Title

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

Journal

Nature Microbiology, 11(7)

ISSN

2058-5276

Authors

Guo, Jiarong

Aroney, Samuel TN

Domínguez-Huerta, Guillermo

et al.

Publication Date

2026-06-29

DOI

10.1038/s41564-026-02391-7

nature microbiology

Article

https://doi.org/10.1038/s41564-026-02391-7

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

Received: 30 May 2025

Accepted: 13 May 2026

Published online: 29 June 2026

Check for updates

Jiarong Guo[1,2,9], Samuel T. N. Aroney[3], Guillermo Domínguez-Huerta[1,2,10], Derek Smith[4], Dean Vik[1,2], Carlos Owusu-Ansah[1,5], Akbar Adjie Pratama[1,11,12], Sergei Solonenko[1], Funing Tian[1,2], Cristina Howard-Varona[1,2], Zhi-Ping Zhong[1,2,6,13], Aminata Fofana[1], Garrett J. Smith[1,2], Suzanne B. Hodgkins[1], Dylan Cronin[1,4], IsoGenie Field Teams 2010–2017, 2019*, EMERGE Coordinators*, Ben J. Woodcroft[3], Gene W. Tyson[3], Virginia I. Rich[1,2,6], Matthew B. Sullivan[1,2,5,6,7] ✉, Simon Roux[8] ✉ & Sarah C. Bagby[4] ✉

Ecosystems are shaped by communities of microorganisms whose niches and impacts depend on functional profiles influenced by gene gains and losses. Culture-based experiments demonstrate that mobile genetic elements (MGEs) can mediate gene flux, but quantitative understanding of these dynamics in natural systems remains limited. Here we develop and apply a systematic, meta-omic framework to investigate MGEs in a complex natural system using an 8-year soil time series collected at Stordalen Mire, in Sweden's thawing permafrost margin. In this climate-critical peatland, we identify ~2.1 million MGE recombinases across 89 microbial phyla and assess ecological distributions, affected functions, past mobility and current activity. This revealed an active mobilome that shapes natural genetic diversity via differential impacts on major phyla and affects a wide range of functions, including metabolic genes involved in carbon flux and nutrient cycling. These findings and this analytic framework suggest avenues towards a better understanding of MGE diversity, activity, mobility and impacts across ecosystems.

Microbiomes are increasingly recognized for driving biogeochemical and nutrient cycling across diverse systems[1–3]. Metagenomic advances have sharpened our understanding of microbial diversity in these systems, from the low resolution of methods such as denaturing gradient gel electrophoresis and 16S amplicon sequencing (reviewed in ref. 4) to metagenome-assembled genomes (MAGs) recovered from shotgun sequencing[5]. Recovering MAGs typically uses metagenome binning approaches that collapse diversity within a population, omitting variants from the consensus[6]. New tools are bringing metagenomic strain-level diversity into focus in intensively studied systems such as the human microbiome, but still struggle in complex or understudied systems such as soils[7–9]. Yet, within-species variation can have profound effects on ecosystem outcomes (for example, ref. 10). Such microdiversity can arise through both mutation and gene flow[11] (reviewed in ref. 12). Understanding the constraints on these processes is a critical step towards understanding the pace and scope of adaptive change accessible to microbiomes in changing environments.

Mobile genetic elements (MGEs) are a major driver of gene flow and genomic rearrangement[13,14]. Recombinase genes enable MGEs to move between genomic loci and can be used to distinguish between several MGE types[15]: insertion sequences and transposons, phage and phage-like elements, conjugative elements and mobility islands,

A full list of affiliations appears at the end of the paper. ✉e-mail: sullivan.948@osu.edu; sroux@lbl.gov; scb126@case.edu

https://doi.org/10.1038/s41564-026-02391-7

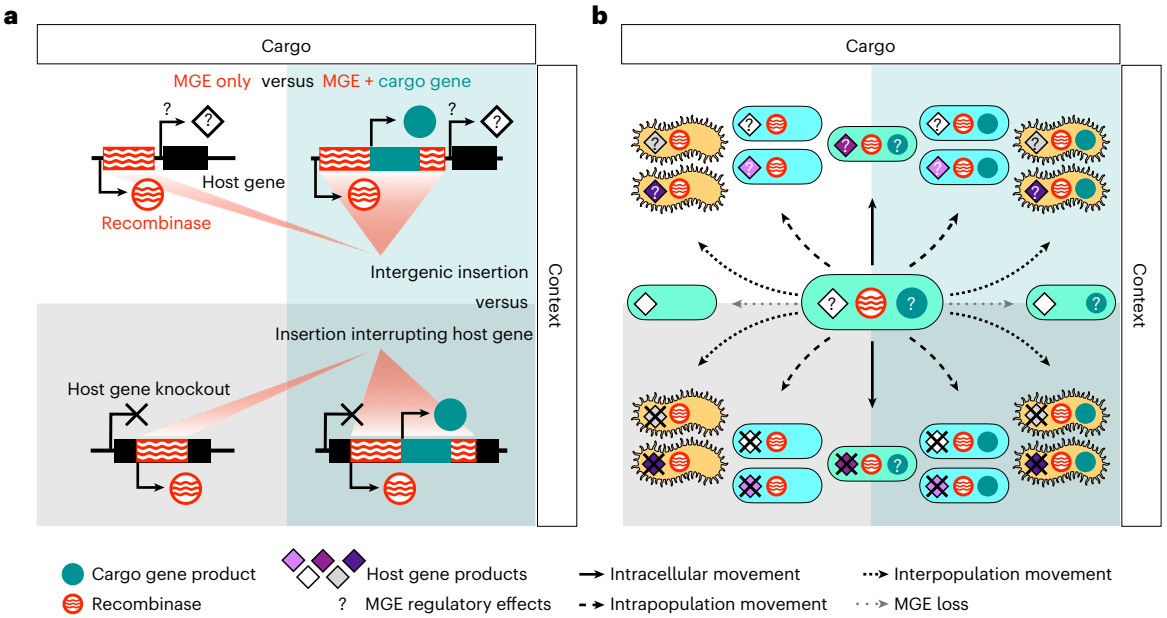


Fig. 1 | MGE cargo and context determine the impact of MGE activity on host coding capacity and population genetic diversity. a, An MGE may or may not carry cargo and may insert either intergenically, potentially altering transcriptional regulation downstream, or into a host gene, causing function loss. The MGE and recombinase gene products are represented by red and white waves, the cargo gene and products are shown in solid teal, the host gene is shown in black and the host gene product is shown with a 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 (represented as diamonds). Intrapopulation movement (long-dashed arrows, blue cells) lets the MGE interrupt or influence the same or different host genes in a new genetic background. Interpopulation movement (short-dashed arrows, yellow cells) lets the MGE interrupt or influence a new suite of genes. MGE loss (grey dotted arrows) may remove or leave behind cargo genes and potentially restore the function of previously disrupted host genes. Not depicted: the capture of host genes as new cargo genes.

and integrons. These types differ in several respects, including their ability to move between cells unaided and the likelihood that, when they move, they carry genes that are not required for MGE mobility but whose expression may alter host metabolism or function ('cargo genes'). MGEs can influence host phenotypes and niches through their cargo and/or the genomic context of MGE insertion (Fig. 1a), causing gains or losses of function and/or regulatory shifts[16,17] 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[18].

Yet, so far, MGEs have not been accessible to comprehensive characterization in a natural, non-host-associated community. With comparatively few reference genomes available in such communities, analysis depends on de novo assembly, but the mobility of MGEs tends to break assembly graphs[19,20]; where binning succeeds, the loss of strain-level variation in consensus sequences hides differences in MGE presence, absence or genomic context. Moreover, few long-term projects sequence deeply enough or often enough to track MGE activity. These structural challenges have made MGEs a persistent unknown in environmental microbiology. Targeted studies have grappled with some aspects of MGE biology, for example, transfers of a given MGE-associated trait (such as antimicrobial resistance[15,21]) or MGE type[22–24], or MGE identification in the genomes of isolates[15]. But so far, there is no comprehensive survey of MGEs in a complex natural system, and thus no quantitative basis for bounding their potential impacts. Addressing this gap requires deeply sequencing the same microbiome repeatedly over short and long timescales.

Long-term ecological study sites provide an opportunity to capture such datasets. The model permafrost peatland Stordalen Mire, Sweden, lies in the rapidly changing discontinuous permafrost margin, where rising temperatures convert the intact palsa habitat to partially thawed bog and fully thawed fen, with concomitant increases in methane efflux from the soil[25–27] (Fig. 2a). Sampling from active-layer soil in these three habitats over almost a decade of peak growing seasons (2010–2017) has yielded 615 short-read metagenomes[28], of which 50 have paired short-read metatranscriptomes. Previous analyses of a subset of this data have identified key taxa and processes[26] and have pointed to the role of functional redundancy and dispersal in allowing each habitat's microbiome to resist environmental change[26,29]. MGE-mediated gene flow could be synergistic with both, but so far, even bounding the potential impacts of MGEs on a natural system has been intractable.

Here, we use the Stordalen Mire dataset to develop and apply the needed analytic approach. We use paired short- and long-read sequencing to determine the expected yields in MGE recovery from short-read data, and use these rates to obtain conservative bounds on the per-genome abundance of several MGE types across different phyla. We assess the functions of genes affected by MGEs, revealing that MGE-mediated gene gain and loss reaches many functions in core nutrient metabolism, with implications for community metabolism and ecosystem outputs. Finally, we present three complementary metatranscriptomic and metagenomic metrics of MGE mobility at different timescales and across MGE types. Collectively, these analyses shed light on the scale and potential impacts of MGE activity in a soil microbiome and provide a set of approaches that will enable quantitative investigation of MGEs in short-read metagenomic data from complex natural systems.

Results

MGE recombinases from permafrost thaw gradient microbiomes

Stordalen Mire 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

https://doi.org/10.1038/s41564-026-02391-7

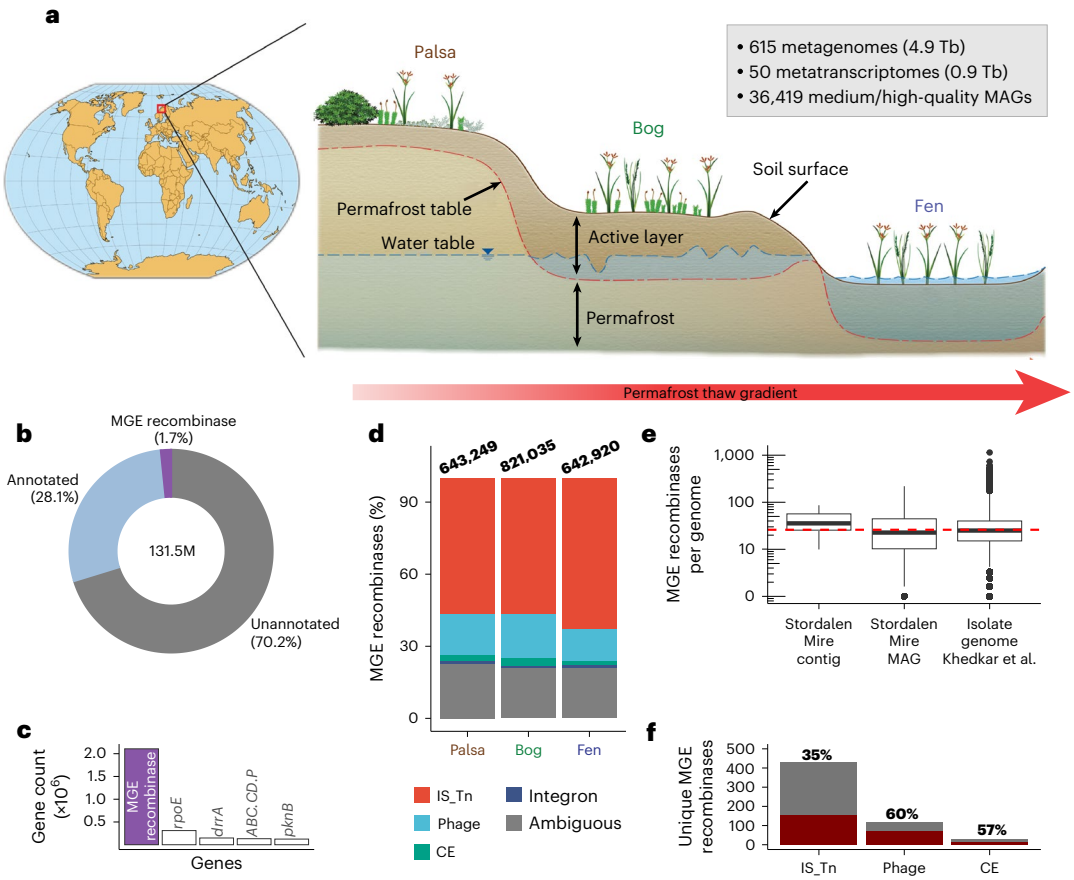


Fig. 2 | The Stordalen Mire model ecosystem and its MGE recombinases. **a**, The long-term study site with large datasets (grey inset) along a palsa–bog–fen thaw gradient. **b**, The MGE recombinase percentage in all genes from Stordalen contigs (≥1,500 bp). **c**, The counts of the most abundant genes: 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/transposon; 'phage', phage and phage-like elements; 'CE', conjugative elements/mobility islands; 'integron', integrons; 'ambiguous', recombinases found in >1 MGE types. **e**, MGE estimates per genome for Stordalen Mire contigs (median 36, $n = 570$) or MAGs (median 23, $n = 5{,}548$) and from cultured isolates from the study by Khedkar et al. (median 25; dashed red line, $n = 76{,}902$). Lower and upper hinges correspond to the first and third quartiles, and the upper and lower whiskers extend from the nearest hinge to the largest (or smallest) value no further than 1.5× the interquartile range from the hinge. Filled circles represent observations beyond this threshold of 1.5× the interquartile range below the first quartile or above the third quartile. **f**, A comparison of MGE recombinase recovery between short-read and hybrid assemblies ($n = 17$), with recombinase matching at 100% protein identity. Grey bars represent the proportion of recombinases found only in hybrid MAGs; red bars represent the proportion of recombinases found in both hybrid and short-read MAGs. Thaw gradient schematic in **a** adapted from ref. 102 under a Creative Commons license CC BY 4.0.

habitats (least to most thawed, respectively). Before our work, ~5.8 terabases of metagenomic and metatranscriptomic sequencing data had been generated from 8 years of field sampling focused on the active layer across habitats (Fig. 2a and Supplementary Table 1). We first identified and quantified all major MGEs in these data via recently developed hidden Markov models (HMMs) and identification protocols that leverage recombinases as MGE hallmark genes[15] (Supplementary Table 2). Importantly, because this approach relies on recombinase detection, it does not fully capture plasmid diversity, but focuses instead on integrative MGEs. We screened 131.5M genes predicted from 30.7M short-read metagenomic contigs, identifying 2,107,204 MGE recombinases (1,130,776 unique protein sequences) on 1,790,672 contigs (Fig. 2b–d, Extended Data Figs. 1a,b and 2a,b, Supplementary Fig. 1 and Supplementary Table 3, 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 Kyoto Encyclopedia of Genes and Genomes (KEGG) gene family, the stress-response sigma factor *rpoE* (2.4 copies per Mbp) (Fig. 2c). This is consistent with previous analyses of cultured microorganisms[15] and reports of recombinase prevalence and abundance in microbiomes[30].

We next classified these abundant recombinases into four previously-established types[15]: (1) integrons, (2) insertion sequences and transposons ('IS_Tn'), (3) phage/phage-like elements ('phage') and (4) conjugative elements and mobility islands ('CE'). This classification facilitates the evaluation of ecological impacts, as 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[15,30] found IS_Tn was dominant in highly abundant MGE pools. In Stordalen Mire, we also found that IS_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) (Methods and Extended Data Fig. 1a). Ecologically, this rank-order pattern of recombinase types 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 and Extended Data Fig. 1b,c). The same pattern largely persisted across phyla for both MAG- and contig-level taxonomic assignments (available for 216,720 and 804,456 MGE recombinases, respectively, assigned to a total of 89 phyla with 93.7% concordance for the subset shared) (Fig. 2a,b, Extended Data Fig. 3a,b and Supplementary Table 3, item A).

https://doi.org/10.1038/s41564-026-02391-7

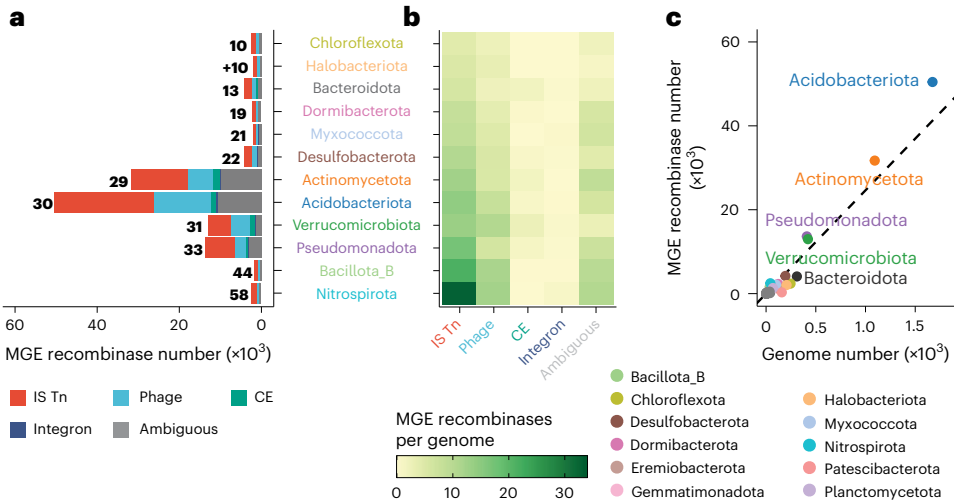


Fig. 3 | MGE recombinase type distribution across microbial host lineages (MAG-based) of the full Stordalen Mire short-read dataset. a, The number of each MGE type found in each phylum, as identified from contigs binned into MAGs (n = 216,720). The taxonomic distribution based on contig-level assignment (n = 804,456) is available in Extended Data Fig. 3, while the remaining MGEs were on contigs too short (<3 kb) to be reliably assigned. The MGE recombinase counts shown for the top 12 phyla (out of 46 total), ordered by the average MGE recombinase number per genome (indicated to the left of each bar with '+' indicating Archaea). **b**, The number of MGE recombinases identified per genome, grouped by MGE type and host phylum. **c**, The number of MGE recombinases (y axis) against the estimated number of genomes per phylum (x axis). The dashed line represents the global phylum-averaged number of MGE recombinases per genome.

