AMERICAN SOCIETY FOR MICROBIOLOGY | mBio®

RESEARCH ARTICLE

Cargo Genes of Tn*7*-Like Transposons Comprise an Enormous Diversity of Defense Systems, Mobile Genetic Elements, and Antibiotic Resistance Genes

Sean Benler,[a] Guilhem Faure,[b] Han Altae-Tran,[b] Sergey Shmakov,[a] Feng Zhang,[b,c,d,e,f] Eugene Koonin[a]

[a]National Center for Biotechnology Information, National Library of Medicine, Bethesda, Maryland, USA
[b]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA
[c]HHMI, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[d]McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[e]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[f]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

ABSTRACT Transposition is a major mechanism of horizontal gene mobility in prokaryotes. However, exploration of the genes mobilized by transposons (cargo) is hampered by the difficulty in delineating integrated transposons from their surrounding genetic context. Here, we present a computational approach that allowed us to identify the boundaries of 6,549 Tn*7*-like transposons. We found that 96% of these transposons carry at least one cargo gene. Delineation of distinct communities in a gene-sharing network demonstrates how transposons function as a conduit of genes between phylogenetically distant hosts. Comparative analysis of the cargo genes reveals significant enrichment of mobile genetic elements (MGEs) nested within Tn*7*-like transposons, such as insertion sequences and toxin-antitoxin modules, and of genes involved in recombination, anti-MGE defense, and antibiotic resistance. More unexpectedly, cargo also includes genes encoding central carbon metabolism enzymes. Twenty-two Tn*7*-like transposons carry both an anti-MGE defense system and antibiotic resistance genes, illustrating how bacteria can overcome these combined pressures upon acquisition of a single transposon. This work substantially expands the distribution of Tn*7*-like transposons, defines their evolutionary relationships, and provides a large-scale functional classification of prokaryotic genes mobilized by transposition.

IMPORTANCE Transposons are major vehicles of horizontal gene transfer that, in addition to genes directly involved in transposition, carry cargo genes. However, characterization of these genes is hampered by the difficulty of identification of transposon boundaries. We developed a computational approach for detecting transposon ends and applied it to perform a comprehensive census of the cargo genes of Tn*7*-like transposons, a large class of bacterial mobile genetic elements (MGE), many of which employ a unique, CRISPR-mediated mechanism of site-specific transposition. The cargo genes encompass a striking diversity of MGE, defense, and antibiotic resistance systems. Unexpectedly, we also identified cargo genes encoding metabolic enzymes. Thus, Tn*7*-like transposons mobilize a vast repertoire of genes that can have multiple effects on the host bacteria.

KEYWORDS Tn*7*, cargo genes, transposases, transposon ends, transposons

Horizontal gene transfer (HGT) between prokaryotic genomes is one of the principal forces shaping prokaryotic genome evolution (1–6). The major routes of HGT include transformation, transduction, conjugation and transposition. Although the contribution of each route to the total number of horizontally transferred genes in a

Editor Stephen J. Giovannoni, Oregon State University

Address correspondence to Eugene Koonin, koonin@ncbi.nlm.nih.gov.

The authors declare no conflict of interest.

This article is a direct contribution from Eugene V. Koonin, a Fellow of the American Academy of Microbiology, who arranged for and secured reviews by Nancy Craig, Johns Hopkins University School of Medicine, and Irina Arkhipova, Marine Biological Laboratory.

Received 21 October 2021
Accepted 4 November 2021
Published 7 December 2021

[This article was published on 7 December 2021 with a byline that misspelled the surname of Feng Zhang. The byline was updated in the current version, posted on 11 February 2022.]

given genome is an outstanding question (7–9), examining the genes mobilized via each pathway offers the opportunity to identify trends universal to all routes. Recent large-scale analyses of horizontally transferred genes have either not considered the molecular pathway of transfer (10–12) or focused on a particular class of mobile genetic elements (MGEs), such as plasmids (13, 14), integrative and conjugative elements (15), or viruses (16–18). However, to the best of our knowledge, the diversity of genes mobilized via transposition has not been comprehensively characterized.

In prokaryotes, transposons that mobilize using a DDE/RNase H family transposase (e.g., insertion sequences) are the most abundant class (19). At a minimum, an autonomous DNA transposon encodes a transposase and conserved transposase binding sites at the 5′ and 3′ ends of the element, where the transposase executes strand cleavage and subsequent transfer of the element from the donor to the acceptor site (20). Some DDE/RNase H transposon families also require genes involved in regulatory and target site selection steps (21), such as the model transposon Tn*7* from *Escherichia coli*. Detailed biochemical dissection of Tn*7* identified five "core" genes that collectively execute transposon excision, target site selection, and integration (22, 23). The Tn*7* transposase TnsB recognizes arrays of binding sites at the left and right ends of the element and mediates 3′ DNA strand cleavage (23, 24). A second transposase, TnsA, mediates cleavage of the 5′ end that results in the complete excision of the element from the donor site (25). Transposition is orchestrated by TnsC, an AAA-ATPase that interposes between the TnsAB proteins bound to the ends of the transposon and the target site-selecting protein (26, 27). The target site is recognized either by TnsD, a site-specific DNA-binding protein, or TnsE, a DNA-binding protein that directs integration into replicating DNA (27, 28).

Recently, it was shown that some Tn*7*-like transposons recognize their target sites via a CRISPR spacer-guided mechanism (29–33). Such CRISPR associated transposons (CASTs) encode either a subtype I-F or subtype I-B Cascade complex or an inactivated type V-K effector (34, 35). Another distinguishing feature of CASTs is an "atypical" repeat and spacer delocalized from the CRISPR array, which form the guide RNAs directing the transposon to the chromosomal target site (29–32). The site selectivity and regulation of transposition by Tn*7* and by the CASTs stand in stark contrast to the nearly random insertion by other DDE-family transposons (19).

In addition to the core genes of Tn*7* and CASTs, other "cargo" genes that are not involved in transposition are often present within the boundaries of transposons (23, 36). Along with the core genes, the cargo genes are mobilized from a donor site to a recipient site, making a substantial contribution to HGT (37, 38). Examination of the cargo carried by Tn*7* and about 50 other Tn*7*-like transposons identified integrons with antibiotic resistance gene cassettes, heavy metal resistance genes, iron-sequestering siderophores, nonribosomal peptide synthases, restriction-modification enzymes, and many other genes of unknown function (36, 38–40). Therefore, a larger-scale analysis of the cargo carried by Tn*7*-like transposons has the potential to illuminate consistent trends in transposon-mediated HGT.

A challenge for the study of integrated MGEs, including Tn*7*-like transposons, is accurate delineation of the 5′ and 3′ boundaries of the element. Several bioinformatic tools delimit integrated MGEs by identifying a local enrichment of MGE-associated gene annotations (e.g., integrases) in a given locus (14, 41–45). However, because the diversity of genes carried by transposons is unknown, this approach is circular and thus of limited utility. Other tools delimit MGE boundaries by aligning sequence reads from the query genome against a reference (46, 47), but selecting a reference genome is nontrivial. Thus, to characterize Tn*7*-like transposons on a large scale, it is highly desirable to develop an approach that is agnostic to gene functions and does not require a closely related reference genome or pangenome.

Here, we present a comprehensive survey of Tn*7*-like transposons in prokaryotic genomes and whole-community metagenomes. The conserved sequences of the transposase binding sites and their unique architecture are shown to carry a signal that is

sufficient to delineate the 5′ and 3′ boundaries of the transposons, unmasking the diverse repertoire of genes mobilized by these elements. The comprehensive dissection of Tn*7*-like transposons enabled by this approach provides insight into prokaryotic HGT by expanding the known phyletic range of Tn*7*-like transposons, assessing the preferred routes and phylogenetic barriers to transposition, and characterizing the diversity of the mobilized genes.

RESULTS

Tn*7*-like transposons are present across diverse bacterial phyla. To investigate the distribution of Tn*7*-like transposons in prokaryotic genomes, hidden Markov models (HMMs) for the core genes of the *E. coli* Tn*7* (*tnsABCDE*) and known Tn*7*-like transposons were used to identify homologs in the database of prokaryotic genomes. Loci were considered candidate Tn*7*-like transposons if they encompassed adjacent open reading frames (ORFs) with significant sequence similarity to the core transposase TnsB and at least one other Tn*7* HMM. The TnsB transposase sequences were employed for phylogenetic reconstruction, given that TnsB is required for transposition, whereas the other subunit of the heteromeric transposase, TnsA, is dispensable (48). Using an iterative procedure to construct the alignments (49), a single TnsB protein sequence alignment was employed to construct a comprehensive phylogenetic tree of Tn*7*-like transposons (Fig. 1A).

The tree partitioned into four well-supported clades of transposons (bootstrap support $>$ 80), and the leaves from each clade were assigned a taxonomic host phylum using an 80% consensus rule. The first clade includes Tn*7*, an experimentally characterized subtype I-B CAST (29), and other related transposons in the phyla *Proteobacteria*, *Firmicutes* and *Chloroflexi*. The second clade is dominated by *Actinobacteria* transposons, followed by transposons in the phyla *Proteobacteria*, *Firmicutes* and *Deinococcus-Thermus*, none of which have been experimentally characterized. The third clade includes the *Proteobacteria* transposons Tn*5053* (50) and Tn*6022* (51), cyanobacterial subtype V-K CASTs (30), and transposons integrated into the genomes of *Deinococcus-Thermus*, *Bacteroidetes*, and *Firmicutes* hosts. The fourth clade represents transposons that are almost exclusively integrated into *Proteobacteria* hosts, including subtype I-F CASTs (32), with only two branches containing *Cyanobacteria* and *Firmicutes* transposons. Across all four clades, a consensus host could not be assigned to individual leaves in several cases, underscoring an even greater degree of transposon mobility between phyla (Fig. 1A). Additional hosts that are less common and were not reported in prior surveys for Tn*7*-like transposons (36, 38) include *Nitrospirae* (8), *Verrucomicrobia* (5), *Acidobacteria* (4), *Fibrobacteres* (4), *Aquificae* (3), *Spirochaetes* (2), *Deferribacteres* (2) and *Planctomycetes* (1) (the numbers in the parentheses refer to the number of host species in each respective taxon; Table S1). No Tn*7*-like transposase homologs were identified in the currently available genomes of any archaea or viruses. Overall, the phylogenetic tree illustrates multiple switches between host phyla in the evolutionary history of Tn*7*-like transposons.

Arrays of transposase binding sites reveal the boundaries of Tn*7*-like transposons. The transposons Tn*7*, Tn*5090*, Tn*5053*, and Tn*552* possess arrays of inverted repeats located subterminal to the 5′ and 3′ boundaries of the transposon (52), also known as the transposon's left end (LE) and right end (RE). The approximately 20-bp inverted repeats include the transposase binding sites and are typically located in intergenic regions (24, 35). We leveraged the presence of multiple inverted repeats to identify the transposon boundaries and discriminate intact transposons from partial ones. All intergenic sequences up to 125 kb from either side of *tnsB* were scanned for the presence of inverted repeat arrays resembling the LE and RE of Tn*7* (see Materials and Methods). Under the premise that phylogenetically close transposases would utilize conserved binding site sequences, any discovered arrays of inverted repeats flanking a given transposase were compared to those found for another, closely related homolog, if available. If the inverted repeat sequences were conserved among $\geq$50% of close transposase homologs and satisfied additional criteria (see Materials and Methods), the arrays were predicted to be the LEs/REs of the respective transposons. The

