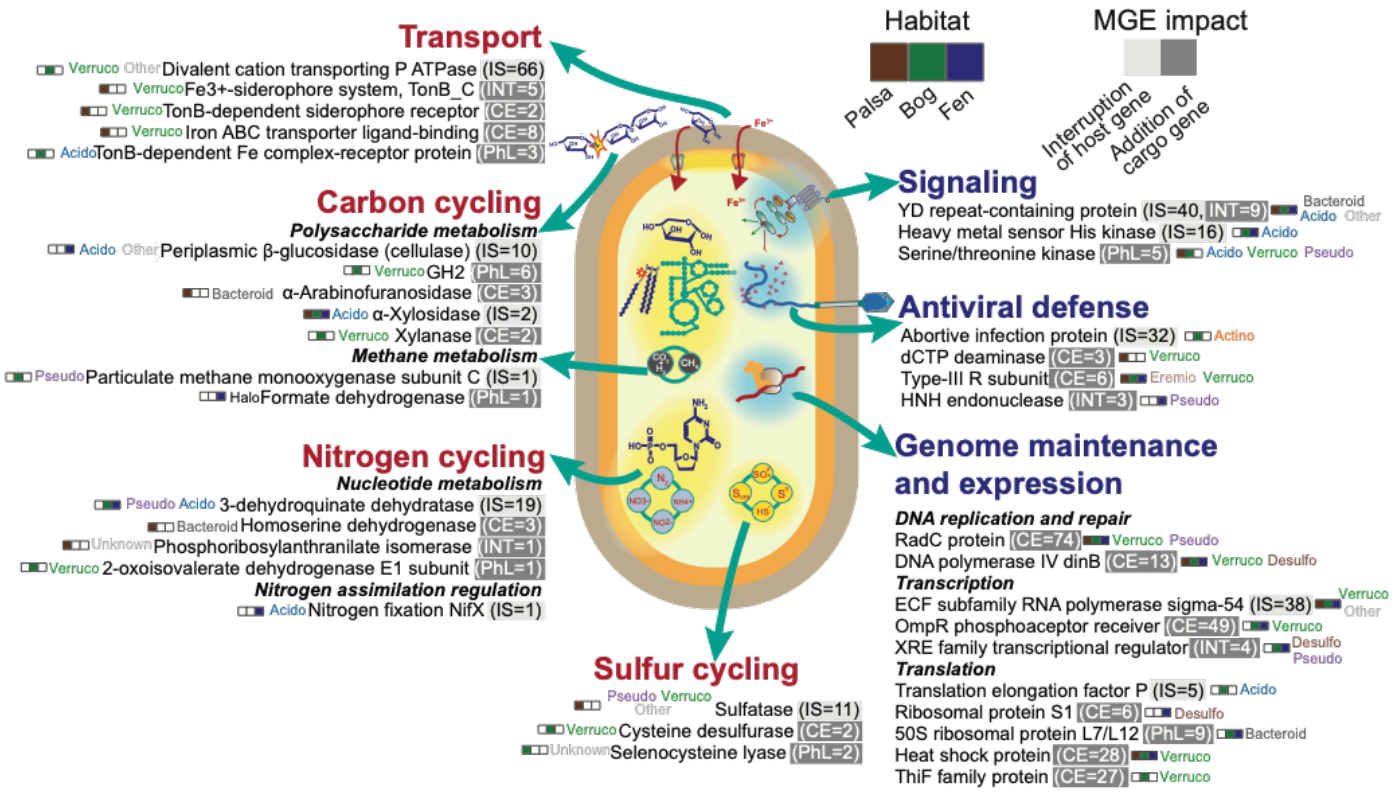

Figure 4. Host functions impacted by MGEs in Stordalen Mire. Metabolic pathways (red text) and regulatory activities (blue text) impacted by MGEs, either through inactivation by IS elements (loss of function by interruption of hots gene) or through transfers across genomes by CEs, Phage-like elements (PhL), and integrons (INT) (gain of function by addition of cargo gene). For each gene impacted, the type and number of MGEs is in parenthesis, and the host microbe and habitat are on the left or right.