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[26] (Fig. 3 and Extended Data Figs. 3 and 4). Echoing isolate-based observations[15], the virtual absence of integrons in the abundant Actinomycetota (Extended Data Fig. 3, ~0.2 per genome in this lineage compared with ~1.0 on average for all lineages) 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 per-genome MGE frequency, previously measured in isolates, would compare with estimates from natural populations, as MGEs may be lost during laboratory cultivation owing to reduced selective pressure in clonal isolates. We quantified MGEs per genome in Stordalen microbiomes via two approaches. First, we used a contig-based estimate, comparing identified MGE recombinase counts with total genome counts estimated using a standard single-copy marker gene approach for each sample[31,32]. This revealed a median of 36 MGE recombinases per genome (Fig. 2e). Separately, we considered only the subset of MGEs that were binned into medium/high-quality MAGs (≥70% completeness, ≤10% contamination) (Supplementary Table 3, 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 the genus level[26,29]. Thus, we interpret the MAG-based estimate as a lower bound, as 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[20,33].

To empirically 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 throughout assembly and binning. 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, Extended Data Figs. 5 and 6, and Supplementary Text). Using this benchmarking to scale our MAG-based counts, we obtain a revised estimate of 38–66 MGE recombinases per genome in situ. We interpret this number as comparable with the contig-based estimate and consistent with the hypothesis that isolates might lose MGEs during laboratory cultivation.

MGE impacts range far beyond biotic social interactions

We next wondered whether MGEs affected genes of ecological importance, and how these genes fit into microbiome activity in Stordalen Mire. Prior genome-resolved metagenomic analyses[26,29] show that intercellular interactions and public goods matter in Stordalen Mire: 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 microorganisms, abundant viral community members harbour auxiliary metabolic genes that probably mediate exopolysaccharide breakdown[27]. Such outward-directed 'social' functions are expected to be affected by MGEs[34–36], as are some stress responses[37], but these are only a subset of microbial activities. Each cell must also manage its own routine 'domestic' processes (for example, genome maintenance, gene expression and nutrient homeostasis), functions that shape physiology and ecosystem outputs but that have been little studied in the mobilome so far. Recently observed plasmid-borne genes affecting domestic activities such as nitrogen fixation[37] highlight the potential for MGEs to influence nutrient cycling directly, but the prevalence of such functions among MGE-affected genes remains unknown.

Thus, we asked which Stordalen Mire functions could be confidently detected as mobilome-affected (that is, 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 (see the Supplementary Text section 'Conservative MGE detection and impacted gene curation'). Notably, this stringent approach excluded all but ~0.8% of our 2.1M MGE recombinases from functional gene analysis, precluding the investigation of MGE-mediated changes in metabolic capacity over time. Annotation identified putative functions for cargo genes within conservative boundaries of 5,288 MGEs (900 CEs, 551 integrons and 3,837 phages) and for host genes interrupted by 11,295 IS_Tns (Methods and Supplementary Fig. 1). Among these, manual curation identified 3,631 MGE-affected genes with high-confidence functional predictions

https://doi.org/10.1038/s41564-026-02391-7

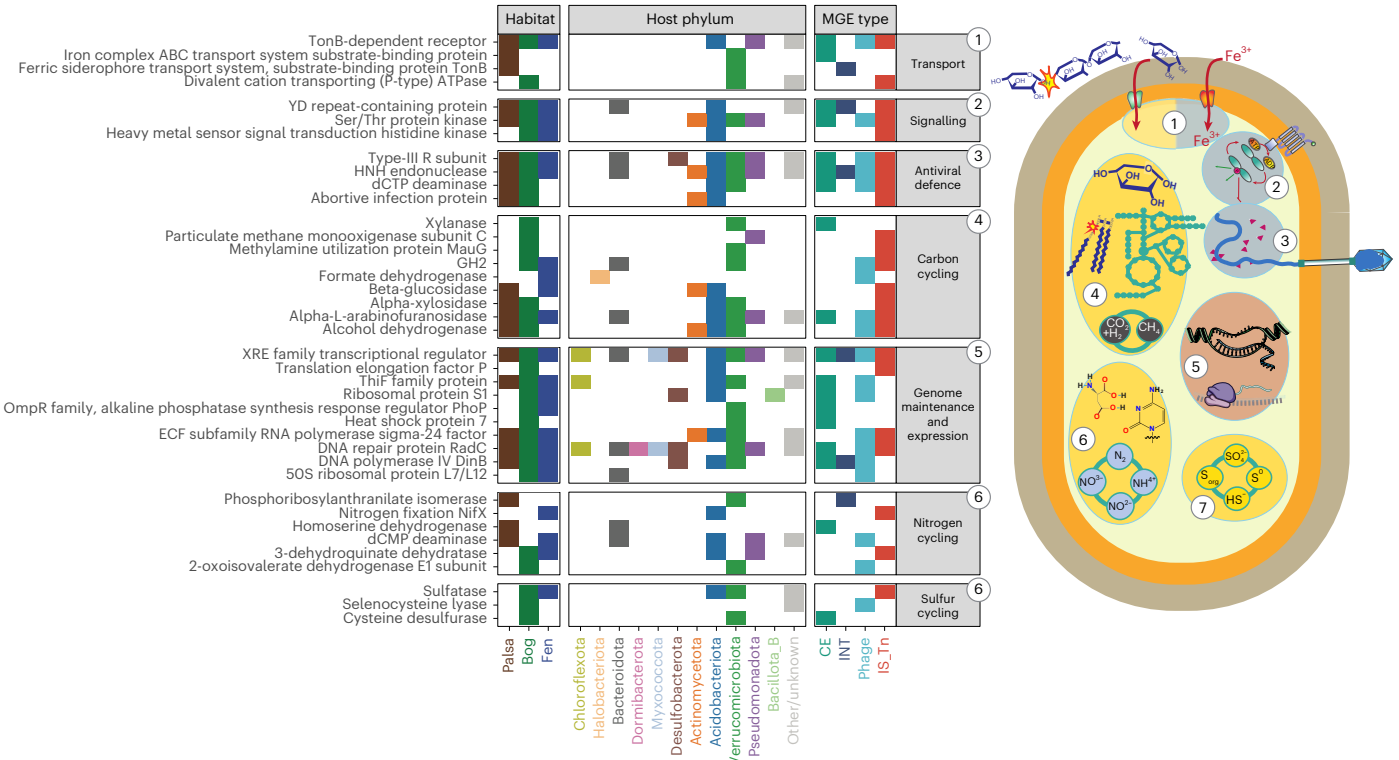


Fig. 4 | Host functions impacted by MGEs in Stordalen Mire. The distribution across habitats, host phyla and MGE types of selected nutrient cycling pathways (the yellow highlights in the cell diagram, groups 1, 4, 6 and 7), housekeeping functions (the red highlights in the cell diagram, group 5), and signalling and social functions (the blue highlights in the cell diagram, groups 1–3) impacted by MGEs whose boundaries could be conservatively established. Coloured boxes indicate positive detection of an MGE-affected function in a habitat, phylum or MGE type. Phyla are ordered from left to right by increasing per-genome MGE count (see Fig. 3). CE, INT (integron) and 'phage' carry impacted functions as cargo genes; IS_Tn interrupt host genes, causing loss of function.

across 1,199 distinct KEGG annotations in diverse categories including carbohydrate metabolism, transport, antiviral defense, signalling, transcription, post-translation modification, lipid metabolism, nucleotide metabolism, DNA repair and antibiotic metabolism (Fig. 4, Supplementary Tables 3 (items C and F) and 4, and Supplementary Text). Here, we highlight two ecosystem-critical processes: carbon cycling and salt stress response.

We uncovered MGEs affecting central carbon metabolism genes (6 carried and 45 interrupted) (Supplementary Table 4), including those encoding enzymes of the tricarboxylic acid (TCA) and reductive TCA cycles and the pentose phosphate shunt. Alcohol dehydrogenases (1 carried and 27 interrupted) and pyruvate metabolism enzymes (5 interrupted) were well represented (Supplementary Table 4); because these enzymes' substrates can be metabolic branchpoints, MGE activity may alter the partitioning of fluxes between pathways with different energy yields. Intriguingly, four of the interrupted alcohol dehydrogenases were in Actinomycetota, where methanotrophy sometimes uses members of this enzyme family[38,39]. We also identified two direct connections to methane metabolism: the key hydrogenotrophic metabolism gene formate dehydrogenase[40] was carried by a phage-like MGE from Halobacteriota (Methanoregula sp003136075), a group of methanogens active at Stordalen Mire[25,26,29], and an IS_Tn-interrupted methane monooxygenase (*pmoC*) subunit in a Pseudomonadota (unbinned Alphaproteobacterial contig). Thus, even within our highly conservative subset of MGE-affected genes, we find 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).

Salt stress is common in permafrost-affected soils owing to brine film formation around soil when water freezes[41]. Cells can respond by modulating membrane stability and permeability, or by synthesizing and/or transporting osmolytes[42]. We found MGEs impacting genes that may contribute to each of these responses. Membrane-tuning cyclic fatty acids vary in abundance between the microbiota of nearby *Sphagnum* (bog) and *Carex* (fen) peats[43], suggesting that the adaptive value of cyclopropane-fatty-acyl-phospholipid synthase (CFAS; two carried and three interrupted; one in bog and four in palsa) may vary with thaw stage. MGEs also affected genes involved in the synthesis or transport of numerous osmolytes, including carnitine (gamma-butyrobetaine hydroxylase, one interrupted), trehalose (malto-oligosyl trehalose synthase, five interrupted), sucrose (sucrose-6F-phosphate phosphohydrolase, one carried), proline (iminopeptidase, one interrupted; proline-specific peptidase, three carried) and glutamate (formiminoglutamate hydrolase, one interrupted; glutamate ABC transporter ATP-binding protein, two carried). However, the involvement of several of these compounds in central metabolism complicates interpretation.

Critically, even though our stringent analysis criteria and the limitations of functional gene annotation in understudied environments together reduced the set of MGE-affected functions analysed to genes affected by 0.02% of the total MGEs identified, we still find abundant examples of mobilome-affected social and domestic functions beyond previous observations[44]. Our analysis adds examples of MGE-impacted functions that influence carbon flow through this climate-critical ecosystem, both directly (exopolysaccharide degradation and central carbon metabolism) and indirectly (via nitrogen, sulfur and/or iron availability to support carbon flux). Moreover, even the small subset of MGE-impacted genes identified here may have large regulatory impacts[45]. For example, we identified many MGEs affecting sigma factors (1 carried and 45 interrupted) of the stress-response sigma-24 (*rpoE*) (Fig. 4) and housekeeping sigma-70 (*rpoD*) families (Supplementary Table 4). Such master regulators' large regulons could

indirectly amplify MGE impacts. Despite prior work's focus on MGE impacts on biotic interactions, in our analysis, ~55% of MGEs affect at least one domestic function, and ~51% of MGE-affected functions are domestic. If this conservative subset is representative of the Stordalen mobilome as a whole, then as sequencing methods, functional annotation and databases improve, we would expect ~1.1M of the full ~2.1M MGE recombinases to influence domestic functions, with important implications for nutrient cycling, as well as microbiome activity, ecology and evolution.

MGE recombinase mobility and activity

To assess the ecological and evolutionary change in Stordalen Mire MGEs, we evaluated their mobility and activity from three different viewpoints: (1) snapshots of metatranscriptomic MGE recombinase expression as a proxy for mobilization underway, (2) genomic neighbourhood change as a marker of completed MGE movement between insertion sites and (3) partial carriage of an MGE by the host population as the result of evolution and selection over longer timescales. Integrating results across these analyses can help highlight individual MGE dynamics. For example, high recombinase expression together with low rates of genomic neighbourhood change for a given MGE could suggest a strong preference among insertion sites. Our overall findings are as follows.

First, we assessed MGE recombinase transcriptional activity by read mapping from 50 available metatranscriptomes against contigs from paired metagenomes (Supplementary Table 5), taking ribosomal protein-encoding gene coverage from read recruitment as a baseline for constitutive transcriptional activity of actively growing cells[46] (Supplementary Table 6). This revealed detectable transcriptional activity in ~3.4–56.4% (median 27.8%) of ribosomal protein-encoding genes and ~0.04–9.1% (median 1.2%) of MGE recombinase genes (Fig. 5a and Supplementary Tables 3 (item G) and 7), making MGE recombinases the second-most-transcribed non-ribosomal genes in Stordalen Mire (Fig. 5a). We found roughly equal levels of transcriptional activity for IS_Tn and phage recombinases, approximately five- to tenfold higher than for integron and CE recombinases (Fig. 5b). Ecologically, IS_Tn recombinase transcription increased significantly with thaw from palsa to fen ($P \leq 0.01$, Wilcoxon rank-sum test) (Fig. 5c and Extended Data Fig. 7b), suggesting an 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 and Extended Data Fig. 7c). The archaeal phylum Halobacteriota was an exception, although neither abundant nor MGE recombinase-rich it 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[25,47], we speculate that their disproportionate MGE recombinase activity might be adaptive.

Second, for MGE recombinases observed multiple times, we assessed the genomic context (300-bp upstream and downstream) surrounding each instance. We interpreted changing context as a proxy for intra-/inter-genomic mobility, either of the recombinase itself or of other recombinase-proximal genes (Methods; see examples of changing neighbourhoods in Extended Data Fig. 8). From the original ~2.1M MGE recombinases, we identified 74,257 MGE recombinase clusters (100% average amino acid identity (AAI)) observed more than once in our dataset. Among these, 0.1–10% showed neighbourhood variation, with rates depending on MGE type (Fig. 5b and Supplementary Table 3, item H). Neighbourhood variation was much more common among IS_Tn recombinases (~10%) than among phage (~1%), integron (0.2%) and CE (0.1%) recombinases. This was consistent among years and habitats, with a larger variation detected across diverse host phyla (up to 37% in Desulfobacterota_G) (Extended Data Fig. 9). Mindful that hypervariable islands in cellular genomes can often be impacted by multiple MGE types, we quantified co-occurring IS_Tn insertions, which could cause the neighbourhood to change while the focal MGE remains stable. Fewer than one in five of our changed-genomic-neighbourhood examples was due to co-occurring IS_Tn insertions (Extended Data Fig. 8). Because our approach only detects neighbourhood variation immediately flanking an 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 (Methods). Where coverage drops significantly within an MGE recombinase, we infer that only a subset of the host population encodes the MGE recombinase. Overall, across MGE types, we found 8,033 recombinase clusters (100% AAI) that had ≥20 kbp of genomic context to support coverage analysis. Of these, ~5–23% showed coverage drops consistent with only a subset of the cellular population having the MGE (Fig. 5b and Supplementary Table 3, item I). These findings were again broadly consistent across years, habitats and well-sampled lineages (Extended Data Fig. 10b–d). In spite of spatio-temporal consistencies, there were notable variations in intrapopulation penetrance between MGE types, with IS_Tn and phage types particularly prone to varying across individuals within a host population (Extended Data Fig. 10d). Because MGE sizes vary, and read mapping for lesser-sampled taxa is noisy, these numbers represent conservative lower bounds on active MGE penetrance into cellular population genomes. Even so, the extent of partial carriage observed highlights MGEs' role in what amounts to large ongoing in situ knockout experiments: among 372 IS_Tn elements that both interrupt a protein-coding gene and were amenable to partial carriage analysis, 81 (~22%) variably interrupt or leave intact host protein-coding genes. Together, these results suggest that MGE penetrance within member genomes of any given population differentially provide individuals access to genetic diversity, while shielding the population from loss of function due to MGE insertion[48,49].