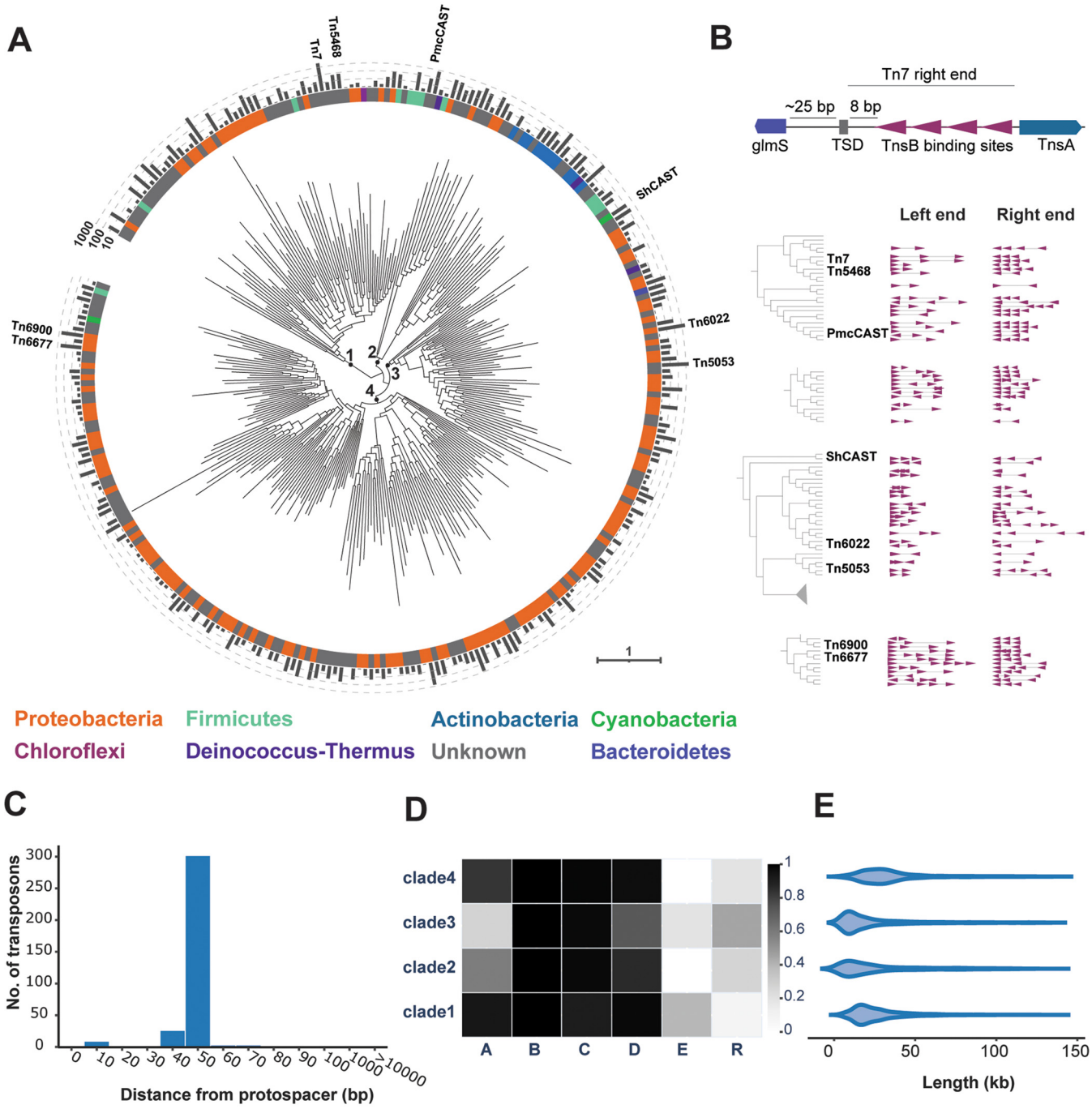

FIG 1 Tn*7*-like transposons are widespread mobile genetic elements across the diversity of bacteria. (A) Phylogenetic tree of the DDE-family TnsB transposase from Tn*7* and related transposons (n = 299 leaves). Deep branches are marked for clarity, and the number of transposases represented by each leaf is plotted on the outer ring. The taxonomic phylum for each leaf is plotted on the inner ring and assigned using the consensus of ≥80% of the transposase sequences; otherwise, the leaf phylum is marked "unknown." (B) The predicted boundaries from selected transposons are displayed for a TnsB subtree, highlighting the fact that related transposases possess similar TnsB binding site architectures. Each binding site is depicted as an arrow, where sites downstream of the transposase were arbitrarily oriented to the positive strand (left end, arrows pointing right) and sites upstream of the transposase are on the negative strand (right end, arrows pointing left). (C) Histogram of distances from the predicted boundary of CASTs to the protospacer-targeted attachment site. (D) Heat map of the presence/absence of Tn*7* core proteins. (E) Length distribution of intact, dereplicated transposons (n = 6,549).

computationally predicted TnsB binding sites were consistent with biochemical data on Tn*7* transposition, where three TnsB binding sites were present at the left end and four at the right end (24) (Fig. 1B). Similar arrays were identified at the (putative) left and right ends of other transposons, despite the phylogenetic distance from Tn*7*

(Fig. 1A). Thus, arrays of inverted repeats, which are candidate TnsB transposase binding sites, are a general feature that unites Tn*7*-like transposons.

To estimate the LE/RE detection accuracy, chromosomes bearing Tn*7* and 17 other Tn*7*-like transposons were analyzed because the precise boundaries of the elements were experimentally or manually identified previously (38). In total, TnsB transposase binding sites could be predicted for 16 of the 18 transposons (Table S1B). The outermost predicted TnsB binding sites were offset from the reported transposon boundary by a median of 12 bp (Table S1B), consistent with the experimentally characterized end architecture of Tn*7* (53) (Fig. 1B). The predicted TnsB binding sites in the left ends of two transposons were more than 1 kb from the reported left boundary (Table S1B). This discrepancy may result from the transposition of one element internal or adjacent to an existing element, yielding multiple left and right ends (37). Considering these two cases and one other as incorrect predictions, the LE/RE detection method correctly identified the subterminal TnsB binding sites present in 13/18 (72%) of the transposons.

To estimate the LE/RE detection accuracy on a larger set of transposons, the predicted boundaries of 523 CASTs (31, 32) were analyzed. It is possible to identify the target site of the CASTs, although transposition has not been demonstrated experimentally, by extracting spacers from the CRISPR array and searching the sequence neighborhood of the transposon for a matching protospacer. CASTs and other Tn*7*-like transposons integrate at a fixed distance downstream from their target sites (27, 54) (Fig. 1B), which is typically ~50 bp in CASTs and likely corresponds to the footprint of the transposition complex (30, 33, 55, 56). In total, the LEs/REs were predicted for 352/523 (67%) of the previously reported CASTs. Almost all predicted integration sites are 50 to 60 bp from the protospacer-containing target site, with only two transposons predicted to be located more than 1 kb away from the attachment site (Fig. 1C). The 171 CASTs that lack both predicted ends could be incompletely sequenced (e.g., the transposon spans multiple contigs), have degenerate TnsB binding sites or LEs/REs that are otherwise undetectable by this approach, and were not investigated further. The results of the CAST analysis indicate that the predicted TnsB binding sites represent the LEs/REs of the transposons with substantial accuracy. Therefore, the genes embedded between the predicted TnsB binding sites in the LE/RE are not simply adjacent to *tnsB* but, rather, are mobilized by Tn*7*-like transposons.

Core gene content of Tn*7*-like transposons. Using the predicted TnsB binding sites to determine the outermost boundaries of the transposons, the nucleotide sequences were extracted from the respective contigs and dereplicated (99% average nucleotide identity across 95% alignment length), resulting in a final set of 6,549 putatively complete Tn*7*-like transposons. The final set contains transposons belonging to all four TnsB clades, with some branches represented by >100 delineated transposons (Table S1C). Besides the universally conserved *tnsB* transposase, which was a minimal requirement of the transposon discovery pipeline, 96% of the identified transposons harbored a *tnsC*-family ATPase (Fig. 1D). The next most conserved gene is the target-site selector *tnsD*/*tniQ* (referred to here as *tnsD*), which was present in 73 to 97% of the transposons in each clade. The PDEXK-family transposase *tnsA* is nearly ubiquitous in clade 1 (92%) but relatively uncommon in clade 3 (28%); this pattern is inverted for the tyrosine and serine superfamily resolvases (*tnsR*) (for a more detailed discussion, see "The replicative transposition pathway of Tn7-like transposons is predicted by the loss of a cut-and-paste transposase and gain of a resolvase," below). Only the transposons in clades 1 and 3 contained a homolog of *tnsE*, which promotes the integration of Tn*7* into replication forks, including those on plasmids (28). A total of 1,298 transposons (20%) were found to be integrated into plasmids, including some lacking an identifiable *tnsE* homolog (Table S1C). This discrepancy could be attributable to multiple factors, including incorrect assignment of a chromosomal contig as a plasmid, the inability to identify weakly conserved *tnsE* homologs, the presence of a *trans*-acting TnsE, or involvement of other, as-yet-unknown plasmid target site-selecting genes. The median length of the transposons ranged from 14 to 29 kb (Fig. 1E), but much larger transposons were detected as well, the longest reaching

140 kb (Table S1C). The cargo genes within the boundaries of the transposons were functionally profiled, as described below.

Tn*7*-like transposons cross horizontal gene transfer barriers. To complement the phylogenetic analysis and identify potential phylogenetic boundaries of horizontal transfer, a similarity matrix was constructed between all pairs of transposons using the number of shared protein clusters. Protein clusters were constructed greedily from ORFs, with the clustering threshold at 80% amino acid sequence identity over 75% of the length of the smaller ORF. Next, the similarity matrix between all transposons was analyzed to identify communities, that is, sets of transposons that shared genes significantly more frequently with each other than with transposons in the rest of the network. Using a hierarchical community detection approach (57), several large communities were identified at lower "resolutions" and then iteratively partitioned into smaller subcommunities at higher resolutions, until a predefined resolution limit was reached. The communities that are stable across resolutions were considered significant, whereas transient communities were discarded, leveraging the network concept known as "persistent homology" (57). To select the predefined resolution limit, which constrains the size of the smallest subcommunities, a range of upper limits was explored, and the limit that resulted in the highest mean community persistence (that is, quality) was chosen to identify the final communities of transposons (Fig. S1). In the resulting hierarchical depiction of the underlying community structure, the communities decrease in size and homogenize at the taxonomic level of phylum with descending levels of the hierarchy (Fig. 2A). The phylum-level taxonomic homogeneity of the communities indicates that recent HGT via transposition between bacterial phyla is limited, in general. Nevertheless, a community of 1,221 promiscuous transposons that overcome this barrier was delineated. In this community, 83% of the transposons originated from *Proteobacteria*, and it could be divided into 15 subcommunities (Fig. 2B). Each subcommunity was labeled with the consensus last common ancestor, such that $\geq$80% of transposons in a subcommunity integrate into hosts that belong to the same taxonomic lineage. Interphylum exchange was found to occur between the *Micrococcales*-dominated subcommunity (*Actinobacteria*) and multiple classes of *Proteobacteria*, with an observable contribution of plasmid-mediated transfer (Fig. 2C). Altogether, for the 15 subcommunities, $\geq$80% of the transposons integrated into hosts from the same phylum but multiple classes (3 communities), the same order but multiple families (3 communities), the same family but multiple genera (5 communities), or the same genus but multiple species (2 communities); the remaining two subcommunities do not have a consensus phylogenetic level (Fig. 2B). We next tested if the communities that encompassed multiple phylogenetically distant hosts were due to a lack of resolution; that is, smaller subcommunities of closely related hosts were actually present but went undetected. Constructing more subcommunities of smaller size did not appreciably change the consensus phylogenetic level of the subcommunities at the lowest levels of the hierarchy (Fig. S1D), indicating that the analysis was not limited by resolution. Instead, the community composition apparently reflected promiscuous HGT between different genera, families, orders, and classes of bacteria via transposons, along with sporadic exchanges between phyla.