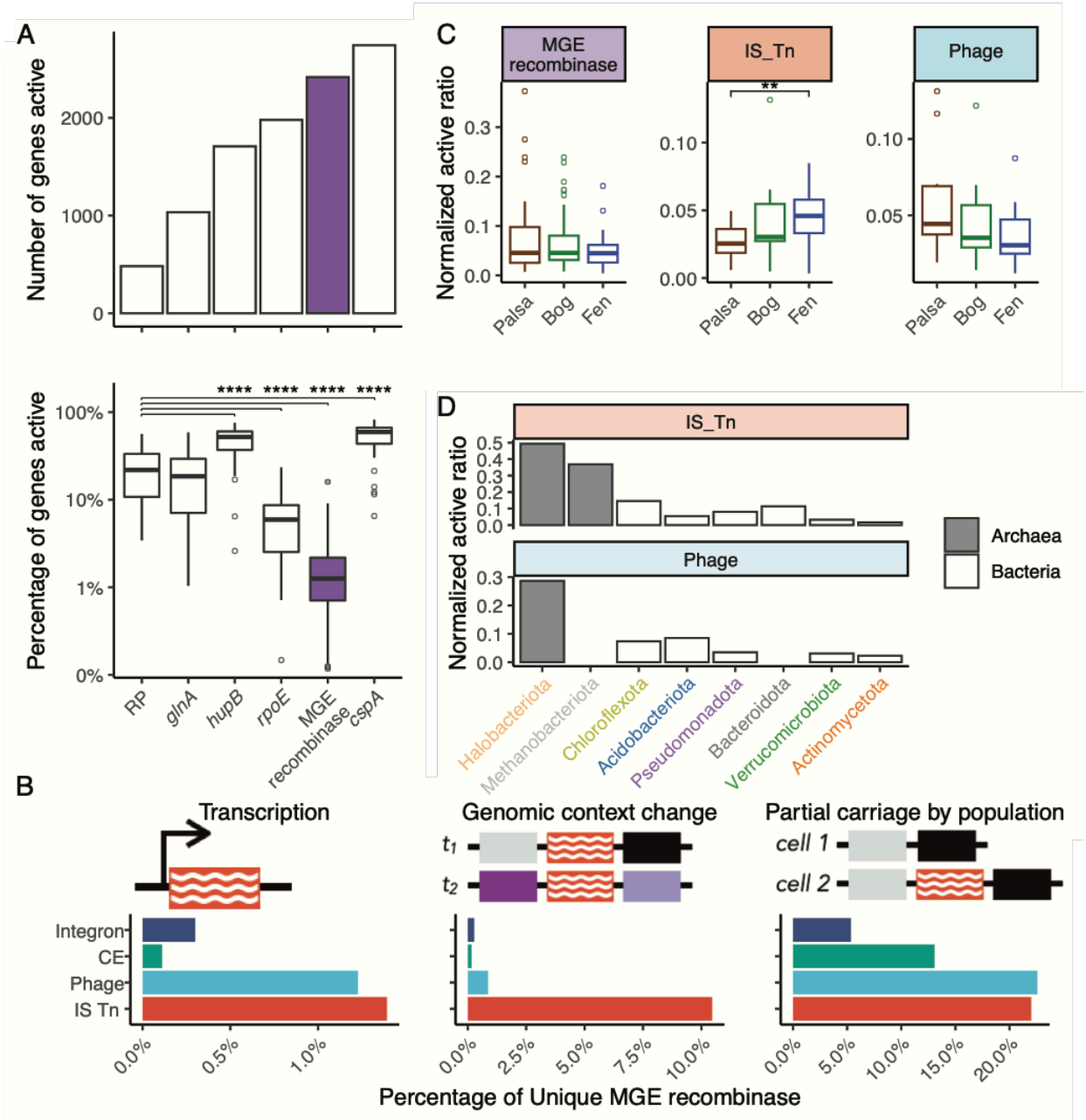

Figure 5. MGE recombinase activity. A. Top panel: Genes detected most frequently as active based on metatranscriptome read mapping. Gene categories from right to left: *cspA*, major cold shock protein gene; MGE recombinase; *rpoE*, RNA polymerase sigma-70 factor gene; *hupB*, DNA binding protein HU-beta gene; *glnA,* glutamine synthetase gene; RP, ribosomal protein genes. Bottom panel: percentage of genes active based on the full set of genes identified in the matching metagenome. Wilcoxon rank-sum test of statistical significance; ****, *P* value≤0.0001. **B**. Three facets of MGE recombinase activity/mobility-- transcription (left), genomic neighborhood stability (middle), partial carriage by host population (right) – are conceptually presented (upper panel) and quantified and summarized per MGE recombinase type (lower panel). The x-axis is the percentage of unique MGE recombinase (100% identity at protein level) detected with activity / mobility within each type. **C**. Comparison of ratio of active MGE

recombinase, “IS_Tn” , and “Phage” recombinases across three habitats, normalized by the ratio of active ribosomal genes for the same sample. **D.** Normalized active ratio of “IS_Tn” and “Phage” recombinase across host phyla with ≥5 active “IS_Tn” or “Phage” recombinases, and ribosomal genes. “CE” and “Integron” recombinases are not shown here due to low counts.