Discussion

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 antimicrobial resistance[15,21]. 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. We find that, well beyond biotic interactions, the mobilome affects functions intrinsic to a wide range of physiologically and ecologically important processes such as carbon cycling and salt stress response. Further, active-layer soil in all three habitats shows persistent MGE activity/mobility, which could be an important force driving host strain variation and adaptation in this thawing permafrost soil system. 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 (that is, transcription and genomic neighbourhood change) (Fig. 5b), highlighting the need for future studies to examine context-dependent MGE effects such as modulation of host gene expression downstream of MGE insertion sites. This persistent activity occurs against a backdrop of overall MGE stability, agreeing with a companion study showing the same microbiomes are stable across the years studied here[29]. Importantly, the limitations of short-read sequencing and assembly could drive systematic underestimates of MGE-mediated variation, through low recovery either of variable MGEs or of rare or fast-disappearing subpopulations/lineages experiencing non-selective excision/recombination events. To assess this, we confirmed and quantitatively evaluated the bias of short-read versus long-read metagenomic MGE recovery, which helped establish our identified mobilome identifications and ecosystem impacts as a lower-bound estimate. However, the framework

https://doi.org/10.1038/s41564-026-02391-7

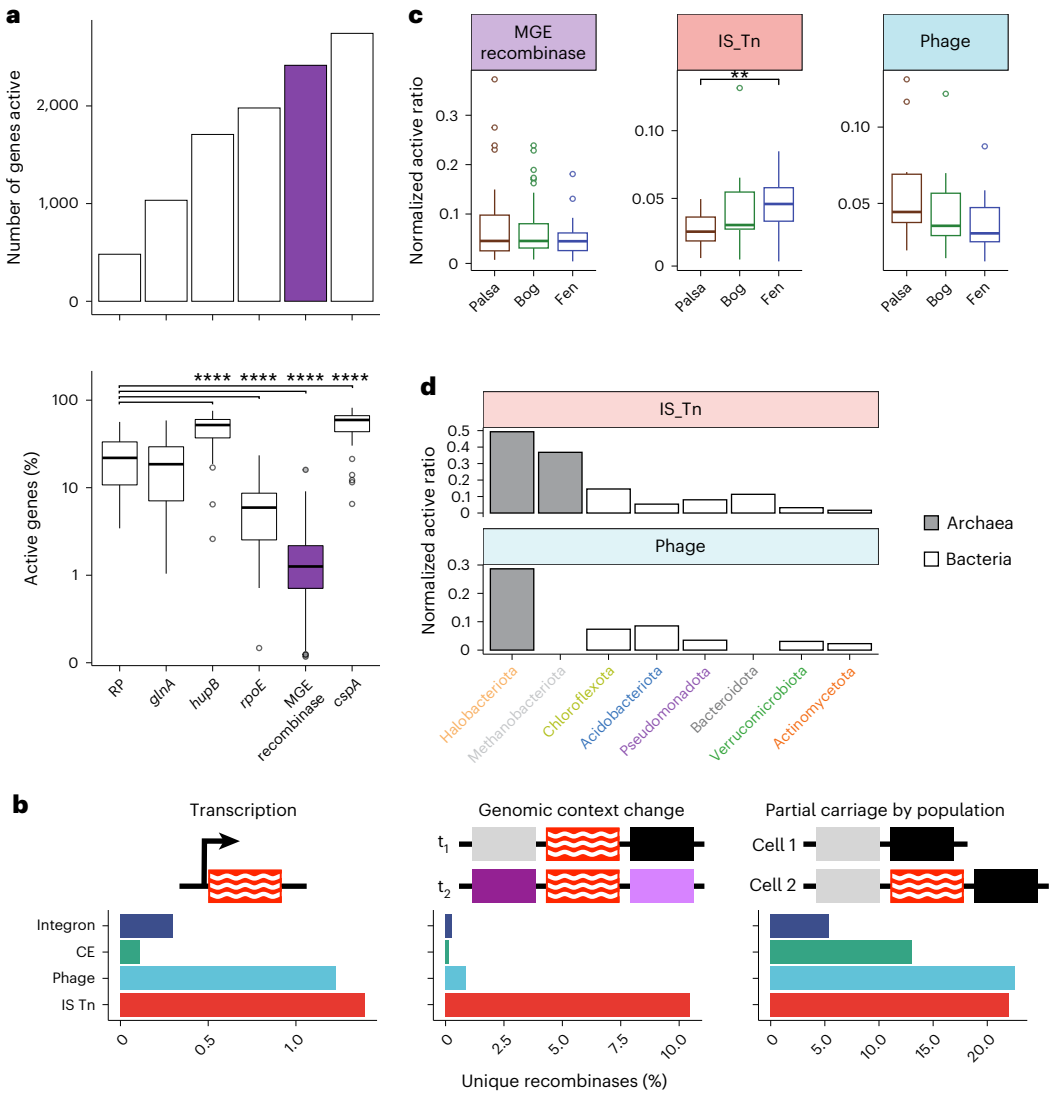


Fig. 5 | MGE recombinase activity. a, Top: genes detected most frequently as 'active' on the basis of metatranscriptome read mapping (number of observations included: references 50, MGE recombinases 192). Gene categories from right to left: *cspA*, major cold shock protein gene (K03704); MGE recombinase; *rpoE*, RNA polymerase sigma factor gene (K03088); *hupB*, DNA binding protein HU-beta gene (K03530); *glnA*, glutamine synthetase gene (K01915); RP, ribosomal protein genes. Bottom: the percentage of active genes based on of the full set of genes identified in the matching metagenome. Wilcoxon rank-sum test of statistical significance, two-sided; ****, *P* value ≤0.0001. Specific significant *P* values from left to right: 5.7×10^{-10}, 3.7×10^{-10}, $<2.22 \times 10^{-16}$ and 9.5×10^{-12}. **b**, Three facets of MGE recombinase activity/mobility—transcription (left), genomic neighbourhood stability (middle) and 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 recombinases (100% identity at the protein level) detected with activity/mobility within each type. For the middle panel schematic, t_1 and t_2 represent two different samples in a time series. **c**, A comparison of ratios of active MGE recombinases, IS_Tn and phage recombinases across three habitats, normalized by the ratio of active ribosomal protein-encoding genes for the same sample (number of observations from left to right: 52, 62 and 78 for MGE recombinases; 15, 14 and 20 for IS_Tn; and 11, 13 and 16 for phage). Wilcoxon rank-sum test of statistical significance, two-sided; **$P = 0.0061$. **d**, The normalized active ratio of IS_Tn and phage recombinases across host phyla with ≥5 active IS_Tn or phage recombinases, and ribosomal protein-encoding genes. CE and integron recombinases are excluded here owing to low counts. For **a** and **c**, the lower and upper hinges correspond to the first and third quartiles, and the upper and lower whiskers extend from the nearest hinge to the largest (or smallest) value no further than 1.5× the interquartile range from the hinge. Filled circles represent observations beyond this threshold of 1.5× the interquartile range below the first quartile or above the third quartile.

and analytics established here are readily transferable to long-read sequencing datasets for future MGE studies.

As improved methods for studying the genomic potential of microorganisms 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[26,31]. Continued advances in sequencing and analytical approaches are increasing this resolution to reveal intrapopulation variation, early speciation and selection[7], with recent MGE typing efforts[15] 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 artefacts and provides empirical baselines and inferred impacts for all major integrative MGEs in a natural soil system. These findings suggest that ~1% of MGE recombinases are active, that genomes contain 36–66 MGE recombinases on average and that more than half of MGEs impact genes affecting basic cell physiology and ecosystem outputs. As advancing technologies bring the mobilome more fully into focus, given the toolkit and baseline

findings presented here, we predict that mapping the mobilome will help reveal subtle alterations that could serve as the signal for early detection of systemic change.

Methods

Data collection, metagenome assembly and binning, gene annotation and taxonomy assignment

Our study site, Stordalen Mire, is a long-term ecological research site in northern Sweden. Large sampling and sequencing efforts since 2010 have yielded ~4.9 Tb of metagenomic sequence data from 615 readsets from 408 bulk soil samples that span space (across the site's three habitat types: n = 188 palsa, n = 200 bog and n = 227 fen; average of ~8 Gb per sample) and time (summer peak growing season, 2010–2017) (Supplementary Table 1) with details described in another parallel study[29], in previous studies[26,27,50] and summarized below. The IsoGenie Consortium and EMERGE Biology Integration Institute have developed extensive resources from Stordalen Mire metagenome data, including assembly across all samples to produce 30,710,180 contigs (≥1,500 bp) and 36,419 medium- to high-quality MAGs (completeness ≥70% and contamination ≤10%; CheckM2 v1.0.2)[51] that clustered into 2,048 populations (95% average nucleotide identity (ANI)) per community standards[52]. Select samples have also been sequenced using metatranscriptomics to assess gene expression (50 spatio-temporally diverse samples; average of ~18 Gb per sample). For this work, we additionally performed paired long- and short-read DNA sequencing on six samples, chosen to be representative of the larger dataset, as a control to assess short-read MGE recovery.

The protocol used for sample preparation and sequencing for short-read metagenomes has been described previously[26,27,50]. In brief, nucleic acids were extracted according to a published protocol[53], except for the MAG source samples from substrate addition incubations ('201607_SubstrateInc_9to19'), which were extracted with the DNeasy PowerSoil Kit (cat. no. 12988-100, Qiagen). Libraries were prepared with either KAPA Hyperprep, TruSeq or Nextera XT kits for short-read metagenomes; SMRTbell Express Template Prep Kit 2.0 for long-read metagenomes; or TruSeq Stranded mRNA kit for metatranscriptomes. Reads were sequenced with either Illumina HiSeq (2000 or 2500), NovaSeq 6000 or NextSeq 500 for short-read metagenomes; PacBio Sequel IIe for long-read metagenomes; or NovaSeq 6000 for metatranscriptomes.

Short-read sequencing data from individual samples were trimmed by Trimmomatic (v.0.36)[54], assembled by SPAdes with the '--meta' option (v.3.12)[55] and then binned by MetaBAT2[56], GroopM2[57] and MaxBin2[58] (Cronin.v2), or an ensemble approach[29]. Those sequenced at Joint Genome Institute (JGI) were processed by JGI's standard assembly and binning pipeline. In brief, reads were quality controlled, filtered and trimmed using the BBTools package[59], assembled with SPAdes with the '--meta' option[55] and binned with MetaBAT2[56]. The bins produced from this effort were combined with 9,406 bins from ref. 26 and 11,120 bins from stable isotope probing experiments with field peat soil and labelled litter (sequenced with NovaSeq at JGI and assembled with MEGAHIT v1.1.3)[60] and then filtered using 70% completeness and 10% contamination CheckM2 (v1.0.2)[51] cut-offs to form the 36,419 MAGs (consisting of 4,193 from Cronin.v1, 27,423 from Cronin.v2, 847 from JGI, 1,543 from ref. 26 and 2,413 from stable isotope probing). Next, bins obtained from the same metagenome were de-duplicated using Galah (95% ANI, commit e9e59a8)[52], and all bins across the entire dataset were further dereplicated at 95% ANI using Galah (commit e9e59a8), producing 2,048 population level clusters. Further, the contigs encoding a recombinase were annotated with DRAM (–min_contig_size 1500)[61] and taxonomically assigned at phylum level with CAT[62], requiring a minimal contig length of 3 kb. The contigs in MAGs were assigned to the MAG taxon (obtained with GTDB-tk)[63], as taxonomic assignments based on whole MAGs are typically more reliable than single-contig assignments.

MGE recombinase identification

Recombinase identification was on the basis of a framework recently developed to robustly identify and classify MGE recombinases from microbial isolate genomes[15]. Here, we applied the same recombinase HMM profiles and suggested cut-offs to identify MGEs in metagenomic assemblies. In brief, open reading frames (ORFs) were predicted from assembled contigs by prodigal (v2.6.3) using the '-p meta' option[64]; then, recombinases were identified among the predicted protein sequences by hmmsearch (v3.3.2)[65] using the 68 recombinase subfamily HMMs (Supplementary Table 2) and the '–cut-ga' option. Identified recombinases were assigned to the respective MGE recombinase type corresponding to the HMM with the best score, following previous recommendations[15].

Annotation-based inclusion/exclusion criteria

MGE recombinase candidates were further quality controlled, as recommended previously[15], by using eggnog-mapper (v2.1.9, default settings)[66]. First, genes annotated as invertase, replication protein, DNA repair, DNA binding, homologous recombination, viral exonucleases or reverse transcriptase (interpreted to be spurious hits to other host genes related to DNA manipulation) were removed, except for the ones also annotated with the Pfam catalytic domains of 'transposase' or 'helicase' that were retained after manual curation. This removed 0.68% of the identified possible recombinases (Supplementary Fig. 1). Second, sequences annotated by eggnog-mapper as CRISPR, defense response to virus or maintenance of DNA repeat elements were removed to filter out spurious hits thought to be similar to Cas family recombinases, but not considered as an MGE recombinase. This removed 0.46% of the identified possible recombinases. Third, we discarded predicted recombinases that contained only a helix–turn–helix (DNA binding) domain, as well as relaxase hits for which Pfam annotates a helix domain, but without the catalytic domain required for recombinase activity. This removed 0.07% of the identified possible recombinases. Fourth, recombinases matching a 'cell' type recombinase HMM (and thus unlikely to be encoded by an MGE) were removed. This removed 2.52% of the identified candidate recombinases, and yielded a final total of 2,107,204 MGE recombinases (Supplementary Table 3, item A). Last, some recombinase subfamilies/HMMs include recombinases from different MGE types, which prevents assignment for that HMM to a specific MGE type. These recombinases with ambiguous origins accounted for 20.51% of the identified recombinases. They were included in the analyses of total MGE recombinases, but excluded in analyses of individual MGE types, which focused on four recombinase-based MGE types: 'IS_Tn' (insertion sequence and transposon), 'CE' (conjugative elements including conjugative element and mobility islands), 'phage' (phage and phage-like elements) and 'integron' (integron) (Supplementary Fig. 1). Note that IS_Tn recombinases integrated into the host chromosome, as opposed to into larger MGEs, cannot be distinguished using our approach and data, as defining the boundaries of the MGE is challenging in short-read assemblies, except for a small fraction of MGEs with curated boundaries (see below). To assess the diversity of these MGE recombinases, we considered recombinase sequences as 'redundant' when they clustered with other sequences at 100% AAI using CD-HIT (v4.8.1)[67] with '-c 1 -g 0 -G 1 -l 30 -d 0 -M 0'. This yielded 1,130,776 clusters that we considered as unique recombinases.

Identification of cargo genes/interrupted genes associated with MGEs

Once MGE recombinases were identified, we surveyed their genomic context to assess potential cargo and/or interrupted genes. This was done using MGE-specific approaches as follows.

Integron. All integron-containing contigs were screened for cargo genes (also known as 'gene cassettes') using integron-finder (v2.0.2)[68] with default parameters. In brief, integron-finder identifies integrons

by using integron-specific integrase HMMs to identify integron integrases (*intI*), and a covariance model to identify the integron attachment site (*attC*), as the two bounding regions of an integron sequence. Only integrons encoding *intI* and at least one *attC* were treated as 'complete' by integron-finder and were retained. There are two types of integron structures that are known in terms of the orientation of *intI* and *attC*: (1) a regular type with *intI* and *attC* on opposite strands and facing outwards and (2) another type with *intI* and *attC* on the same strand, called an inverted integrase integron[68]. Of the 686 'complete' integrons we identified, 681 integrons also met these orientation expectations and the 5 that did not were not considered further. The gene cassettes delimited by *intI* and *attCs* in these integrons were treated as cargo genes associated with integrons[69], resulting in integron regions with various lengths (minimum 983 bp, median 2,310 bp, mean 3,092 bp and maximum 14,195 bp). Problematically, most contigs (97.4% or 25,953 out of 26,634) encoding Khedkar's 'integron' type recombinases were not considered complete by integron-finder. To test this, we also replaced the integron integrase HMM in the integron-finder tool with that from the study by Khedkar et al.[15] and found it produced nearly identical results and did not yield additional instances of predicted complete integrons. One explanation could be that *attC*, not integron integrase, is the limitation, as the flanking *attC* sequences are identified using a database of known *attC* sequences and these could miss novel *attC* sequences used by integrons in natural microbiomes. Given this, we interpret the large discrepancy between Khedkar's and integron-finder's integrons to represent one of the following scenarios: (1) degraded or domesticated integrons that lack *attC* sites[70,71], (2) integrons that lack *attC* sites owing to fragmented assembly or (3) intact and probably complete integrons that have novel *attC* sites. Thus our analyses will underestimate the extent of cargo genes carried by integrons in Stordalen Mire.

CE. CEs are made of three key components—a relaxosome (includes the CE-type recombinase), a conjugation couple protein and a type IV secretion system (T4SS)—that can be used as hallmark genes that are more conserved than other CE genes[72]. To identify cargo genes associated with CE, the CE hallmark genes were identified in CE recombinase-encoding contigs by conjscan implemented in MacSyFinder (v2.0rc7) with '–models CONJScan/Plasmids all–db-type gembase' and '–models CONJScan/Chromosome all–db-type gembase' with gene protein sequences as input[72,73]. Conjscan specifically establishes eight types of CEs using hallmark genes, which are core across all eight CE types, and accessory genes that are diagnostic for specific CE types. CEs identified by conjscan from Stordalen Mire sequence data were those that matched a relaxase, a couple protein (an ATPase that couples MGE with the T4SS), a T4SS ATPase and an additional structural gene from T4SS. For the purposes of identifying cargo genes, matches of one model, MOB, were removed, as it only requires the relaxase and thus does not identify the boundaries to establish a complete CE. As a result, we identified 1,017 near-complete CEs. As recommended, we required the distance between the hallmark and accessory genes to be ≤30 genes, except for the relaxase, which could be as much as 60 genes away from any other hallmark and accessory genes[72]. In a small number (13 of 1,017) of instances conjscan identified >1 CE in a region. For these, we manually separated the identified hallmark and accessory genes into subgroups according to gene proximity and required each subgroup to include all hallmark genes (relaxase, conjugation couple protein and T4SS ATPase) and at least one accessory T4SS gene. Once such 'complete' CEs were identified, the genes between the leftmost and rightmost hallmark and accessory genes were assumed to be cargo genes carried by the CE (except for hallmark and accessory genes themselves). This resulted in CE regions with various lengths (minimum 7,907 bp, median 16,290 bp, mean 23,383 bp and maximum 81,148 bp). Overall, 1.6% of the 62,660 CE-type recombinase genes identified by our analyses were considered near-complete CE by conjscan. Again, this was not a function of improved HMMs as we tested replacing the relaxase HMMs in conjscan by those from Khedkar et al.[15] and found it produced similar results. Instead we interpret this large discrepancy to suggest that most CEs in these data are degraded, fragmented from poor assembly or cannot be formally detected as complete because of limitation and bias in hallmark gene databases.