The transposon communities were further analyzed to determine if membership was driven by core or cargo genes. Pairs of transposons carrying identical cargo were common, including those from different clades. For example, two *Bacillus cereus* transposons from clades 1 and 2 showed minimal sequence similarity between their core proteins but carry nearly identical arsenate resistance operons and because of that belong to the same community (Fig. S2). Conversely, pairs of transposons in the same community could be found that mobilized entirely different cargo. For example, two transposons integrated in the same *Pseudomonas monteilii* genome shared no protein clusters, apart from the core proteins (Fig. S2), highlighting the idea that closely related transposons can transport unrelated cargo. These contrasting examples demonstrate that community membership is defined by both core and cargo genes; furthermore,

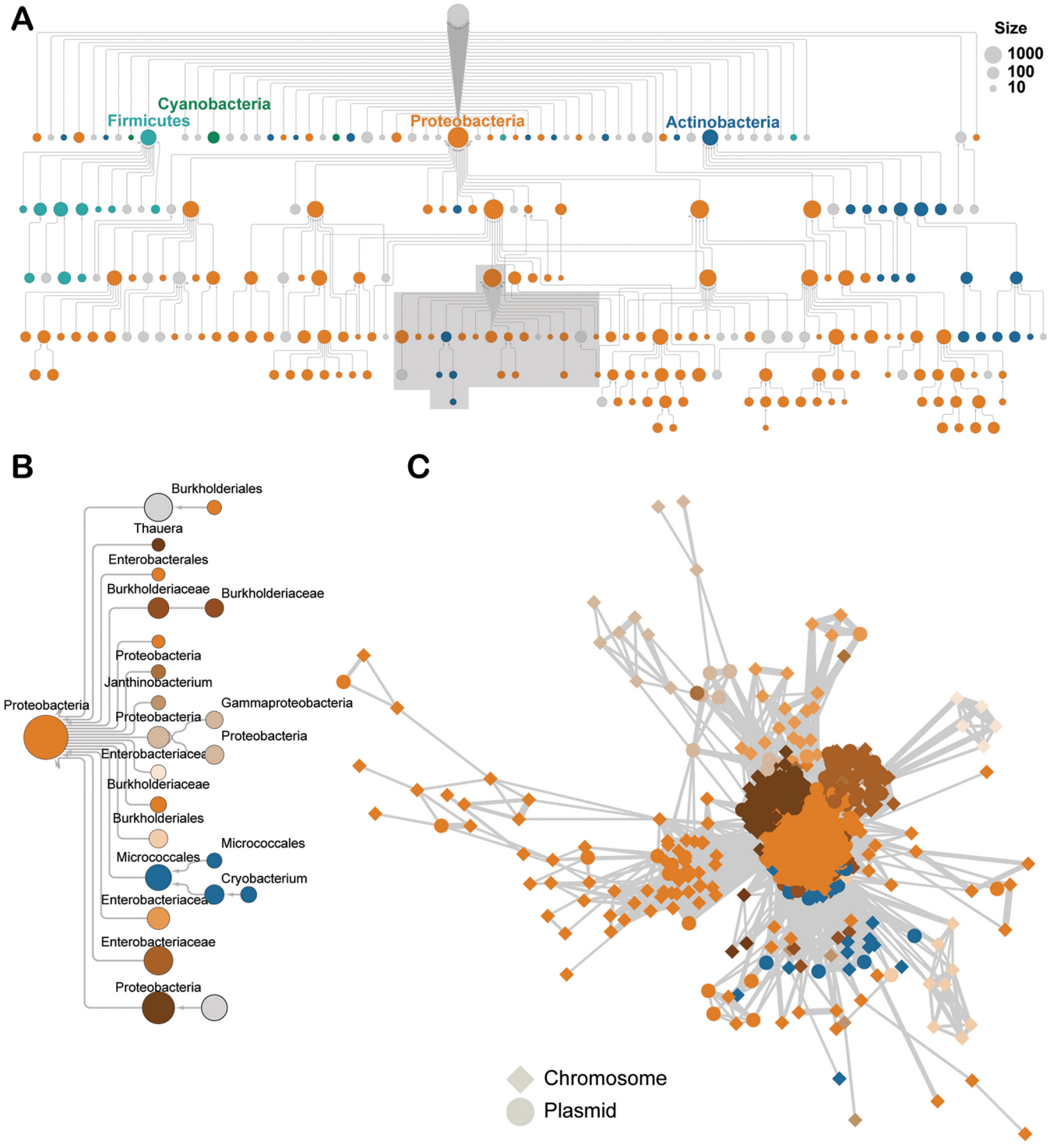

FIG 2 Tn*7* transposons disseminate genes across deep phylogenetic divides. (A) Hierarchical network of transposon communities (nodes), where each edge signifies the division of one community into another, smaller subcommunity. All communities are scaled by the total number of members and are colored if ≥80% of the members belong to the same phylum. (B) Expansion of the area shaded in gray, with subcommunities recolored and labeled with their lowest common ancestor (80% consensus), if present. (C) Network of 1,221 transposons (nodes) connected with edges that are weighted by the similarity between the respective two transposons. The shape of the node indicates if the transposon is inserted into the host's chromosome or a plasmid.

these findings emphasize that the cargo is decoupled from the transposition core machinery.

Functional repertoire of Tn*7* cargo includes other mobile genetic elements. To examine the functional repertoire of the cargo, the subset of the Tn*7*-like transposons

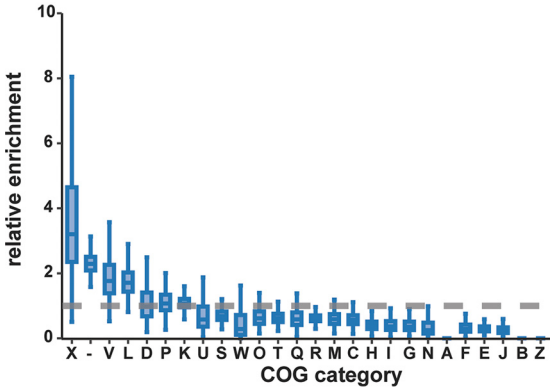

FIG 3 Tn*7*-like transposons mobilize a distinct repertoire of genes. A subset of transposons in completely sequenced bacteria (n = 739) was used to calculate the enrichment of individual COG categories in the transposons relative to randomly sampled genomic loci. The three enriched categories are "mobilome" (X), "defense" (V), and "replication, recombination, and repair" (L), as well as ORFs without significant similarity to a COG (indicated by a dash).

found to be integrated into completely sequenced bacterial genomes (n = 739) were compared against randomly selected loci of equal lengths. All ORFs were annotated against the Clusters of Orthologous Groups (COG) database (58) and binned into the COG functional classes to estimate the relative frequency of each functional class in both gene sets. Obviously, homologs of the Tn*7* core genes drive the enrichment of the mobilome COG class in transposons relative to random chromosomal loci (Fig. 3). However, other MGEs also contribute to the enrichment of the mobilome genes. The most common MGEs identified within Tn*7*-like transposons are insertion sequences (ISs), particularly those of the IS*3* family, that are present in ~30% of the transposons (Fig. S3A). In some cases, two IS elements of the same family flank one or several genes from both sides within a Tn*7*-like transposon. Such proximity between two ISs of the same family can enable both to mobilize in tandem as a composite, cargo-carrying transposon (19). Two examples of clinical importance are ISs that flank phosphoethanolamine transferases *mcr-9.1* or *mcr-3* (Fig. S3B) that reduce the affinity of lipid A to colistin, a "last resort" antibiotic (59). The *mcr-3*-carrying transposon is nested in a Tn*7*-like transposon borne on a plasmid in *E. coli*, highlighting the broad mobile potential of these genes. A third example of potential ecological relevance was found in *Nitrospira moscoviensis*, where two IS21-family transposons flank cytochrome P460 genes that are involved in nitrogen cycling (60), all within the boundaries of a Tn*7*-like transposon (Fig. S3C). Finally, resolvase-carrying transposons unrelated to Tn*7* were also identified as cargo (Fig. S3D). Thus, one of the major sources of the Tn*7*-like transposon cargo is other MGEs, some of which carry their own cargo.

Defense systems associated with Tn*7*-like transposons. The next most highly enriched set of genes, apart from genes of unknown function, comes from the defense COG class. A search of the cargo for genes involved in biological conflicts uncovered numerous, diverse innate immune systems, the most common and widespread being restriction-modification modules (Table 1). The next most frequent defense system includes the OLD (overcoming lysogenization defect)-family nuclease, which contains a TOPRIM (topoisomerase-primase) domain (61) and interferes with the replication of phage lambda (62). The OLD nuclease is typically associated with a UvrD-family helicase (Fig. S4), jointly constituting the Gabija system, which confers immunity against multiple phages (63). Another widely mobilized phage defense system is the cyclic oligonucleotide-based antiphage signaling system (CBASS) (64), which includes a nucleotidyltransferase that recognizes phage proteins and mediates programmed cell death through various effectors (65). Other, less abundant innate immune systems with diverse mechanisms were identified as well (63, 66–69) (Table 1). The transposons also mobilize single-component defense systems that confer protection against multiple

TABLE 1 Tn*7*-like transposons mobilize diverse immune systems

Immune mechanism	System	Component(s)[a]	No. of Tns[b]	Phylum(a)[c]
Innate	Restriction-modification	Restriction endonucleases (COG0610, COG4096, COG3440, COG1002, COG3421, COG1715, COG3587, COG4636, COG4804, COG1403, COG3183, COG4127, COG4748) methylases (COG0286), specificity subunits (COG0732, COG4268, COG1401)	947	Pro, Act, Fir, Cy, Chl, Spi, Bac
	Gabija	OLD nuclease (COG3593)	338	Pro, Fir, Act, Bac
	Sirtuin	Sirtuin2 (cl00195)	242	Pro, Act, Fir, Cy
	CBASS	Nucleotidyltransferase (cd05400)	216	Pro, Act, Fir
	Thoeris	ThsB (pfam08937)	61	Pro, Act, Fir
	Retron	RT (cd01646)	50	Pro, Act, Fir
	Hachiman	Hama (pfam08878)	43	Pro, Fir, Spi
	BREX	BrxC (NF033441)	40	Pro, Act, Fir, Cy, Nit, DT
	Wadjet	JetD (COG4924)	34	Pro, Act
	Kiwa	KwaB (pfam16162)	23	Pro, Act
	Nucleotide depletion	dGTPase (PRK05318), dCTPase (cd01286)	21	Pro, Fir
	PIWI	PIWI REase (pfam18154)	19	Pro, Act, Fir
	DND	DndE (pfam08870), DndB (pfam14072)	14	Pro, Bac, Fir, Cy
	Zorya	ZorA (pfam01618)	3	Pro
	Viperin	Viperin (TIGR04278)	1	Pro
Toxin-antitoxin	ImmA	ImmA (COG2856, pfam06114)	295	Pro, Act, Bac, Fir, Cy
	PIN	PIN (cl28905)	251	Pro, Act, Fir, Cy
	ParE	ParE (cl21503)	120	Pro, Act, Cy, DT
	RES	Res (cl02411)	95	Pro, Act, Fir
	RelA-SpoT	RelA (cd05403, COG1669, smart00954, cd05399, pfam14907, COG2357, pfam04607, pfam01909, cd07749)	70	Pro, Act, Bac, Fir, Cy
	HipA	HipA (cl28916)	51	Pro, Act, Cy
	AbiC	AbiC (pfam14355)	48	Pro, Fir
	HEPN	HEPN (cl00824)	37	Pro, Fir, Cy
	AbiF	AbiF (pfam07751, COG4823)	23	Pro, Spi
	AbiE	AbiEii (pfam08843, COG4849)	19	Pro, Act
	Fic/Doc	Doc (COG3654, TIGR02613)	7	Pro, Cy
	Barnase	BrnT (pfam04365, COG2929)	5	Pro, Cy
Adaptive	Minimal I-F	Cas6_I-F, Csy3_I-F, Csy2_I-F	304	Pro
	Minimal V-K	Cas12k	78	Cy
	Minimal I-B	Multiple	9	Cy
	Type III	Cas10	2	Pro
	Type I-B	Multiple	1	Cy

[a]Accession numbers in parentheses can be located at the NCBI Conserved Domain Database (CDD) (https://www.ncbi.nlm.nih.gov/cdd).
[b]The number of transposons mobilizing each immune system is tabulated based on the listed components.
[c]Pro, *Proteobacteria*; Act, *Actinobacteria*; Fir, *Firmicutes*; Cy, *Cyanobacteria*; Chl, *Chloroflexi*; Spi, *Spirochaetes*; Bac, *Bacteroides*; DT, *Deinococcus-Thermus*; Nit, *Nitrosomonas*.

double-stranded-DNA (dsDNA) phages, including enzymes of the sirtuin superfamily (69), viperins (70), "defensive" reverse transcriptases (69), and enzymes that deplete the pool of nucleotides available for phage reproduction (71). Although these previously described innate immune systems jointly represent only 2% of the cargo genes, they are carried by 26% of the transposons. It appears likely that unknown defense systems are lurking among the cargo genes with currently unknown functions.

Mobile genetic elements often encompass toxin-antitoxin (TA) gene pairs, which frequently colocalize with innate immune systems (72, 73) and can function as defense systems themselves, typically eliciting dormancy or programmed cell death (74–76). Tn*7*-like transposons mobilize a variety of toxin effectors, including PIN family ribonucleases, RelA/SpoT-like nucleotidyltransferases, ParE-family mRNA interferases, ADP-ribosyltransferases, and HipA-family kinases (Table 1). Several abortive infection (AI) systems first identified in lactococci were also identified (77), implicating transposons in horizontal transfer of these genes. In multiple phyla, the toxins are associated with a

cognate antitoxin and/or a defense system (Fig. S5). Thus, TAs and the functionally similar AI systems substantially augment the repertoire of innate immune systems carried by Tn*7*-like transposons.