References

1. T. W. Crowther, J. van den Hoogen, J. Wan, M. A. Mayes, A. D. Keiser, L. Mo, C. Averill, D. S. Maynard, The global soil community and its influence on biogeochemistry. *Science*, doi: 10.1126/science.aav0550 (2019).
2. L. Paoli, H.-J. Ruscheweyh, C. C. Forneris, F. Hubrich, S. Kautsar, A. Bhushan, A. Lotti, Q. Clayssen, G. Salazar, A. Milanese, C. I. Carlström, C. Papadopoulou, D. Gehrig, M. Karasikov, H. Mustafa, M. Larralde, L. M. Carroll, P. Sánchez, A. A. Zayed, D. R. Cronin, S. G. Acinas, P. Bork, C. Bowler, T. O. Delmont, J. M. Gasol, A. D. Gossert, A. Kahles, M. B. Sullivan, P. Wincker, G. Zeller, S. L. Robinson, J. Piel, S. Sunagawa, Biosynthetic potential of the global ocean microbiome. *Nature* **607**, 111–118 (2022).
3. S. Nayfach, Z. J. Shi, R. Seshadri, K. S. Pollard, N. C. Kyrpides, New insights from uncultivated genomes of the global human gut microbiome. *Nature* **568**, 505–510 (2019).
4. P. J. Turnbaugh, M. Hamady, T. Yatsunenko, B. L. Cantarel, A. Duncan, R. E. Ley, M. L. Sogin, W. J. Jones, B. A. Roe, J. P. Affourtit, M. Egholm, B. Henrissat, A. C. Heath, R. Knight, J. I. Gordon, A core gut microbiome in obese and lean twins. *Nature* **457**, 480–484 (2009).
5. M. Haudiquet, J. M. de Sousa, M. Touchon, E. P. C. Rocha, Selfish, promiscuous and sometimes useful: how mobile genetic elements drive horizontal gene transfer in microbial populations. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **377**, 20210234 (2022).
6. M. G. Durrant, M. M. Li, B. A. Siranosian, S. B. Montgomery, A. S. Bhatt, A Bioinformatic Analysis of Integrative Mobile Genetic Elements Highlights Their Role in Bacterial Adaptation. *Cell Host Microbe* **27**, 140-153.e9 (2020).
7. S. R. Partridge, S. M. Kwong, N. Firth, S. O. Jensen, Mobile Genetic Elements Associated with Antimicrobial Resistance. *Clin. Microbiol. Rev.* **31**, e00088-17 (2018).
8. S. Khedkar, G. Smyshlyaev, I. Letunic, O. M. Maistrenko, L. P. Coelho, A. Orakov, S. K. Forslund, F. Hildebrand, M. Luetge, T. S. B. Schmidt, O. Barabas, P. Bork, Landscape of mobile genetic elements and their antibiotic resistance cargo in prokaryotic genomes. *Nucleic Acids Res.* **50**, 3155 (2022).
9. J. Zhang, Y. Xu, M. Wang, X. Li, Z. Liu, D. Kuang, Z. Deng, H.-Y. Ou, J. Qu, Mobilizable plasmids drive the spread of antimicrobial resistance genes and virulence genes in Klebsiella pneumoniae. *Genome Med.* **15**, 106 (2023).
10. D. J. Rankin, E. P. C. Rocha, S. P. Brown, What traits are carried on mobile genetic elements, and why? *Heredity* **106**, 1–10 (2011).