Insertion sequences/transposons. To identify IS_Tn recombinase-interrupted genes, up to 3,000 bp upstream and downstream of IS_Tn recombinases were collected from each contig, requiring a minimal length of at least 50 bp on each end and at least 500 bp across the two ends combined (with 1,063,800 IS_Tn recombinases in total meeting the requirements), and then aligned by minimap2 (v.2.22, default settings)[74] to the coding sequence of representatives of protein clusters (PCs) obtained by MMseqs2 at 95% identity (v4.8.1)[75] from all contigs (≥1,500 bp) used in this study. The corresponding coding sequence was considered interrupted if the following four conditions were met: (1) if upstream and downstream alignments combined covered ≥50% of the coding sequence of the reference PC; (2) if there is a gap between upstream and downstream alignments of the coding sequence, the gap is ≤10% of the coding sequence length; (3) if there is an overlap between upstream and downstream alignments on the coding sequence, the overlap is ≤10% of the shorter alignment; and (4) if there are multiple alignments either upstream or downstream of the coding sequence, the pair with largest coverage of coding sequence is chosen. This identified 5,342 PCs putatively interrupted by a total of 12,332 IS_Tn-type recombinases.

Phage/phage-like elements. All contigs encoding 'phage' recombinases were screened for temperate phage by VirSorter2 (v2.2.3) following VS2 SOP step 1 (v3.0)[76,77], and the predicted phage sequences trimmed by checkV (default setting, version 0.7.0)[78]. Next, the checkV trimmed sequences encoding a 'phage' recombinase were considered as phage/phage-like regions. To identify cargo genes, that is, host(-like) genes encoded on phage/phage-like elements, we opted for a conservative approach in which we reduced the dataset for 'cargo gene inspection' to include only phage/phage-like regions with one 'phage' recombinase, and use the leftmost and rightmost checkV, called 'viral gene' or 'phage' recombinases, as the boundaries. This identified 4,277 temperate phages with 'phage' recombinases and at least one extra gene other than the two at the ends. Meanwhile, 347 phage/phage-like elements were removed for having more than one 'phage' recombinase, in order to avoid cases where multiple contiguous temperate phages or degraded ones yielded a single prophage prediction, which could lead to host genes being mistaken as cargo genes. This resulted in 3,930 conserved phage/phage-like regions (minimum 1,386 bp, median 7,293 bp, mean 13,779 bp and maximum 352,302 bp).

Annotations of cargo/interrupted genes. Insertion sequence or transposon interrupted genes and genes within other MGEs were functionally annotated using KEGG and Pfam databases in DRAMv for phages/phage-like elements or DRAM for other MGEs (v1.2.3,–min_contig_size 1500)[61], and further evaluated by comparison with best hits in NCBI NR (downloaded on 4 October 2020) using DIAMOND (v2.0.8.146, blastp bitscore ≥50)[79]. A subset of the genes of specific interest to the Stordalen Mire thawing permafrost system were selected and manually curated for display (Supplementary Table 4) including transport, carbon/nitrogen/sulfur cycling, signalling, antiviral defense, salt stress response and genome maintenance and expression.

Evaluation of short-read assembly and binning artefacts for recovering MGE recombinases using long reads

Assembly and binning of paired long reads and short reads. Soil samples (six total) were collected at the Stordalen Mire site from 2019 from two depths (centimetres below ground) across three habitats

(palsa, bog and fen). DNA extraction and Illumina NovaSeq sequencing was performed as described above. In addition, DNA sequencing was performed from the same DNA extractions using the PacBio Sequel IIe HiFi sequencing platform at JGI. PacBio data were processed at JGI to form filtered circular consensus sequencing reads. Assemblies were generated through two methods: short-only and hybrid short-read plus long-read (herein, hybrid). Short-only was assembled with metaSPAdes v3.15.4[55] using Aviary v0.5.3[80] with default parameters. Hybrid assembly was performed using Aviary v0.5.3 with default parameters. This involved a step-down procedure with long-read assembly through metaFlye v2.9-b1768[81], followed by short-read polishing by Racon v1.4.3[82], Pilon v1.24[83] and then Racon again. Next, reads that did not map to high-quality metaFlye contigs were hybrid assembled with metaSPAdes v3.15.4 and binned out with MetaBAT2 v2.1.5[56]. For each bin, the reads within the bin were hybrid assembled using Unicycler v0.4.8[84]. The high-coverage metaFlye contigs and Unicycler contigs were then combined to form the assembly fasta file. Genome recovery was performed using Aviary v0.5.3 with samples chosen for differential abundance binning by Bin Chicken v0.4.2[85] using SingleM metapackage S3.0.5[86]. This involved initial read mapping through CoverM v0.6.1[87] using minimap2 v2.18[74] and binning by MetaBAT v2.1.5[6], MetaBAT2 v2.1.5[56], VAMB v3.0.2[88], SemiBin v1.3.1[89], Rosella v0.4.2[90], CONCOCT v1.1.0[91] and MaxBin2 v2.2.7[58]. Genomes were analysed using CheckM2 v1.0.2[51] and clustered at 95% ANI using Galah v0.4.0[52].

Assessment of MGE recombinase recovery by short-read assembly and binning. We considered the high-quality hybrid MAGs as the best available references and used these to assess the recovery of MGE recombinases in short-read contigs and MAGs. A total of 17 high-quality MAGs (completeness ≥95%, contaminant ≤5% and at most 100 contigs) from hybrid assembly were selected and paired with 17 MAGs recovered from short-read assembly through the 95% ANI clusters (the highest ANI were chosen if there were many) for comparison. The short-read MAGs included 5 medium quality (≥70% completeness, ≤10% contamination) and 12 high quality (≥95% completeness, ≤5% contamination).

Evaluation of these 17 MAGs were carried out as follows. First, each pair of hybrid MAG and short-read MAG were compared in terms of total number of MGE recombinases, as well as unique and shared MGE recombinases at 100% AAI by CD-HIT (v4.8.1) with '-c 1 -g 0 -G 1 -l 30 -d 0 -M 0' (ref. 67). Second, we evaluated the impact of sequencing depth, assembly and binning to assess which step was most problematically leading to under-detection of MGE recombinase identification in short reads. Initially, we mapped short-read assembled contigs (sheared into 500-bp fragments) to hybrid MAGs from the same sample using CoverM v0.6.1, requiring each read to be 75% aligned with 95% identity. Third, these data were used to assign MGE recombinases in the 17 hybrid MAGs into one of three categories: (1) those recovered by short-read MAGs (termed 'covered'); (2) those not fully covered by short-read contigs, that is, those missed by short-read assembly (termed 'missed assembly'); and (3) those fully covered by short-read contigs, that is, those recovered by short-read assembly but missed by short-read binning (termed 'missed binning'). Fourth, the coverage of each gene was generated by dirseq v0.4.3[26] using the BAM file from above and used to assess the relative abundances of each category as a metric for the impact on MGE-representation in the short-read assemblies.

With a better understanding for which step was failing most often for MGE recombinase recovery in short-read assemblies, we sought next to evaluate what factor(s) might be causing the failed capture events. First, we evaluated sequencing depth, as low coverage might cause breakages in short-read assembly graphs. This was done by mapping trimmed short reads to hybrid MAGs using CoverM v0.6.1, requiring each read to be 75% aligned with 95% identity, calculating the read coverage of each gene by dirseq v0.4.3 and comparing coverages among the three categories of MGE recombinases. Second, we compared the genomic neighbourhood number among the three categories, reasoning that perhaps indels were an issue. Genomic neighbourhood number was calculated by clustering genomic neighbourhoods of each unique MGE recombinase (see below) at 90% ANI. To test this, nucleotide diversity for each base position was calculated as 1 − [(number of 'A' bases/total bases)^2 + (number of 'C' bases/total bases)^2 + (number of 'T' bases/total bases)^2 + (number of 'G' bases/total bases)^2 + (number of 'deletion' bases/total bases)^2 + (number of 'insertion' bases/total bases)^2], adopting the nucleotide diversity definition from inStrain[7] with the addition of deletion and insertion variants. The variant count at the base position was generated by pysamstats (v1.1.2; https://github.com/alimanfoo/pysamstats) with the '–type=variation' option. Third, we assessed the nucleotide diversity of each MGE recombinase, reasoning that perhaps elevated nucleotide diversity could lead to short-read assembly graph breakages. To test this, for each of the 17 MAG pairs, MGE recombinases only identified in hybrid MAG were compared with short-read assembled contigs to track whether they were assembled by short reads, and then whether they were binned in MAGs from other MAG clusters. The tracking of MGE recombinases in contigs was accomplished with 'MMseqs map' (v13.45111), with default settings, and then filtered for hits with 100% identity in the protein sequence[75].

MGE recombinase per genome estimation from contigs and MAGs

To estimate the MGE recombinase number per genome from contigs, the median count of ribosomal genes was first used to estimate the number of genomes in each assembly. To do this, ribosomal genes were extracted from contig annotations by DRAM[61] using KEGG IDs (Supplementary Table 6), and the average number of recombinase genes per genome was calculated by dividing the total number of recombinases in the sample by the estimated total number of genomes. We next linked individual recombinase-encoding contigs to MAGs built from the same sample. To that end, we used exact string searching to identify MGE-recombinase-encoding contigs in the dereplicated MAGs from the same sample. Through this process, 106,017 out of a total 665,700 (15.9%) MGE-recombinases-encoding contigs ≥3 kbp could be linked to a MAG. To account for partial MAG recovery, the number of MGE recombinases per genome estimated from MAGs was calculated as the number of MGE recombinases in a given MAG divided by MAG completeness.

Genomic neighbourhoods of MGE recombinases and mobility evaluation

MGE mobility was assessed by comparing the genomic neighbourhoods of individual recombinases. For this analysis, we limited the input data to recombinases that occurred ≥2 times in the total dataset (assembled contigs per sample). This was achieved by clustering the recombinases at 100% AAI using CD-HIT v4.8.1 with '-c 1 -g 0 -G 1 -l 30 -d 0 -M 0' (ref. 67), only keeping the clusters with at least two recombinases. For each MGE recombinase, 300 bp upstream and downstream of the recombinase were considered as its immediate genomic neighbourhoods, excluding those near the contig ends (<300 bp). For each recombinase cluster, a pairwise comparison of neighbourhoods was performed among all member recombinases using both neighbourhoods (that is, upstream versus upstream and downstream versus downstream) using BLAST+ (v2.8.1+) with '-task blastn'[92] and a bitscore ≥50. The ANI for each neighbourhood comparison was calculated as the 'sequence identity from blastn output' multiplied by the 'alignment coverages'. For each comparison, we labelled the neighbourhood with the higher ANI value as 'high_ani' and the neighbourhood with lower ANI as 'low_ani'. For some MGEs, we expect the neighbourhood on the MGE side of the recombinase to be part of the MGE, and thus remain identical after integration into a new locus, so we used 'low_ani' values (that is, the lower ANI between the upstream and downstream region of the pair) to evaluate whether the neighbourhood of an MGE changed or not. An MGE recombinase pair was considered to be in different genomic

locations if 'low_ani' was <1%, and a recombinase cluster was treated as active if one pairwise comparison showed different neighbourhoods. This approach only used immediate neighbourhoods (within 300 bp) as a requirement, as neighbourhoods with longer ranges would remove more recombinases from this analysis owing to the short average contig length. However, as a sensitivity analysis, we also tested the above with 600-bp and 1,000-bp neighbourhoods. These analyses showed that the size of the genomic neighbourhood (300–1,000 bp) had very little impact on the fraction of MGE recombinases identified as mobile via this changing genomic neighbourhood metric (Extended Data Fig. 9).

MGE recombinase activity by metatranscriptome

MGE activity was assessed via documenting which MGE recombinases were detectably transcribed in metatranscriptomes. Leveraging 81 paired metatranscriptomes and metagenomes, we evaluated MGE recombinase transcriptional activity by mapping metatranscriptome reads to contigs assembled from metagenomes using transcriptM v0.3.1[93] with ORF predictions from DRAM (see above) as the '–gff' input, which also removed tRNA, rRNA, tmRNA and PhiX from the reads. To ensure the quality of RNA sequencing data (reverse stranded), we only analysed 58 pairs with ≥80% of reads mapped to the coding reverse strand of ORFs in contigs, based on dirseq v0.4.3[26] with BAM and GFF files from above as inputs (Supplementary Table 5). We only considered contigs ≥10 kbp to avoid spurious ORF predictions from short contigs. Next, we defined genes with ≥90% of gene length covered as expressed using pydirseq v0.1.0-alpha with BAM and GFF files from above as inputs[94]. To normalize for the differences in sequencing depth and overall activity of the microbial community across samples, we used the average percentage of ribosomal genes (Supplementary Table 6) detected as active as the baseline activity to normalize other genes (termed 'normalized active ratio'), as ribosomal genes are expected to be constitutively expressed in active cells. This 'normalized active ratio' thus provides an estimation of the number of genes detected as expressed within a functional category (for example, recombinases) relative to an estimated total number of genomes detected as transcriptionally active in the same sample.

MGE recombinase partial carriage by host population

As a result of MGE movement within the host population and an equilibrium of the co-evolution of MGEs and their hosts[95], we noticed that some MGE recombinases were present in only a fraction of the contigs assigned to a given host population (termed 'partial carriage'). To quantitatively evaluate this, trimmed metagenomic reads were mapped to contigs from the same sample using CoverM v0.6.1 with default settings[87], the resulting BAM file was then filtered (requiring ≥90% of each read aligned with ≥95% identity) and used as the input for 'samtools depth' (v1.17) with the '-a' option to get the read depth coverage at each position[96]. The first and last 150 bp of each contig were not considered, as read depth coverage near the ends are expected to be artificially lower. We defined partial carriage as cases where the coverage of the MGE recombinase was at least one standard deviation lower than the larger one (requiring a minimal coverage of 1×) of the average coverages of the upstream and downstream contig regions (Cohen's $d \geq 1$). To ensure enough host regions were considered for reliable mapping-based estimation of genome coverage, the MGE recombinase total set was filtered to 23,647 MGE recombinases with ≥20 kb upstream and downstream. Note that sensitivity analyses revealed similar patterns when analysing 2.5-kbp instead of 20-kbp regions (Extended Data Fig. 10), which suggests that this conservative cut-off did not substantially influence the detection of partial carriage in these data.

We expect that this method should give us a lower-bound estimate of partial carriage, as it is not able to detect cases for which (1) the hosts have low read coverage (rare), (2) the carriage ratio is lower than but close to 1, (3) host genome coverage has a wide range of variation in different genomic locations due to strain variation within the population or (4) the MGE recombinase is on the edge of a contig.

Statistics

All statistics and visualization were performed in R v4.2.1[97]. The R packages and versions used are available via GitHub[98]. For statistical significance between groups, the Wilcoxon rank-sum test was used for two groups and Kruskal Wallis test was used for >2 groups as implemented in 'compare_means' in ggpubr package v0.6.0, with *P* values adjusted by Benjamini–Hochberg adjustment for multiple comparisons.

Reporting summary

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

Data availability

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 Supplementary Table 1. 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 via Zenodo at https://doi.org/10.5281/zenodo.13984630 (ref. 99) (main set of 36,419 MAGs) and at https://doi.org/10.5281/zenodo.14207415 (ref. 100) (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 listed in Supplementary Table 8. Processed data, including tables for MGE recombinase identification, activity by metatranscriptome, mobility by genomic neighbourhood change and partial carriage, and MGE-impacted functions are available through the supplementary tables and a large file data package available via Zenodo at https://doi.org/10.5281/zenodo.20061508 (ref. 101); this record gives source data for figure generation.

Code availability

The code produced in this study, including code for figure generation, is also available via Zenodo at https://doi.org/10.5281/zenodo.20061508 (ref. 101), which is an archive of the latest version of the project's repository available via GitHub at https://github.com/emerge-bii/mobilome-submit-202407.

References

1. Crowther, T. W. et al. The global soil community and its influence on biogeochemistry. *Science* **365**, eaav0550 (2019).
2. Paoli, L. et al. Biosynthetic potential of the global ocean microbiome. *Nature* **607**, 111–118 (2022).
3. Nayfach, S., Shi, Z. J., Seshadri, R., Pollard, K. S. & Kyrpides, N. C. New insights from uncultivated genomes of the global human gut microbiome. *Nature* **568**, 505–510 (2019).
4. Nocker, A., Burr, M. & Camper, A. K. Genotypic microbial community profiling: a critical technical review. *Microb. Ecol.* **54**, 276–289 (2007).
5. Nielsen, H. B. et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* **32**, 822–828 (2014).
6. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).
7. Olm, M. R. et al. inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nat. Biotechnol.* **39**, 727–736 (2021).
8. Quince, C. et al. STRONG: metagenomics strain resolution on assembly graphs. *Genome Biol.* **22**, 214 (2021).

