

Why does a gene cassette keep the company of an orphan reverse transcriptase? Pollutibacter on five continents

Chulmin Pyo

pyo@ai3.kr

AI3 Discovery <https://orcid.org/0009-0003-1439-0740>

Research Article

Keywords: reverse transcriptase, gene cassette, Bacteroidota, Pollutibacter, orphan lineage, comparative genomics

Posted Date: October 20th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11420391/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Why does a gene cassette keep the company of an orphan reverse transcriptase? Pollutibacter on five continents

Chulmin Pyo

AI3 Discovery, Republic of Korea

Corresponding author: discovery@ai3.kr

Abstract

Several bacterial reverse transcriptases (RTs) protect cells from phage, yet some fit no known family. Screening 189,630 genomes, we found an RT that none of our comparisons (known RT types, trees, predicted structures) places in a known family (nearest published-library RT at most 34.6% identical), which we call ESOL RT. It sits beside a gene cassette in 16 genomes, chosen for carrying it, from lake, river, pond and other waters in nine countries on five continents, all in one Bacteroidota genus, Pollutibacter. The cassette has a conserved upstream element and four genes at fixed positions, including a MIDAS-motif protein and a predicted metalloprotease-like one lacking H[EDQ]xxH. At 27 further Pollutibacter loci, 97 of 99 observable positions beside the RT hold a cassette part, both exceptions at one locus. No observable copy of the four neighbouring genes is disrupted, codon models suggest purifying selection, and long reads from other freshwater samples span up to six modules. In a post hoc search of 794 genomes, the element lies upstream of an RT at 18 of 19 hits, the 19th cut by a contig end. Yet relatives of the -1 and $+3$ genes occur as an adjacent pair with no RT detected nearby in 54 of 149 genera carrying a $+3$ -type gene, so why does the cassette keep the RT's company? The pairing is observed; two questions remain: did the RT join an existing module, and does it work with it, as DRT10 is reported to with an RNA and a protein partner?

1 Introduction

Reverse transcriptases (RTs) are common in bacteria, where they occur as group II introns, retrons, diversity-generating retroelements (DGRs), the UG/Abi families (RT groups of unknown function and of abortive-infection systems) and several other groups [1–4]. A survey of bacterial genomes grouped these RTs into 17 main groups [1], the UG/Abi family alone has since been divided into 42 groups [4], and RTs linked to CRISPR–Cas systems into 15 clades [5]; the counts depend on the classification used and are not additive. Many of these RTs serve host functions: retrons protect cells from phage [6], with their accessory genes predicted and classified computationally [7], and several UG/Abi and defence-associated RT (DRT) groups protect cells too [4,8], DGRs diversify target proteins [9], and RT–Cas1 fusions acquire CRISPR spacers from RNA [10].

Four preprints from three author teams (Sánchez-Nieto et al. 2026; Tang et al. 2025 and Ramirez et al. 2026, which share a senior author; Chen et al. 2026) describe DRT10 (UG17) as a tripartite system of a structured non-coding RNA (ncRNA), a SLATT (SMODS and SLOG-associating two-transmembrane) effector protein and an RT, and Sánchez-Nieto et al. and Tang et al. report all three components as

37 required for defence in the *E. coli* systems they tested [11–14]. Wang et al. (Cell 2026; author
38 manuscript) report selective synthetic lethality between DRT3 and loss of RecBCD (Δ recB) and
39 between DRT9 and loss of ExoI (Δ sbcB), and that the phage RecBCD inhibitor Gam triggers DRT3-
40 mediated abortive infection [8]. van den Berg et al. (bioRxiv preprint, 2025), working with
41 *Pseudomonas aeruginosa* gene clusters, describe bacterial systems that protect against phage as
42 highly modular, with sensing, signal-transmission and effector components frequently exchanged
43 among gene clusters [15].

44 In Bacteroidota, Toro (mSystems 2026) reports that 170 of 232 type XI-like retron candidates of the
45 SPIRE (Searchable planetary-scale microbiome REsource) representative collection are in
46 Bacteroidota, 167 of the 232 (72%) being low-confidence calls, and also computationally predicted
47 candidate msr/msd-like ncRNA motifs (msr and msd are the two elements of the retron ncRNA)
48 upstream of the RT in subsets of loci of two candidate lineages, with no experimental validation
49 reported [16]. We make no claim that the RT described here belongs to these retron-like candidates or
50 to the UG17 subtypes, and we describe it by comparison, composition and co-occurrence only.

51 Showing that an RT lies outside the known types is hard. RT sequences diverge deeply: a similarity
52 search cannot say whether a divergent RT is a remote member of a known family or a separate
53 lineage, and trees built from the RT core saturate. A claim that an RT is unassigned also depends on
54 the libraries that were checked: during this project, candidate groups that looked unassigned
55 matched published families that were missing from our library. We therefore treat unassigned as a
56 statement about specified comparisons and not as a claim that a lineage is new, and we ask a second
57 question that does not depend on the answer: what stands beside such an RT, and how far can that be
58 shown from public sequence data?

59 We began with a census of RT loci in 189,630 bacterial species-representative genomes of the Genome
60 Taxonomy Database (GTDB) release R232, in which 32 of 98 candidate groups remained unassigned
61 against the published libraries then available, and one of the 98 stood out for recurring neighbours of
62 the same kinds. We call its RT ESOL RT, after the working label ESOL (“enigmatic standalone ORF
63 lineage”), which carries no claim that the RT is a lineage; whether it is a separate lineage or a
64 divergent member of a known one stays open, and we describe it as unassigned. The RT keeps
65 company: we call the neighbourhood the ESOL cassette, an upstream non-coding element and
66 protein slots at positions -1 , $+1$, $+2$ and $+3$ relative to the RT open reading frame (ORF), defined by
67 position alone, with no claim of mobility or function. Throughout, ESOL RT means the single RT that
68 stands beside the cassette, which the figures label the cassette RT.

69 We describe an RT that none of our comparisons assigns to a named lineage, the cassette beside it, the
70 16 water samples (lake, river, pond and other waters in nine countries on five continents; the
71 continents are derived from the country field) from which the cassette-bearing metagenome-
72 assembled genomes (MAGs) come, and the Bacteroidota genus *Pollutibacter*, in which we find it
73 within the range we searched. The centre of the work is a puzzle: the cassette's parts also occur
74 without an RT in other genera, yet in the single genus where the cassette is found they keep an RT's
75 company, so we ask why. We tested four alternative explanations. A known RT type: no gate assigns
76 it (0 of 9 repeat pairs DGR-type on the contigs of 12 of the 14 members, 0 of 17 members with a group
77 II maturase domain, nearest published-library RT at most 34.6% identical among 30 ESOL queries). A
78 decaying relic: no observable module copy is disrupted. An assembly artefact: among reads from

79 samples and studies that gave no genome to our reference, 36 cross a junction between adjacent
80 cassette modules (11 in the frozen layer, which was fixed after the first hits were seen, and 25 in post
81 hoc layers added later). A detection gap: in a reference-pool emulation matched on identity and not
82 on alignment coverage, the unchanged tool thresholds called 49 of 49 known units but 0 of 16 cassette
83 RT units, a contrast that is not evidence in either direction, so this explanation stays open. What
84 remains after these tests is a puzzle that we state as two separate questions, whether the RT joined a
85 module that already existed and whether it works with that module, and for each we name the
86 observation that would decide. We state what the evidence shows (the cassette's parts, their
87 arrangement and intact state, and the company the RT keeps) and what it leaves open (where that
88 company comes from and what it does), and no wet-laboratory result is reported here, although wet-
89 laboratory experiments are under way.

90 **2 Results**

91 A reverse transcriptase that none of our comparisons assigns to a known family keeps the same
92 company of genes (Figure 1) in 16 genomes from 16 water samples (lake, river, pond and other
93 waters) in 9 countries on five continents, yet all 16 belong to one genus, *Pollutibacter*. The cassette is
94 therefore found across five continents but within a narrow lineage: within the range we searched it
95 lies in one genus, *Pollutibacter*, whose close relatives count as one lineage in any test of independent
96 recurrence (Figure 2), and the open question is why this RT and this company stay together. The
97 results follow the logic of the question: what the RT is, what stands beside it, whether that company
98 is real and intact, where in the world it turns up and in which genus, and why it might be there; Table
99 1 defines the terms we use.

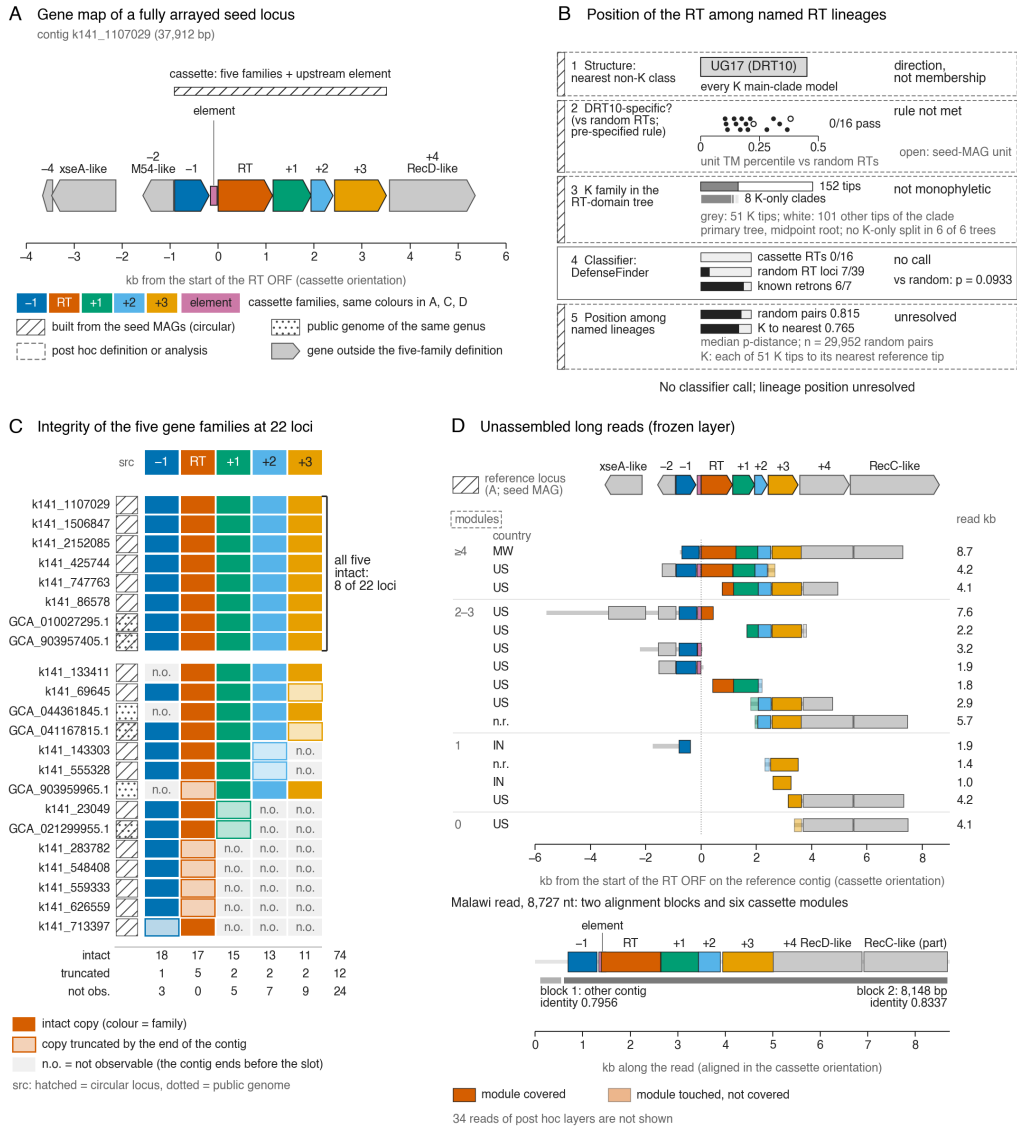


Figure 1. The cassette RT and the gene families beside it: a gene map, the position of the RT among named RT lineages, integrity of the gene families at assembled loci and cassette-specific sequence on unassembled long reads. A, Gene map of one seed MAG locus whose contig spans the whole array (contig k141_1107029, 37,912 bp), drawn from the annotation table in the cassette orientation, with positions in kb from the start of the RT ORF. Hatching marks objects of the 16 seed MAGs from which the family models were built, so the genes of this locus are circular by construction, and the 16 MAGs were selected for carrying the cassette and belong to one genus. B, The RT compared with named RT lineages by five routes, with the result of each; dashed outlines mark post hoc definitions or analyses and hatching marks comparison sets built from the seed MAGs. C, State of the five gene families (−1, RT, +1, +2, +3) at 22 loci of one genus, 16 seed MAG loci and 6 public genomes, of which 20 loci are flagged circular in Supplementary Data S1 (the 16 seed MAG loci and 4 of the public genomes; hatched in the src column): of 110 slots, 74 hold an intact copy, 12 a copy truncated by the end of the contig and 24 are not observable because the contig ends before the slot. D, Unassembled long reads: each of the 15 reads of the frozen layer is drawn on the cassette coordinates of the reference contig to which its main alignment block aligns, anchored at the start of the RT ORF and grouped by the number of cassette modules covered by that

114 block; one read of 7.6 kb is placed in the 2–3 group because a second alignment block on another contig, not
115 drawn, adds a fourth module, so four of the 15 reads cover four or more modules in Figure 7C and
116 Supplementary Data S2 (country codes: MW, Malawi; US, United States; IN, India; n.r., not recorded).

117 Alt text: Four-panel figure on the cassette RT. A, a gene map of one seed locus on contig k141_1107029
118 (37,912 bp), with an axis in kb from the start of the RT ORF and genes from xseA-like to RecD-like;
119 the five families are coloured and the hatched cassette bar marks objects built from the seed MAGs
120 (circular). B, five boxed comparisons of the RT with named RT lineages, with results from “direction,
121 not membership” to “unresolved”; dashed outlines mark post hoc analyses. C, a grid of 22 loci by
122 five families: 74 intact, 12 truncated, 24 not observable. D, 15 long reads grouped by modules
123 covered: ≥ 4 , 2–3, 1 and 0.