To identify transposons that mobilize adaptive immune systems, CRISPR arrays were predicted and the cargo genes were scanned for significant sequence similarity to *cas* genes. In this survey, 410 transposons were shown to harbor at least one *cas* gene or a CRISPR array; of these, all but 14 belong to one of three subclades in the TnsB phylogenetic tree (Fig. S7). The location of CASTs in the TnsB tree demonstrates that Tn*7*-like transposons exapted CRISPR-Cas systems for target site selection on at least three independent occasions, in agreement with previous observations (34, 35). The LEs/REs of one experimentally characterized CAST (CAST I-B2) (29) could not be predicted automatically, raising the possibility that improvements in transposon delineation could reveal novel CASTs. Additional transposons scattered across the TnsB tree contain individual components of CRISPR-Cas systems or, less frequently, a complete system (Table 1; Fig. S6). Specifically, one *Geminocystis* sp. transposon is closely related to subtype I-B CASTs but carries *cas2* next to a CRISPR array and a *cas1-cas4* fusion that together make up the adaptation module, which is normally absent in CASTs (Fig. S6). Interestingly, this bacterium also hosts two subtype V-K CASTs which are integrated at different loci. Two other transposons encode a type III CRISPR-Cas system with the hallmark *cas10* effector (Fig. S6). Overall, Tn*7*-like transposons infrequently possess a complete, fully functional adaptive immune system but often carry a minimal suite of CRISPR-Cas genes that functions as the target site selection machinery.

The replicative transposition pathway of Tn*7*-like transposons is predicted by the loss of a cut-and-paste transposase and gain of a resolvase. The final COG class that is enriched in the Tn*7*-like transposon cargo is "replication, recombination, and repair" and, more specifically, genes encoding enzymes of the tyrosine and serine integrase/resolvase superfamilies. Enzymes from both superfamilies are encoded by diverse MGEs and catalyze site-specific rearrangements in DNA that are involved in transposon relocation from a donor to an acceptor site (20). For example, Tn*3* and Tn*21* transposition involves a cointegrate intermediate that is resolved into two separate molecules by a serine and a tyrosine superfamily resolvase, respectively (78). Because resolvases can be functionally involved in MGE mobility, they neither fit the strict definition of "cargo" nor can be clearly classified into the "replication, recombination, and repair" or "mobilome" COG functional category. Therefore, the genetic context and phyletic distribution of the serine and tyrosine superfamily enzymes were examined separately to better ascertain their functional roles.

In the collection of Tn*7*-like transposons analyzed here, serine and tyrosine resolvases are prevalent, but not universally conserved, in clades 2 and 3. Clade 3 includes Tn*5053* and other transposons that possess either a serine or a tyrosine resolvase, whereas clade 2 transposons are almost exclusively associated with tyrosine resolvases (Fig. S7). Both these clades contain branches of transposons that lack homologs of *tnsA*, which is a transposase essential for cut-and-paste transposition in Tn*7* (79) (Fig. 1; Fig. S7). The absence of *tnsA* combined with the presence of a serine or a tyrosine resolvase suggests that these transposons do not mobilize via the cut-and-paste mechanism. Instead, their transposition likely proceeds through a cointegrate intermediate that needs to be resolved prior to integration. Transposons that lack a cognate resolvase can still mobilize via a cointegrate intermediate, where resolution is achieved by resolvase in *trans* or by the host *recA*-mediated homologous recombination (48, 78, 80), perhaps explaining the lack of resolvases in some of these transposons. Together, these findings predict that replicative transposition is the principal pathway employed by distinct branches of transposons within these two clades, in which case the resolvases are not cargo but rather core components of the transposition machinery.

Numerous other transposons outside these two branches also harbor serine resolvases, but their dispersal across the TnsB tree and presence of *tnsA* homologs does not support a functional role during replicative transposition (Fig. S7). In some cases, the serine resolvase is encoded as part of a nested transposon (Fig. S3C). Here, the resolvase is a

core component of the Tn3-family transposon but appears to be cargo for the larger, Tn*7*-like transposon, although it cannot be ruled out that the resolvase functions during replicative transposition of both transposons. In other cases, the serine resolvase does not appear to be part of a nested transposon, suggesting that it is carried as cargo only by the Tn*7*-like transposon and might perform other functions, such as DNA inversion.

Cargo genes involved in antibiotic resistance, biosynthesis, and central carbon metabolism. To characterize the cargo genes that belong to COG functional classes that are not enriched or even are depleted (Fig. 3), the specific metabolic or functional pathways were tabulated for each individual COG. Few complete pathways were carried by any of the Tn*7*-like transposons, but 37% of the transposons harbor at least one gene implicated in various aspects of cellular physiology, ranging from amino acid catabolism to nucleotide and lipid biosynthesis (Fig. 4). The most common, albeit incomplete, pathway is folate biosynthesis (555 transposons, category H), which includes the genes *sul1* and *dfrA1*, encoding, respectively, dihydropteroate synthase and dihydrofolate reductase, which confer resistance to sulfonamide antibiotics (81). Similarly, the mercury resistance gene *merA* (COG1249, category E) is responsible for the apparent commonality of the glycine cleavage pathway. To explore the larger pool of antibiotic resistance genes, which belong to various COG categories, the cargo was searched against a database of genes with experimentally demonstrated roles in antibiotic and xenobiotic resistance (82). Genes with heavy metal detoxification activity, including mercury, were found to be abundant, as well as antibiotic resistance genes that confer protection against the aminoglycoside, sulfonamide, and beta-lactam classes of antibiotics (Fig. S8). Thus, Tn*7*-like transposons commonly mobilize genes of biological and clinical relevance.

To elucidate how antibiotic and stress-resistance genes are captured by transposons, the cargo was examined for the presence of integrons. Integrons are MGEs that site-specifically incorporate "cassettes" of DNA and constitute a notable component of the cargo of Tn*7* and other families of transposons because of their role in aggregating antibiotic resistance genes (83, 84). Altogether, 10% of the Tn*7*-like transposons harbor an intact or partial integron, collectively carrying 2,334 integron gene cassettes (i.e., ORFs with a 3′ *attC* site). The cassettes are dominated by *qacE* (334 cassettes) and *dfrA1* (157 cassettes), which confer resistance to ammonium-based disinfectants and trimethoprim, respectively. Integrons carrying homologs of genes that provide resistance to aminoglycosides, beta-lactams, phenicols, sulfonamides, macrolides, lincosamides, quinolones, and bleomycins were also captured by Tn*7*-like transposons, illuminating the role of integrons in expanding the diversity of the antibiotic and stress resistance cargo.

The transposons mobilizing genes involved in biosynthetic pathways and central carbon metabolism were analyzed next. The most common biosynthetic pathways in transposons are those for fatty acids and menaquinones, including COG1028 (short-chain alcohol dehydrogenase), COG1960 (acyl coenzyme A [acyl-CoA] dehydrogenase), COG0183 (acetyl-CoA acetyltransferase), and COG0596 (MenH-family esterase). These enzymes appear in a variety of genetic contexts, making their specific roles and end products difficult to infer, but generally are involved in the synthesis or degradation of "auxiliary" compounds. Additionally, multiple transposons carry genes involved in central metabolic processes, including glycolysis, gluconeogenesis, the pentose phosphate pathway, and the tricarboxylic acid cycle (Fig. 4). The genetic context of these central carbon metabolism genes includes other genes involved in sugar catabolism and transport (Fig. S9). Overall, although the components of core and auxiliary metabolic pathways are not common cargo compared to their representation in bacterial genomes, their presence broadly reveals the potential for Tn*7*-like transposons to mobilize genes that modulate cellular metabolism.

DISCUSSION

A search of prokaryotic genomes initiated with HMMs for the core Tn*7* proteins involved in transposition recovered thousands of predicted transposons in diverse bacterial phyla, underscoring the mobility of these elements. The phylogeny of TnsB demonstrates the relationship between Tn*7* and other well-characterized transposons,

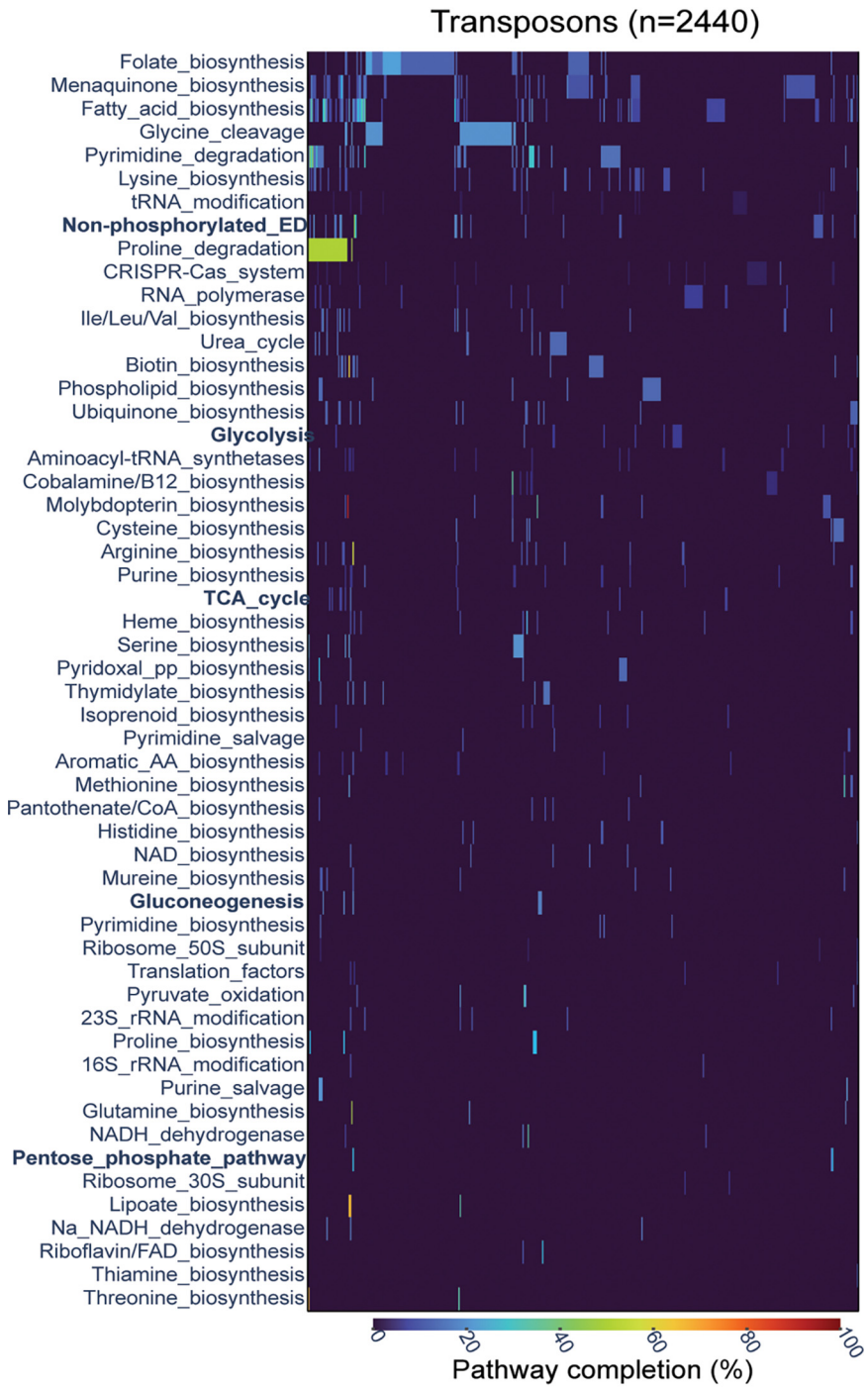

FIG 4 Metabolic pathways mobilized by Tn*7*-like transposons. The *y* axis lists the COG pathways, sorted by the pathways that are found in the most transposons to least. Each column corresponds to an individual transposon, with the color of the cell proportional to the completeness of the respective pathway. Central carbon metabolism pathways discussed in the text are in bold.

including Tn*5053* and the recently described CASTs. Besides the phylogenetic coherence of the transposase, all these transposons possess similar architectures of inverted repeats at their left and right ends that mark the boundaries of the transposon. Moreover, within these boundaries, most encode a transposition-regulating ATPase (*tnsC*) and a target site-selecting protein (*tnsD*) flanking the core transposase. These shared features are the hallmarks of a vast, distinct group of MGE informally dubbed Tn*7*-like transposons (38, 85).