11. E. P. C. Rocha, D. Bikard, Microbial defenses against mobile genetic elements and viruses: Who defends whom from what? *PLoS Biol.* **20**, e3001514 (2022).
12. A. Dragoš, A. J. C. Andersen, C. N. Lozano-Andrade, P. J. Kempen, Á. T. Kovács, M. L. Strube, Phages carry interbacterial weapons encoded by biosynthetic gene clusters. *Curr. Biol. CB* **31**, 3479-3489.e5 (2021).
13. S. Benler, G. Faure, H. Altae-Tran, S. Shmakov, F. Zhang, E. Koonin, Cargo Genes of Tn7-Like Transposons Comprise an Enormous Diversity of Defense Systems, Mobile Genetic Elements, and Antibiotic Resistance Genes. *mBio* **12**, e02938-21.
14. J. J. Kerkvliet, A. Bossers, J. G. Kers, R. Meneses, R. Willems, A. C. Schürch, Metagenomic assembly is the main bottleneck in the identification of mobile genetic elements. *PeerJ* **12**, e16695 (2024).

15. F. Maguire, B. Jia, K. L. Gray, W. Y. V. Lau, R. G. Beiko, F. S. L. Brinkman, Metagenome-assembled genome binning methods with short reads disproportionately fail for plasmids and genomic Islands. *Microb. Genomics* **6**, mgen000436 (2020).
16. S. Roux, J. R. Brum, B. E. Dutilh, S. Sunagawa, M. B. Duhaime, A. Loy, B. T. Poulos, N. Solonenko, E. Lara, J. Poulain, S. Pesant, S. Kandels-Lewis, C. Dimier, M. Picheral, S. Searson, C. Cruaud, A. Alberti, C. M. Duarte, J. M. Gasol, D. Vaqué, Tara Oceans Coordinators, P. Bork, S. G. Acinas, P. Wincker, M. B. Sullivan, Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537**, 689–693 (2016).
17. J. Tamayo-Leiva, J. Alcorta, F. Sepúlveda, S. Fuentes-Alburquenque, J. I. Arroyo, J. E. González-Pastor, B. Díez, Structure and dispersion of the conjugative mobilome in surface ocean bacterioplankton.
18. Y. Yi, S. Liu, Y. Hao, Q. Sun, X. Lei, Y. Wang, J. Wang, M. Zhang, S. Tang, Q. Tang, Y. Zhang, X. Liu, Y. Wang, X. Xiao, H. Jian, A systematic analysis of marine lysogens and proviruses. *Nat. Commun.* **14**, 1–15 (2023).
19. P. Prentki, B. Teter, M. Chandler, D. J. Galas, Functional promoters created by the insertion of transposable element IS1. *J. Mol. Biol.* **191**, 383–393 (1986).
20. A. Lipszyc, M. Szuplewska, D. Bartosik, How Do Transposable Elements Activate Expression of Transcriptionally Silent Antibiotic Resistance Genes? *Int. J. Mol. Sci.* **23**, 8063 (2022).
21. A. J. Weisberg, J. H. Chang, Mobile Genetic Element Flexibility as an Underlying Principle to Bacterial Evolution. *Annu. Rev. Microbiol.* **77**, 603–624 (2023).
22. S. S. Finks, J. B. H. Martiny, Plasmid-Encoded Traits Vary across Environments. *mBio* **14**, e0319122 (2023).
23. C. K. McCalley, B. J. Woodcroft, S. B. Hodgkins, R. A. Wehr, E.-H. Kim, R. Mondav, P. M. Crill, J. P. Chanton, V. I. Rich, G. W. Tyson, S. R. Saleska, Methane dynamics regulated by microbial community response to permafrost thaw. *Nature* **514**, 478–481 (2014).
24. B. J. Woodcroft, C. M. Singleton, J. A. Boyd, P. N. Evans, J. B. Emerson, A. A. F. Zayed, R. D. Hoelzle, T. O. Lamberton, C. K. McCalley, S. B. Hodgkins, R. M. Wilson, S. O. Purvine, C. D. Nicora, C. Li, S. Frolking, J. P. Chanton, P. M. Crill, S. R. Saleska, V. I. Rich, G. W. Tyson, Genome-centric view of carbon processing in thawing permafrost. *Nature* **560**, 49–54 (2018).