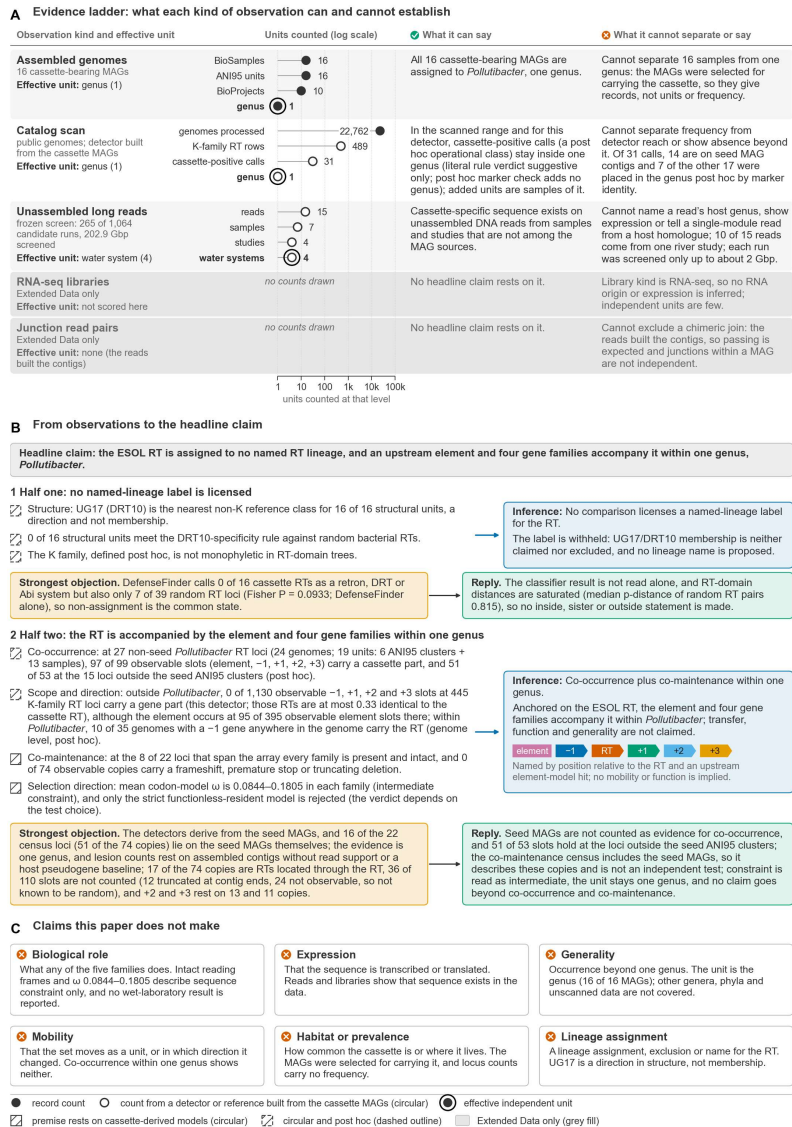


Figure 2. Evidence ladder for the ESOL RT and its cassette: what each kind of observation can say, what it cannot separate, and how the headline claim follows from it. In panel A each ladder gives the counting units of one kind of observation on a log scale, with counts kept per layer and never added across layers; cassette-positive calls are a post hoc operational class (a -1 gene on the strand of a K-family RT); filled symbols are record counts, open symbols are counts made with a detector or reference built from the cassette-bearing MAGs, and a ring marks the effective independent unit. The two grey-filled rungs without counts are marked Extended Data only: the RNA-seq counts are shown in the Extended Data and in Section 2.6, no headline claim rests on them, and the junction read pairs are shown in the Extended Data only. Panel B follows the headline claim through a chain for each part of it, each with the strongest objection and our reply beneath it; hatched squares mark premises that depend on models built from the cassette-bearing MAGs, and dashed outlines mark post hoc premises. Panel C lists six claims that this paper does not make: biological role, expression, generality beyond one genus, mobility as a unit, habitat or prevalence, and a lineage assignment or name.

Alt text: Three-panel evidence figure. A, a table of observation kinds with units counted on a log scale from 1 to 100k, what each can say and what it cannot separate; genus (1) is the effective unit for assembled genomes and the catalog scan, water systems (4) for long reads, and two rows marked Extended Data only have a grey fill and no counts: the RNA-seq counts are shown in an Extended Data figure and in Section 2.6, and no headline claim rests on them. Filled dots are record counts, open circles are counts from a detector or reference built from the cassette MAGs, and rings mark the effective independent unit. B, premises leading to the headline claim, each half with a strongest objection and a reply; hatched squares mark circular premises, dashed outlines post hoc ones. C, six boxed claims this paper does not make.

2.1 An orphan by our comparisons: an RT that none of them assigns to a known family

We found the RT by asking what the known RT families leave behind: of 122,323 RT loci in 189,630 bacterial species-representative genomes of GTDB R232 [17], 109,524 matched a known RT family, 8,178 could not be measured and 4,621 reached a library entry but no family (diverged); 3,750 of these formed 98 candidate groups, 32 of which the census partition left unassigned, and one of the 98, ESOL, stood out for recurring neighbouring genes of the same kinds (Supplementary Table S1, table H; Supplementary Table S1, table I).

Table 1. Terms for the RT, the cassette and the sets of related RTs.

Term	Refers to	Defined by
ESOL RT	The RT that anchors the ESOL cassette (called the cassette RT in some panels and legends)	The RT encoded beside the cassette in the cassette-bearing MAGs
ESOL cassette	The arrangement around that RT	Position alone: an upstream non-coding element and protein slots at -1, +1, +2 and +3 relative to the RT; a RecD-like gene at +4 is noted and not counted as a cassette part
Module (part)	One position-defined part of the ESOL cassette: the upstream element or the protein at -1, +1, +2 or +3	Position alone; "module" and "part" are used for the same thing and imply no joint movement, joint origin or function
Cassette-bearing MAG (seed MAG)	One of the 16 MAGs from which the cassette models were built	Selected because the assembly carries the cassette contig of the study's module table; a selection and not a detection, so circular; 14 of the 16 are also cassette-positive calls
Cassette-positive call	A K-family RT locus with a -1 gene on its strand in the catalogue scans (31 of the 489 K-family RT rows of the nucleotide-window catalogue; Figure 8, Methods); called cassette-type in the supplementary tables	A -1 gene on the strand of a K-family RT: a post hoc operational class that does not require the upstream element or the +1, +2 and +3 genes, so a call is not a complete cassette
Cassette RT clade	RTs closely related to the ESOL RT	Identity to the ESOL RT; in the range we searched, confined to one genus, <i>Pollutibacter</i>
K family (K-family RTs)	A wider post hoc set of related RTs, mostly distant relatives, that includes the ESOL RT; not a clade	A profile score and a paired-identity rule; K is an arbitrary label of our analysis tables and not a lineage name
ESOL RT group	RTs that hit the ESOL RT profile in our census	Profile hits above a score and length threshold

The sets differ in breadth and are not alternative names for one thing: the ESOL RT is a single RT, the cassette RT clade is its close relatives, and the K family is a much wider post hoc set that includes both; the ESOL RT group and the K family were defined separately and overlap by identifier mapping only.

Against a library of retron, UG/Abi and Toro 2014 RTs [1,4,7] the nearest identity of the 30 ESOL queries (14 census-group members, 3 further genome members and 13 metagenomic RTs) has a maximum of 34.6% (31.7% for the census-group members alone), and against 306 retron-RT candidates assigned to Bacteroidota in a recent metagenome survey (39% of them high-confidence) [16] no comparison places an ESOL member in a retron group (Supplementary Table S2, table M).

Trees and predicted structures give no better answer: in sequence trees the K-family RTs, a wider post hoc set of related RTs that includes the ESOL RT (Table 1), are not monophyletic; in predicted structures their main clade points toward UG17 (DRT10) as a direction and not a membership; and the position of the RT in sequence trees is unresolved (Figure 3). The nearness to DRT10 is not

specific: against a random bacterial RT null no structural unit passes the pre-set rule (the rule on the template-modelling score, TM-score), although the positive-control map is dominated by retrons and the rule has low sensitivity.

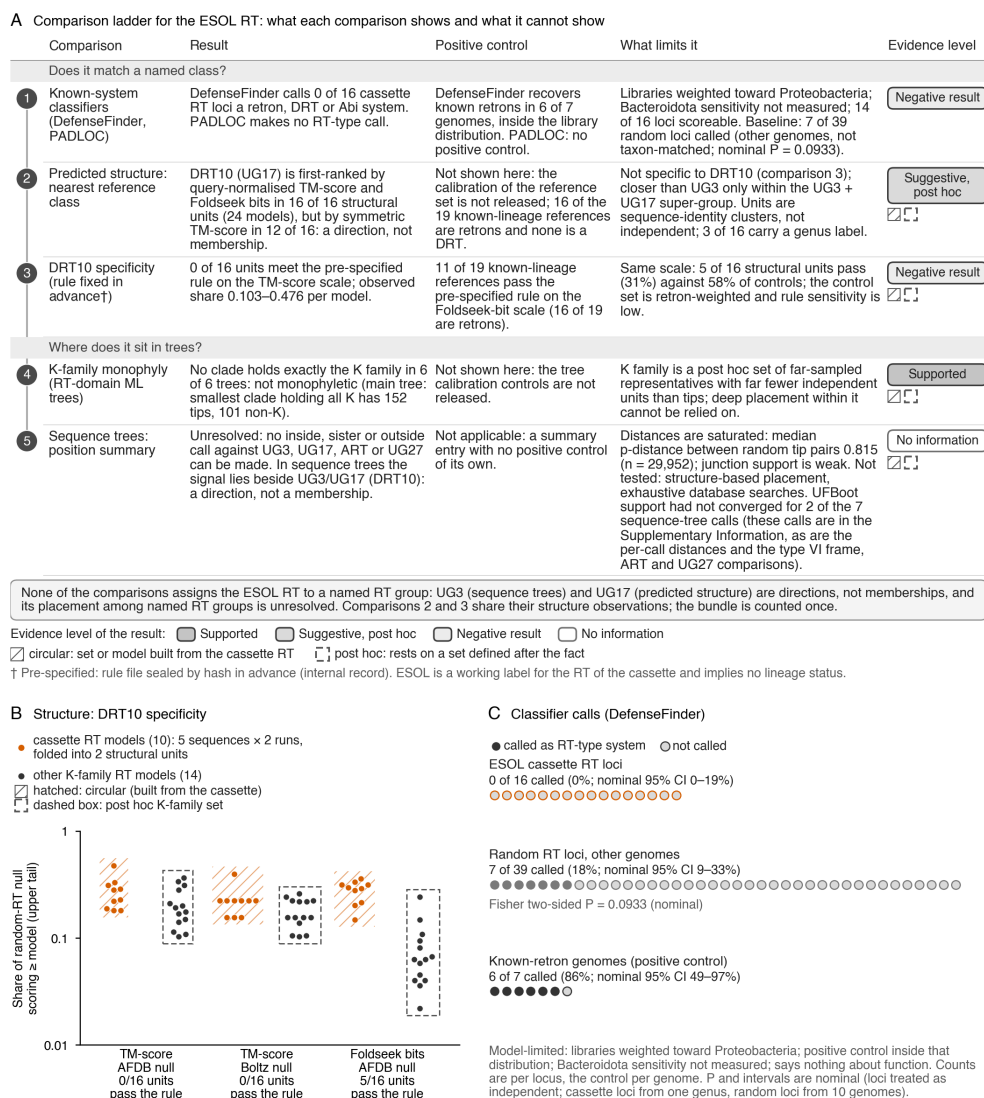


Figure 3. What each comparison shows and cannot show about the position of the ESOL RT among named RT groups. Panel A lists five numbered entries, four comparisons of the ESOL RT, the RT of the 16 cassette RT loci, with named RT groups and a fifth that summarises its sequence-tree position, and gives for each entry its result, its positive control where this figure shows one, and what limits it. Each entry carries the evidence level of the result it rests on, with hatching where the result is circular and a dashed outline where it rests on a set defined after the fact. The K family, to which the cassette RT belongs by a post hoc definition, is not monophyletic in RT-domain maximum-likelihood trees. None of the 16 structural units of the K family meets the DRT10-specificity rule on the two TM-score scales (panel B); on the secondary Foldseek-bit scale 5 of 16 units pass, none of them a cassette-lineage unit. DefenseFinder calls none of the 16 cassette RT loci a retron, DRT or Abi system (panel C), and PADLOC makes no RT-type call either. In sequence trees the signal lies beside UG3 and UG17 (DRT10), and in predicted structures UG17 is the nearest non-K class, each a direction and not a membership.

Alt text: Three-panel figure on the position of the ESOL RT among named RT groups. A, a ladder of five numbered entries in two groups, “Does it match a named class?” and “Where does it sit in trees?”, each giving result, positive control, what limits it and an evidence-level badge (Negative result, Suggestive, post hoc, Supported or No information); hatched squares mark circular entries and dashed squares post hoc ones. B, dot plots on a log axis from 0.01 to 1 for three scales, on which 0/16, 0/16 and 5/16 units pass the rule; cassette RT models are orange. C, rows of circles for classifier calls: 0 of 16, 7 of 39 and 6 of 7.

A DRT10 reading has two sides, and we give both. In a post hoc stratification of K-family RT loci in GTDB R232 representative proteomes by nearest reference class (assigned from RT sequence identity alone), a same-strand SLATT-type gene beside the RT was far more frequent at loci whose nearest class is UG17 (DRT10) than at other K-family units, and it was not seen in matched null windows or beside non-K RTs; the K-family list is itself derived from the cassette, and the literature set of DRT10 RTs, which is defined by SLATT association, gives no independent support. Against this reading, no SLATT-type gene was detected beside the few assessable loci and genus units of the K main clade that contains the cassette (a low-power negative), no SLATT-family protein was found in the proteomes of the cassette-bearing MAGs, and the ESOL RT shows no C-terminal serine or tyrosine priming residue of the kind that preprints report for DRT10 RTs [12–14], although absence in incomplete assemblies does not exclude a divergent partner. These are descriptive comparisons within one genus: they do not place the cassette RT in DRT10 and do not show that it works with such a module. So the RT is an orphan: the predicted structures of the main K clade sit nearer UG17 than UG3, but no gate, tree or structure assigns the RT to a named lineage, and we infer no new lineage from them (Question 1); the company it keeps is the better clue. Further detail on the census group, the post hoc sister-candidate comparisons and the upstream element is in Supplementary Results S1.1.

2.2 The RT keeps company: an upstream element and four genes at fixed positions

Many RT systems are encoded together with their partners, as DRT10 is with its ncRNA and SLATT protein, so we asked what stands beside the orphan, and found a company. We define the ESOL cassette by position alone: an upstream non-coding element and protein slots at –1, +1, +2 and +3 relative to the RT, drawn for one fully arrayed seed locus in Figure 1A and set out module by module in Figure 4 and Table 2 (the component cards are in Supplementary Table S3), with a RecD-like gene at +4 that we do not count among the modules. Around the RT stand a conserved upstream element, a protein at –1 that carries a MIDAS (metal-ion-dependent adhesion site) motif in every –1 ORF we could evaluate, proteins at +1 and +2, a protein predicted to fold like a metalloprotease at +3, and a RecD-like helicase at +4 whose Walker A/B motifs are intact in every island copy we could evaluate (in the 8 *Pollutibacter* genomes with a complete +4 ORF it is the only RecD-like protein found in the assembly; a second copy on an unassembled contig is not excluded). Among the complete ORFs of the 16 seed contigs, none of the +1, +2 or +3 proteins matches an entry of the Pfam, NCBIfam or CDD libraries, which is a result at the level of the annotation libraries and not a statement about their folds (the third layer of Figure 4 and Table 2, 3Di, is a Foldseek search of predicted structures in the 3Di structure alphabet, and its codon-model columns use the site-wise test FEL, the fixed effects likelihood method, and the recombination screen GARD, the genetic algorithm for recombination detection).