Tn*7*-like transposons contain a plastic repertoire of genes, even among the genes that are directly involved in transposition. A prominent example of such plasticity is *tnsA*, where point mutations in the catalytic site abolish cut-and-paste transposition but do not impede replicative transposition (48, 86). Multiple branches of Tn*7*-like transposons lack *tnsA* altogether (Fig. S7), evincing the dispensability of this gene. Typically, the *tnsA*-lacking transposons gain a tyrosine- or serine-superfamily resolvase (*tnsR*) that most likely mediates replicative transposition. The target site-selecting gene *tnsE* is absent in most clades of transposons, although the actual spread of this gene might be underestimated due to the weak sequence conservation. Similarly, the CRISPR-Cas target site-selecting systems are relatively rare, but additional CASTs might be discovered with improvements in transposon delineation. The evolutionary gain and loss of *tnsAER* and the specialized CRISPR-Cas systems illustrate the modularity of the transposition machinery of the Tn*7*-like transposons.

The cargo genes carried by Tn*7*-like transposons exhibit an even greater plasticity. To survey the biological functions of the cargo, a comparison against genes located outside the transposons was undertaken. An unexpected outcome of this analysis was that the most highly enriched function, the mobilome, was not explained by the presence of core Tn*7* proteins alone but also by substantial contributions from other MGEs, specifically, IS elements. The enrichment of ISs suggests a stepwise process of gene capture by Tn*7*-like transposons, beginning with the formation of a compound transposon by two ISs at a locus outside a Tn*7*-like element, followed by the compound transposon jumping into the boundaries of the element, yielding a nested transposon. Such nesting of MGEs has been previously observed in individual transposons and other MGEs (37, 80, 87–89). The results presented here confirm that MGEs are a widespread, prominent source of the cargo carried by Tn*7*-like transposons. Although insertion of MGEs into chromosomes can be deleterious (46), the cargo region of Tn*7*-like transposons is likely to be a "safe haven" for sequence insertion and is therefore prone to the accretion of other MGEs. Accretion of MGEs apparently complements other mechanisms proposed to be involved in the capture of genes by transposons, such as homologous recombination (78, 90), culminating in the diverse and seemingly haphazard repertoire of cargo genes observed here. The relative contribution of MGE accretion versus homologous recombination to the diversity of cargo was not addressed here, but this question is now tractable with the database of transposons compiled in this and other studies (91).

One of the principal forces limiting HGT in prokaryotes is the cost of assimilating new genes into existing, regulated networks (2, 92–94). Integration itself can result in fitness consequences for the host cell, for example, as a result of disrupting and hence inactivating a gene (46). Tn*7*-like transposons appear to integrate into "safe sites," such as downstream of tRNAs, using dedicated target site-selecting proteins, leaving the attachment site intact and thus preempting a potential deleterious effect for the host cell (95). Beyond the consequences of integration itself, the introduction of horizontally acquired genes into an existing network can affect host fitness. A case in point is the antibiotic resistance genes mobilized by hundreds of Tn*7*-like transposons identified in this study, many of which were captured via integrons. The additional burden of the integron cassettes is offset by endowing the cell with the ability to survive in the presence of antibiotics (87). Analogously, a similar fraction of the Tn*7*-like transposons mobilize a diverse compendium of genes that confer resistance to bacteriophages (Table 1). Although possession of defense systems comes at a high cost due to autoimmunity and other effects (94, 96–99), their enrichment in Tn*7*-like transposons implies that the cost is superseded by the benefit of protection against phage predation. The concurrent circulation of antibiotic and phage resistance genes is also observed in other MGEs (100), highlighting how the cargo of a single MGE can confer immunity to both agents and redresses the cost of harmonization with the existing cellular network. An additional twist on the subject of defense systems in Tn*7*-like transposons is the incorporation of CRISPR arrays that are apparently involved in competition with other MGEs (34).

The metabolic cargo genes carried by Tn*7*-like transposons offer a unique perspective on the energetic demands of their hosts. Intuitively, these genes would reflect the

diverse niches occupied by each host, which span a variety of habitats, from the open ocean to the human gut. For example, a *Nitrospira moscoviensis* Tn*7*-like transposon carries a P460 cytochrome involved in the oxidation of hydroxylamine to nitrous oxide (60), divulging the importance of the ammonia oxidation pathway for these soil-dwelling bacteria. The widespread transfer of *sul1* and *dfrA1* involved in the synthesis of folic acid is likely driven by environmental exposure to antibiotics that target this pathway. Transposons in the genera *Vibrio*, *Marinobacter*, *Bacillus*, *Rhizobium*, and others harbor genes encoding enzymes of central carbon metabolism pathways, again prompting the question of how these "core" metabolic genes harmonize with the existing network upon introduction into a new host genome. The identification of these cargo-carrying transposons will facilitate experimental efforts aimed at this question.

The Tn*7*-like transposons carry an expansive repertoire of genes that are functionally distinct from the rest of the host genome, raising the question of how certain genes are captured and maintained whereas others are either never captured or are purged from the elements. The evidence presented here implies that any gene can, at one point, be present within a transposon, if only ephemerally. Mechanistically, this notion is supported by the prominence of ISs and compound transposons that vectorize genes (19, 101) into Tn*7*-like transposons (Fig. S3), conceivably providing the transposon access to any gene in the host genome. Under this premise, the question becomes what selective forces result in the nonuniform distribution of the functions of cargo genes (Fig. 3). On one hand, antiphage and antibiotic resistance genes appear to enhance the fitness of the host cell and favor their retention in the transposons. Furthermore, these genes as well as TA modules can make the host cell addicted to the transposons, such that loss of a transposon results in cell death. On the other hand, the observation that certain functions are rarely represented in the cargo (for example, information processing [COG category J]) implies that their continual presence in transposons is unfavorable to the host and hence to the transposons. Selection against these genes might manifest for a variety of reasons and drive their loss from the transposons. For example, and in particular, in the case of translation system components, expression of the respective genes from a transposon might result in a deleterious disruption of the stoichiometry of the protein complexes involved in these processes (102, 103).

MATERIALS AND METHODS

Identification of Tn*7*-like transposons. Multiple-sequence alignments (MSAs) of the five core genes from *E. coli* Tn*7* (*tnsABCDE*) were collected from the Pfam database using the following accession numbers: TnsA, PF08721 and PF08722; TnsB, PF00665; TnsC, PF11426 and PF05621; TnsD, PF15978; TniQ, PF06527; TnsE, PF18623. Additional MSAs of Tn*7*-like core proteins were generated from CASTs, and all MSAs were converted to HMMs with HMMer (v. 3.1b2). For the purpose of a comprehensive search for Tn*7*-like transposons, a nucleotide database encompassing the NCBI whole-genome shotgun (WGS) database and the MG-RAST database of metagenomes (104) was prepared. The nucleotide sequences were input into Prodigal (v. 2.6.3) (105) for ORF prediction. The collection of ORFs was queried with the Tn*7* HMMs, using model-specific gathering cutoffs for the Pfam HMMs and the following cutoffs for the custom HMMs: -T 25 –domT 25 –incT 25 –incdomT 25. Any ORFs that produced a hit but were located less than 3 kb from the contig boundary were discarded to remove incomplete transposons. Additional heuristics were employed to select candidate Tn*7*-like transposons, using the following criteria: (i) a hit to the TnsB HMM with a bitscore of $\geq$60, (ii) a hit to at least one other Tn*7* HMM, and (iii) two hits in a putative operon, which is operationally defined as two codirected ORFs separated by less than 50 bp of noncoding sequence. All loci satisfying these criteria were considered candidate Tn*7*-like transposons, and the putative TnsB orthologs were subjected to phylogenetic analysis.

The taxonomic information for each contig harboring a Tn*7*-like transposon was extracted from the NCBI taxonomy database using the Entrez suite of command line tools. If a sequence was not indexed in the database, it was assigned to the domain *Bacteria*, given that Tn*7*-like transposons have not so far been identified in genomes of archaea or viruses (see Results). Contigs were scanned using ViralVerify (45) and the default database of HMMs to classify contigs as viral, chromosomal, or plasmid.

Phylogenetic analysis. A phylogenetic tree of the DDE-family TnsB transposase was constructed using a previously described approach (49). Briefly, the transposase ORFs were first clustered at 80% amino acid sequence identity over 75% of the length of the shorter ORF, and then, the representative sequences were reclustered at 50% identity using MMSeqs2 (v. 12-113e3) (106). The secondary cluster members were aligned using MUSCLE (107) and compared to one another using HHSearch (108). The HHSearch similarity scores were used to construct an unweighted pair group method with arithmetic mean (UPGMA) dendrogram. The dendrogram was used to guide the pairwise alignment of clusters

with HHalign (108). Any clusters that could not be aligned using this approach were discarded. The single resulting alignment was filtered to remove partial sequences and sites with more than 50% gaps and homogeneity lower than 0.1 (109). An approximate maximum-likelihood tree was constructed from the filtered alignment using FastTree2 (110) with the Whelan-Goldman models of amino acid evolution and gamma-distributed site rates. Next, the tree was ultrametricized and branches were collapsed if the phylogenetic distance between them was less than 1. A single representative sequence was selected arbitrarily from each collapsed branch. The representatives were extracted from the main alignment and input to IQ-Tree (111) for phylogenetic reconstruction with parameters set to perform the aBayes branch test (112), ultrafast bootstrap approximation (113), and automatic model selection (114), which selected LG+F+R10 as the best model. The tree was visualized using the Interactive Tree Of Life (115).

Delineating Tn*7*-like transposon boundaries. For a contig harboring a Tn*7*-like transposon, all intergenic sequences longer than 50 bp were extracted into a single FASTA-formatted file. Coding sequences were excluded given that the boundaries of most Tn*7*-like transposons do not overlap a predicted ORF (34, 35). The intergenic sequences were used to construct a Markov model of the AT/GC content of the contig using the fasta-get-markov script provided in the MEME suite of tools (v. 5.3.0) (116). The Markov model was used as a background file for all motif detection steps described in the subsequent paragraph. As Tn*7*-like transposons up to 117 kb in length have been reported (35), all intergenic sequences up to 125 kb on either side of a putative *tnsB* homolog were collected and stored in a separate file. An all-versus-all BLASTn search was executed with the following parameters: -word_size 4 -max_hsps 100 -evalue 100. The left and right ends of *E. coli* Tn*7* contain 22-bp inverted repeats (24), so the output was filtered for inverted repeats 15 to 40 bp in length, with no more than two gaps and 6 mismatches. Two intergenic sequences were considered candidate left and right ends if at least one of these sequences was located less than 20 kb from the start codon of the *tnsB* transposase and the two sequences shared two or more inverted repeats. Each unique combination of two intergenic sequences satisfying these criteria was input into MEME to detect nucleotide motifs.

The motif discovery was initiated with the following command line options: -mod anr -nmotifs 10 -minw 15 -maxw 20 -minsites 4 -maxsites 6 -revcomp -markov_order 0 -evt 0.1. This combination of parameters was selected on the basis that Tn*7* contains three TnsB binding sites at the left end and four at the right end (24). The output was filtered to remove motifs with *P* values greater than 10^{-5}, present in only one of the two input sequences, and/or all located on the same strand. All motifs from all pairs of sequences that passed these filters were collected. The resulting set of motifs, representing putative TnsB binding sites, was used to search the 125-kb window of intergenic sequences around *tnsB* using the program FIMO (117), constrained with a false-discovery rate of 0.1 (*q* value). The same filtering approach for the BLASTn output described above was applied to the FIMO output, with the following additional criteria: (i) the spacing between any two instances of a motif must be ≤75 bp, (ii) the total length of all motif instances on a sequence must be ≤120 bp, (iii) the motif cannot be present on 5 or more intergenic sequences, and (iv) the product of the motif's estimated false-discovery rate (combined *q* value) must be less than 0.01. Criteria 1 and 2 were enforced to match the spacing and combined length of TnsB binding sites in Tn*7* (24), whereas criteria 4 and 5 were applied to remove common, weakly conserved nucleotide motifs. The single motif with the lowest combined *q* value satisfying all of these criteria was selected as the motif that best represents the binding site of the respective input transposase.

The individual motifs from closely related *tnsB* homologs were next compared to one another. The motifs were compiled into a single file if the phylogenetic distance between the respective transposases was less than 1 (see the preceding section for details on the construction of the TnsB tree). Each file was input individually to TomTom (118) for an all-versus-all motif alignment, using a minimum alignment length of 10 and the distance metric set as Pearson. Any motifs that did not align with ≥50% of the motifs present in the file were discarded; the remainder were collected into a single, nonredundant motif data set. All of the TnsB-centered windows of intergenic sequences were subjected to a final, competitive search against the motif data set using FIMO. The best motif was selected using the FIMO filtering criteria described above, and the outermost nucleotide coordinates of the motifs were recorded as the boundaries of the transposon.