9. Blanco-Míguez, A. et al. Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat. Biotechnol.* **41**, 1633–1644 (2023).
10. Loman, N. J. et al. A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of shiga-toxigenic *Escherichia coli* O104:H4. *J. Am. Med. Assoc.* **309**, 1502–1510 (2013).
11. Crits-Christoph, A., Olm, M. R., Diamond, S., Bouma-Gregson, K. & Banfield, J. F. Soil bacterial populations are shaped by recombination and gene-specific selection across a grassland meadow. *ISME J.* **14**, 1834–1846 (2020).
12. Van Rossum, T., Ferretti, P., Maistrenko, O. M. & Bork, P. Diversity within species: interpreting strains in microbiomes. *Nat. Rev. Microbiol.* **18**, 491–506 (2020).
13. Haudiquet, M., de Sousa, J. M., Touchon, M. & Rocha, E. P. C. Selfish, promiscuous and sometimes useful: how mobile genetic elements drive horizontal gene transfer in microbial populations. *Philos. Trans. R. Soc. Lond. B* **377**, 20210234 (2022).
14. Durrant, M. G., Li, M. M., Siranosian, B. A., Montgomery, S. B. & Bhatt, A. S. A bioinformatic analysis of integrative mobile genetic elements highlights their role in bacterial adaptation. *Cell Host Microbe* **27**, 140–153 (2020).
15. Khedkar, S. et al. Landscape of mobile genetic elements and their antibiotic resistance cargo in prokaryotic genomes. *Nucleic Acids Res.* **50**, 3155–3168 (2022).
16. Prentki, P., Teter, B., Chandler, M. & Galas, D. J. Functional promoters created by the insertion of transposable element IS1. *J. Mol. Biol.* **191**, 383–393 (1986).
17. Lipszyc, A., Szuplewska, M. & Bartosik, D. How do transposable elements activate expression of transcriptionally silent antibiotic resistance genes?. *Int. J. Mol. Sci.* **23**, 8063 (2022).
18. Weisberg, A. J. & Chang, J. H. Mobile genetic element flexibility as an underlying principle to bacterial evolution. *Annu. Rev. Microbiol.* **77**, 603–624 (2023).
19. Kerkvliet, J. J. et al. Metagenomic assembly is the main bottleneck in the identification of mobile genetic elements. *PeerJ* **12**, e16695 (2024).
20. Maguire, F. et al. Metagenome-assembled genome binning methods with short reads disproportionately fail for plasmids and genomic islands. *Microb. Genom.* **6**, mgen000436 (2020).
21. Partridge, S. R., Kwong, S. M., Firth, N. & Jensen, S. O. Mobile genetic elements associated with antimicrobial resistance. *Clin. Microbiol. Rev.* **31**, e00088–17 (2018).
22. Roux, S. et al. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537**, 689–693 (2016).
23. Tamayo-Leiva, J. et al. Structure and dispersion of the conjugative mobilome in surface ocean bacterioplankton. *ISME Commun.* **4**, ycae059 (2024).
24. Yi, Y. et al. A systematic analysis of marine lysogens and proviruses. *Nat. Commun.* **14**, 6013 (2023).
25. McCalley, C. K. et al. Methane dynamics regulated by microbial community response to permafrost thaw. *Nature* **514**, 478–481 (2014).
26. Woodcroft, B. J. et al. Genome-centric view of carbon processing in thawing permafrost. *Nature* **560**, 49–54 (2018).
27. Emerson, J. B. et al. Host-linked soil viral ecology along a permafrost thaw gradient. *Nat. Microbiol.* **3**, 870–880 (2018).
28. Sun, C. L. et al. Virus ecology and 7-year temporal dynamics across a permafrost thaw gradient. *Environ. Microbiol.* **26**, e16665 (2024).
29. Cronin, D. et al. Stable states in an unstable landscape: microbial resistance at the front line of climate change. Preprint at *bioRxiv* https://doi.org/10.1101/2025.02.07.636677 (2025).
30. Aziz, R. K., Breitbart, M. & Edwards, R. A. Transposases are the most abundant, most ubiquitous genes in nature. *Nucleic Acids Res.* **38**, 4207–4217 (2010).
31. He, C. et al. Genome-resolved metagenomics reveals site-specific diversity of episymbiotic CPR bacteria and DPANN archaea in groundwater ecosystems. *Nat. Microbiol.* **6**, 354–365 (2021).
32. Roux, S. et al. Ecology and molecular targets of hypermutation in the global microbiome. *Nat. Commun.* **12**, 3076 (2021).
33. Zablocki, O. et al. VirION2: a short- and long-read sequencing and informatics workflow to study the genomic diversity of viruses in nature. *PeerJ* **9**, e11088 (2021).
34. Rankin, D. J., Rocha, E. P. C. & Brown, S. P. What traits are carried on mobile genetic elements, and why?. *Heredity* **106**, 1–10 (2011).
35. Dragoš, A. et al. Phages carry interbacterial weapons encoded by biosynthetic gene clusters. *Curr. Biol.* **31**, 3479–3489 (2021).
36. Benler, S. et al. Cargo genes of Tn7-like transposons comprise an enormous diversity of defense systems, mobile genetic elements, and antibiotic resistance genes. *mBio* **12**, e0293821 (2021).
37. Finks, S. S. & Martiny, J. B. H. Plasmid-encoded traits vary across environments. *mBio* **14**, e0319122 (2023).
38. van Spanning, R. J. M. et al. Methanotrophy by a *Mycobacterium* species that dominates a cave microbial ecosystem. *Nat. Microbiol.* **7**, 2089–2100 (2022).
39. Kambara, H. et al. First isolation of a methanotrophic *Mycobacterium* reveals ammonia- and pH-tolerant methane oxidation. *Appl. Environ. Microbiol.* **91**, e0079625 (2025).
40. Abdul Halim, M. F., Fonseca, D. R., Niehaus, T. D. & Costa, K. C. Functionally redundant formate dehydrogenases enable formate-dependent growth in *Methanococcus maripaludis*. *J. Biol. Chem.* **300**, 105550 (2024).
41. Ernakovich, J. G. et al. Microbiome assembly in thawing permafrost and its feedbacks to climate. *Glob. Change Biol.* **28**, 5007–5026 (2022).
42. Sleator, R. D. & Hill, C. Bacterial osmoadaptation: the role of osmolytes in bacterial stress and virulence. *FEMS Microbiol. Rev.* **26**, 49–71 (2002).
43. Borga, P., Nilsson, M. & Tunlid, A. Bacterial communities in peat in relation to botanical composition as revealed by phospholipid fatty acid analysis. *Soil Biol. Biochem.* **26**, 841–848 (1994).
44. Benler, S. & Koonin, E. V. Recruitment of mobile genetic elements for diverse cellular functions in prokaryotes. *Front. Mol. Biosci.* **9**, 821197 (2022).
45. Sharon, I. et al. Viral photosynthetic reaction center genes and transcripts in the marine environment. *ISME J.* **1**, 492–501 (2007).
46. Franzosa, E. A. et al. Relating the metatranscriptome and metagenome of the human gut. *Proc. Natl Acad. Sci. USA* **111**, E2329–E2338 (2014).
47. Mondav, R. et al. Discovery of a novel methanogen prevalent in thawing permafrost. *Nat. Commun.* **5**, 3212 (2014).
48. Bernheim, A. & Sorek, R. The pan-immune system of bacteria: antiviral defence as a community resource. *Nat. Rev. Microbiol.* **18**, 113–119 (2020).
49. Cordero, O. X., Ventouras, L.-A., DeLong, E. F. & Polz, M. F. Public good dynamics drive evolution of iron acquisition strategies in natural bacterioplankton populations. *Proc. Natl Acad. Sci. USA* **109**, 20059–20064 (2012).
50. McGivern, B. B. et al. Microbial polyphenol metabolism is part of the thawing permafrost carbon cycle. *Nat. Microbiol.* **9**, 1454–1466 (2024).
51. Chklovski, A., Parks, D. H., Woodcroft, B. J. & Tyson, G. W. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat. Methods* **20**, 1203–1212 (2023).
52. Aroney, S. T. N., Camargo, A. P., Tyson, G. W. & Woodcroft, B. J. Galah: more scalable dereplication for metagenome assembled genomes. *Zenodo* https://doi.org/10.5281/zenodo.10526086 (2024).

53. Li, Y.-F., Isogenie 1 Project Coordinators, Rich, V. & Mondav, R. Modified nucleic acid extraction protocol for humic-rich permafrost peat samples. *protocols.io* https://www.protocols.io/view/modified-nucleic-acid-extraction-protocol-for-humi-csvywe7w (2024).
54. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
55. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
56. Kang, D. D. et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
57. Imelfort, M. et al. GroopM: an automated tool for the recovery of population genomes from related metagenomes. *PeerJ* **2**, e603 (2014).
58. Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605–607 (2016).
59. Bushnell, B., Rood, J. & Singer, E. BBMerge – accurate paired shotgun read merging via overlap. *PLoS ONE* **12**, e0185056 (2017).
60. Li, D., Liu, C. M., Luo, R., Sadakane, K. & Lam, T. W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
61. Shaffer, M. et al. DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Res.* **48**, 8883–8900 (2020).
62. von Meijenfeldt, F. A. B., Arkhipova, K., Cambuy, D. D., Coutinho, F. H. & Dutilh, B. E. Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biol.* **20**, 217 (2019).
63. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2020).
64. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
65. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
66. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
67. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
68. Néron, B. et al. IntegronFinder 2.0: identification and analysis of integrons across bacteria, with a focus on antibiotic resistance in *Klebsiella*. *Microorganisms* **10**, 700 (2022).
69. Boucher, Y., Labbate, M., Koenig, J. E. & Stokes, H. W. Integrons: mobilizable platforms that promote genetic diversity in bacteria. *Trends Microbiol.* **15**, 301–309 (2007).
70. Cury, J., Jové, T., Touchon, M., Néron, B. & Rocha, E. P. Identification and analysis of integrons and cassette arrays in bacterial genomes. *Nucleic Acids Res.* **44**, 4539–4550 (2016).
71. Loot, C. et al. Integron cassettes integrate into bacterial genomes via widespread non-classical attG sites. *Nat. Microbiol.* **9**, 228–240 (2024).
72. Cury, J., Abby, S. S., Doppelt-Azeroual, O., Néron, B. & Rocha, E. P. C. Identifying conjugative plasmids and integrative conjugative elements with CONJscan. *Methods Mol. Biol.* **2075**, 265–283 (2020).
73. Abby, S. S., Néron, B., Ménager, H., Touchon, M. & Rocha, E. P. C. MacSyFinder: a program to mine genomes for molecular systems with an application to CRISPR-Cas systems. *PLoS ONE* **9**, e110726 (2014).
74. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
75. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
76. Guo, J. et al. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* **9**, 37 (2021).
77. Guo, J., Vik, D., Pratama, A. A., Roux, S. & Sullivan, M. Viral sequence identification SOP with VirSorter2. *protocols.io* https://www.protocols.io/view/viral-sequence-identification-sop-with-virsorter2-bwm5pc86 (2021).
78. Nayfach, S. et al. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* **39**, 578–585 (2021).
79. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
80. Newell, R. J. P. et al. Aviary: hybrid assembly and genome recovery from metagenomes with Aviary. *Zenodo* https://doi.org/10.5281/zenodo.10806928 (2024).
81. Kolmogorov, M. et al. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat. Methods* **17**, 1103–1110 (2020).
82. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
83. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).
84. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput. Biol.* **13**, e1005595 (2017).
85. Aroney, S. T. N., Newell, R. J. P., Tyson, G. W. & Woodcroft, B. J. Bin Chicken: targeted metagenomic coassembly for the efficient recovery of novel genomes. *Nat. Methods* **22**, 2516–2524 (2025).
86. Woodcroft, B. J. Default SingleM reference 'metapackage' data. *Zenodo* https://doi.org/10.5281/zenodo.7130825 (2022).
87. Aroney, S. T. N. et al. CoverM: read alignment statistics for metagenomics. *Bioinformatics* **41**, btaf147 (2025).
88. Nissen, J. N. et al. Improved metagenome binning and assembly using deep variational autoencoders. *Nat. Biotechnol.* **39**, 555–560 (2021).
89. Pan, S., Zhu, C., Zhao, X.-M. & Coelho, L. P. A deep siamese neural network improves metagenome-assembled genomes in microbiome datasets across different environments. *Nat. Commun.* **13**, 2326 (2022).
90. Newell, R. J. P., Tyson, G. W. & Woodcroft, B. J. Rosella: metagenomic binning using UMAP and HDBSCAN. *Zenodo* https://doi.org/10.5281/zenodo.10460259 (2024).
91. Alneberg, J. et al. Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144–1146 (2014).
92. Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
93. Sternes, P. R. & Tyson, G. W. TranscriptM: a metatranscriptome bioinformatics pipeline including metagenome contamination correction. *Zenodo* https://doi.org/10.5281/zenodo.11157285 (2024).
94. Guo, J. pydirseq: a tool to estimate per gene depth and breadth coverage from bam files. *Zenodo* https://doi.org/10.5281/zenodo.12683197 (2024).
95. Koonin, E. V., Makarova, K. S., Wolf, Y. I. & Krupovic, M. Evolutionary entanglement of mobile genetic elements and host defence systems: guns for hire. *Nat. Rev. Genet.* **21**, 119–131 (2020).
96. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
97. R Core Team. R: a language and environment for statistical computing (R Foundation for Statistical Computing, 2014).

98. Guo, J. et al. Collection of R and python scripts. *GitHub* https://github.com/emerge-bii/mobilome-submit-202407 (2026).
99. Aroney, S., Cronin, D., Bagby, S., Hodgkins, S. & NSF EMERGE Biology Integration Institute. Metagenome-assembled genomes from Stordalen Mire, Sweden (MAGs v2). *Zenodo* https://doi.org/10.5281/zenodo.13984630 (2024).
100. Aroney, S. & Woodcroft, B. J. Metagenome-assembled genomes from Stordalen Mire, Sweden (2019) (MAGs from long-read, short-read, & hybrid assemblies). *Zenodo* https://doi.org/10.5281/zenodo.14207415 (2024).
101. Guo, J., Bagby, S., Sullivan, M. & Roux, S. Mobilome paper supplementary large file data package. *Zenodo* https://doi.org/10.5281/zenodo.20061508 (2026).
102. Chang, K.-Y. et al. Large carbon cycle sensitivities to climate across a permafrost thaw gradient in subarctic Sweden. *Cryosphere* https://doi.org/10.5194/tc-13-647-2019 (2019).

Acknowledgements

We thank the Ohio Supercomputer Center for providing computing resources. We thank B. Bolduc for help with data validation and O. Zablocki for help with preparation of the paper. We thank Abisko Scientific Research station (ANS) for their support to the field team.

Author contributions

R.K.V. is the representative for the IsoGenie Field Teams and V.I.R. is the representative for the EMERGE Coordinators. J.G., S.R., M.B.S. and S.C.B. conceptualized, designed and coordinated the study. S.T.N.A. and J.G. performed long-read and short-read comparison. S.T.N.A., A.F. and J.G. performed MGE recombinase host lineage analyses. G.D.-H., D.S., D.V., S.C.B., A.A.P., G.J.S., F.T., C.H.-V., S.S., Z.-P.Z., C.O.-A. and J.G. performed MGE-impacted function analyses. D.C., S.T.N.A. and S.B.H. performed metagenome assembly, binning and data curation. J.G., S.R., M.B.S. and S.C.B. drafted the paper with contributions from B.J.W., G.W.T. and V.I.R. All authors read, edited, commented on and approved the final paper.

Funding

This work is supported by the following sources of funding: the National Science Foundation, Biology Integration Institutes Program (award no. 2022070 to V.I.R.) and the DOE BER (award no. DE-SC0023307 to M.B.S.). Portions of this work were supported by the BER Support Science Proposal 503530 (proposal doi: 10.46936/10.25585/60001148 to V.I.R.) and used resources at the DOE Joint Genome Institute (https://ror.org/04xm1d337) and the Environmental Molecular Sciences Laboratory (https://ror.org/04rc0xn13), which are DOE Office of Science User Facilities. Both facilities are sponsored by the Office of Biological and Environmental Research and operated under Contract Nos. DE-AC02-05CH11231 (JGI) and DE-AC05-76RL01830 (EMSL). We thank the Swedish Polar Research Secretariat and the Swedish Infrastructure for Ecosystem Science (SITES) for the support of the work done at the Abisko Scientific Research Station; SITES is currently supported by the Swedish Research Council's grant 4.3-2021-00164. Work conducted at LLNL was conducted under the auspices of the US Department of Energy under Contract DE-AC52-07NA27344.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41564-026-02391-7.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41564-026-02391-7.

Correspondence and requests for materials should be addressed to Matthew B. Sullivan, Simon Roux or Sarah C. Bagby.