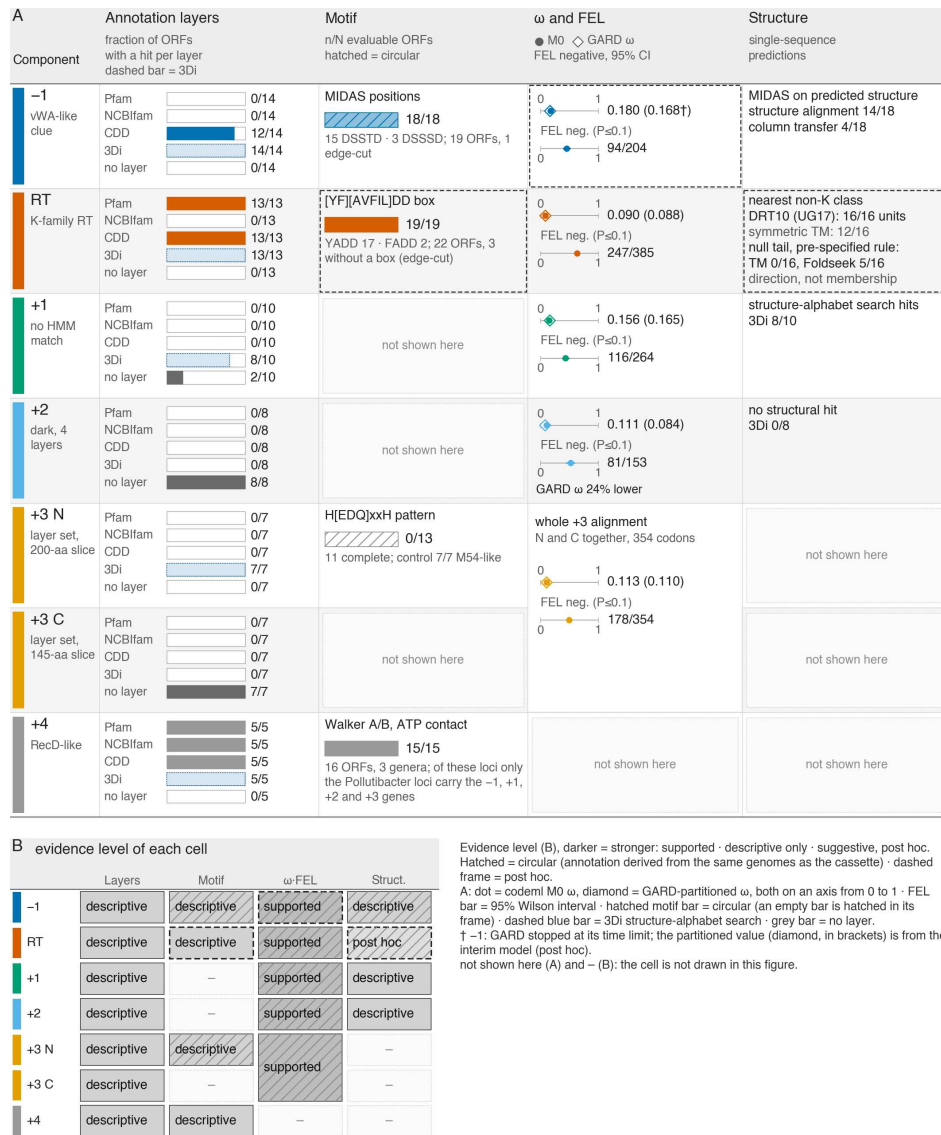


Figure 4. Component card mosaic of the ESOL cassette, with one row for each of the -1, RT, +1 and +2 genes, the N and C parts of the +3 gene and the RecD-like +4 gene. Panel A gives for each row the annotation layers searched, the motif counts, the codon-model ω with FEL and the structure-level reading. Panel B gives the evidence level of every cell of panel A with its circular and post hoc marks. Evidence levels are shown by grey shading that is darker for a stronger level (three levels: supported, descriptive only, and suggestive, post hoc), hatching marks a circular value, and a dashed frame marks a value that rests on a post hoc result. In panel A the dashed frames mark the RT structure cell (the K family is a post hoc set), the RT catalytic box (the pattern was widened after two copies had been seen) and the -1 ω cell, whose partitioned value (+) comes from an interim recombination model. In panel A a dashed bar is the 3Di structure-alphabet search, the frame of an empty motif bar is hatched when its value is circular, and a cell marked not shown here is not drawn in this figure.

Alt text: Two-panel component card mosaic of the ESOL cassette. A, a table with one row per component (-1, RT, +1, +2, +3 N, +3 C and +4) and four columns: annotation layers, with bars for the fraction of open reading frames with a hit, such as 12/14; motif counts, such as 19/19; codon-model ω with FEL, drawn as a dot (M0) and a diamond (GARD ω), for example 0.180 (0.168†) for -1; and

single-sequence structure predictions. B, a grid of the evidence level of each cell (descriptive, supported or post hoc), in darker grey for stronger levels, with hatching for circular values. Dashed frames mark post hoc results: the RT motif and structure cells and the -1ω cell.

Table 2. Composition of the ESOL cassette by position relative to the RT open reading frame (ORF), with the RecD-like +4 gene shown as auxiliary (not a module) and two neighbouring genes outside the cassette for orientation.

Position	Gene family (annotation layers Pfam / CDD / 3Di)	Protein length range in aa (n ORFs)	Observable copies (intact / total)	Motif results (n/N)	Codon-model omega range
Upstream element	element-model hit (nucleotide profile); no ORF assessed	not applicable (no ORF); element-model hit 138–141 nt in each of 16 seed windows (the call at the seed loci of the co-occurrence census gives 48–141 nt, one hit being cut by the window, in Supplementary Table S3)	not applicable (no ORF)	not applicable	not applicable (no ORF)
–1	vWA-like (clue); Pfam 0/14, CDD 12/14, 3Di 14/14	198–239 (14 of 15)	18/18 (22 loci)	MIDAS D-S-S-[T/S]-D 18/18 (DSSTD 15, DSSSD 3)	0.168–0.180 (15 seq.)
RT (anchor)	RT (RVT_1); Pfam 13/13, CDD 13/13, 3Di 13/13	377–425 (12 of 16)	17/17 (22 loci)	aromatic-x-DD box 19/19 (YADD 17, FADD 2; pattern widened after the FADD copies were seen)	0.088–0.090 (16 seq.)
+1	no family assigned; Pfam 0/10, CDD 0/10, 3Di 8/10	261–294 (10 of 11)	15/15 (22 loci)	not testable (no reference motif)	0.156–0.165 (14 seq.)
+2	no family assigned; Pfam 0/8, CDD 0/8, 3Di 0/8	152–163 (8 of 10)	13/13 (22 loci)	not testable (no reference motif)	0.084–0.111 (12 seq.)
+3, N-terminal part	metzincin-like (clue); Pfam 0/7, CDD 0/7, 3Di 7/7	355–372 (7 of 8; whole +3 ORF)	11/11 (22 loci; whole +3 ORF)	H[EDQ]xxH absent 12/12 in the N-terminal test (absent 13/13 over the full ORF)	0.110–0.113 (11 seq.; whole +3 alignment)
+3, C-terminal part	no family assigned; Pfam 0/7, CDD 0/7, 3Di 0/7	355–372 (7 of 8; whole +3 ORF)	11/11 (22 loci; whole +3 ORF)	not applicable	0.110–0.113 (11 seq.; whole +3 alignment)
+4, RecD-like (auxiliary, not a module)	RecD-like; Pfam 5/5, CDD 5/5, 3Di 5/5	592–622 (5 of 6)	8/8 (22 loci)	Walker A/B and ATP-contact class intact 8/8	0.084–0.086 (8 seq.)
Adjacent genes outside the cassette (slot –2 M54-like; slot –3 xseA-like)	–2 M54-like (Pfam 5/5, CDD 5/5, 3Di 5/5); –3 xseA-like (clue only; no layer search)	–2: 208–213 (5 ORFs); –3: not recorded	not recorded	–2: H[EDQ]xxH present 7/7 (positive control); –3: not recorded	not recorded

The positions are operational and assign no activity or biological role, and a term such as vWA-like or metalloprotease-like is a clue and not an assignment; the xseA-like term is an annotation label read from a library model and was not tested. Observable copies and motif counts refer to the assembled loci of one genus. The columns refer to different sets, so their denominators differ (for the cassette positions, length ranges, layer counts and the element length come from the annotation of the 16 seed contigs, observable copies from the 22 loci), and the ω range spans the unpartitioned and the recombination-partitioned codeml M0 estimates; the lower end of the -1 range (0.168) is the partitioned value of an interim recombination model and is post hoc. The calls of every ORF in the three layers are in Supplementary Table S3, file SI-04_layer_calls_per_orf.

In an alignment of 15 upstream sequences from other genera and from metagenomic assemblies, the most conserved stretch of 60 consecutive columns lies 25 to 84 positions upstream of the RT start, with a mean frequency of the most common base of 0.77 against 0.72 over all 141 columns (Extended Data Figure 1). The identity of the element is unresolved (Question 6) and its form description is not diagnostic; in the scans of host-lineage genomes described in Supplementary Results S1.2, element-plus-RT pairs are found in 20 genomes, 15 species and 4 genera (Supplementary Table S4, table A20), a count that comes from those scans and not from the 794-genome search described later.

Every evaluable -1 ORF carries the MIDAS (metal-ion-dependent adhesion site) motif (18 of 18): in 14 the five positions were read by structure mapping onto ESMFold-predicted models [18] and in 4 by transfer of the aligned columns from those mappings (Extended Data Figure 2). The $+3$ protein is predicted to fold like a metalloprotease, yet the canonical HExxH motif (searched as H[EDQ]xxH) is absent from all 13 of its ORFs (11 of them complete), whereas the same motif is present in all 7 Peptidase_M54-like ORFs from the same loci and in human MMP1; the absence is a property of the whole $+3$ family (nearly all free-standing $+3$ -like sequences also lack it: 2 of 127, one per genus, carry it; Supplementary Table S5, table T10, rows H-8 and H-9), and whether the protein is an active protease is an open question (Question 4). A profile-to-profile search did not support the $+3$ N-terminal domain as a catalytic-site-loss variant of an M54 peptidase (it sees only relationships about as close as that between the -1 gene and vWA-type proteins), and a post hoc comparison of predicted structures resembles an archaemetzincin-type fold, so the $+3$ protein is described by a predicted-structure resemblance and not by an assigned family. In a pre-set structure-level comparison, the cassette -1 and $+3$ proteins scored below the lower quartile of free-to-free pairs of RT-free -1 -type and $+3$ -type proteins, so both pre-set rules were only partly met, although the scores exceeded those of unrelated folds. Further detail on the upstream element's structure and covariance-model scan, the catalytic and helicase motif integrity and the structure annotation of the -1 protein is in Supplementary Results S1.2.

2.3 The company is found with the RT inside one genus, and nowhere else we can see

If the company is a unit of the RT, it should stand beside the RT wherever the RT occurs, so we anchored on the RT and counted what stands in each slot, first inside Pollutibacter, then outside it, and then from the side of the parts (Figure 5). The RT does not stand alone in Pollutibacter: at 27 further RT loci in 24 genomes, beyond the 16 seed assemblies, 97 of 99 observable neighbour slots (upstream element, -1 , $+1$, $+2$, $+3$; a slot beyond a contig end is not scored) carry the cassette part and the 2 absences sit at a single locus; after removing every genome of the ANI95 clusters (clusters of genomes by average nucleotide identity, ANI; Section 4.2) that also contain a seed MAG, the 15

284 remaining loci (14 genomes, 14 units) still carry the element in 9 of 9, -1 in 7 of 7, +1 in 14 of 15, +2 in
285 11 of 12 and +3 in 10 of 10 (post hoc). The locus with the two absences has an RT only 0.388 identical
286 to the cassette RT; its upstream element, -1 and +3 positions are unobservable, and hidden deletions
287 could not be checked.

288 Outside *Pollutibacter*, for this detector and this scanned range, no cassette part is found beside a K-
289 family RT: among 445 K-family RT loci, 0 of 334 observable -1, 0 of 305 +1, 0 of 258 +2 and 0 of 233 +3
290 slots carry a cassette part (0 of 1130 slots in all). The outside zero is not evidence of absence: nearly all
291 of the 445 outside loci have an RT of low identity to the cassette RT, so for those loci the zero cannot
292 be told apart from the detector failing to reach so low an identity. Two profile detectors that do not
293 use the cassette profile hidden Markov models (HMMs) find the +3-type protein at none of the 445
294 outside loci and at none of 1,127 control windows around non-K RVT_1 RTs; the leave-one-out family
295 profile detects every real +3-type pair-family member it was tested on in the two identity bands (18 of
296 18 and 39 of 39; Supplementary Table S4, table R01; Supplementary Table S4, table R02), so for the +3
297 part the zero is an informative absence for these detectors, these loci and these windows only. Hits of
298 the upstream-element model do occur beside K-family RTs outside *Pollutibacter* (95 of 395 outside
299 loci with an observable element window), but at those 95 loci none of the observable -1, +1, +2 or +3
300 slots carries a cassette part (none of 54, 59, 45 and 39 observable slots), so the element-plus-RT layer
301 and the cassette-part layer are counted separately.

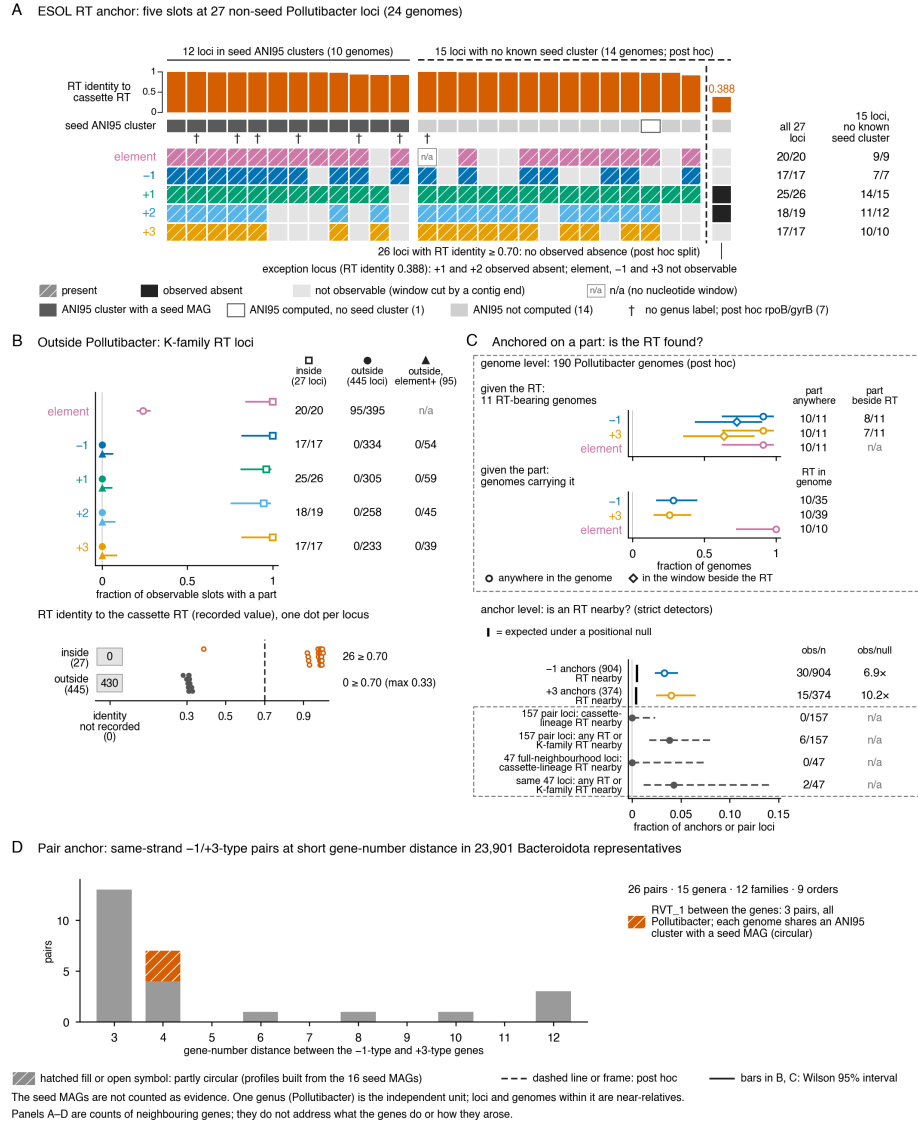


Figure 5. The companions of the ESOL RT are counted inside Pollutibacter (A), outside it (B), from the side of the parts (C) and in a scan of -1/+3-type pairs (D), and the layers are shown side by side and not summed. Hatched fills and open symbols mark values that are partly circular because the detector profiles were built from the 16 seed MAGs, dashed lines and frames mark post hoc analyses, and the horizontal bars of panels B and C are Wilson 95 percent intervals. (A) At 27 non-seed Pollutibacter loci (24 genomes) the observable slots carry the upstream element in 20 of 20, the -1 gene in 17 of 17, +1 in 25 of 26, +2 in 18 of 19 and +3 in 17 of 17. (B) Outside Pollutibacter, for this detector and this scanned range, no cassette part is found beside a K-family RT: 0 of 334 observable -1, 0 of 305 +1, 0 of 258 +2 and 0 of 233 +3 slots at 445 loci carry one. The strip under panel B gives the recorded identity to the cassette RT of each locus: 26 of the 27 loci inside Pollutibacter reach 0.70 or more, none of the 445 loci outside does (maximum 0.33), and 430 of the outside loci have a recorded identity of 0, which can mean not reported. (C) In Pollutibacter genome units (190 genomes: 186 GTDB R232 Pollutibacter genomes and 4 unlabelled near-relatives; post hoc), the 11 RT-bearing genomes carry a strict -1 gene, a strict +3 gene and the element somewhere in the genome in 10, 10 and 10 of them. At the anchor level a strict -1 or +3 gene has an RT of any kind nearby in 30 of 904 and 15 of 374 anchors, 6.9 and 10.2 times a positional null, and none of 157 re-detected -1/+3 pair loci (0 of 47 with a full neighbourhood) has a strict cassette-lineage RT (any