Annotation of genes carried by Tn*7*-like transposons. The nucleotide sequence of each transposon was extracted from the respective contig using the predicted left and right boundaries. The ORFs were predicted using Prodigal in metagenomic mode (v. 2.6.3) (105) and clustered at 80% amino acid identity across 75% of the length of the shorter ORF using MMseqs2 (v. 12-113e3) (106). MSAs from the NCBI conserved domain database (v. 3.19) (119) were used to query the representative ORFs of each cluster using PSI-BLAST (120) with an E value cutoff of 0.01. The ORFs that did not produce a significant PSI-BLAST match were subjected to an additional round of annotation. Each of these ORFs was queried against the UniProt database clustered at 30% identity (constructed in June 2020; available at http://wwwuser.gwdg.de/~compbiol/uniclust/) with HHblits (108), enforcing the requirement that any database sequence align with ≥20% of the query. The resulting alignments were used for a second iteration of the search and/or terminated if the number of effective sequences in the alignment was greater than 10. The MSAs were then used to query HH-suite-formatted databases of alignments from the PDB and NCBI conserved domain database, accepting hits with probability greater than 90. Any annotations assigned to the representative ORF of a given cluster were transferred to each member of the cluster.

The annotations obtained as described above were revised in the following cases to be consistent with field-specific nomenclature. For experimentally characterized antibiotic resistance genes, representative ORFs were scanned using AMRFinderPlus (v. 3.10.15) (82) with default settings. For *cas* genes, representative ORFs were searched using PSI-BLAST against a database of multiple sequence alignments

obtained from a recent survey (121). For insertion sequences, the annotations were manually updated to the family-level designation according to the ISFinder database (122).

Noncoding RNAs were annotated using Infernal (v. 1.1.3) (123) against the Rfam database (v. 14.5) (124) using the model-specific bit score gathering threshold as a cutoff for significance. CRISPR arrays were detected using Minced (v. 0.4.2; https://github.com/ctSkennerton/minced) with the default settings. Integron *attC* sites were annotated using IntegronFinder (v. 2.0) (125). Annotations were displayed graphically using custom scripts and clinker (v. 0.0.21) (126).

Transposon dereplication, gene sharing network, and community detection. The nucleotide sequences of the transposons were dereplicated at 99% average nucleotide identity across 95% of the contig length using dRep (127) and associated dependencies (128, 129). The protein clusters from dereplicated transposons were used to calculate a similarity matrix as previously described (130). The similarity matrix was input into Hidef (v. 1.0.0) (57) to construct a hierarchical network of gene-sharing communities. Hidef requires a maximum resolution limit, which dictates the total number of communities and their size. To optimize this parameter, the similarity matrix was scanned with the maximum resolution parameter incremented in steps of 5, up to 1,000. At each step, the output was parsed to assign transposons to their smallest community. The total number, size, persistence, and last common ancestor (LCA) of each community were tabulated. The LCA was defined using an 80% consensus rule, such that $\geq$80% of the transposons in a community have the same ancestor. A maximum resolution size of 140 yielded the highest mean community persistence without major changes to the phylogenetic level of the LCA (see Results), so the final network was constructed using this resolution and a persistence cutoff of 20; all other parameters were left as default. The LCA of each community in the final network was retabulated using the same 80% consensus rule, but without assigning transposons to their single, smallest subcommunity, so that the taxonomic makeup of communities at all hierarchical levels could be determined. Networks were visualized using Cytoscape (v. 3.7.2) (131).

Functional annotation of transposon genes using COGs. The enrichment or depletion of COG functional categories for the subset of transposons in completely sequenced bacterial genomes were calculated as described previously (132). Briefly, the ORFs in these genomes were annotated by comparison to the COG database of multiple sequence alignments (58) using the PSI-BLAST parameters described above. The COG categories corresponding to the transposon-borne ORFs, or a randomly selected genomic locus of identical length, were extracted and summed. The relative frequency of each COG category in the transposons versus the genomic loci was calculated from a random selection of 100 pairs of transposons and their genomic equivalents, iterated 1,000 times.

The COG annotations of transposon-borne ORFs were extracted and assigned to their respective COG pathways. The completeness of the pathway was calculated as the number of COGs from the pathway represented in the transposon divided by the total number of COGs constituting the pathway. Pathways were discarded if represented in fewer than 10 transposons.

Data availability. All sequences are publicly available from the NCBI and MG-RAST databases. The nucleotide coordinates of all transposons are provided in Table S1C. FASTA-formatted files, annotations for all transposon open reading frames, and source data for Fig. 1 and 3 are available at https://ftp.ncbi.nih.gov/pub/yutinn/benler_2021/Tn7/source_data/. The code for delineating Tn*7*-like transposon boundaries is publicly available at https://github.com/sean-bam/Tnacity.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

TABLE S1, XLSX file, 4.9 MB.

FIG S1, EPS file, 2.6 MB.

FIG S2, EPS file, 0.8 MB.

FIG S3, EPS file, 2.3 MB.

FIG S4, EPS file, 2.1 MB.

FIG S5, EPS file, 1.2 MB.

FIG S6, EPS file, 1 MB.

FIG S7, EPS file, 1.4 MB.

FIG S8, EPS file, 1.4 MB.

FIG S9, EPS file, 1.8 MB.

ACKNOWLEDGMENT

Funding for this project was provided by the Intramural Research Program of the National Institutes of Health (National Library of Medicine).

REFERENCES

1. Puigbò P, Lobkovsky AE, Kristensen DM, Wolf YI, Koonin EV. 2014. Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes. BMC Biol 12:66. https://doi.org/10.1186/s12915-014-0066-4.

2. Sela I, Wolf YI, Koonin EV. 2019. Selection and genome plasticity as the key factors in the evolution of bacteria. Phys Rev X 9:e031018. https://doi.org/10.1103/PhysRevX.9.031018.

3. Koonin EV. 2015. The turbulent network dynamics of microbial evolution and the statistical tree of life. J Mol Evol 80:244–250. https://doi.org/10.1007/s00239-015-9679-7.
4. Treangen TJ, Rocha EPC. 2011. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. PLoS Genet 7:e1001284. https://doi.org/10.1371/journal.pgen.1001284.
5. Takeuchi N, Kaneko K, Koonin EV. 2014. Horizontal gene transfer can rescue prokaryotes from Muller's ratchet: benefit of DNA from dead cells and population subdivision. G3 (Bethesda) 4:325–339. https://doi.org/10.1534/g3.113.009845.
6. Nelson-Sathi S, Sousa FL, Roettger M, Lozada-Chávez N, Thiergart T, Janssen A, Bryant D, Landan G, Schönheit P, Siebers B, McInerney JO, Martin WF. 2015. Origins of major archaeal clades correspond to gene acquisitions from bacteria. Nature 517:77–80. https://doi.org/10.1038/nature13805.
7. Popa O, Landan G, Dagan T. 2017. Phylogenomic networks reveal limited phylogenetic range of lateral gene transfer by transduction. ISME J 11:543–554. https://doi.org/10.1038/ismej.2016.116.
8. Halary S, Leigh JW, Cheaib B, Lopez P, Bapteste E. 2010. Network analyses structure genetic diversity in independent genetic worlds. Proc Natl Acad Sci U S A 107:127–132. https://doi.org/10.1073/pnas.0908978107.
9. McInnes RS, McCallum GE, Lamberte LE, van Schaik W. 2020. Horizontal transfer of antibiotic resistance genes in the human gut microbiome. Curr Opin Microbiol 53:35–43. https://doi.org/10.1016/j.mib.2020.02.002.
10. Oliveira PH, Touchon M, Cury J, Rocha EPC. 2017. The chromosomal organization of horizontal gene transfer in bacteria. Nat Commun 8:841. https://doi.org/10.1038/s41467-017-00808-w.
11. Yaffe E, Relman DA. 2020. Tracking microbial evolution in the human gut using Hi-C reveals extensive horizontal gene transfer, persistence and adaptation. Nat Microbiol 5:343–353. https://doi.org/10.1038/s41564-019-0625-0.
12. Song W, Wemheuer B, Zhang S, Steensen K, Thomas T. 2019. MetaCHIP: community-level horizontal gene transfer identification through the combination of best-match and phylogenetic approaches. Microbiome 7:36. https://doi.org/10.1186/s40168-019-0649-y.
13. Redondo-Salvo S, Fernández-López R, Ruiz R, Vielva L, de Toro M, Rocha EPC, Garcillán-Barcia MP, de la Cruz F. 2020. Pathways for horizontal gene transfer in bacteria revealed by a global map of their plasmids. Nat Commun 11:3602. https://doi.org/10.1038/s41467-020-17278-2.
14. Antipov D, Raiko M, Lapidus A, Pevzner PA. 2019. Plasmid detection and assembly in genomic and metagenomic data sets. Genome Res 29:961–968. https://doi.org/10.1101/gr.241299.118.
15. Cury J, Touchon M, Rocha Eduardo PC. 2017. Integrative and conjugative elements and their hosts: composition, distribution and organization. Nucleic Acids Res 45:8943–8956. https://doi.org/10.1093/nar/gkx607.
16. Roux S, Páez-Espino D, Chen I-MA, Palaniappan K, Ratner A, Chu K, Reddy TBK, Nayfach S, Schulz F, Call L, Neches RY, Woyke T, Ivanova NN, Eloe-Fadrosh EA, Kyrpides NC. 2021. IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. Nucleic Acids Res 49:D764–D775. https://doi.org/10.1093/nar/gkaa946.
17. Breitbart M, Bonnain C, Malki K, Sawaya NA. 2018. Phage puppet masters of the marine microbial realm. Nat Microbiol 3:754–766. https://doi.org/10.1038/s41564-018-0166-y.
18. Benler S, Yutin N, Antipov D, Rayko M, Shmakov S, Gussow AB, Pevzner P, Koonin EV. 2021. Thousands of previously unknown phages discovered in whole-community human gut metagenomes. Microbiome 9:78. https://doi.org/10.1186/s40168-021-01017-w.
19. Siguier P, Gourbeyre E, Varani A, Ton-Hoang B, Chandler M. 2015. Everyman's guide to bacterial insertion sequences, p 555–590. *In* Chandler M, Gellert M, Lambowitz AM, Rice PA, Sandmeyer SB (ed), Mobile DNA III. ASM Press, Washington, DC.
20. Craig NL. 2015. A moveable feast: an introduction to mobile DNA, p 1–39. *In* Chandler M, Gellert M, Lambowitz AM, Rice PA, Sandmeyer SB (ed), Mobile DNA III. ASM Press, Washington, DC.
21. Hickman AB, Dyda F. 2015. Mechanisms of DNA transposition, p 529–553. *In* Chandler M, Gellert M, Lambowitz AM, Rice PA, Sandmeyer SB (ed), Mobile DNA III. ASM Press, Washington, DC.
22. Waddell CS, Craig NL. 1988. Tn7 transposition: two transposition pathways directed by five Tn7-encoded genes. Genes Dev 2:137–149. https://doi.org/10.1101/gad.2.2.137.
23. Peters JE, Craig NL. 2001. Tn7: smarter than we thought. Nat Rev Mol Cell Biol 2:806–814. https://doi.org/10.1038/35099006.
24. Arciszewska LK, Drake D, Craig NL. 1989. Transposon Tn7: cis-acting sequences in transposition and transposition immunity. J Mol Biol 207:35–52. https://doi.org/10.1016/0022-2836(89)90439-7.
25. Sarnovsky RJ, May EW, Craig NL. 1996. The Tn7 transposase is a heteromeric complex in which DNA breakage and joining activities are distributed between different gene products. EMBO J 15:6348–6361. https://doi.org/10.1002/j.1460-2075.1996.tb01024.x.
26. Stellwagen AE, Craig NL. 1997. Gain-of-function mutations in TnsC, an ATP-dependent transposition protein that activates the bacterial transposon Tn7. Genetics 145:573–585. https://doi.org/10.1093/genetics/145.3.573.
27. Bainton RJ, Kubo KM, Feng J-N, Craig NL. 1993. Tn7 transposition: target DNA recognition is mediated by multiple Tn7-encoded proteins in a purified in vitro system. Cell 72:931–943. https://doi.org/10.1016/0092-8674(93)90581-A.
28. Parks AR, Li Z, Shi Q, Owens RM, Jin MM, Peters JE. 2009. Transposition into replicating DNA occurs through interaction with the processivity factor. Cell 138:685–695. https://doi.org/10.1016/j.cell.2009.06.011.
29. Saito M, Ladha A, Strecker J, Faure G, Neumann E, Altae-Tran H, Macrae RK, Zhang F. 2021. Dual modes of CRISPR-associated transposon homing. Cell 184:2441–2453.e2418. https://doi.org/10.1016/j.cell.2021.03.006.
30. Strecker J, Ladha A, Gardner Z, Schmid-Burgk JL, Makarova KS, Koonin EV, Zhang F. 2019. RNA-guided DNA insertion with CRISPR-associated transposases. Science 365:48–53. https://doi.org/10.1126/science.aax9181.
31. Hsieh S-C, Peters JE. 2021. Tn7-CRISPR-Cas12K elements manage pathway choice using truncated repeat-spacer units to target tRNA attachment sites. bioRxiv https://doi.org/10.1101/2021.02.06.429022.
32. Petassi MT, Hsieh S-C, Peters JE. 2020. Guide RNA categorization enables target site choice in Tn7-CRISPR-Cas transposons. Cell 183:1757–1771.e1718. https://doi.org/10.1016/j.cell.2020.11.005.
33. Klompe SE, Vo PLH, Halpin-Healy TS, Sternberg SH. 2019. Transposon-encoded CRISPR–Cas systems direct RNA-guided DNA integration. Nature 571:219–225. https://doi.org/10.1038/s41586-019-1323-z.
34. Faure G, Shmakov SA, Yan WX, Cheng DR, Scott DA, Peters JE, Makarova KS, Koonin EV. 2019. CRISPR–Cas in mobile genetic elements: counter-defence and beyond. Nat Rev Microbiol 17:513–525. https://doi.org/10.1038/s41579-019-0204-7.
35. Peters JE, Makarova KS, Shmakov S, Koonin EV. 2017. Recruitment of CRISPR-Cas systems by Tn7-like transposons. Proc Natl Acad Sci U S A 114:E7358–E7366. https://doi.org/10.1073/pnas.1709035114.
36. Parks AR, Peters JE. 2007. Transposon Tn7 is widespread in diverse bacteria and forms genomic islands. J Bacteriol 189:2170–2173. https://doi.org/10.1128/JB.01536-06.
37. Peters JE, Fricker AD, Kapili BJ, Petassi MT. 2014. Heteromeric transposase elements: generators of genomic islands across diverse bacteria. Mol Microbiol 93:1084–1092. https://doi.org/10.1111/mmi.12740.
38. Parks AR, Peters JE. 2009. Tn7 elements: engendering diversity from chromosomes to episomes. Plasmid 61:1–14. https://doi.org/10.1016/j.plasmid.2008.09.008.
39. Hamidian M, Hall RM. 2021. Dissemination of novel Tn7 family transposons carrying genes for synthesis and uptake of fimsbactin siderophores among Acinetobacter baumannii isolates. Microb Genom 7:mgen000548. https://doi.org/10.1099/mgen.0.000548.
40. Aprile F, Heredia-Ponce Z, Cazorla FM, Vicente A, Gutiérrez-Barranquero JA, Kivisaar M. 2021. A large Tn*7*-like transposon confers hyperresistance to copper in *Pseudomonas syringae* pv. syringae. Appl Environ Microbiol 87:e02528-20. https://doi.org/10.1128/AEM.02528-20.
41. Nayfach S, Camargo AP, Schulz F, Eloe-Fadrosh E, Roux S, Kyrpides NC. 2021. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nat Biotechnol 39:578–585. https://doi.org/10.1038/s41587-020-00774-7.
42. Saak CC, Dinh CB, Dutton RJ. 2020. Experimental approaches to tracking mobile genetic elements in microbial communities. FEMS Microbiol Rev 44:606–630. https://doi.org/10.1093/femsre/fuaa025.
43. Cury J, Abby SS, Doppelt-Azeroual O, Néron B, Rocha EPC. 2020. Identifying conjugative plasmids and integrative conjugative elements with CONJscan, p 265–283. *In* de la Cruz F (ed), Horizontal gene transfer: methods and protocols. Springer, New York, NY.
44. Liu M, Li X, Xie Y, Bi D, Sun J, Li J, Tai C, Deng Z, Ou H-Y. 2019. ICEberg 2.0: an updated database of bacterial integrative and conjugative elements. Nucleic Acids Res 47:D660–D665. https://doi.org/10.1093/nar/gky1123.