25. J. B. Emerson, S. Roux, J. R. Brum, B. Bolduc, B. J. Woodcroft, H. B. Jang, C. M. Singleton, L. M. Solden, A. E. Naas, J. A. Boyd, S. B. Hodgkins, R. M. Wilson, G. Trubl, C. Li, S. Frolking, P. B. Pope, K. C. Wrighton, P. M. Crill, J. P. Chanton, S. R. Saleska, G. W. Tyson, V. I. Rich, M. B. Sullivan, Host-linked soil viral ecology along a permafrost thaw gradient. *Nat. Microbiol.* **3**, 870–880 (2018).
26. R. K. Aziz, M. Breitbart, R. A. Edwards, Transposases are the most abundant, most ubiquitous genes in nature. *Nucleic Acids Res.* **38**, 4207–4217 (2010).
27. C. He, R. Keren, M. L. Whittaker, I. F. Farag, J. A. Doudna, J. H. D. Cate, J. F. Banfield, Genome-resolved metagenomics reveals site-specific diversity of episymbiotic CPR bacteria and DPANN archaea in groundwater ecosystems. *Nat. Microbiol.* **6**, 354–365 (2021).
28. S. Roux, B. G. Paul, S. C. Bagby, S. Nayfach, M. A. Allen, G. Attwood, R. Cavicchioli, L. Chistoserdova, R. J. Gruninger, S. J. Hallam, M. E. Hernandez, M. Hess, W.-T. Liu, T. A. McAllister, M. A. O'Malley, X. Peng, V. I. Rich, S. R. Saleska, E. A. Eloe-Fadrosh, Ecology and molecular targets of hypermutation in the global microbiome. *Nat. Commun.* **12**, 3076 (2021).
29. O. Zablocki, M. Michelsen, M. Burris, N. Solonenko, J. Warwick-Dugdale, R. Ghosh, J. Pett-Ridge, M. B. Sullivan, B. Temperton, VirION2: a short- and long-read sequencing and informatics workflow to study the genomic diversity of viruses in nature. *PeerJ* **9**, e11088 (2021).
30. M. L. Coleman, M. B. Sullivan, A. C. Martiny, C. Steglich, K. Barry, E. F. DeLong, S. W. Chisholm, Genomic Islands and the Ecology and Evolution of Prochlorococcus. *Science*, doi: 10.1126/science.1122050 (2006).
31. K. J. Romanowicz, B. C. Crump, G. W. Kling, Genomic evidence that microbial carbon degradation is dominated by iron redox metabolism in thawing permafrost. *ISME Commun.* **3**, 1–11 (2023).

32. M. Pester, K.-H. Knorr, M. W. Friedrich, M. Wagner, A. Loy, Sulfate-reducing microorganisms in wetlands – fameless actors in carbon cycling and climate change. *Front. Microbiol.* **3**, 19769 (2012).
33. J. Limpens, F. Berendse, C. Blodau, J. G. Canadell, C. Freeman, J. Holden, N. Roulet, H. Rydin, G. Schaepman-Strub, Peatlands and the carbon cycle: from local processes to global implications – a synthesis. *Biogeosciences* **5**, 1475–1491 (2008).
34. R. AminiTabrizi, R. M. Wilson, J. D. Fudyma, S. B. Hodgkins, H. M. Heyman, V. I. Rich, S. R. Saleska, J. P. Chanton, M. M. Tfaily, Controls on Soil Organic Matter Degradation and Subsequent Greenhouse Gas Emissions Across a Permafrost Thaw Gradient in Northern Sweden. *Front. Earth Sci.* **8**, 557961 (2020).
35. N. Tal, R. Sorek, SnapShot: Bacterial immunity. *Cell* **185**, 578-578.e1 (2022).
36. S. Koskiniemi, J. G. Lamoureux, K. C. Nikolakakis, C. t'Kint de Roodenbeke, M. D. Kaplan, D. A. Low, C. S. Hayes, Rhs proteins from diverse bacteria mediate intercellular competition. *Proc. Natl. Acad. Sci.* **110**, 7032–7037 (2013).
37. S. Rakoff-Nahoum, M. J. Coyne, L. E. Comstock, An ecological network of polysaccharide utilization among human intestinal symbionts. *Curr. Biol. CB* **24**, 40–49 (2014).