RT: 6 of 157 and 2 of 47); the pair-locus rows are post hoc (dashed). (D) Over 23,901 GTDB R232 Bacteroidota species representatives, the pre-specified strict detectors find 26 same-strand $-1/+3$ -type pairs (15 genera, 12 families, 9 orders), of which 3 have an RVT_1 domain between the genes.

Alt text: Four-panel figure counting the companions of the ESOL RT. A, a grid of 27 *Pollutibacter* loci by element, -1 , $+1$, $+2$ and $+3$ slots, in a group of 12 loci in seed clusters and a post hoc group of 15, with hatched cells present, black cells observed absent and pale grey cells not observable; totals include 20/20 for the element and 25/26 for $+1$. B, outside *Pollutibacter*, 445 loci give 0 of 334 observable -1 slots with a part, against 17/17 inside, and a strip of recorded identities marks 430 outside loci as “identity not recorded (0)”. C, a post hoc genome-level block and anchor-level rows, with four dashed-frame pair-locus rows (0/157, 6/157, 0/47, 2/47). D, a bar chart of 26 pairs by gene-number distance, from 3 to 12.

Within *Pollutibacter* the parts are carried by the RT-bearing genomes more often than the RT is carried by the part-bearing genomes: of 11 RT-bearing genomes among 190 examined, 10 carry a strict -1 gene, whereas of 35 genomes with a strict -1 gene only 10 carry the RT, and $+3$ shows the same asymmetry (10 of 39 genomes with a $+3$ gene carry the RT). These genome-level counts come from the same genomes as the anchored census, so they are one observation counted again and not independent evidence. Further detail on the census scope, the stratification by RT identity, the detector reach and control windows, and the genome-level coexistence count is in Supplementary Results S1.3.

2.4 No copy we can see is disrupted, and the codon models are consistent with purifying constraint

A cassette that is decaying should carry broken copies, and one that is kept should not, so we scored every observable copy of the five families for frameshifts, premature stops and truncating deletions and asked how strongly selection constrains their codons (Figure 6; Extended Data Figure 3, Extended Data Figure 4 and Extended Data Figure 5 give variants and controls). Across every observable module instance, the assembled contigs show no frameshift, premature stop or truncating deletion; the observed defect counts sit far below the neutral-drift expectation, and the detection positive control passed, so the strict form of the decay prediction is rejected (Figure 6; Supplementary Data S1). This rejection rests on an exact calculation of premature stop codons on the observed trees and holds only while the stop frequency exceeds 0.29 to 1.10 per cent depending on the series (Supplementary Table S6, file SI-10_A6_break_even_stop_frequency): with the stop frequency set by the observed coding composition (no stop codon observed; 0.022 to 0.074 per cent) the probability that every leaf is free of premature stop codons is 0.35 to 0.75 in all 20 cells and nothing is rejected, so the calculation gives no information on a model in which stop gain is suppressed by selection or by a mutation bias that the composition does not capture. In post hoc sensitivity checks, dropping any one of the 10 studies behind the 16 MAG loci leaves at least 48 of the 57 observable neighbour-module copies, none of them disrupted (two-sided exact upper limit 0.0740 at 48), and keeping one locus per study together with the 6 public-genome loci leaves 37 to 46 copies, none disrupted (upper limit 0.0949 at 37); these checks change the denominator and not the count of disrupted copies.

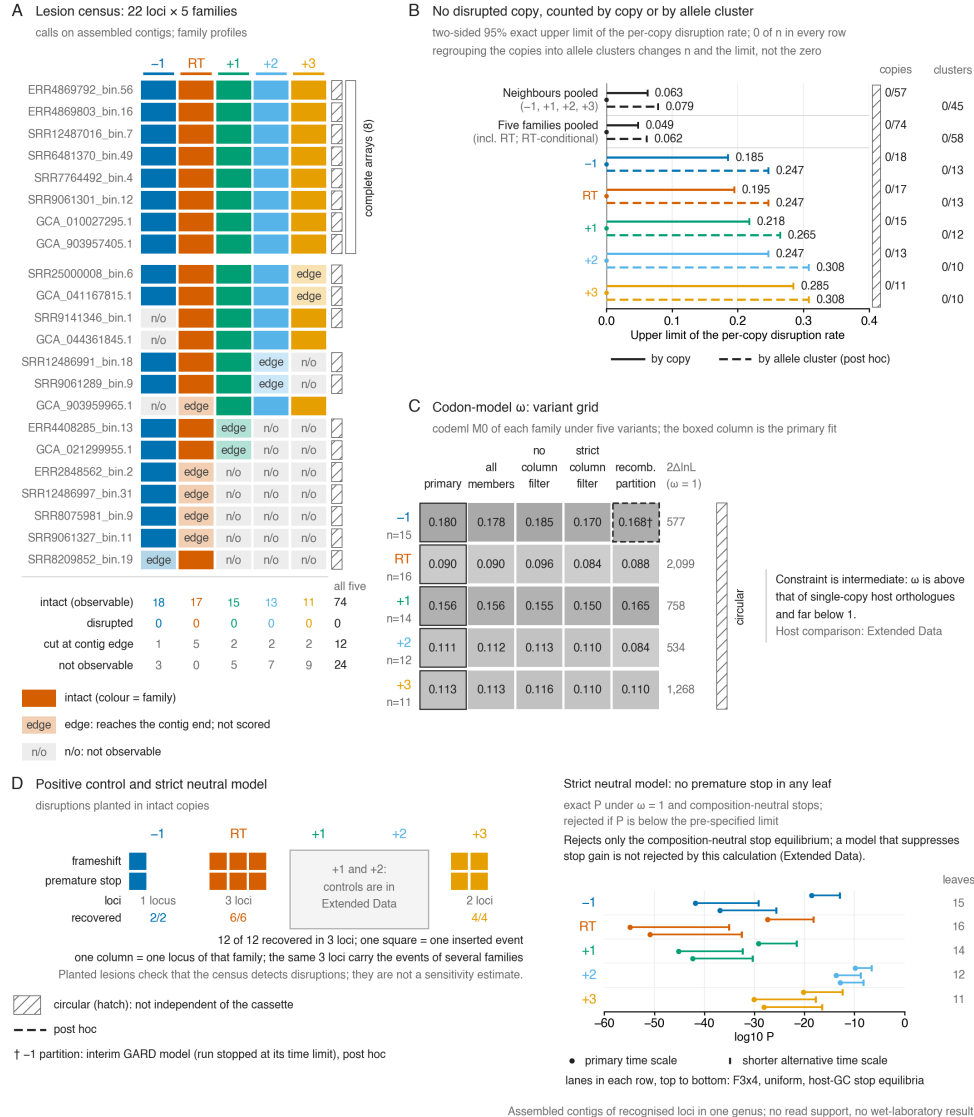


Figure 6. No frameshift, premature stop or truncating deletion is seen in any observable copy of the five cassette families at 22 assembled loci of one genus, and the codon-model ω (codeml M0) of the five families lies between 0.084 and 0.185 in each of the five variants shown in C. A, Each of 22 loci (16 MAG loci and six public-genome loci; rows, complete arrays first) is scored for the five cassette families (columns) as intact, cut by the contig end (edge) or not observable (n/o). B, Exact two-sided 95% upper limits for the per-copy disruption rate with 0 of n copies disrupted, drawn solid by copy and dashed by allele cluster (post hoc). C, Codon-model ω (codeml M0) of each family for the ANI95 representative set with alignment columns kept at more than 50% occupancy (trim50; primary, boxed, with its number of sequences n at left), all members, no column filter, gap-free columns only (strict) and GARD recombination partitions (codon-length-weighted mean of segment-wise ω ; for -1 the partitions come from an interim GARD model, because the run stopped at its time limit, and are post hoc, †). Hatching in A, B and C marks circular items, because the family profiles were built from the cassette MAGs; in A the mark is on the 20 of 22 loci flagged circular in Supplementary Data S1 (the 16 seed MAG loci and 4 public genomes), and the bracket marks the 8 complete arrays. D, Left, a six-frame detector recovered 12 of 12 planted lesions, which checks the detector wiring and gives no estimate of its sensitivity at other positions, and the controls for +1 and +2 are in Extended Data Figure 5. D, Right, an exact pruning calculation (not a simulation) on

the observed tree of each series gives the probability P that every leaf is free of premature stop codons at every site under the strict neutral model ($\omega = 1$ for every non-synonymous change, stop gain included, stop codons at a composition-neutral equilibrium; the 30 cell values are in the source data of this figure).

Alt text: Four-panel figure on lesions in the five cassette families. A, a grid of 22 loci by -1 , RT, $+1$, $+2$ and $+3$, coloured where intact, with edge for cut by the contig end and n/o for not observable; a bracket marks complete arrays (8); totals 74 intact, 0 disrupted, 12 cut at contig edge and 24 not observable. B, bars of upper limits of the per-copy disruption rate, from 0.049 to 0.308, solid by copy and dashed (post hoc) by allele cluster. C, codon-model ω for five variants, from 0.084 to 0.185, with a dashed frame on 0.168† for the -1 recomb. partition. D, 12 of 12 planted lesions recovered, and log₁₀ P intervals from -60 to 0 .

In codon models the ω of every module lies far from the neutral value in the purifying direction, which is consistent with purifying selection, with negative sites at many codons and, in FEL without synonymous-rate variation, at most one positive site (in -1 , unpartitioned only; with synonymous-rate variation at most two per family), and the $+3$ family, which lacks the catalytic H[EDQ]xxH motif, shows the same direction while its function stays undetermined (Supplementary Table S6; Extended Data Figure 3). In a post hoc check of the FEL output for the $+3$ alignment, the synonymous rate α equals exactly 1.000 at 349 of 354 sites in each of two HyPhy builds, so site-wise synonymous rates are not resolved there and its negative calls are β below 1.000 under α fixed at 1.000; the cause of the α pattern is undetermined. The constraint is intermediate.

2.5 Single unassembled long reads carry several modules of the cassette

Sequence seen only in assemblies could be an artefact of assembly, so we asked whether the cassette exists on single unassembled long reads from public sequencing runs, and whether such reads could be chimeras (Figure 7; Extended Data Figure 6, panel D). In the frozen layer of public freshwater long-read data, 15 unassembled reads from 7 samples in 4 studies match the cassette; none of these samples or studies contributed a genome to the reference, a reversed-sequence decoy produced no hit, the reads were re-checked with BLASTN, minimap2 and strict protein blastx, and their host genus was not determined. Pooled over all strata, single public long reads cover all six cassette modules and reach beyond the cassette on the recC/recD side; these post hoc strata were chosen after hits were seen and do not fit random-sample statistics.

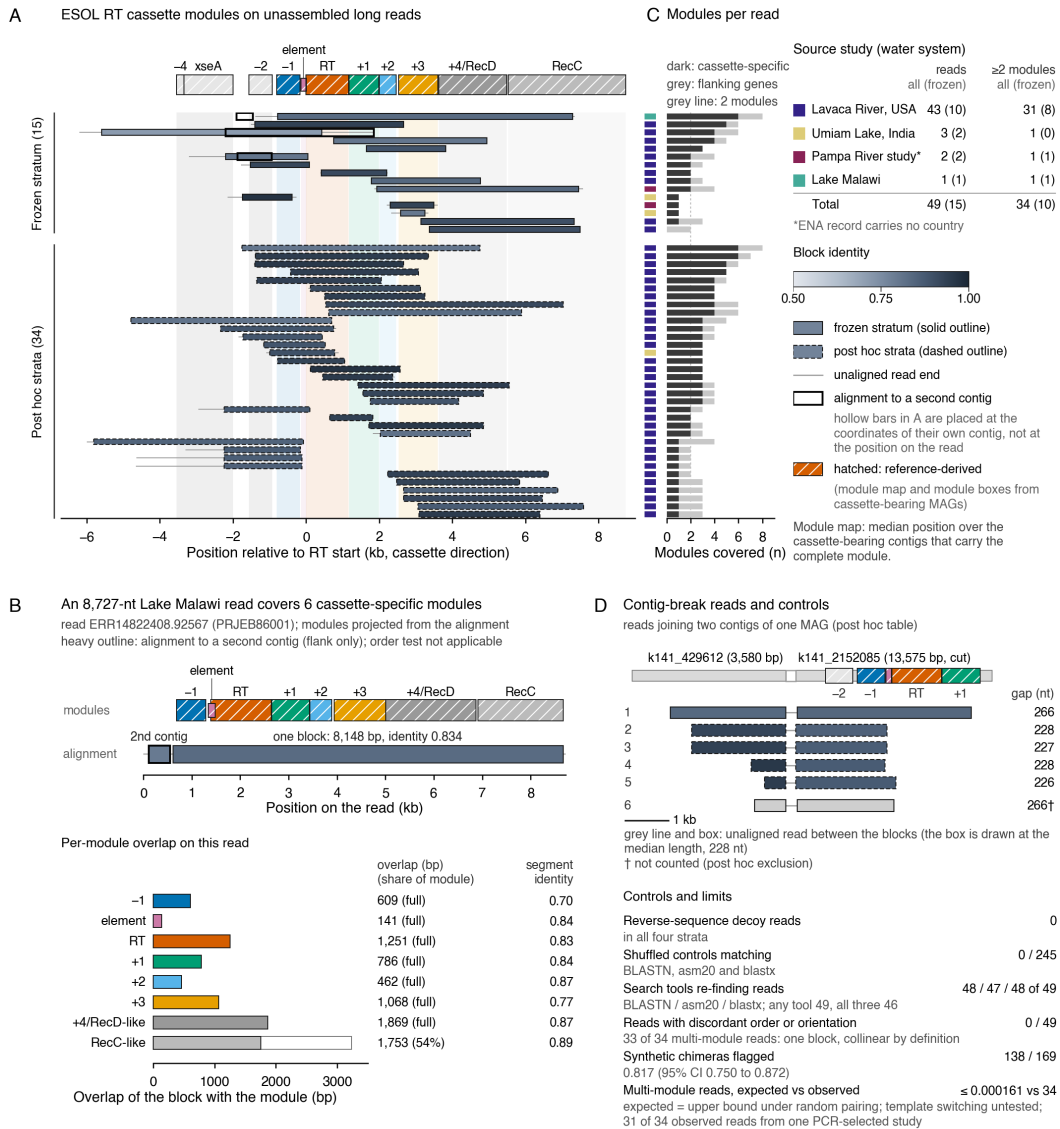


Figure 7. The ESOL cassette on single, unassembled long reads. (A) Each row is one of 49 public Oxford Nanopore reads (15 from the frozen screen and 34 from post hoc strata) whose retained alignment blocks lie on 8 reference contigs from cassette-bearing metagenome-assembled genomes (MAGs), drawn in cassette direction with positions in kb relative to the RT start and each block coloured by its identity. The module map above the rows and the module boxes in B and D are hatched because they are annotations of the reference contigs, which derive from the cassette-bearing MAGs. (B) The 8,727-nt Lake Malawi read has one main alignment block of 8,148 bp (counted in alignment columns; 8,062 nt of the read against 8,067 bp of the reference) at 0.834 identity that covers the six cassette-specific modules (colours), the +4/RecD-like gene and part of the RecC-like gene (grey). (C) Bars give the cassette-specific modules (dark) and the flanking genes (grey) covered by each read in the row order of A, and the coloured chips and the table give the source study, the number of reads and the number of multi-module reads, for all strata and for the frozen stratum (in brackets). (D) Reads that join two contigs of one MAG are drawn with the unaligned stretch between their blocks (post hoc table), and the controls and limits report the reverse-sequence decoy reads, the shuffled-read controls, the re-finding of the reads by three search tools, the reads with discordant order or orientation, the synthetic chimeras flagged and the expected against the observed number of multi-module reads.