Peer review information *Nature Microbiology* thanks Ellie Harrison, Gilda Varliero and Lyle Whyte for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional 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 (QUT), Translational Research Institute, Woolloongabba, Queensland, Australia. [4]Department of Biology, Case Western Reserve University, Cleveland, OH, USA. [5]Biophysics Graduate Program, The Ohio State University, Columbus, OH, USA. [6]Byrd Polar and Climate Research Center, The Ohio State University, Columbus, OH, USA. [7]Department of Civil, Environmental, and Geodetic Engineering, The Ohio State University, Columbus, OH, USA. [8]DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [9]Present address: Division of Intramural Research, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA. [10]Present address: Centro Oceanográfico de Málaga (IEO-CSIC), Malaga, Spain. [11]Present address: Institute of Biodiversity, Ecology and Evolution, Aquatic Geomicrobiology, Friedrich Schiller University Jena, Jena, Germany. [12]Present address: Cluster of Excellence Balance of the Microverse, Friedrich Schiller University Jena, Jena, Germany. [13]Present address: State Key Laboratory of Tibetan Plateau Earth System, Environment, and Resources, Institute of Tibetan Plateau Research, Chinese Academy of Sciences, Beijing, China. *Lists of authors and their affiliations appear at the end of the paper. ✉e-mail: sullivan.948@osu.edu; sroux@lbl.gov; scb126@case.edu

IsoGenie Field Teams 2010–2017, 2019

Darya Anderson[14], Kathryn Bennett[15], Sky Dominguez[14], Maria Florencia Fahnestock[15], Moira Hough[16], Joachim Jansen[17], Rhiannon Mondav[18], Apryl Perry[19], Gareth Trubl[1,20] & Ruth K. Varner[19,21]

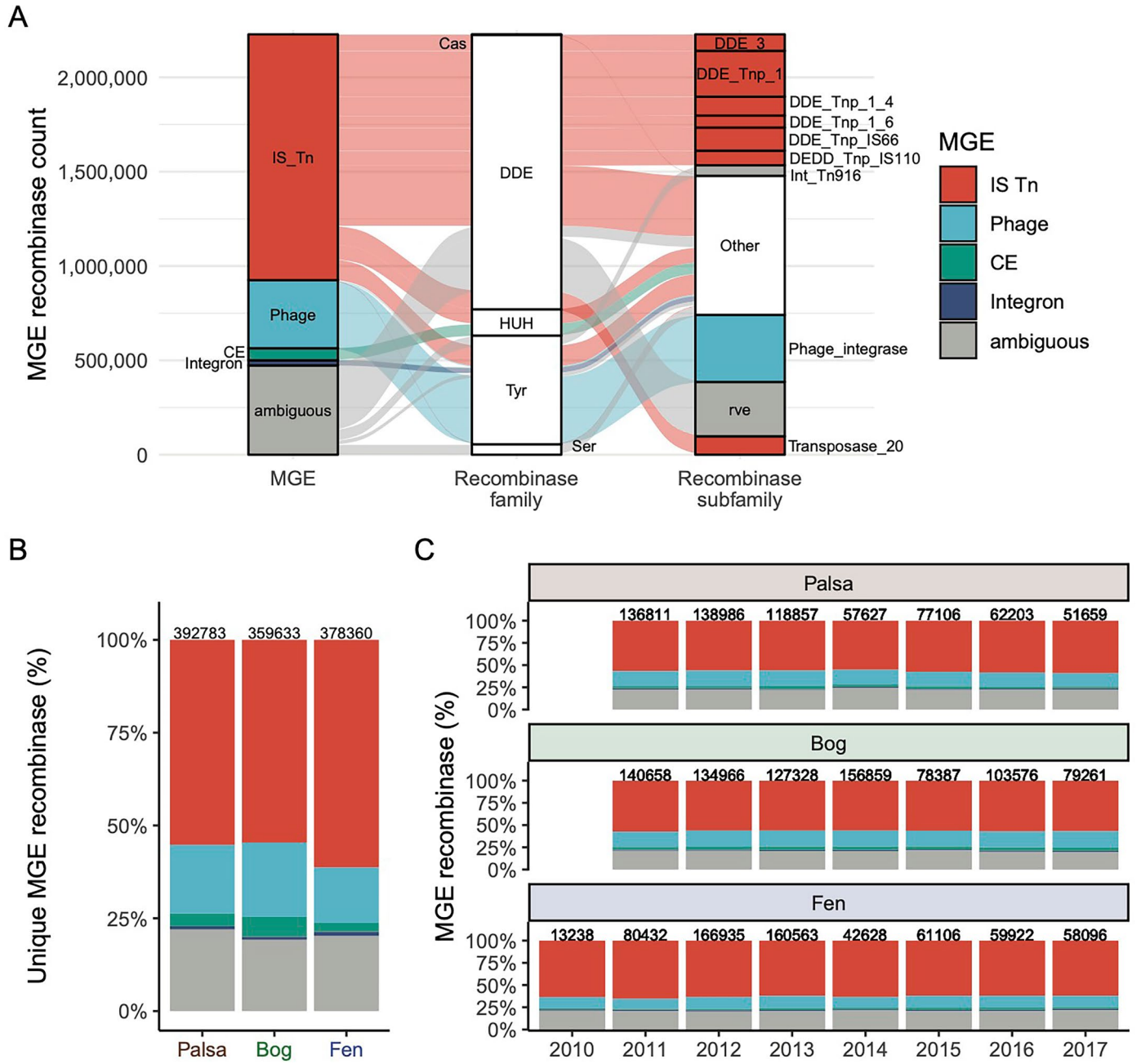
[14]Department of Environmental Science, University of Arizona, Tucson, AZ, USA. [15]Department of Earth Sciences, University of New Hampshire, Durham, NH, USA. [16]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, USA. [17]Department of Geological Sciences, Stockholm University, Stockholm, Sweden. [18]Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, University of Queensland, Brisbane, Queensland, Australia. [19]Department of Earth Sciences and Earth Systems Research Center, Institute for the Study of Earth, Oceans, and Space, University of New Hampshire, Durham, NH, USA. [20]Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, CA, USA. [21]Department of Geological Sciences and Bolin Centre for Climate Research, Stockholm University, Stockholm, Sweden.

EMERGE Coordinators

Sarah C. Bagby[4], Patrick Crill[21], Jessica Gilman Ernakovich[22], Regis Ferriere[16,23], Suzanne B. Hodgkins[1], Mike Ibba[24], Virginia I. Rich[1,2,6], Matthew B. Sullivan[1,2,5,6,7], Malak M. Tfaily[14], Gene W. Tyson[3], Ruth K. Varner[19,21], Ben J. Woodcroft[3] & Ahmed A. Zayed[1]

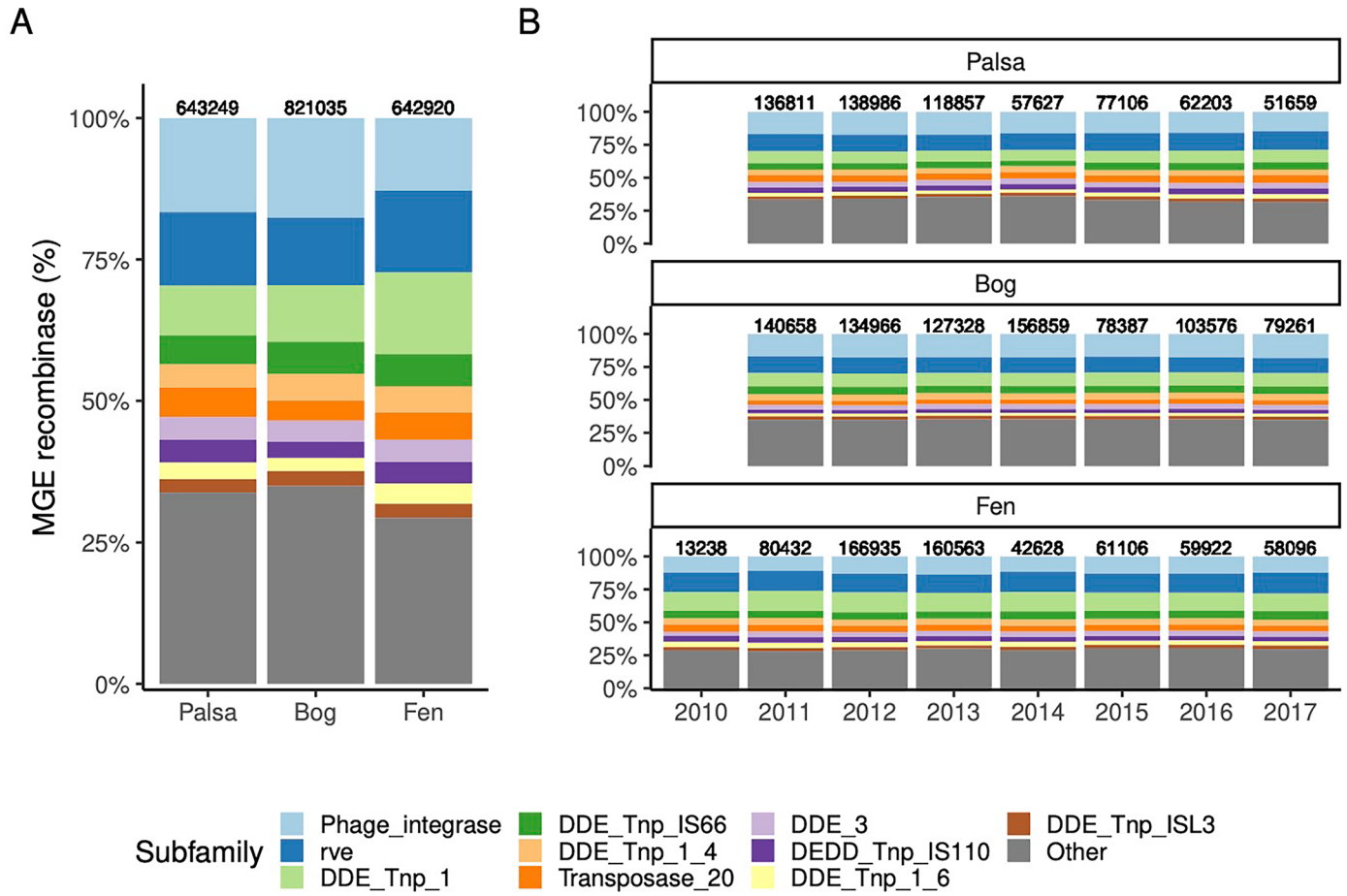
[22]Department of Natural Resources and the Environment, University of New Hampshire, Durham, NH, USA. [23]Institut de Biologie de l'Ecole Normale Supérieure, Université Paris Sciences & Lettres, Paris, France. [24]Schmid College of Science and Technology, Chapman University, Orange, CA, USA.

https://doi.org/10.1038/s41564-026-02391-7



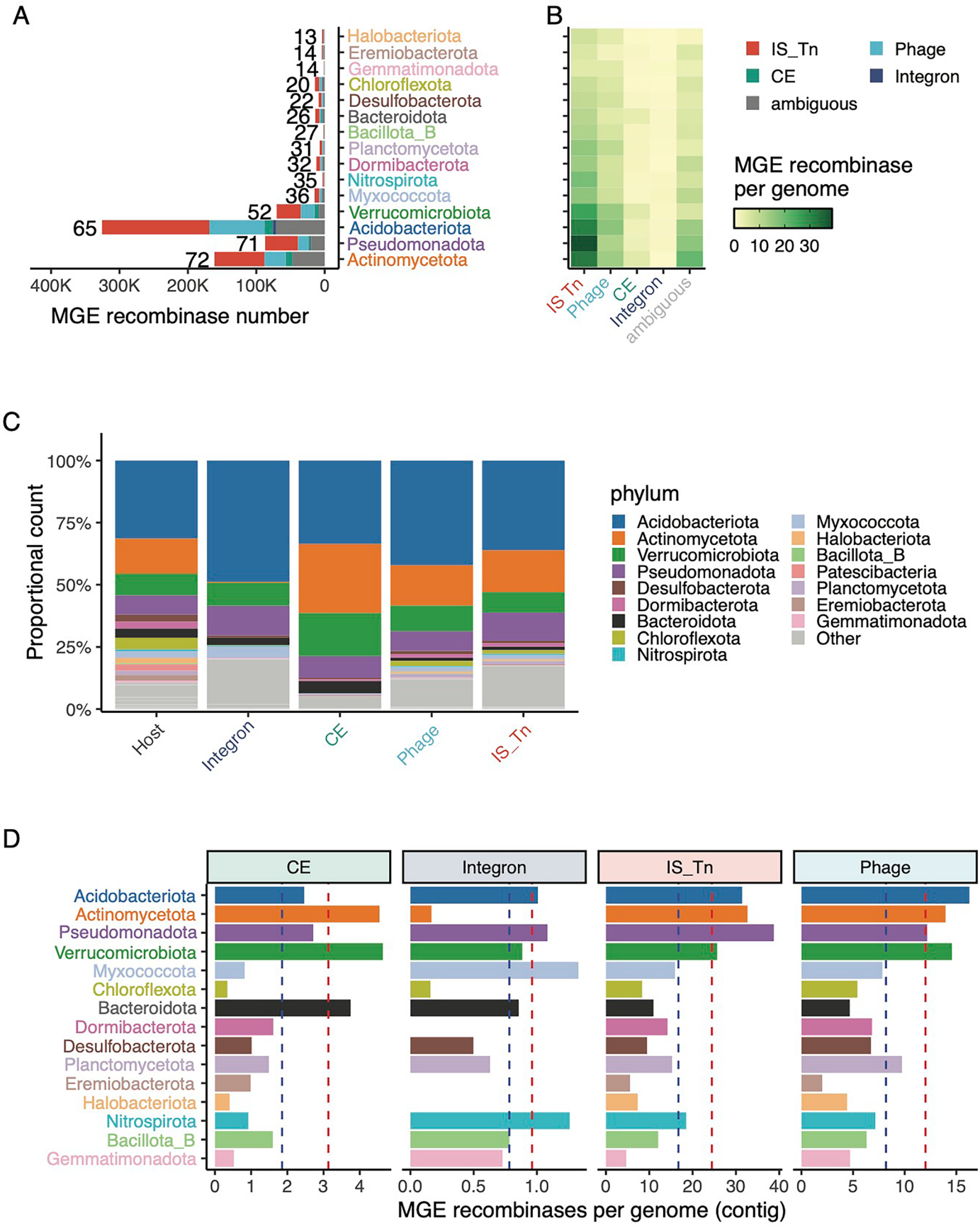
Extended Data Fig. 1 | Overview of MGE recombinase distribution from Stordalen Mire. A. Distribution of contig-identified MGE recombinases by MGE type, as used in the main text, as well as MGE recombinase family and subfamily (see a full list described in table S2). **B**. Unique MGE recombinases (clustered by 100% AAI) by habitat. Numbers above bars are the number of unique MGE recombinases. **C**. Total MGE recombinase distribution by habitat and year. Numbers above bars are the number of total MGE recombinases.

https://doi.org/10.1038/s41564-026-02391-7



Extended Data Fig. 2 | MGE recombinase distribution based on subfamily. A. MGE recombinase subfamily distribution by habitat. Numbers above bars are the number of total MGE recombinases. **B**. MGE recombinase subfamily distribution by habitat and year.

https://doi.org/10.1038/s41564-026-02391-7

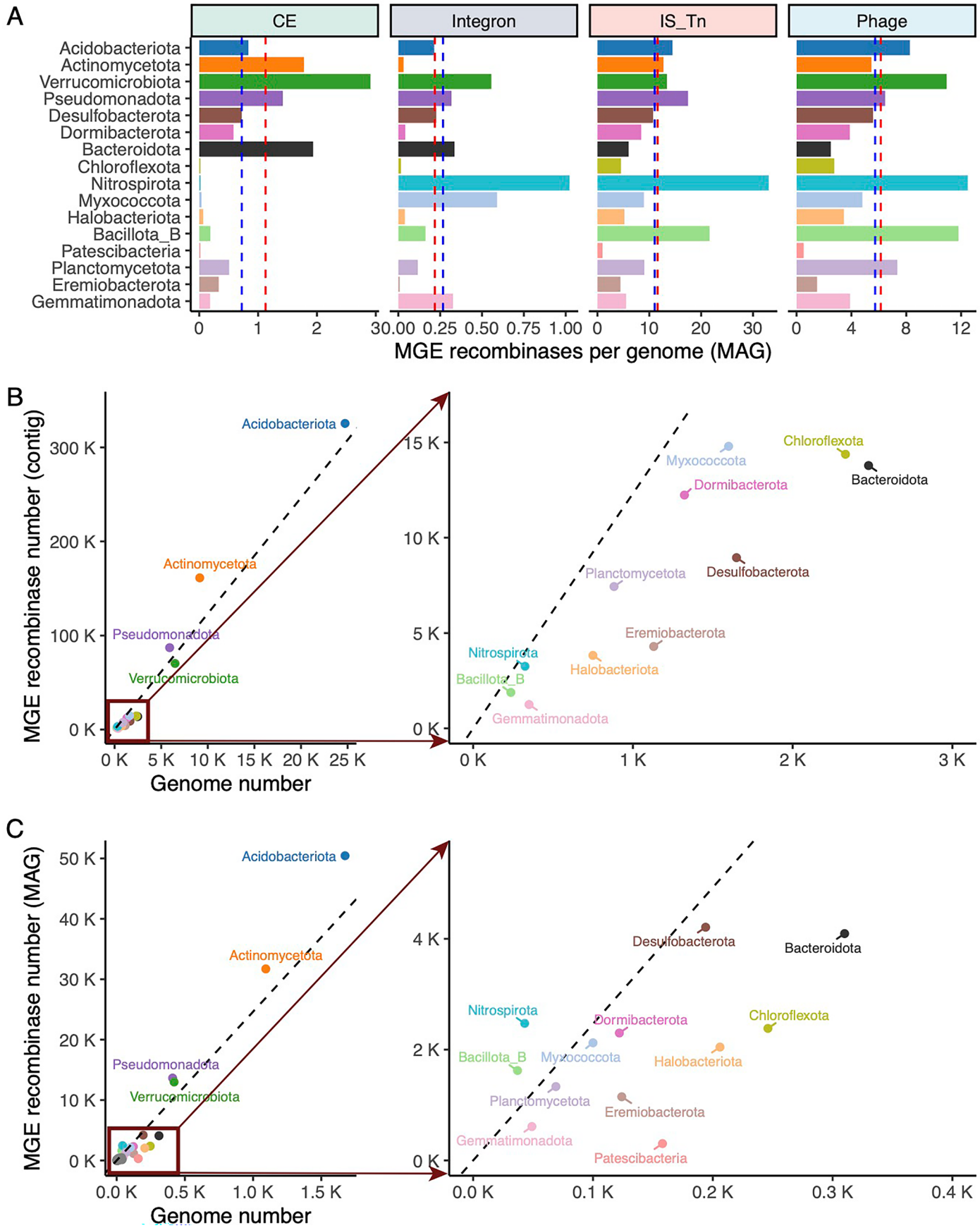


Extended Data Fig. 3 | See next page for caption.

https://doi.org/10.1038/s41564-026-02391-7

Extended Data Fig. 3 | MGE host lineage based on MGE recombinases identified from contigs. MGE recombinases identified on contigs (≥3 kb) were evaluated for taxonomic patterns as a complement to MAG-based analyses in Fig. 3. **A.** Number of each MGE type found in each of the top 15 phyla (out of 89 in total), accounting for 84.4% of total MGE recombinases. Phyla are ordered by MGE recombinase count per genome (the number to the left of each bar), with names colored per Fig. 3. The genome number is estimated from ribosomal genes (see Methods). **B.** Number of MGE recombinases identified per genome grouped by MGE type and host phylum. **C.** Phylum distribution of the overall microbial community (estimated from ribosomal gene encoding contigs) and MGE recombinase types. **D.** MGE recombinase number per genome per phylum. Dashed red lines are the average MGE recombinase number per genome across all phyla, and dashed blue lines are the average across the phyla shown in the plot.