38. J. J. Morris, R. E. Lenski, E. R. Zinser, The Black Queen Hypothesis: Evolution of Dependencies through Adaptive Gene Loss. *mBio* **3**, 10.1128/mbio.00036-12 (2012).
39. M. F. Abdul Halim, D. R. Fonseca, T. D. Niehaus, K. C. Costa, Functionally redundant formate dehydrogenases enable formate-dependent growth in *Methanococcus maripaludis*. *J. Biol. Chem.* **300**, 105550 (2024).
40. B. A. Lerch, D. A. Smith, T. Koffel, S. C. Bagby, K. C. Abbott, How public can public goods be? Environmental context shapes the evolutionary ecology of partially private goods. *PLOS Comput. Biol.* **18**, e1010666 (2022).
41. B. J. Kapili, S. E. Barnett, D. H. Buckley, A. E. Dekas, Evidence for phylogenetically and catabolically diverse active diazotrophs in deep-sea sediment. *ISME J.* **14**, 971–983 (2020).
42. R. Hidese, H. Mihara, N. Esaki, Bacterial cysteine desulfurases: versatile key players in biosynthetic pathways of sulfur-containing biofactors. *Appl. Microbiol. Biotechnol.* **91**, 47–61 (2011).
43. Exploring the selenium-over-sulfur substrate specificity and kinetics of a bacterial selenocysteine lyase. *Biochimie* **182**, 166–176 (2021).
44. S. Benler, E. V. Koonin, Recruitment of Mobile Genetic Elements for Diverse Cellular Functions in Prokaryotes. *Front. Mol. Biosci.* **9**, 821197 (2022).
45. E. A. Franzosa, X. C. Morgan, N. Segata, L. Waldron, J. Reyes, A. M. Earl, G. Giannoukos, M. R. Boylan, D. Ciulla, D. Gevers, J. Izard, W. S. Garrett, A. T. Chan, C. Huttenhower, Relating the metatranscriptome and metagenome of the human gut. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E2329-2338 (2014).
46. R. Mondav, B. J. Woodcroft, E.-H. Kim, C. K. McCalley, S. B. Hodgkins, P. M. Crill, J. Chanton, G. B. Hurst, N. C. VerBerkmoes, S. R. Saleska, P. Hugenholtz, V. I. Rich, G. W. Tyson, Discovery of a novel methanogen prevalent in thawing permafrost. *Nat. Commun.* **5**, 1–7 (2014).
47. I. L. Brito, S. Yilmaz, K. Huang, L. Xu, S. D. Jupiter, A. P. Jenkins, W. Naisilisili, M. Tamminen, C. S. Smillie, J. R. Wortman, B. W. Birren, R. J. Xavier, P. C. Blainey, A. K. Singh, D. Gevers, E. J. Alm, Mobile genes in the human microbiome are structured from global to individual scales. *Nature* **535**, 435–439 (2016).
48. A. Bernheim, R. Sorek, The pan-immune system of bacteria: antiviral defence as a community resource. *Nat. Rev. Microbiol.* **18**, 113–119 (2020).
49. O. X. Cordero, L.-A. Ventouras, E. F. DeLong, M. F. Polz, Public good dynamics drive evolution of iron acquisition strategies in natural bacterioplankton populations. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 20059–20064 (2012).
50. M. R. Olm, A. Crits-Christoph, K. Bouma-Gregson, B. A. Firek, M. J. Morowitz, J. F. Banfield, inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nat. Biotechnol.* **39**, 727–736 (2021).

51. D. M. Bickhart, M. Kolmogorov, E. Tseng, D. M. Portik, A. Korobeynikov, I. Tolstoganov, G. Uritskiy, I. Liachko, S. T. Sullivan, S. B. Shin, A. Zorea, V. P. Andreu, K. Panke-Buisse, M. H. Medema, I. Mizrahi, P. A. Pevzner, T. P. L. Smith, Generating lineage-resolved, complete metagenome-assembled genomes from complex microbial communities. *Nat. Biotechnol.* **40**, 711–719 (2022).
52. T. Chang, G. S. Gavelis, J. M. Brown, R. Stepanauskas, Genomic representativeness and chimerism in large collections of SAGs and MAGs of marine prokaryoplankton. *Microbiome* **12**, 126 (2024).