Alt text: Four-panel figure of the cassette on unassembled long reads. A, one horizontal bar per read for 49 reads, drawn against a hatched module map on an axis from -6 to 8 kb relative to the RT start; 15 frozen-stratum reads have solid outlines and 34 post hoc reads dashed outlines, shaded by block identity from 0.50 to 1.00. B, one 8,727-nt Lake Malawi read with one main block of 8,148 bp (plus a short block on a second contig), identity 0.834, covering six cassette-specific modules, and bars of per-module overlap. C, bars of modules covered per read and a table of source studies, total 49 (15) reads. D, six contig-break reads with gaps of 266, 228, 227, 228, 226 and 266+ nt, and a table of controls.

Random joining of two molecules predicts at most 1.61×10^{-4} multi-module reads (reads covering two or more modules) even if every cassette read were a chimera, yet public long-read data contain 34, 10 of them in the frozen stratum (Figure 7; Supplementary Data S2, file SI-12_RANDOM_PAIRING_TOTALS). The comparison holds only under random pairing, and non-random pairing in a PCR-selected library was not tested. The long reads come from 9 runs in 4 water systems, and one PCR-selected study supplies 31 of the 34 multi-module reads and 43 of the 49 cassette reads (Supplementary Data S2, file SI-12_RANDOM_PAIRING_BY_RUN). Under the pre-specified strict variant, which counts every read aligned to two contigs of the same MAG, in a reference that holds more contigs than the 16 cassette-bearing ones, as at odds with it, the share at odds is 2 of the 34 multi-module reads (0.059), 5 of the 46 reads covering two or more modules of any class (cassette-specific, host-adjacent or the -2 and -4 flank genes; 0.109) and 5 of all 49 reads (0.102); this is a sensitivity analysis and not a finding about chimeras, and all 5 reads come from 4 samples of one study (Supplementary Data S2, OBS_READS). Further detail on the multi-module reads, the junction-crossing reads from samples and studies outside the reference, and the frozen-stratum comparison with random pairing is in Supplementary Results S1.5.

2.6 Traces in public RNA and in culture DNA: present, but RNA and DNA are not told apart

Public RNA-sequencing runs from Lake Erie, from a Beijing culture lineage and from Sandusky Bay and Wyoming extensions made after the result was seen contain reads matching the cassette window under a counting rule fixed beforehand, which shows that matching sequence is present in those libraries, not that the cassette is transcribed (Extended Data Figure 6, panels A and B). In a post hoc full-depth remapping of one public freshwater metatranscriptome run (SRR6466487), done after a first targeted signal, we recovered three distinct fragments (six reads) of high identity to the cassette and no decoy reads, which shows that cassette-like sequence is present in that run and nothing more; the full-depth rescan of the same run under the counting rule of Supplementary Table S7 counts six matching fragments, and the two counts are not comparable (Supplementary Table S8); these are distinct molecules from one sample, not independent samples, sites or lineages. In post hoc strand-specific total-RNA libraries from one lake series (Beaver Lake), cassette-specific reads lie on the same strand as host mRNA, an imbalance that double-stranded DNA contamination alone would not produce, but this cannot tell RNA from single-stranded DNA, the bias depends on the library kit and is also seen in other runs, and the evidence comes from one lake series. In the unstranded libraries of the Beijing series, a single culture series from one study, the density of cassette reads relative to host reads does not exceed what DNA alone would give, so public sequencing reads cannot currently tell whether cassette-matching reads derive from RNA or from DNA.

455 In DNA from a laboratory culture series of a single study (4 whole-genome shotgun runs, WGS), a
 456 separate assembly carries one cassette-clade RT locus on a 4,566 nt contig whose five module
 457 proteins, including the RT, are each 100.0% identical in amino acid sequence to those of a seed MAG
 458 from run SRR9061301 (99.93% nucleotide identity over the module contig; post hoc comparison;
 459 Supplementary Table S7, table T12); this is an existence observation from one study and one genus,
 460 the detectors were built from the seed MAGs, and no host, origin or function is assigned.

461 Screens of public long-read RNA runs are uninformative in either direction: the candidate runs were
 462 dominated by amplicon libraries and small datasets, and almost none contained reads from the host
 463 lineage, so a lack of cassette reads in them cannot separate not seen from not there. No RNA
 464 sequencing runs were found in the 10 studies that supplied the cassette-bearing MAGs (a lookup in
 465 the European Nucleotide Archive, ENA, at the time of the search), and no RNA run shared a
 466 BioSample with the source sample of any sampled cassette locus, so expression was not tested from
 467 public data here, and these lookups do not show that no RNA data exist elsewhere. So the public data
 468 we could screen cannot say whether the cassette is transcribed; the next question is where in the
 469 world the cassette has been found. Further detail on the rationale for searching public RNA data, the
 470 pond RNA read search and the archive lookups of RNA runs for source studies and samples is in
 471 Supplementary Results S1.6.

472 **2.7 Where in the world: lake, river and pond waters on five continents, and one genus**

473 Where does the company occur? We turned to the public records of the genomes and reads that carry
 474 it (Figure 8). Cassette-bearing genomes turn up in lake, river and pond waters in 9 countries on 5
 475 continents: 16 genomes from 16 different samples in 10 submission projects (Canada, Switzerland,
 476 Korea, Chile, Rwanda, Israel, the USA, India and China; the continent is derived from the recorded
 477 country) (Figure 8). Formal GTDB-Tk classification places all 16 cassette MAGs, which were selected
 478 because their assembly carries the cassette contig (14 of them are also called cassette-positive by the
 479 catalogue scan; Section 4.2; Supplementary Table S9, MAG_SOURCES), in *Pollutibacter*, and the 16
 480 are distinct ANI-level units from 16 separate samples, only half of them high-quality assemblies.
 481 These records show where the cassette-bearing genomes were found and not how common the
 482 cassette is.

483 The free text of none of the 16 records names soil, sediment, a host, wastewater or marine water,
 484 which describes these records and says nothing about other habitats. In the wider *Pollutibacter* sets
 485 the records are not uniformly environmental-sample records: 114 of the 190 distinct BioSamples
 486 behind the genomes examined carry 'sample type = metagenomic assembly', against none of the 16
 487 records, so for the wider sets the country, coordinate and environment text come from MAG
 488 submission records and the originating sample records were not followed. Because the genus
 489 background is already fresh water (152 of the 186 GTDB R232 genomes of *Pollutibacter* carry a
 490 freshwater label from a dictionary reading that failed its manual pre-check, so every habitat
 491 statement here is post hoc), the data cannot test whether the cassette prefers fresh water to the other
 492 habitats of the genus (Extended Data Figure 7).

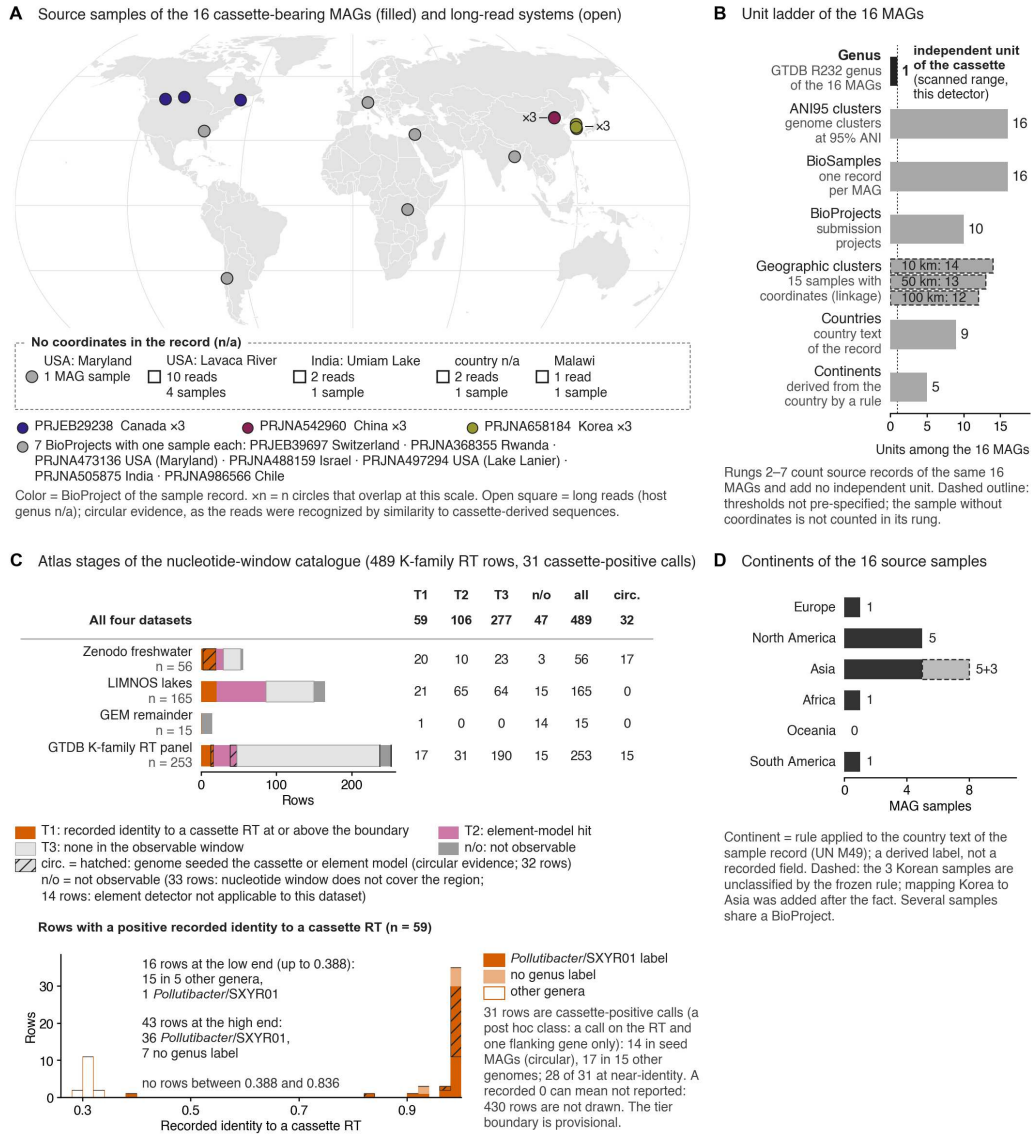


Figure 8. Source records of the 16 cassette-bearing MAGs (map, unit ladder and continents) and the 489 K-family RT rows of the nucleotide-window catalogue by tier. A, Source samples of the 16 cassette-bearing MAGs are drawn at the coordinates given in their public BioSample records (15 of 16 samples), one circle per sample, colored by BioProject for the three BioProjects that contribute three samples each and gray for the seven BioProjects that contribute one (10 BioProjects). The Maryland sample has no coordinate pair and is shown in the dashed strip below the map, together with the four long-read water systems, which have no coordinates in the records we hold. B, The 16 MAGs are counted at seven grains: one genus, ANI95 genome clusters, BioSamples, BioProjects, geographic clusters of the 15 samples with coordinates, countries and continents. C, Upper, the 489 K-family RT rows of the nucleotide-window catalogue (a row is one K-family RT gene in one source table; 31 of the rows are cassette-positive calls, a post hoc operational class defined as a -1 gene on the RT strand and not the whole arrangement, and all 31 lie in T1) are shown by dataset and tier: T1, a recorded identity to a cassette RT at or above the tier boundary (59 rows, 15 of them in other genera at 0.33 identity or less, so a tier is not a clade); T2, an element-model hit upstream of the RT without T1 (106 rows); T3, neither of these in an observable window (277 rows); and 47 rows not observable (n/o); circ. marks rows from the seed MAGs or from genomes on which the detectors are circular (32). Lower, the recorded identity to a cassette RT of the 59 rows that have one, coloured by genus label; rows with a recorded identity of 0, which can mean not reported, are not drawn,

509 and the tier boundary is provisional. D, The 16 cassette-bearing MAG samples are counted by the continent
510 derived from the country text of their BioSample records, which gives five named continents; the three Korean
511 samples are unclassified by the frozen rule, and mapping Korea to Asia was added after the fact (dashed
512 segment, Asia 5+3).

513 Alt text: Four-panel figure on the sources of the 16 cassette-bearing MAGs. A, a world map of source
514 samples coloured by BioProject (Canada ×3, China ×3, Korea ×3, grey for seven single-sample
515 projects), with open squares for long reads and a dashed box for samples without coordinates. B, bars
516 of units: genus 1, ANI95 clusters 16, BioSamples 16, BioProjects 10, countries 9, continents 5, and
517 dashed geographic clusters of 14, 13 and 12 at 10, 50 and 100 km. C, bars of 489 K-family RT rows by
518 tier, T1 59, T2 106, T3 277 and n/o 47, with a histogram of 59 recorded identities. D, continents:
519 Europe 1, North America 5, Asia 5+3 (dashed part: three Korean samples), Africa 1, Oceania 0, South
520 America 1.

521 Within the range we searched and for this detector, every cassette-bearing genome we can place is
522 *Pollutibacter*, which is not evidence that the cassette is absent elsewhere: is the pairing genus-bound,
523 or just unseen outside it? Further detail on the matching public long reads, the source sample records
524 and the catalogue scan loci is in Supplementary Results S1.7.