https://doi.org/10.1038/s41564-026-02391-7

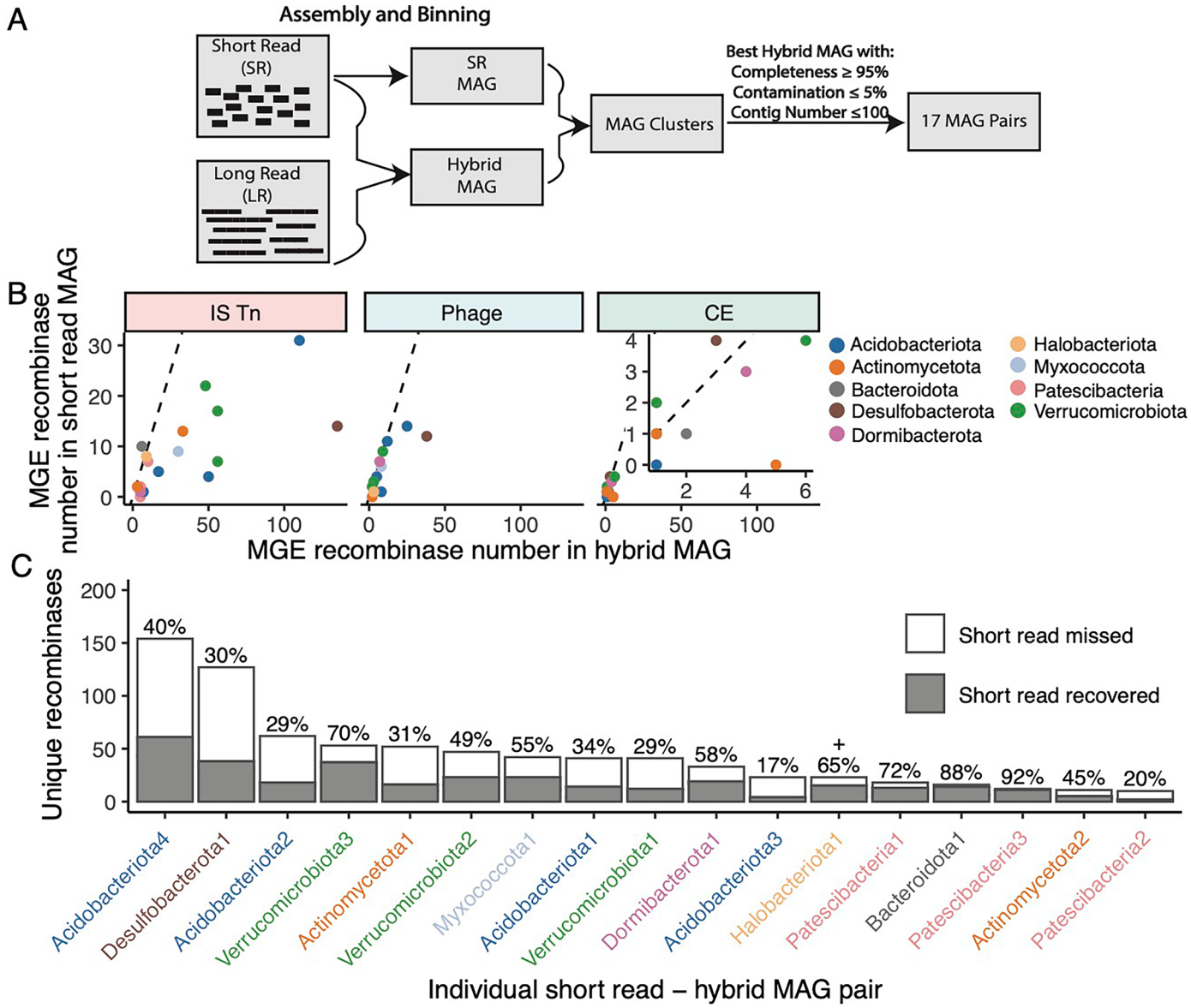


Extended Data Fig. 4 | See next page for caption.

https://doi.org/10.1038/s41564-026-02391-7

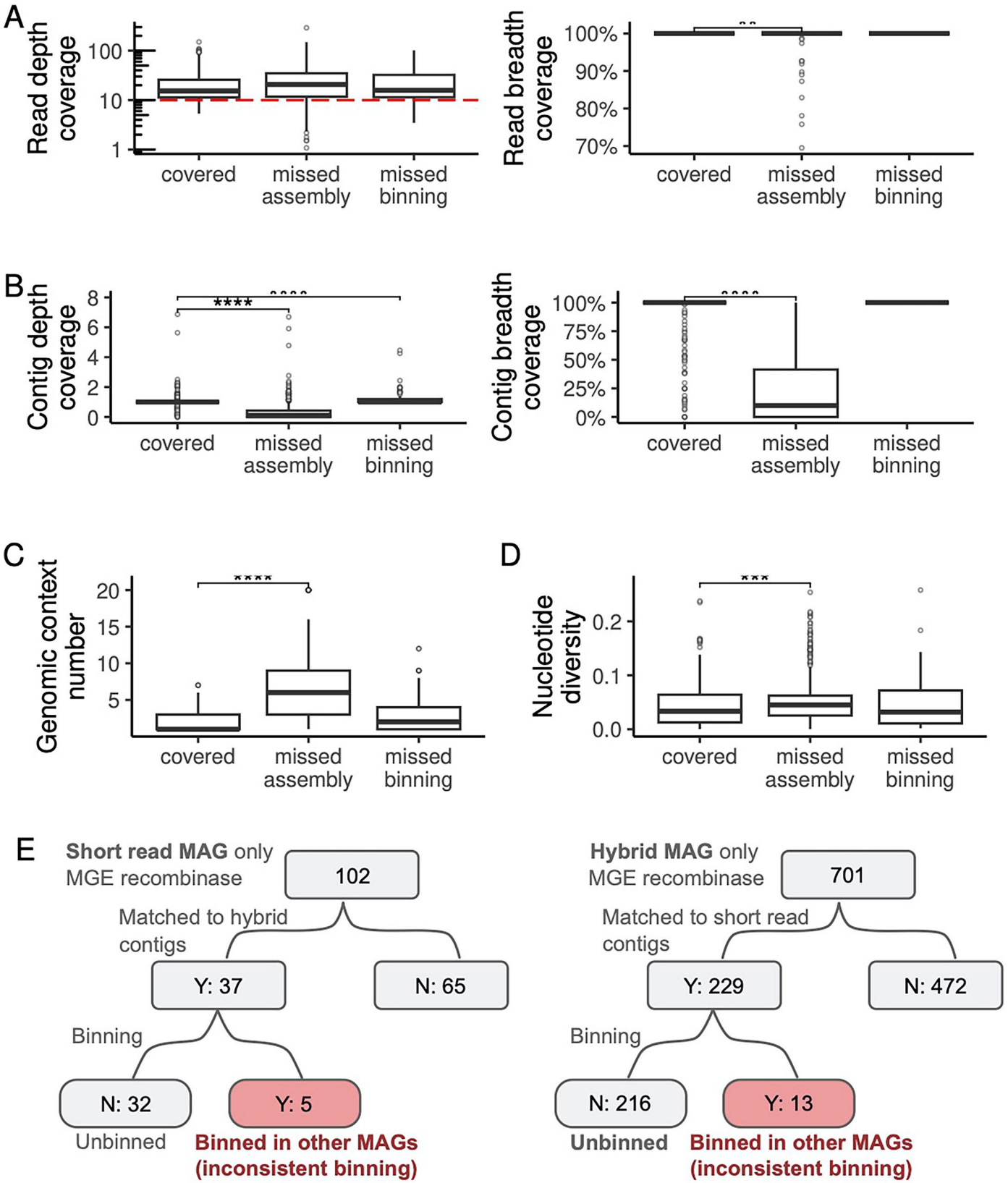
Extended Data Fig. 4 | MGE recombinase and genome number distribution across phyla. A. MGE recombinase number per genome per phylum from MAGs. Dashed red lines are the average MGE recombinase number per genome across all phyla, and dashed blue lines are the average across the phyla shown in the plot. **B**. Number of MGE recombinases against number of genomes per phylum using contigs. Dashed line represents the global average number of MGE recombinases per genome. The genome number is estimated from ribosomal genes. Right panel enlarges the area boxed in the left panel. **C**. Number of MGE recombinases against number of genomes per phylum using MAGs. Right panel enlarges the area boxed in the left panel.

https://doi.org/10.1038/s41564-026-02391-7



Extended Data Fig. 5 | A six-sample assessment of short-read assembly and binning artifacts for recovering MGE recombinases. A. Because short-read assemblies can struggle with hypervariable and/or repeat regions of genomes, six metagenomic samples were re-sequenced using high-fidelity long-read sequencing to develop a set of high-quality MAGs (≥95% completeness, ≤5% contamination, and <100 contigs) that could be compared for recombinase signal to short-read assembled MAGs from the same species cluster (≥95% ANI). A total of 17 such pairs were identified. **B.** Recombinase-centric patterns in recovery, where each point represents the total recombinase number identified in the short-read versus matching hybrid MAG. Dashed lines, 1:1 ratio. The inset in CE expands the dense region at lower left. **C.** Lineage-centric patterns in MGE recombinase recovery. MAG name color matches color in panel **B**. Numbers above bars are the percentages of total unique MGE recombinases recovered by short-read MAGs relative to the total.

https://doi.org/10.1038/s41564-026-02391-7

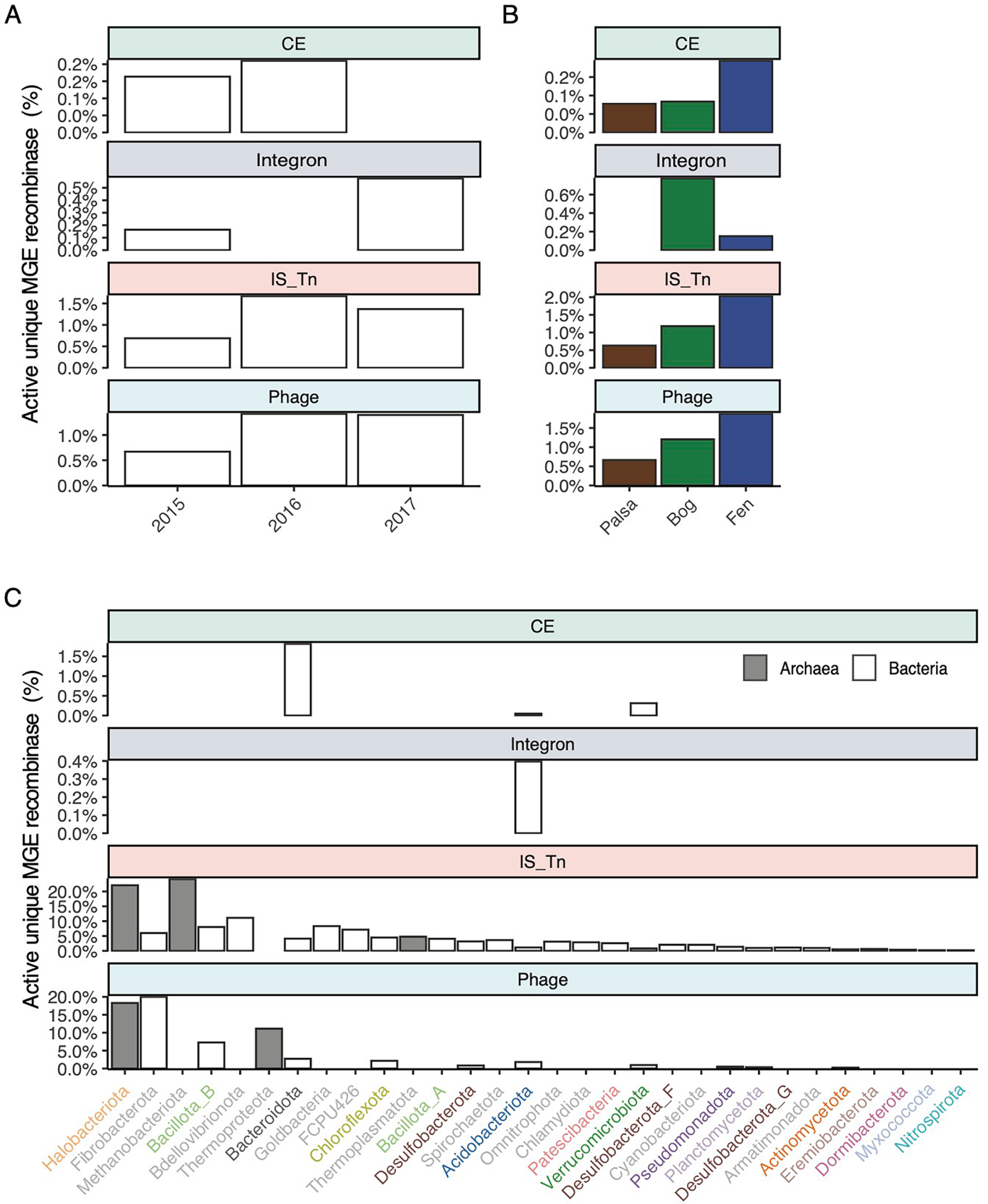


Extended Data Fig. 6 | See next page for caption.

https://doi.org/10.1038/s41564-026-02391-7

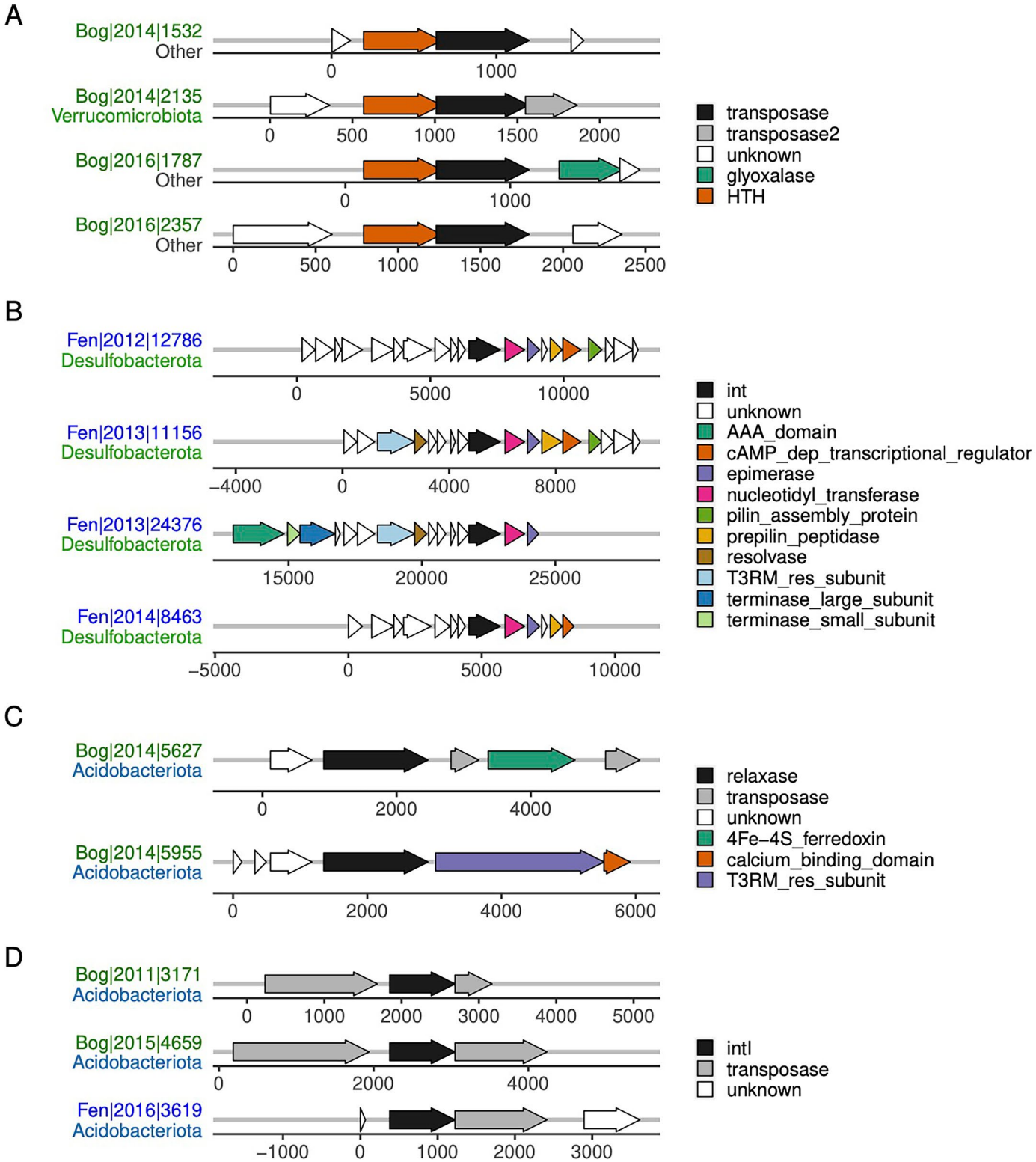
Extended Data Fig. 6 | Technical assessment for why short-read MAGs miss some MGE recombinases. In all panels, 'covered' represents MGE recombinases assembled and binned both in short-read and hybrid MAGs, 'missed assembly' represents those in hybrid MAGs missing in short-read MAGs and short-read contigs, 'missed binning' represents those in hybrid MAGs missing in short-read MAGs, but captured in short-read contigs (statistical significance per Fig.1). **A**. Short reads mapped to the full hybrid MAGs dataset of MGE recombinases evaluated by sequencing depth (left panel, vertical coverage) and breadth (right panel, horizontal coverage). The red dashed line at y = 10 (left panel) approximates the minimal depth coverage needed for genome assembly considering the normal distribution of per bp position of depth coverages. Wilcoxon rank-sum test of statistical significance; two-sided; P value = 0.0053. **B**. Short-read-contig-based depth (left panel) and breadth (right panel) coverage per MGE recombinase to assess whether MGE recombinases were missed due to assembly or binning. Wilcoxon rank-sum test of statistical significance; two-sided; specific significant p-values from left to right: <2.22e-16, and 1.1e-05, and <2.22e-16. **C**. To evaluate whether highly variable MGE-recombinase-caused genomic neighborhood context could prevent assembly, we quantified such variation as genomic context number via a clustering approach (see Methods). Wilcoxon rank-sum test of statistical significance; two-sided; P value < 2.22e-16. **D**. Nucleotide variation (pi, see Methods) across three MGE recombinase 'capture scenarios'. Wilcoxon rank-sum test of statistical significance; two-sided; P value = 0.00099. **E**. Logic workflow to establish the fraction of short- (left panel) or hybrid- (right panel) MAG-only recombinases that are mis-binned in other MAGs. Each step is a yes or no answered question, and numbers represent MGE recombinase counts in our dataset. For boxplots in panels **A**, **B**, and **D**, the number of observations included in the plots is as follows: covered=236, missed assembly=606, missed binning=95. For panel **C**, the number of observations is covered=102, missed assembly=321, missed binning=45. For all boxplots, the lower and upper hinges correspond to the first and third quartiles, and the upper and lower whiskers extend from the nearest hinge to the largest (or smallest) value no further than 1.5 times the interquartile range from the hinge. Filled circles represent observations beyond this threshold of 1.5 times the interquartile range below the first quartile or above the third quartile.