45. Antipov D, Raiko M, Lapidus A, Pevzner PA. 2020. MetaviralSPAdes: assembly of viruses from metagenomic data. Bioinformatics 36:4126–4129. https://doi.org/10.1093/bioinformatics/btaa490.
46. Durrant MG, Li MM, Siranosian BA, Montgomery SB, Bhatt AS. 2020. A bioinformatic analysis of integrative mobile genetic elements highlights their role in bacterial adaptation. Cell Host Microbe 27:140–153.E149. https://doi.org/10.1016/j.chom.2019.10.022.
47. Jiang X, Hall AB, Xavier RJ, Alm EJ. 2019. Comprehensive analysis of chromosomal mobile genetic elements in the gut microbiome reveals phylum-level niche-adaptive gene pools. PLoS One 14:e0223680. https://doi.org/10.1371/journal.pone.0223680.
48. May EW, Craig NL. 1996. Switching from cut-and-paste to replicative Tn7 transposition. Science 272:401–404. https://doi.org/10.1126/science.272.5260.401.
49. Wolf YI, Kazlauskas D, Iranzo J, Lucía-Sanz A, Kuhn JH, Krupovic M, Dolja VV, Koonin EV. 2018. Origins and evolution of the global RNA virome. mBio 9:e02329-18. https://doi.org/10.1128/mBio.02329-18.
50. Kholodii GY, Mindlin SZ, Bass IA, Yurieva OV, Minakhina SV, Nikiforov VG. 1995. Four genes, two ends, and a res region are involved in transposition of Tn5053: a paradigm for a novel family of transposons carrying either a mer operon or an integron. Mol Microbiol 17:1189–1200. https://doi.org/10.1111/j.1365-2958.1995.mmi_17061189.x.
51. Hamidian M, Hall RM. 2011. AbaR4 replaces AbaR3 in a carbapenem-resistant Acinetobacter baumannii isolate belonging to global clone 1 from an Australian hospital. J Antimicrob Chemother 66:2484–2491. https://doi.org/10.1093/jac/dkr356.
52. Kamali-Moghaddam M, Sundström L. 2001. Arrayed transposase-binding sequences on the ends of transposon Tn5090/Tn402. Nucleic Acids Res 29:1005–1011. https://doi.org/10.1093/nar/29.4.1005.
53. Arciszewska LK, Craig NL. 1991. Interaction of the Tn7-encoded transposition protein TnsB with the ends of the transposon. Nucleic Acids Res 19:5021–5029. https://doi.org/10.1093/nar/19.18.5021.
54. Rao JE, Miller PS, Craig NL. 2000. Recognition of triple-helical DNA structures by transposon Tn7. Proc Natl Acad Sci U S A 97:3936–3941. https://doi.org/10.1073/pnas.080061497.
55. Shen Y, Gomez-Blanco J, Petassi MT, Peters JE, Ortega J, Guarné A. 2021. Structural basis for DNA targeting by the Tn7 transposon. bioRxiv https://doi.org/10.1101/2021.05.24.445525.
56. Park J-U, Tsai AW-L, Mehrotra E, Petassi MT, Hsieh S-C, Ke A, Peters JE, Kellogg EH. 2021. Structural basis for target site selection in RNA-guided DNA transposition systems. Science 373:768–774. https://doi.org/10.1126/science.abi8976.
57. Zheng F, Zhang S, Churas C, Pratt D, Bahar I, Ideker T. 2021. HiDeF: identifying persistent structures in multiscale 'omics data. Genome Biol 22:21. https://doi.org/10.1186/s13059-020-02228-4.
58. Galperin MY, Wolf YI, Makarova KS, Vera Alvarez R, Landsman D, Koonin EV. 2021. COG database update: focus on microbial diversity, model organisms, and widespread pathogens. Nucleic Acids Res 49:D274–D281. https://doi.org/10.1093/nar/gkaa1018.
59. Grégoire N, Aranzana-Climent V, Magréault S, Marchand S, Couet W. 2017. Clinical pharmacokinetics and pharmacodynamics of colistin. Clin Pharmacokinet 56:1441–1460. https://doi.org/10.1007/s40262-017-0561-1.
60. Caranto JD, Vilbert AC, Lancaster KM. 2016. Nitrosomonas europaea cytochrome P460 is a direct link between nitrification and nitrous oxide emission. Proc Natl Acad Sci U S A 113:14704–14709. https://doi.org/10.1073/pnas.1611051113.
61. Aravind L, Leipe DD, Koonin EV. 1998. Toprim—a conserved catalytic domain in type IA and II topoisomerases, DnaG-type primases, OLD family nucleases and RecR proteins. Nucleic Acids Res 26:4205–4213. https://doi.org/10.1093/nar/26.18.4205.
62. Sironi G. 1969. Mutants of Escherichia coli unable to be lysogenized by the temperate bacteriophage P2. Virology 37:163–176. https://doi.org/10.1016/0042-6822(69)90196-2.
63. Doron S, Melamed S, Ofir G, Leavitt A, Lopatina A, Keren M, Amitai G, Sorek R. 2018. Systematic discovery of antiphage defense systems in the microbial pangenome. Science 359:eaar4120. https://doi.org/10.1126/science.aar4120.
64. Millman A, Melamed S, Amitai G, Sorek R. 2020. Diversity and classification of cyclic-oligonucleotide-based anti-phage signalling systems. Nat Microbiol 5:1608–1615. https://doi.org/10.1038/s41564-020-0777-y.
65. Cohen D, Melamed S, Millman A, Shulman G, Oppenheimer-Shaanan Y, Kacen A, Doron S, Amitai G, Sorek R. 2019. Cyclic GMP–AMP signalling protects bacteria against viral infection. Nature 574:691–695. https://doi.org/10.1038/s41586-019-1605-5.
66. Goldfarb T, Sberro H, Weinstock E, Cohen O, Doron S, Charpak-Amikam Y, Afik S, Ofir G, Sorek R. 2015. BREX is a novel phage resistance system widespread in microbial genomes. EMBO J 34:169–183. https://doi.org/10.15252/embj.201489455.
67. Wang L, Jiang S, Deng Z, Dedon PC, Chen S. 2019. DNA phosphorothioate modification—a new multi-functional epigenetic system in bacteria. FEMS Microbiol Rev 43:109–122. https://doi.org/10.1093/femsre/fuy036.
68. Burroughs AM, Iyer LM, Aravind L. 2013. Two novel PIWI families: roles in inter-genomic conflicts in bacteria and Mediator-dependent modulation of transcription in eukaryotes. Biol Direct 8:13. https://doi.org/10.1186/1745-6150-8-13.
69. Gao L, Altae-Tran H, Böhning F, Makarova KS, Segel M, Schmid-Burgk JL, Koob J, Wolf YI, Koonin EV, Zhang F. 2020. Diverse enzymatic activities mediate antiviral immunity in prokaryotes. Science 369:1077–1084. https://doi.org/10.1126/science.aba0372.
70. Bernheim A, Millman A, Ofir G, Meitav G, Avraham C, Shomar H, Rosenberg MM, Tal N, Melamed S, Amitai G, Sorek R. 2021. Prokaryotic viperins produce diverse antiviral molecules. Nature 589:120–124. https://doi.org/10.1038/s41586-020-2762-2.
71. Tal N, Millman A, Stokar-Avihail A, Fedorenko T, Leavitt A, Melamed S, Yirmiya E, Avraham C, Amitai G, Sorek R. 2021. Antiviral defense via nucleotide depletion in bacteria. bioRxiv. https://doi.org/10.1101/2021.04.26.441389.
72. Makarova KS, Wolf YI, Koonin EV. 2013. Comparative genomics of defense systems in archaea and bacteria. Nucleic Acids Res 41:4360–4377. https://doi.org/10.1093/nar/gkt157.
73. Makarova KS, Wolf YI, Koonin EV. 2009. Comprehensive comparative-genomic analysis of type 2 toxin-antitoxin systems and related mobile stress response systems in prokaryotes. Biol Direct 4:19. https://doi.org/10.1186/1745-6150-4-19.
74. Chopin M-C, Chopin A, Bidnenko E. 2005. Phage abortive infection in lactococci: variations on a theme. Curr Opin Microbiol 8:473–479. https://doi.org/10.1016/j.mib.2005.06.006.
75. Harms A, Brodersen DE, Mitarai N, Gerdes K. 2018. Toxins, targets, and triggers: an overview of toxin-antitoxin biology. Mol Cell 70:768–784. https://doi.org/10.1016/j.molcel.2018.01.003.
76. Page R, Peti W. 2016. Toxin-antitoxin systems in bacterial growth arrest and persistence. Nat Chem Biol 12:208–214. https://doi.org/10.1038/nchembio.2044.
77. Labrie SJ, Moineau S. 2007. Abortive infection mechanisms and prophage sequences significantly influence the genetic makeup of emerging lytic lactococcal phages. J Bacteriol 189:1482–1487. https://doi.org/10.1128/JB.01111-06.
78. Nicolas E, Lambin M, Dandoy D, Galloy C, Nguyen N, Oger CA, Hallet B. 2015. The Tn3-family of replicative transposons, p 693–726. *In* Chandler M, Gellert M, Lambowitz AM, Rice PA, Sandmeyer SB (ed), Mobile DNA III. ASM Press, Washington, DC.
79. Peters JE. 2015. Tn7, p 647–667. *In* Chandler M, Gellert M, Lambowitz AM, Rice PA, Sandmeyer SB (ed), Mobile DNA III. ASM Press, Washington, DC.
80. Minakhina S, Kholodii G, Mindlin S, Yurieva O, Nikiforov V. 1999. Tn5053 family transposons are res site hunters sensing plasmidal res sites occupied by cognate resolvases. Mol Microbiol 33:1059–1068. https://doi.org/10.1046/j.1365-2958.1999.01548.x.
81. Huovinen P, Sundström L, Swedberg G, Sköld O. 1995. Trimethoprim and sulfonamide resistance. Antimicrob Agents Chemother 39:279–289. https://doi.org/10.1128/AAC.39.2.279.
82. Feldgarden M, Brover V, Haft DH, Prasad AB, Slotta DJ, Tolstoy I, Tyson GH, Zhao S, Hsu C-H, McDermott PF, Tadesse DA, Morales C, Simmons M, Tillman G, Wasilenko J, Folster JP, Klimke W. 2019. Validating the AMRFinder tool and resistance gene database by using antimicrobial resistance genotype-phenotype correlations in a collection of isolates. Antimicrob Agents Chemother 63:e00483-19. https://doi.org/10.1128/AAC.00483-19.
83. Escudero JA, Loot C, Nivina A, Mazel D, Rice P, Craig N. 2015. The integron: adaptation on demand. Microbiol Spectr 3:MDNA3-0019-2014. https://doi.org/10.1128/microbiolspec.MDNA3-0019-2014.
84. Ramírez MS, Quiroga C, Centrón D. 2005. Novel rearrangement of a class 2 integron in two non-epidemiologically related isolates of *Acinetobacter baumannii*. Antimicrob Agents Chemother 49:5179–5181. https://doi.org/10.1128/AAC.49.12.5179-5181.2005.
85. Oppon JC, Sarnovsky RJ, Craig NL, Rawlings DE. 1998. A Tn*7*-like transposon is present in the *glmUS* region of the obligately chemoautolithotrophic