525 **2.8 Two pieces that meet in one genus: the $-1/+3$ pair recurs without the RT, and the element is** 526 **found with it**

527 The reverse question matters as much: do the parts need the RT to exist? We asked whether the -1
528 and $+3$ genes occur as a pair with no RT hit in the surrounding window, and whether anything lies
529 upstream of the pair that could mark a shared control (Extended Data Figure 8). Not within this
530 detector's range: the $-1/+3$ gene pair occurs with no RT in its window in 54 of 149 anchor genera,
531 those that carry a $+3$ -type gene, of the confirmatory population (pre-specified rule met), and “strong”
532 attaches to the existence of the recurrence only, in a biased, non-random subset of annotated public
533 Bacteroidota genomes, with function and origin untested (Extended Data Figure 8; Supplementary
534 Table S10).

535 A stricter scan asked the same question of all 23,901 Bacteroidota species representatives of GTDB
536 R232 with detectors fixed beforehand: it found 26 same-strand $-1/+3$ -type gene pairs (15 genera, 12
537 families, 9 orders; 23 of the pairs lie in 14 genera outside the cassette-lineage genera), and the 3 pairs
538 with an RVT_1 domain between the two genes all lie in *Pollutibacter*, in the cassette arrangement.
539 These counts belong to these detectors, which derive from the cassette (broader profiles give 55 pairs
540 in 26 genera; Supplementary Table S4, table B02 (config_id), detector variant 4), and the scan is not an
541 independent confirmation of the single-genus result. In a post hoc extension of the adjacent-pair
542 recurrence to this larger spacing, the genera outside the cassette-lineage genera exceed the 1.26
543 expected from a rotation null by a factor of 11.1, but the null was calibrated only at short distances,
544 many of the pairs have a member in a further adjacent pair, and genera are not independent units;
545 restricting to pairs without such a partner keeps 9 genera outside the cassette lineage (10 pairs).

546 The upstream element behaves the other way round. In a post hoc search anchored on the element in
547 794 genomes (all 190 *Pollutibacter* genomes, representatives of the element-bearing genera, and two
548 negative groups of 150 and 139 genomes; this selection overlaps only in part with the genomes

549 scanned in Supplementary Results S1.2; Supplementary Table S4, table A13), 19 element hits were
550 called: 18 lie upstream of an RT (any RT hit, not only the cassette RT) in sense orientation and 1 has
551 its RT side cut by a contig end, so no observable element hit lacked an RT after a rule added once the
552 results had been seen (before that rule, 1 did), and the 289 genomes of the negative groups gave no
553 hit. The element model derives from the cassette, and the population is centred on element-bearing
554 genera, so this describes these genomes and this detector.

555 The RT-free pairs are laid out differently from the cassette. Counted again from the genomes, of 172
556 strong $-1/+3$ -type pairs with no RT of any kind in the window, 151 have the two genes adjacent,
557 with a median gap of 24 nt from the -1 -type to the $+3$ -type gene, whereas at the 30 main cassette loci
558 (the 31 cassette-positive calls of the nucleotide-window catalogue less one locus known from a
559 protein set only, which has no flanking sequence) the -1 to RT gap has a median of 189 nt; an overlap
560 of exactly 4 nt is rare between the paired genes (9 of 151) and usual between $+1$ and $+2$ (17 of 19), and
561 the nominal P values describe the layout and do not test an origin (Supplementary Table S10, sheet
562 F2; Supplementary Table S11, file SI-07_spacing_all_loci).

563 In a post hoc comparison added after the primary results had been seen, flanking sequence next to
564 RT-free pairs never matched flanking sequence next to cassettes in related *Pollutibacter* genomes (0 of
565 627 flank comparisons; 0 of the 27 between genomes at 91.96 to 99.71 per cent genome-wide
566 nucleotide identity and 0 of the 10 within one species-level unit), whereas the same method matched
567 flanks in 15 of 17 pair-to-pair and 22 of 22 cassette-to-cassette comparisons at those identities
568 (Supplementary Table S10, table T). This suggests that in these genomes the pairs and the cassettes sit
569 at different sites, and three genomes carry both, each on different contigs, but the 27 comparisons are
570 24 unordered genome pairs (25 combinations of a pair-bearing and a cassette-bearing genome)
571 formed by 6 genomes with a pair and 11 with a cassette, 26 of them compare one flank only, and the
572 Fisher P values that treat comparisons as independent overstate the evidence. The one pre-specified
573 comparison of neighbouring gene families (255 combinations) found a shared family in 29 against a
574 null mean of 1.35, but 26 of the 29 come from one small protein family, the count is 0 with a narrower
575 window or a stricter identity cut-off, and without that family (post hoc) 3 combinations remain
576 against 1.39, so the pre-specified result is neither evidence that pairs and cassettes share flanking sites
577 nor evidence against it.

578 The host tree gives no direction either. On the GTDB R232 Bacteroidota species tree of 23,901 tips the
579 RT-free adjacent pair is carried by 362 tips and the cassette arrangement by 3 tips, all in *Pollutibacter*;
580 the pair state clusters on the tree more than a label permutation does (265 changes against a
581 permutation mean of 360.3; Supplementary Table S4, table O13), but ancestral-state reconstruction
582 depends on the model (the ordered model gives the pair state to the *Pollutibacter* ancestor with
583 probability 0.68 and the all-rates-different model with 0.98, whereas the hidden-rate model, which
584 has the lowest Akaike information criterion (AIC), gives no pair state 0.93), and only 1 independent
585 genus-level tip carries the cassette arrangement. In gene trees of the 22 RT-free pair genes of
586 *Pollutibacter*, the 6 cassette genes of public *Pollutibacter* genomes and the pair members of other
587 genera, the cassette genes form a separate branch with ultrafast bootstrap support 100 in both the -1 -
588 type and the $+3$ -type tree, a branch that also holds 5 *Pollutibacter* genes of neither class (median
589 distance from a cassette gene to a free-pair gene of *Pollutibacter* 6.17 substitutions per site for the -1 -
590 type gene against 1.27 among those free-pair genes, and 3.41 against 1.43 for the $+3$ -type gene;

Supplementary Table S4, table O07); distances in the saturated range are not informative about timing, so the ratios show that the cassette genes sit on long, separate branches and not how long ago the branches separated, and they do not say which came first. One reading is that the cassette consists of two pieces that are found apart elsewhere, an element beside an RT and a -1/+3 pair, and that *Pollutibacter* is the only genus in which we find the two together; nothing we ran says which came first (Question 3). Further detail on the same-strand pairs in other genera, the upstream palindrome, the junction repeat scans and the gene-tree and structure-based direction analyses is in Supplementary Results S1.8.

2.9 Is this orphan unusual, and how far can the search be trusted?

The ESOL RT is more isolated by nearest identity than the candidates of an untargeted hunt that did not target it, although the alignments differ in coverage (Supplementary Results S1.9): its 30 queries (13 of them metagenomic RTs; 29 with a published-library hit) have a median nearest identity to published RTs of 29% (maximum 35%), against cluster medians of 32% to 50% for the 9 of 15 unassigned candidate clusters of our own hunt, scored with DefensePredictor [19], that lay beside retron lineages; the other 6 reduced to known types.

Our own pipeline is nearly blind to lineages that have a published relative: of 37 published RT lineages removed from its libraries, it re-found only 4 (11%; confidence interval 4 to 25%; Supplementary Table S12, file SI-28_removed_lineage_recall). The pipeline therefore finds a lineage with no relative in the published libraries but tends to classify a close relative of a published lineage as known, so the absence of other candidates is not evidence that none exist. Further detail on the untargeted searches, the neighbour-domain hunt, the metagenome-assembled genome comparison and the sequence-archive clustering is in Supplementary Results S1.9.

2.10 Two pre-set predictions: a contrast at matched identity, and a tree test that told us nothing

In an emulation of existing detection tools, matched for amino-acid identity, profiles rebuilt from a reference pool called 49 of 49 known UG17, UG3 and UG8 units (matched to the 24.8 to 33.3 per cent identity range of the 12 full-length cassette RTs to their closest reference in the pool; the matched units lie at 26.9 to 33.2 per cent; 2 of the 49 are UG17 units) but 0 of 16 cassette RT units (Supplementary Table S2, table Q1). This is a reference-pool emulation within one genus, matched on identity alone, and not a holdout run of the tools themselves. The match was on identity and not on alignment: the matched known members align to their closest reference with a median query coverage of about 98 per cent (Supplementary Table S2, table Q1), against 45.2 to 70.7 per cent for the 12 full-length cassette RTs, so the calls the tools would make at the alignment distance the cassette RTs show cannot be tested with known members at all, and the 49 of 49 versus 0 of 16 contrast is not evidence in either direction. In a post hoc exploratory analysis, known UG17 members, UG17 being the nearest reference lineage of 11 of the 12 full-length cassette RTs, were cut to the profile columns that the cassette RTs align to: 0 of 336 were called at the shipped threshold, against 335 of 336 full-length members, so at the coverage the cassette RTs show a missing call is not readable as a property of their RT lineage (Supplementary Table S2, table Q4).

629 The gene trees of the 4 neighbouring cassette families did not meet the pre-specified congruence
630 condition against the cassette RT tree in 4 of 4 families, so under the pre-specified rule the congruence
631 condition is not met for any of them. At the pre-specified threshold the tree-consistency test therefore
632 yields no information on whether the neighbouring families share the RT's descent: its 4-of-4 negative
633 result is neither a refutation nor support for shared descent, because its own positive control failed
634 and because, in a post hoc control that compared two column halves of the same gene, where shared
635 descent holds by construction, the same rule still returned a supported conflict in 7 of 15 replicates
636 (Supplementary Table S12, test T15). Further detail on the pre-set predictions, the reference-pool
637 profile comparison and the gene trees of the neighbouring cassette families is in Supplementary
638 Results S1.10.

639 **2.11 Hypotheses the data leave open, and the observations that would decide them**

640 Question 1 — What is the RT, in name? Every known-type gate failed to place the ESOL RT: 0 of 9
641 repeat pairs on the contigs of 12 of the 14 census-group members were DGR-type, 0 of 17 members
642 carried a maturase domain, 0 of the 8 census-group loci with annotated neighbours had a Cas
643 domain nearby, and the nearest library identity of the 30 ESOL queries, 13 of them metagenomic RTs,
644 was at most 34.6% (median 29.3% over the 29 queries with a hit; 31.7% and 28.5% for the 14 census-
645 group members alone; Supplementary Table S2, table M and Supplementary Table S2, table N); in
646 sequence trees it cannot be placed in a UG3/UG17-adjacent group or as a relative of the array-
647 associated RTs (ART) [20] or of UG27 [4] because distances are saturated and the joins are weak, so
648 the result is unassigned (Extended Data Figure 9).

649 Question 2 — Does the cassette depend on a host nuclease? Wang et al. (Cell 2026, author
650 manuscript) [8] report selective synthetic lethality between DRT3 and loss of RecBCD and between
651 DRT9 and loss of ExoI; by analogy this motivates the question for a cassette whose +4 part is RecD-
652 like, but it is not support: no experimental test exists, and our genome-association tests of general
653 RecBCD- and SbcB-dependence rules were negative or uninformative (see Limitations).

654 Question 3 — Why did none of the 1,130 observable slots beside the 445 K-family RT loci outside
655 Pollutibacter carry a cassette part for this detector and range, while the parts recur without an RT in
656 other genera? Under hypothesis 1, sampling and detector reach limit a one-genus result: the zero
657 outside the genus cannot be separated from detector failure and the true distribution may be wider
658 than the atlas. Under hypothesis 2, the combination arose once and stayed in Pollutibacter while the
659 parts coexist elsewhere without the RT. So far nothing gives such a history a direction: reconstruction
660 of the state of the Pollutibacter ancestor on the Bacteroidota species tree depends on the model,
661 rooting of the -1 gene tree with outgroups was untestable under the pre-specified saturation rules
662 (Extended Data Figure 10), and flanking sequence next to RT-free pairs never matched flanking
663 sequence next to cassettes in related Pollutibacter genomes (0 of 27 post hoc comparisons, whereas
664 the same method matched flanks in 22 of 22 cassette-to-cassette comparisons), which suggests
665 different sites without saying which came first. Outside Pollutibacter, in a post hoc read-level re-
666 examination of three consecutive ORFs of a Ferruginibacter RT island (an ORF upstream of the RT,
667 the RT and a RecD) that are identical between a Ferruginibacter genome and a CAINEC01 genome
668 from different studies and samples, the public bin assemblies of the two studies contain exactly four
669 contigs with the full identical block: three from mesocosm bins that are nearly identical to one

670 another and so form one clonal lineage, and one short CAINEC01 contig that appears in no other bin
671 and whose source cannot be attributed, so a second host is neither established nor excluded and no
672 direction or donor is inferred.

673 Question 4 — Is the +3 protein catalytically active? The +3 protein lacks the canonical zinc motif
674 HExxH (searched as H[EDQ]xxH) in every recovered ORF, yet codon models place its gene in the
675 purifying direction, so whether that constraint reflects a function stays open; those ORFs are close
676 relatives within one genus, the absence is not specific to the cassette, and composition-matched
677 shuffles would lack the motif in a set of that size about half the time (0.44 to 0.51, depending on
678 whether the N-terminal region or the whole ORF is shuffled; Supplementary Table S5, table T10,
679 rows H-8 and H-9), so this describes sequence and does not test activity, and active and inactive both
680 remain open.

681 Question 5 — Does the +3 protein bind the palindrome? It has not been tested here. The palindrome
682 (core CAAGCTTG, upstream of the start of the -1 gene in the free -1/+3 unit) met the pre-set rule as
683 written in held-out genera, but the genera are not independent lineages (most of the held-out genera
684 in orders unseen in training are Chitinophagales), strict sites occur in 78 of 103 genera and not in all,
685 its non-coding position is an exploratory call, and co-inheritance of a mobile locus is not excluded, so
686 binding and co-inheritance both remain open.

687 Question 6 — What is the element — a retron msr/msd pair, a non-coding RNA of another kind, or
688 something unannotated? Toro (2026) reports computationally predicted msr/msd-like ncRNA motifs
689 upstream of the RT in subsets of loci of two candidate type XI-like lineages, with no experimental
690 validation; no identity is asserted between that element and ours, and the distance of our element to
691 catalogued retron msr/msd pairs is not resolved here. Further detail on the RecBCD pairing and
692 partner comparisons, the detector-reach and rooting analyses, the read-level re-examination of one
693 island and the protein structure predictions is in Supplementary Results S1.11.

694 **3 Discussion**

695 **3.1 What we found, and what it might mean**

696 Of 122,323 RT loci in 189,630 genomes, one candidate group of divergent RTs stood out for recurring
697 neighbours of the same kinds, and the cassette built from them is found beside the RT in 16 genomes,
698 chosen for carrying it, from 16 separate samples in nine countries on five continents (the continent is
699 derived from the recorded country). The central claim is composition and co-occurrence: the
700 assemblies that define the cassette's clade come from 16 water samples on five continents (six from
701 lakes, three from one river, two from ponds and five described only as freshwater, surface or aquatic
702 water), and the free text of none of the 16 records names soil, sediment, a host, wastewater or marine
703 water; within the range searched, the RT of that clade, which none of our comparisons assigns to a
704 named lineage, occurs in one genus, beside the cassette in the genomes that define it and at further
705 loci that played no part in defining it. At 27 further loci of that genus the slots beside the RT hold the
706 cassette parts in 97 of 99 observable cases (the two gaps lie at one locus whose RT is only 0.388
707 identical to the cassette RT), the copies we can see are whole and show a purifying direction of

708 selection, and public long reads that span two or more cassette-specific modules are far more
709 numerous than random pairing of two molecules allows; the parts also recur without an RT in other
710 genera (Figure 5 to Figure 7; Extended Data Figure 8).