https://doi.org/10.1038/s41564-026-02391-7



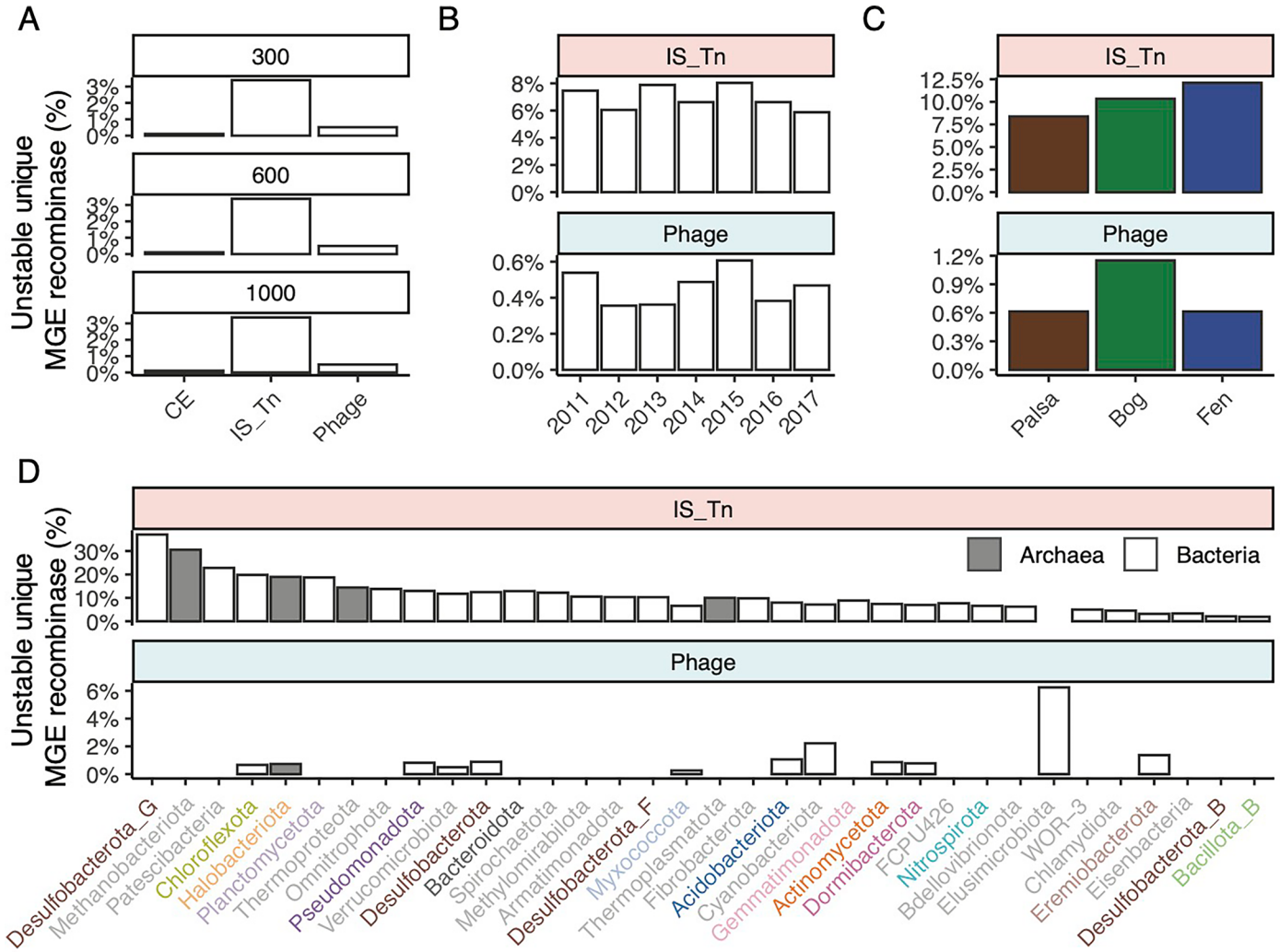
Extended Data Fig. 7 | MGE recombinase activity accessed by metatranscriptomics. This figure complements Fig. 5. Panels **A**–**C** all show MGE recombinase metatranscriptomic activity based on active OTU (clustered at 100% identity in protein sequence) percentage. Each OTU with ≥1 active MGE recombinase is considered as active. **A**. MGE recombinase metatranscriptomic activity within different years from 2015 to 2017; **B**. within each habitat; **C**. within host phyla (only those with ≥10 OTUs are shown; the X axis is ordered by the average of active OTU percentage across MGE recombinase types).

https://doi.org/10.1038/s41564-026-02391-7



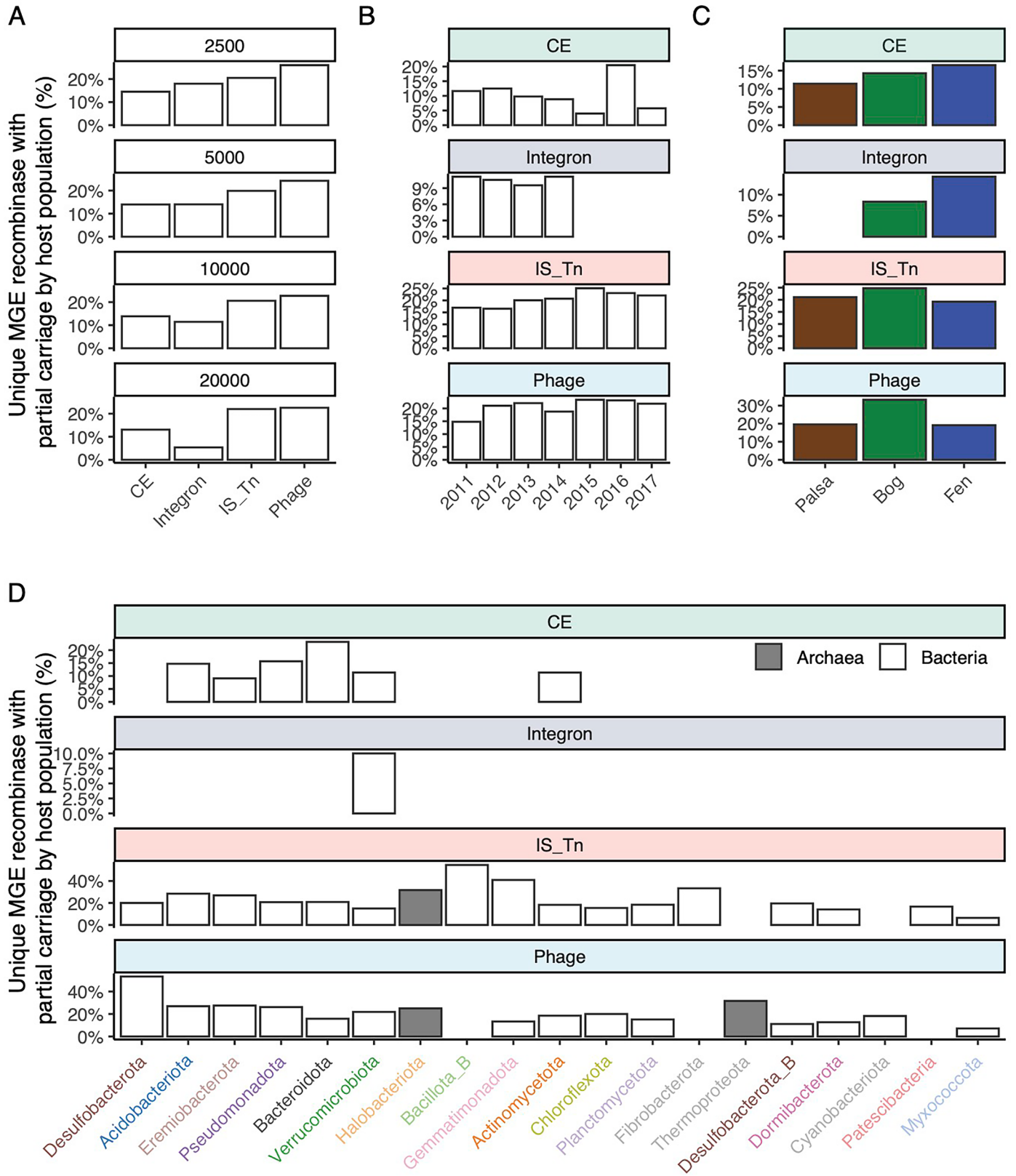
Extended Data Fig. 8 | Example of unstable genomic neighborhoods of MGE recombinases. The labels at the left of genome maps give habitat, year, and contig length (separated by '|'), followed by host phyla. **A.** IS_Tn type (transposase); **B.** Phage type (int, short for integrase); **C.** CE type (relaxase), and **D.** Integron type (intI) recombinase.

https://doi.org/10.1038/s41564-026-02391-7



Extended Data Fig. 9 | MGE recombinase genomic neighborhood stability. To measure MGE recombinase genomic neighborhood stability, MGE recombinases were clustered at 100% AAI, and neighborhoods of 300 bp up/down-stream of recombinase were extracted. If different neighborhoods (<1% ANI) are detected within a cluster, the recombinase cluster is considered unstable. The proportion of unstable MGE recombinase OTUs were summarized **A**. within each MGE type across neighborhood length cut-offs ranging from 300 to 1000 bp; **B**. within each year; **C**. within each habitat; and **D**. within host phyla (only those with ≥10 OTUs are shown; the X axis is ordered by the average across MGE recombinase types). CE and integron recombinases are excluded here due to lack of data.

https://doi.org/10.1038/s41564-026-02391-7



Extended Data Fig. 10 | MGE recombinase partial carriage by microbial populations. In contigs with ≥20 kb upstream and downstream of an MGE recombinase, partial carriage was assessed as a significant drop in coverage from the surrounding contig to within the MGE recombinase gene. MGE recombinase OTUs (100% AAI) with ≥1 partially carried MGE recombinase were considered partially carried OTUs. The percentage of partially carried MGE recombinase OTUs were measured **A.** within each MGE type across neighborhood length cut-offs ranging from 2500 to 20,000 bp; **B.** within each year; **C.** within each habitat; and **D.** within host phyla (only those with ≥10 OTUs are shown; the X axis is ordered by the average across MGE recombinase types).

nature portfolio

Corresponding author(s): Matthew B. Sullivan, Simon Roux, Sarah C. Bagby

Last updated by author(s): Jun 8, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used to collect data.
Data analysis	Trimmomatic (v.0.36), SPAdes (v.3.12) with --meta option, GroopM2, and MaxBin2 (Cronin.v2), BBTools, MEGAHIT (v1.1.3), CheckM2 (v1.0.2), Galah (commit e9e59a8 and v0.4.0), CAT (minimal contig length 3 kb), GTDB-tk), prodigal (v2.6.3) using the "-p meta" option, hmmsearch (v3.3.2) using the 68 recombinase subfamily HMMs (table S2) and the "--cut-ga" option, eggnog-mapper (v2.1.9, default settings), CD-HIT (v4.8.1) with "-c 1 -g 0 -G 1 -l 30 -d 0 -M 0" and "-c 1 -g 0 -G 1 -l 30 -d 0 -M 0" integron-finder (v2.0.2) with default parameters, MacSyFinder (v2.0rc7) with "--models CONJScan/Plasmids all --db-type gembase" and "--models CONJScan/Chromosome all --db-type gembase", minimap2 (v.2.22, default settings), MMseqs2 (v4.8.1) at 95% identity, VirSorter2 (v2.2.3) following VS2 SOP step 1 (v3.0), checkV (default setting, version 0.7.0), DRAMv/DRAM (v1.2.3, --min_contig_size 1500), DIAMOND (v2.0.8.146, blastp bit-score ≥50) metaSPAdes v3.15.4 using Aviary v0.5.3, metaFlye v2.9-b1768, Racon v1.4.3, Pilon v1.24, MetaBAT v.2.1.5, MetaBAT2 v2.1.5, Unicycler v0.4.8, Bin Chicken v0.4.2, SingleM metapackage S3.0.5, CoverM v0.6.1, minimap2 v2.18, VAMB v3.0.2, SemiBin v1.3.1, Rosella v0.4.2, CONCOCT v1.1.0 and MaxBin2 v2.2.7 pysamstats (v1.1.2) with "--type=variation", MMseqs map (v13.45111, with default settings) BLAST+ (v2.8.1+) with "-task blastn" and bitscore ≥50, transcriptM v0.3.1, dirseq v0.4.3, pydirseq v0.1.0-alpha, samtools depth (v1.17) with "-a" option

R v4.2.1 (108), with ggpubr package v0.6.0 and other packages and versions given in https://github.com/emerge-bii/mobilome-submit-202407

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

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/15102623). The code produced in this study can be found in the project's GitHub repository (https://github.com/emerge-bii/mobilome-submit-202407).

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	No human participants, data, or biological material were involved in this study.
Reporting on race, ethnicity, or other socially relevant groupings	No human participants, data, or biological material were involved in this study.
Population characteristics	No human participants, data, or biological material were involved in this study.
Recruitment	No human participants, data, or biological material were involved in this study.
Ethics oversight	No human participants, data, or biological material were involved in this study.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Our study used established hidden Markov models (32) to identify mobile genetic element (MGE) recombinases in metagenomic data collected from depth-resolved soil samples from three distinct thaw-stage habitats at Stordalen Mire: palsa (overlying intact permafrost; n = 188 core sections), bog (partially thawed permafrost; n = 200), and fen (fully thawed, n = 227). We assessed MGE recovery from short- vs. long-read sequence data; estimated MGE occurrence rates across host phyla, MGE types, and habitats; assessed gene functions affected by MGEs; and used metatranscriptomic (n = 50 samples) and metagenomic data to investigate rates of MGE movement. (32) 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).
Research sample	Triplicate peat cores (11-cm diameter) were collected from palsa, bog, and fen habitats from Stordalen Mire in peak growing season annually 2010-2017 and 2019. Cores were sectioned by depth and processed as described previously (28). All samples were collected under appropriate permits and with knowledge and in some cases the assistance of the staff of the Åbisko Scientific Research Station (ANS).

(28) 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).

Sampling strategy	No a priori sample size calculations were performed before sampling. Sample sizes balance the need for replication with the need to minimize disruption to a long-term ecological study site.
Data collection	Samples were collected and environmental metadata measured by the IsoGenie/EMERGE field teams 2010-2019. Details of field team members and other metadata are publicly available in the EMERGE database, https://emerge-db.asc.ohio-state.edu. Nucleic acids were extracted as described previously (28). Samples were sequenced by Illumina HiSeq (2000 or 2500), NovaSeq 6000, or NextSeq 500 for short-read metagenomes; PacBio Sequel IIe for long-read metagenomes; or NovaSeq 6000 for metatranscriptomes. (28) 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).
Timing and spatial scale	Samples were collected in peak growing seasons of 2010-2017 and 2019. Details of sampling dates and times, GPS locations, active layer depth, and environmental metadata are publicly available in the EMERGE database, https://emerge-db.asc.ohio-state.edu.
Data exclusions	As described in the manuscript, conservative curation criteria were applied to exclude (1) sequence data that fell below quality or length thresholds and (2) annotation results that fell below quality thresholds.
Reproducibility	The study was observational, using sequence data from annual sampling over 7 years.
Randomization	Samples were grouped based on the habitat in which they were collected, as defined by the plant community at the surface of the core. Core images are available in the EMERGE database, https://emerge-db.asc.ohio-state.edu/.
Blinding	Blinding was not relevant to this study.

Did the study involve field work? ☒ Yes ☐ No

Field work, collection and transport

Field conditions	The study incorporates data from 408 bulk soil samples collected over 9 field seasons. Sampling metadata are publicly available in the EMERGE database, https://emerge-db.asc.ohio-state.edu/.
Location	Stordalen Mire (68.35°N, 19.05°E), southeast of Abisko Scientific Research Station (ANS), is a peat plateau (altitude 351 m) in the discontinuous permafrost zone of northern Sweden.
Access & import/export	All sampling was in accordance with Abisko Scientific Research Station (ANS) policy. Transport of samples to our team's laboratories in the United States and Australia was in accordance with applicable national regulations and permitting processes.
Disturbance	Sampling sites were accessed via existing boardwalk to minimize disruption. Any excess material was returned promptly to core holes.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	No seed stocks were collected in this study.
Novel plant genotypes	No novel plant genotypes were generated in this study.
Authentication	No seed stocks or novel plant genotypes were involved in this study.