bacterium *Thiobacillus ferrooxidans*. J Bacteriol 180:3007–3012. https://doi.org/10.1128/JB.180.11.3007-3012.1998.
86. Vo PLH, Acree C, Smith ML, Sternberg SH. 2021. Unbiased profiling of CRISPR RNA-guided transposition products by long-read sequencing. Mob DNA 12:13. https://doi.org/10.1186/s13100-021-00242-2.
87. Liebert CA, Hall RM, Summers AO. 1999. Transposon Tn21, flagship of the floating genome. Microbiol Mol Biol Rev 63:507–522. https://doi.org/10.1128/MMBR.63.3.507-522.1999.
88. Krupovic M, Makarova KS, Wolf YI, Medvedeva S, Prangishvili D, Forterre P, Koonin EV. 2019. Integrated mobile genetic elements in Thaumarchaeota. Environ Microbiol 21:2056–2078. https://doi.org/10.1111/1462-2920.14564.
89. Sota M, Endo M, Nitta K, Kawasaki H, Tsuda M. 2002. Characterization of a class II defective transposon carrying two haloacetate dehalogenase genes from *Delftia acidovorans* plasmid pUO1. Appl Environ Microbiol 68:2307–2315. https://doi.org/10.1128/AEM.68.5.2307-2315.2002.
90. Partridge SR. 2011. Analysis of antibiotic resistance regions in Gram-negative bacteria. FEMS Microbiol Rev 35:820–855. https://doi.org/10.1111/j.1574-6976.2011.00277.x.
91. Ross K, Varani AM, Snesrud E, Huang H, Alvarenga DO, Zhang J, Wu C, McGann P, Chandler M, Gottesman S. 2021. TnCentral: a prokaryotic transposable element database and web portal for transposon analysis. mBio 12:e02060-21.
92. Sorek R, Zhu Y, Creevey CJ, Francino MP, Bork P, Rubin EM. 2007. Genome-wide experimental determination of barriers to horizontal gene transfer. Science 318:1449–1452. https://doi.org/10.1126/science.1147112.
93. Bershtein S, Serohijos AWR, Bhattacharyya S, Manhart M, Choi J-M, Mu W, Zhou J, Shakhnovich EI. 2015. Protein homeostasis imposes a barrier on functional integration of horizontally transferred genes in bacteria. PLoS Genet 11:e1005612. https://doi.org/10.1371/journal.pgen.1005612.
94. Iranzo J, Cuesta JA, Manrubia S, Katsnelson MI, Koonin EV. 2017. Disentangling the effects of selection and loss bias on gene dynamics. Proc Natl Acad Sci U S A 114:E5616–E5624. https://doi.org/10.1073/pnas.1704925114.
95. Peters JE. 2019. Targeted transposition with Tn7 elements: safe sites, mobile plasmids, CRISPR/Cas and beyond. Mol Microbiol 112:1635–1644. https://doi.org/10.1111/mmi.14383.
96. Cui L, Bikard D. 2016. Consequences of Cas9 cleavage in the chromosome ofEscherichia coli. Nucleic Acids Res 44:4243–4251. https://doi.org/10.1093/nar/gkw223.
97. Heussler GE, Cady KC, Koeppen K, Bhuju S, Stanton BA, O'Toole GA. 2015. Clustered regularly interspaced short palindromic repeat-dependent, biofilm-specific death of Pseudomonas aeruginosa mediated by increased expression of phage-related genes. mBio 6:e00129-15. https://doi.org/10.1128/mBio.00129-15.
98. Koonin EV, Zhang F. 2017. Coupling immunity and programmed cell suicide in prokaryotes: life-or-death choices. Bioessays 39:e201600186. https://doi.org/10.1002/bies.201600186.
99. Pleška M, Qian L, Okura R, Bergmiller T, Wakamoto Y, Kussell E, Guet CC. 2016. Bacterial autoimmunity due to a restriction-modification system. Curr Biol 26:404–409. https://doi.org/10.1016/j.cub.2015.12.041.
100. LeGault KN, Hays SG, Angermeyer A, McKitterick AC, Johura F-T, Sultana M, Ahmed T, Alam M, Seed KD. 2021. Seed KD: temporal shifts in antibiotic resistance elements govern phage-pathogen conflicts. Science 373:eabg2166. https://doi.org/10.1126/science.abg2166.
101. Toleman MA, Bennett PM, Walsh TR. 2006. IS*CR* elements: novel gene-capturing systems of the 21st century? Microbiol Mol Biol Rev 70:296–316. https://doi.org/10.1128/MMBR.00048-05.
102. Veitia RA. 2004. Gene dosage balance in cellular pathways: implications for dominance and gene duplicability. Genetics 168:569–574. https://doi.org/10.1534/genetics.104.029785.
103. Veitia RA, Potier MC. 2015. Gene dosage imbalances: action, reaction, and models. Trends Biochem Sci 40:309–317. https://doi.org/10.1016/j.tibs.2015.03.011.
104. Wilke A, Bischof J, Gerlach W, Glass E, Harrison T, Keegan KP, Paczian T, Trimble WL, Bagchi S, Grama A, Chaterji S, Meyer F. 2016. The MG-RAST metagenomics database and portal in 2015. Nucleic Acids Res 44:D590–D594. https://doi.org/10.1093/nar/gkv1322.
105. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119. https://doi.org/10.1186/1471-2105-11-119.
106. Steinegger M, Söding J. 2018. Clustering huge protein sequence sets in linear time. Nat Commun 9:2542. https://doi.org/10.1038/s41467-018-04964-5.
107. Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797. https://doi.org/10.1093/nar/gkh340.
108. Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, Söding J. 2019. HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics 20:473. https://doi.org/10.1186/s12859-019-3019-7.
109. Yutin N, Makarova KS, Mekhedov SL, Wolf YI, Koonin EV. 2008. The deep archaeal roots of eukaryotes. Mol Biol Evol 25:1619–1630. https://doi.org/10.1093/molbev/msn108.
110. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2—approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490. https://doi.org/10.1371/journal.pone.0009490.
111. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32:268–274. https://doi.org/10.1093/molbev/msu300.
112. Anisimova M, Gil M, Dufayard J-F, Dessimoz C, Gascuel O. 2011. Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. Syst Biol 60:685–699. https://doi.org/10.1093/sysbio/syr041.
113. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: improving the ultrafast bootstrap approximation. Mol Biol Evol 35:518–522. https://doi.org/10.1093/molbev/msx281.
114. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods 14:587–589. https://doi.org/10.1038/nmeth.4285.
115. Letunic I, Bork P. 2021. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res 49:W293–W296. https://doi.org/10.1093/nar/gkab301.
116. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME suite: tools for motif discovery and searching. Nucleic Acids Res 37:W202–W208. https://doi.org/10.1093/nar/gkp335.
117. Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. Bioinformatics 27:1017–1018. https://doi.org/10.1093/bioinformatics/btr064.
118. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. 2007. Quantifying similarity between motifs. Genome Biol 8:R24. https://doi.org/10.1186/gb-2007-8-2-r24.
119. Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Marchler GH, Song JS, Thanki N, Yamashita RA, Yang M, Zhang D, Zheng C, Lanczycki CJ, Marchler-Bauer A. 2020. CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Res 48:D265–D268. https://doi.org/10.1093/nar/gkz991.
120. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402. https://doi.org/10.1093/nar/25.17.3389.
121. Makarova KS, Wolf YI, Iranzo J, Shmakov SA, Alkhnbashi OS, Brouns SJJ, Charpentier E, Cheng D, Haft DH, Horvath P, Moineau S, Mojica FJM, Scott D, Shah SA, Siksnys V, Terns MP, Venclovas Č, White MF, Yakunin AF, Yan W, Zhang F, Garrett RA, Backofen R, van der Oost J, Barrangou R, Koonin EV. 2020. Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. Nat Rev Microbiol 18:67–83. https://doi.org/10.1038/s41579-019-0299-x.
122. Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M. 2006. ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res 34:D32–D36. https://doi.org/10.1093/nar/gkj014.
123. Nawrocki EP, Eddy SR. 2013. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29:2933–2935. https://doi.org/10.1093/bioinformatics/btt509.
124. Kalvari I, Nawrocki EP, Ontiveros-Palacios N, Argasinska J, Lamkiewicz K, Marz M, Griffiths-Jones S, Toffano-Nioche C, Gautheret D, Weinberg Z, Rivas E, Eddy SR, Finn RD, Bateman A, Petrov AI. 2021. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. Nucleic Acids Res 49:D192–D200. https://doi.org/10.1093/nar/gkaa1047.
125. Cury J, Jové T, Touchon M, Néron B, Rocha EP. 2016. Identification and analysis of integrons and cassette arrays in bacterial genomes. Nucleic Acids Res 44:4539–4550. https://doi.org/10.1093/nar/gkw319.

126. Gilchrist CLM, Chooi Y-H. 2021. clinker and clustermap.js: automatic generation of gene cluster comparison figures. Bioinformatics 37:2473–2475. https://doi.org/10.1093/bioinformatics/btab007.

127. Olm MR, Brown CT, Brooks B, Banfield JF. 2017. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J 11:2864–2868. https://doi.org/10.1038/ismej.2017.126.

128. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun 9:5114. https://doi.org/10.1038/s41467-018-07641-9.

129. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. 2016. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol 17:132. https://doi.org/10.1186/s13059-016-0997-x.

130. Makarova KS, Timinskas A, Wolf YI, Gussow AB, Siksnys V, Venclovas Č, Koonin EV. 2020. Evolutionary and functional classification of the CARF domain superfamily, key sensors in prokaryotic antivirus defense. Nucleic Acids Res 48:8828–8847. https://doi.org/10.1093/nar/gkaa635.

131. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–2504. https://doi.org/10.1101/gr.1239303.

132. Shmakov SA, Utkina I, Wolf YI, Makarova KS, Severinov K, Koonin EV. 2020. CRISPR arrays away from cas genes. CRISPR J 3:535–549. https://doi.org/10.1089/crispr.2020.0062.