711 Sánchez-Nieto et al. report in a preprint that no viable transformants carrying the DRT10 system
712 were recovered in an *E. coli* Δ sbcB strain, that is, that DRT10 activity is toxic in the absence of SbcB
713 (ExoI) [11]; like the DRT3 and DRT9 reports, this is a source of hypotheses for a cassette with a RecD-
714 like part and not an explanation of the pairing.

715 The puzzle is why the RT keeps this company: in *Pollutibacter* the company is present in 97 of 99
716 observable slots, yet beside 445 K-family RTs outside the genus the detectors find none, and at that
717 distance they could not see it (detection near zero for nearly all of the 445). Profile detectors that
718 reach lower find no +3-type protein beside those RTs either (none of 445 loci), which is an informative
719 absence for that part, so the zero is not only a failure of reach. Meanwhile relatives of the -1 and +3
720 genes occur as an adjacent pair with no RT in the window in 54 of the 149 genera of the confirmatory
721 population that carry a +3-type gene; the existence of this RT-free repeat is supported under the pre-
722 specified rule, against a positional null that is not calibrated (Supplementary Results S1.8), though its
723 function and origin are untested. The element, by contrast, was found upstream of an RT at all of the
724 observable element hits of a post hoc search, so the RT-free pair and the element beside an RT look
725 like two pieces that exist apart and that we find together only in *Pollutibacter* (Question 3).

726 Two questions follow, and they are separate. Origin: did the RT join a module that already existed
727 without it, or did the unit arise together? The paired parts recur without an RT in a non-random
728 subset of annotated public Bacteroidota genomes, but the direction of gain or loss cannot be told
729 without polarity-resolved trees. Function: does the RT work with its company, as preprints describe
730 for DRT10, a tripartite system of a structured ncRNA, a SLATT protein and an RT? Nothing here tests
731 it, and the answers are independent, since an RT that joined an older module could have become
732 essential to it, and an RT that arrived with the unit could still be an incidental passenger.

733 What we do not claim is function, a mobility mechanism, a vertical or horizontal origin, or generality
734 beyond the genus and range searched, and these are the questions that the tests below are for. The
735 strongest objection is that detectors built from the cassette MAGs re-find what they were built from:
736 hits in genomes that did not enter the models are new observations, and the checks probe taxon
737 dependence, but sensitivity cannot be judged with the same models, so some circularity remains.

738 Each reading we leave open comes with an observation that would end it (Questions 1 to 6; Extended
739 Data Figure 9): a second genus that carries the cassette parts beside a K-family RT, a structure-based
740 placement that resolves the RT into a known lineage at strong support, outgroup-rooted trees for
741 which polarity can be tested, a 5'-end map of the RT transcript and an activity assay of the +3 N-
742 terminal domain. A wet-laboratory programme is under way to run these tests: constructs, primers
743 and controls are designed and checked in silico (Supplementary Note S1), and no wet-laboratory
744 result is reported here. Further detail on the related literature, the flanking-sequence comparisons in
745 related genomes and the search for unassigned candidate clusters is in Supplementary Discussion
746 S2.1.

3.2 Limitations

Within the range we surveyed, the cassette occurs in one genus, *Pollutibacter*, in genomes from 16 separate samples in 10 projects and nine countries. Because these genomes are close relatives, they count as one lineage in any test of independent recurrence, so the data bear on the existence and arrangement of the cassette but not yet on its host range or on how often it arose. That the RT is unassigned is a statement about the comparisons and libraries listed here, the cassette is defined by position alone, and the 16 seed MAGs and every detector built from them are circular. Nor can the genome arrangement say which came first: at only 6 of the 30 main cassette loci (the 31 cassette-positive calls of the nucleotide-window catalogue less the one known from a protein set only; Section 2.8) is the sequence beside the locus informative on both sides, only one genome that lacks the cassette is a close relative of a cassette genome and all 6 of its flank hits end at a contig edge, and no empty cassette site and no filled pair site was observed in the pre-specified flank-mapping analysis, so it cannot say whether the RT joined or left a module.

The pre-specified part detectors lose sensitivity at identities well above those of most RTs outside the genus, so an absence of a cassette part at lower identity is uninformative, and the one-genus observation is bounded by detector reach as well as by biology (Question 3). Two further profile detectors reach lower for the +3 part than for the -1 part, and they find the +3-type protein at none of the 445 outside loci, so the outside zero is informative for +3 within these detectors and loci and limited to detector reach for -1; every detector family still derives from the seed MAGs of one genus, so a variant unlike the seed would not be seen. The estimate of detector reach depends on one model of sequence evolution and on detectors derived from the cassette, and no model-independent check has reproduced it.

No wet-laboratory result is reported, so nothing here shows what the cassette or its parts do: the intact neighbouring genes and the purifying direction of the codon models are consistent with constraint but are not evidence of function. Our genome-association tests of a general RecBCD-dependence rule did not support one or were uninformative: K-carrying genomes occur in genera with no RecBCD at all, and the pre-set test of RecC companionship, run with an element-based K definition, was uninformative. Under a rule fixed before the check, no cultured isolate qualifies as a candidate for follow-up work: every distinct genome in which the upstream element was called at a K-positive RT locus within the scanned range is a metagenome-assembled genome, which describes genomes in which a seed-dependent element call was made, in that range only, and does not show whether cultured strains carry the element or the cassette. Further detail on the circularity of the part detectors, prevalence estimates, false-negative rates for absence calls and long-read detection limits is in Supplementary Discussion S2.2.

4 Methods

4.1 Census data, RT detection and clustering

The census is the search for RT loci in 189,630 bacterial species-representative genomes with the Pfam RVT_1 model, followed by clustering and the known-type gates, and the later hunt is the untargeted search over the protein collection of 199,923 genomes; both are described in the Results. In both

786 stages the independent unit of the cassette observations is the genus, and the census group and the
787 K-family loci are related only by identifier correspondence, for which no sequence was aligned.

788 The census covered the 189,630 bacterial representatives that have an NCBI assembly (archaea were
789 not included, and 171 further bacterial representatives have no retrievable assembly). The later hunt
790 read the GTDB tarball of predicted proteins for all species representatives, 199,923 genomes (189,801
791 bacterial and 10,122 archaeal), which contains every census genome. The Genomes from Earth's
792 Microbiomes (GEM) catalogue was used as the published protein tarball with Prodigal coordinates.

793 RTs were called with the Pfam RVT_1 model at the gathering threshold (HMMER [21], Pfam [22])
794 with a protein length filter.

795 Clusters were formed at 50% identity with 80% coverage (DIAMOND-based clustering for the GTDB
796 hunt and MMseqs2 [23] for GEM). The published-library reference combined the RT sets of retrons
797 and UG/Abi systems [4,7] and the bacterial RT classification of Toro and Nisa-Martínez [1].

798 Every candidate cluster was checked against the following gates and the result was recorded as a
799 column in the candidate table. G1 (DGR) looked for the bAvd Pfam model [22] and for
800 template/variable repeat pairs with adenine-biased mismatches [9]. G2 (group II intron) looked for
801 maturase and HNH domains, the YADD motif and a gene-splitting arrangement. G3 (published
802 library) took the nearest identity to the library. G4 (Toro retrons) repeated this against the 306
803 Bacteroidota retron-RT candidates and is limited to that phylum. G5 (CRISPR-Cas1) looked for a
804 Cas1 domain fused to the RT or among its neighbours. G6 (defence-system names) matched the
805 nearest library label against the names in DefenseFinder and PADLOC and works at the level of
806 names only.

807 4.2 Data sources, denominators and sample provenance

808 We scanned public genome catalogs for K-family reverse transcriptase (RT) loci, and we report
809 coverage as genomes processed and never as completeness. We call a locus cassette-positive when a
810 -1 gene lies on the strand of a K-family RT, which is a post hoc operational class and not a clade, and
811 which does not require the upstream element or the +1, +2 and +3 genes.

812 We computed ANI with skani 0.3.2 and joined two genomes into one ANI95 unit when their ANI was
813 at least 95 per cent and the larger of their two alignment fractions was at least 50 per cent, by single
814 linkage.

815 We count a set of genomes and their source samples at up to seven rungs, the genus, the ANI95 unit,
816 the BioSample, the BioProject, the coordinate cluster, the country and the continent, and we never
817 add counts of different rungs. For the cassette the independent unit is the genus, and the counts at
818 the other rungs describe source records of the same genomes. The number of ANI95 units depends
819 on the skani mode and the alignment-fraction rule, and the 16 seed MAGs are samples selected
820 because they carry the locus, so they must not be used to estimate carriage. The scan called a -1 gene
821 on the RT strand in 14 of the 16 seed MAGs and a K-family RT in all 16; in the other two MAGs the
822 call could not be made, because the RT is the first ORF of its contig in one and the upstream ORF is
823 cut by the contig end in the other, and because the detector was built from these MAGs the 14 is a
824 plumbing check and not a measure of sensitivity.

4.3 Locus definition, cassette-type call and atlas states

Each component of a locus is one cell with one of four states: present, absent but observable, not observable, or not applicable (NA). We report "not detected" only for cells that are absent but observable, and it means not found by this detector inside the observable window.

4.4 Detection models

We detect the modules around K-family RTs with profile HMMs built from the cassette's own ORFs in 16 cassette-bearing MAGs. K-family RT candidates and the calls around them follow fixed rules, which are given with the models, seed lists, thresholds and tool versions in Supplementary Table S13 and Supplementary Data S3, and every call is reported as a result in the deposited tables.

4.5 Motifs, structure prediction and structure-based placement

An ORF cut by a contig end was excluded from motif calls and counted as not evaluable, never as absent.

For the -1 gene we took the metal-ion-dependent adhesion site (MIDAS) positions from the I domain of human ITAL (UniProt P20701), D162, S164, S166, T231 and D264, which are D137, S139, S141, T206 and D239 in chain A of PDB 1T0P. A -1 ORF was called MIDAS-retained when the five residues fell in the classes D or E, S or T, S or T, S or T, and D or E, MIDAS-partial when three or four did, and MIDAS-lost when two or fewer did.

The K unit is the smallest clade containing the ESOL RT tips, and the K clade is the largest clade reached by climbing from the K unit while each added sibling clade consists of at least 50% K-family tips.

4.6 Defense-system classifier comparison and the gap test

An ESOL RT counted as called when either tool placed it in a Retron, DRT or Abi (AbiA, AbiK, AbiP2) system.

To ask whether the missing calls reflect the RT lineage or a loss of sensitivity of the classifier profiles at the distance of the ESOL RTs from known RTs, we ran a gap test with a rule that we sealed before any call, recovery or holdout result of the test was seen.

4.7 Intactness census and selection analyses

The intactness census asks whether the coding sequences of the cassette modules carry a frameshift, a premature stop codon or a truncating deletion in assembled contigs. It covers 22 loci, the 16 loci of the MAG bins that carry the cassette and 6 loci in 6 public *Pollutibacter* genomes, one for each of the 6 ANI95 units that hold a public genome whose RT has at least 50% identity to an ESOL RT (the locus atlas lists 13 such public loci in 11 genomes; the 7 loci in 5 further genomes of the same units are not part of the census), and the six slots RT, -1 , $+1$, $+2$, $+3$ and $+4$ at each locus; the RecD-like $+4$ slot is assessed as an annotation slot but is not counted as a cassette part (Supplementary Data S1, CENSUS_SLOTS genome_origin).

4.8 Long-read screen, chimera baseline and read-level controls

We looked for reads of the ESOL cassette in public sequencing runs with four kinds of analysis: a screen of long-read metagenome runs, screens of long-read transcriptome runs, rescans of short-read transcriptome and matched metagenome runs, and read-level controls on the reads that these screens returned. The long-read observation is the presence of cassette-sequence reads in the frozen sample; it is not the presence of an assembled contig, and the host genus of a read was not determined. The independent unit of these observations is the water system and not the sample, and one study supplies many of the samples.

The frozen sample holds 265 processed runs, 263 of the freshwater stratum and 2 of the sediment stratum, with 202.9 Gbp of reads, and 799 candidate runs were not reached.

4.9 Co-occurrence census, pair-anchor scan, part recurrence, palindromes and outgroup trees

The co-occurrence census (named co-travel in the file and column names of the deposited tables) asks, for single loci, genomes and genera, how often the parts of the ESOL cassette are found together with a K-family RT, and in which direction the association runs. It is a descriptive census that re-aggregates existing detector tables after the fact, and it detects no new loci except in supplementary scans of genomes that the existing tables did not cover.

The pair-anchor scan re-ran a count of same-strand $-1/+3$ -type gene pairs and of RVT_1 genes between their members, with detectors, ranges, an RT definition and null models fixed beforehand, and it reports what the new detectors give whether or not this matches the count made before sealing (Supplementary Table S10, sheet F2).

The universe was the 23,901 GTDB R232 species representatives of the phylum Bacteroidota.

To ask whether -1 and $+3$ gene pairs without an RT recur in public Bacteroidota genomes, the analysis fixed its detectors, population, null and decision rules before any pair was counted.

4.10 Pre-specification and sealing

Before an analysis was run, a rule file was written. A rule file was sealed by recording its SHA-256 hash and the time; the stamps are internal records and not third-party evidence of the time, and changes made after results were seen are marked as such. The protocol record compiles the rule files of the study; each rule file was sealed by hash before the analysis that it governs, the compiled record was itself sealed later, and its date and SHA-256 digest are given in the guide to the Supplementary Information.

4.11 Software and public catalogues

Sequences were aligned with MAFFT [24] and trees inferred with IQ-TREE [25]; covariance models were built and searched with Infernal [26], 15 Rfam models [27] were also searched, and consensus structures used RNAalifold from ViennaRNA [28]. Predicted structures were searched with Foldseek [29], geNomad [30] was run with default thresholds, and the classifiers were DefenseFinder [31] and PADLOC [32]. Gene coordinates for the GEM proteins came from Prodigal [33], and clusters were formed with DIAMOND [34] for the GTDB hunt and MMseqs2 [23] for GEM. Among the public catalogues we searched were the GEM catalogue [35], Logan [36] and SPIRE [37]. ANI was computed

with skani [38]; codon-model ω was estimated with codeml of PAML 4 [39], site-wise selection with HyPhy [40] (FEL [41]) and recombination breakpoints with GARD [42]. Long reads were re-checked with minimap2 [43] and BLAST+ [44]. Taxonomy followed GTDB release R232 [17,45], and the cassette MAGs were classified with GTDB-Tk [46].

Declarations

Data availability. The public sequence data analysed in this work are identified by accession in Supplementary Table S9, and the locus, component and annotation tables that underlie every figure and supplementary table are provided in Supplementary Data S4 and Supplementary Data S5. The nucleotide sequences of the 16 seed assemblies (the cassette-bearing MAGs) are deposited in Zenodo (DOI 10.5281/zenodo.23346665); every other locus is identified by public accession and coordinates in Supplementary Data S4 and Supplementary Data S5. The detector models, decision thresholds, seed lists, build commands and tool versions are given in Supplementary Table S13 and Supplementary Data S3. The Supplementary Tables are provided as Excel files and the Supplementary Notes as PDF files with the submission, the Supplementary Data are distributed in the Zenodo record only, and the Zenodo record also holds the tables and notes.

Supplementary material. The Supplementary Information opens with a guide to its items; the items that the sections above do not cite are named here. Supplementary Table S14 holds the long-read screening of public sequencing runs, the freshwater enrichment of the K-family, the genus habitat records and the pond time series. Supplementary Table S15 holds the map of literature precedents and the account of how pre-emption was handled. Supplementary Table S16 lists the statements that this work does not make. Supplementary Table S17 holds the source sample records of the cassette-bearing genomes. Supplementary Table S18 gives the known-type gates applied to the census group of ESOL and their results. Supplementary Table S19 gives the composition of the ESOL cassette by position relative to the RT open reading frame, with the RecD-like +4 gene shown as auxiliary and two neighbouring genes outside the cassette for orientation. Supplementary Note S2 holds the dossier of the open questions. Extended Data Figure 11 shows the array coordinates, the spacing signature and the xseA-like to RecD-like interval of the cassette. Extended Data Figure 12 shows the reference RT-domain tree in which the post hoc K-family set is split into several clades. Extended Data Figure 13 shows the +1/+2 pair scan under varying thresholds, the recovery of known members by the same detectors and the pre-specified rule outcomes of the branches that closed without a direction. Extended Data Figure 14 shows the sensitivity of the search pipeline. Extended Data Figure 15 shows the search routes and the known-type gates.

Licence. The data are released under a Creative Commons Attribution 4.0 licence.

Use of AI assistance. The AI3 Discovery scientific-discovery harness was used together with AI tools (Claude and GLM models); the author supervised all of it and is responsible for all content.

Experimental work. No wet-laboratory result is reported in this study. Wet-laboratory experiments to test the predictions are under way, and their design is set out in Supplementary Note S1.

Competing interests. The author declares no competing interests.

Funding. This research received no specific funding and was self-funded.

937 Author contributions. C.P. conceived the study, directed the analyses and the drafting, and is
938 responsible for the content.

939 References

- 940 1. Toro N, Nisa-Martínez R. Comprehensive phylogenetic analysis of bacterial reverse transcriptases. *PLoS*
941 *One* 2014;9:e114083. <https://doi.org/10.1371/journal.pone.0114083>
- 942 2. Zimmerly S, Wu L. An Unexplored Diversity of Reverse Transcriptases in Bacteria. *Microbiol Spectr*
943 2015;3:MDNA3-0058-2014. <https://doi.org/10.1128/microbiolspec.MDNA3-0058-2014>
- 944 3. González-Delgado A, Mestre MR, Martínez-Abarca F, Toro N. Prokaryotic reverse transcriptases: from
945 retroelements to specialized defense systems. *FEMS Microbiol Rev* 2021;45:fuab025. [https://doi.org/](https://doi.org/10.1093/femsre/fuab025)
946 [10.1093/femsre/fuab025](https://doi.org/10.1093/femsre/fuab025)
- 947 4. Mestre MR, Gao LA, Shah SA, López-Beltrán A, González-Delgado A, Martínez-Abarca F, et al. UG/ Abi: a
948 highly diverse family of prokaryotic reverse transcriptases associated with defense functions. *Nucleic Acids*
949 *Res* 2022;50:6084-6101. <https://doi.org/10.1093/nar/gkac467>
- 950 5. Toro N, Martínez-Abarca F, Mestre MR, González-Delgado A. Multiple origins of reverse transcriptases
951 linked to CRISPR-Cas systems. *RNA Biol* 2019;16:1486-1493. [https://doi.org/10.1080/](https://doi.org/10.1080/15476286.2019.1639310)
952 [15476286.2019.1639310](https://doi.org/10.1080/15476286.2019.1639310)
- 953 6. Millman A, Bernheim A, Stokar-Avihail A, Fedorenko T, Voichek M, Leavitt A, et al. Bacterial Retrons
954 Function In Anti-Phage Defense. *Cell* 2020;183:1551-1561.e12. [https://doi.org/10.1016/](https://doi.org/10.1016/j.cell.2020.09.065)
955 [j.cell.2020.09.065](https://doi.org/10.1016/j.cell.2020.09.065)
- 956 7. Mestre MR, González-Delgado A, Gutiérrez-Rus LI, Martínez-Abarca F, Toro N. Systematic prediction of
957 genes functionally associated with bacterial retrons and classification of the encoded tripartite systems.
958 *Nucleic Acids Res* 2020;48:12632-12647. <https://doi.org/10.1093/nar/gkaa1149>
- 959 8. Wang M, Yoneyama K, Žedaveinytė R, Ishikawa J, Tang S, Le HC, et al. Coordinated RNA- and protein-
960 templated synthesis of double-stranded DNA by a dual reverse transcriptase immune system. *Cell*
961 2026;189:4997-5011.e11. <https://doi.org/10.1016/j.cell.2026.07.012>
- 962 9. Wu L, Gingery M, Abebe M, Arambula D, Czornyj E, Handa S, et al. Diversity-generating retroelements:
963 natural variation, classification and evolution inferred from a large-scale genomic survey. *Nucleic Acids Res*
964 2018;46:11-24. <https://doi.org/10.1093/nar/gkx1150>
- 965 10. Silas S, Mohr G, Sidote DJ, Markham LM, Sanchez-Amat A, Bhaya D, et al. Direct CRISPR spacer
966 acquisition from RNA by a natural reverse transcriptase-Cas1 fusion protein. *Science* 2016;351:aad4234.
967 <https://doi.org/10.1126/science.aad4234>
- 968 11. Sánchez-Nieto E, Wilkinson ME, Millán V, Wenardy C, Cabrerizo D, Martínez-Abarca F, et al. Host
969 exonuclease SbcB and a phage-encoded SSB-like protein control activation of the DRT10 reverse
970 transcriptase defense system. *bioRxiv* (preprint) 2026. <https://doi.org/10.64898/2026.04.29.721553>
- 971 12. Tang S, Ramirez JL, Mestre MR, Zhang DJ, Wang M, Wiegand T, et al. Antiviral reverse transcriptases reveal
972 the evolutionary origin of telomerase. *bioRxiv* (preprint) 2025. [https://doi.org/10.1101/](https://doi.org/10.1101/2025.10.16.682844)
973 [2025.10.16.682844](https://doi.org/10.1101/2025.10.16.682844)
- 974 13. Ramirez JL, He Q, Wiegand T, Lampe GD, Wang J, Tang S, et al. Mechanism of tandem-repeat DNA
975 synthesis by an antiviral reverse transcriptase. *bioRxiv* (preprint) 2026. [https://doi.org/10.64898/](https://doi.org/10.64898/2026.06.29.735271)
976 [2026.06.29.735271](https://doi.org/10.64898/2026.06.29.735271)

14. Chen S, Zhang S, Zhang Y, Fu P, Mao T, Zhan Y, et al. Defense-associated reverse transcriptases synthesize DNA in a telomerase-like manner. *bioRxiv* (preprint) 2026. <https://doi.org/10.64898/2026.10.08.757509>
15. van den Berg DE, Costa AR, Esser JQ, Muralidharan A, van den Bossche H, Brouns SJJ. Discovery of phage defense systems through component modularity networks. *bioRxiv* (preprint) 2025. <https://doi.org/10.1101/2025.09.30.679545>
16. Toro N. Landscape of retron diversity across the SPIRE microbial metagenome resource reveals candidate novel type XI-like lineages. *mSystems* 2026;e00732-26. <https://doi.org/10.1128/msystems.00732-26>
17. Parks DH, Chuvochina M, Rinke C, Mussig AJ, Chaumeil PA, Hugenholtz P. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res* 2022;50:D785-D794. <https://doi.org/10.1093/nar/gkab776>
18. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 2023;379:1123-1130. <https://doi.org/10.1126/science.ade2574>
19. DeWeirdt PC, Mahoney EM, Laub MT. DefensePredictor: A machine learning model to discover prokaryotic immune systems. *Science* 2026;392:eadv7924. <https://doi.org/10.1126/science.adv7924>
20. Yoon PH, Athukoralage JS, Ameisen E, Kauderer-Abrams E, Perry NT, Durrant MG. Autonomous AI agents discover reverse transcriptases with tandem repeat arrays. *bioRxiv* (preprint) 2026. <https://doi.org/10.64898/2026.10.05.755809>
21. Eddy SR. Accelerated Profile HMM Searches. *PLoS Comput Biol* 2011;7:e1002195. <https://doi.org/10.1371/journal.pcbi.1002195>
22. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: The protein families database in 2021. *Nucleic Acids Res* 2021;49:D412-D419. <https://doi.org/10.1093/nar/gkaa913>
23. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* 2017;35:1026-1028. <https://doi.org/10.1038/nbt.3988>
24. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 2013;30:772-780. <https://doi.org/10.1093/molbev/mst010>
25. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* 2020;37:1530-1534. <https://doi.org/10.1093/molbev/msaa015>
26. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 2013;29:2933-2935. <https://doi.org/10.1093/bioinformatics/btt509>
27. Kalvari I, Nawrocki EP, Ontiveros-Palacios N, Argasinska J, Lamkiewicz K, Marz M, et al. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res* 2021;49:D192-D200. <https://doi.org/10.1093/nar/gkaa1047>
28. Lorenz R, Bernhart SH, Höner Zu Siederdisen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. *Algorithms Mol Biol* 2011;6:26. <https://doi.org/10.1186/1748-7188-6-26>
29. van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, et al. Fast and accurate protein structure search with Foldseek. *Nat Biotechnol* 2024;42:243-246. <https://doi.org/10.1038/s41587-023-01773-0>
30. Camargo AP, Roux S, Schulz F, Babinski M, Xu Y, Hu B, et al. Identification of mobile genetic elements with geNomad. *Nat Biotechnol* 2024;42:1303-1312. <https://doi.org/10.1038/s41587-023-01953-y>

- 1019 31. Tesson F, Hervé A, Mordret E, Touchon M, d'Humières C, Cury J, et al. Systematic and quantitative view of
1020 the antiviral arsenal of prokaryotes. *Nat Commun* 2022;13:2561. [https://doi.org/10.1038/s41467-022-](https://doi.org/10.1038/s41467-022-30269-9)
1021 [30269-9](https://doi.org/10.1038/s41467-022-30269-9)
- 1022 32. Payne LJ, Todeschini TC, Wu Y, Perry BJ, Ronson CW, Fineran PC, et al. Identification and classification of
1023 antiviral defence systems in bacteria and archaea with PADLOC reveals new system types. *Nucleic Acids*
1024 *Res* 2021;49:10868-10878. <https://doi.org/10.1093/nar/gkab883>
- 1025 33. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition
1026 and translation initiation site identification. *BMC Bioinformatics* 2010;11:119. [https://doi.org/10.1186/](https://doi.org/10.1186/1471-2105-11-119)
1027 [1471-2105-11-119](https://doi.org/10.1186/1471-2105-11-119)
- 1028 34. Buchfink B, Reuter K, Drost HG. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat*
1029 *Methods* 2021;18:366-368. <https://doi.org/10.1038/s41592-021-01101-x>
- 1030 35. Nayfach S, Roux S, Seshadri R, Udwy D, Varghese N, Schulz F, et al. A genomic catalog of Earth's
1031 microbiomes. *Nat Biotechnol* 2021;39:499-509. <https://doi.org/10.1038/s41587-020-0718-6>
- 1032 36. Chikhi R, Lemane T, Loll-Krippelbein R, Montoliu-Nerin M, Raffestin B, Camargo AP, et al. Logan:
1033 Planetary-Scale Genome Assembly Surveys Life's Diversity. *bioRxiv preprint (version 2) 2025*. [https://](https://doi.org/10.1101/2024.07.30.605881)
1034 doi.org/10.1101/2024.07.30.605881
- 1035 37. Schmidt TSB, Fullam A, Ferretti P, Orakov A, Maistrenko OM, Ruscheweyh HJ, et al. SPIRE: a Searchable,
1036 Planetary-scale mIcrobiome REsource. *Nucleic Acids Res* 2024;52:D777-D783. [https://doi.org/10.1093/](https://doi.org/10.1093/nar/gkad943)
1037 [nar/gkad943](https://doi.org/10.1093/nar/gkad943)
- 1038 38. Shaw J, Yu YW. Fast and robust metagenomic sequence comparison through sparse chaining with skani.
1039 *Nat Methods* 2023;20:1661-1665. <https://doi.org/10.1038/s41592-023-02018-3>
- 1040 39. Yang Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol Biol Evol* 2007;24:1586-1591. [https://](https://doi.org/10.1093/molbev/msm088)
1041 doi.org/10.1093/molbev/msm088
- 1042 40. Kosakovsky Pond SL, Poon AFY, Velazquez R, Weaver S, Hepler NL, Murrell B, et al. HyPhy 2.5—A
1043 Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies. *Mol Biol Evol* 2020;37:295-
1044 299. <https://doi.org/10.1093/molbev/msz197>
- 1045 41. Kosakovsky Pond SL, Frost SDW. Not so different after all: a comparison of methods for detecting amino
1046 acid sites under selection. *Mol Biol Evol* 2005;22:1208-1222. <https://doi.org/10.1093/molbev/msi105>
- 1047 42. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW. Automated phylogenetic detection
1048 of recombination using a genetic algorithm. *Mol Biol Evol* 2006;23:1891-1901. [https://doi.org/10.1093/](https://doi.org/10.1093/molbev/msl051)
1049 [molbev/msl051](https://doi.org/10.1093/molbev/msl051)
- 1050 43. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;34:3094-3100. [https://](https://doi.org/10.1093/bioinformatics/bty191)
1051 doi.org/10.1093/bioinformatics/bty191
- 1052 44. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and
1053 applications. *BMC Bioinformatics* 2009;10:421. <https://doi.org/10.1186/1471-2105-10-421>
- 1054 45. Parks DH, Chaumeil PA, Mussig AJ, Rinke C, Chuvochina M, Hugenholtz P. GTDB release 10: a complete
1055 and systematic taxonomy for 715 230 bacterial and 17 245 archaeal genomes. *Nucleic Acids Res*
1056 2026;54:D743-D754. <https://doi.org/10.1093/nar/gkaf1040>
- 1057 46. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk v2: memory friendly classification with the
1058 genome taxonomy database. *Bioinformatics* 2022;38:5315-5316. [https://doi.org/10.1093/](https://doi.org/10.1093/bioinformatics/btac672)
1059 [bioinformatics/btac672](https://doi.org/10.1093/bioinformatics/btac672)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryInformation.pdf](#)
- [SupplementaryNoteS1.pdf](#)
- [SupplementaryNoteS2.pdf](#)
- [SupplementaryTableS1.xls](#)
- [SupplementaryTableS2.xls](#)
- [SupplementaryTableS3.xls](#)
- [SupplementaryTableS4.xls](#)
- [SupplementaryTableS5.xls](#)
- [SupplementaryTableS6.xls](#)
- [SupplementaryTableS7.xls](#)
- [SupplementaryTableS8.xls](#)
- [SupplementaryTableS9.xls](#)
- [SupplementaryTableS10.xls](#)
- [SupplementaryTableS11.xls](#)
- [SupplementaryTableS12.xls](#)
- [SupplementaryTableS13.xls](#)
- [SupplementaryTableS14.xls](#)
- [SupplementaryTableS15.xls](#)
- [SupplementaryTableS16.xls](#)
- [SupplementaryTableS17.xls](#)
- [SupplementaryTableS18.xls](#)
- [SupplementaryTableS19.xls](#)