

Supplementary Information — 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

This document accompanies the main text. Section S1 holds the supporting detail and the qualifications that belong to each Results section, in the order and under the headings of the main text; Section S2 does the same for the Discussion and the Limitations; Section S3 describes the data sources and denominators, the analyses with their public tools and the counting conventions. Supplementary Tables, Data and Notes are cited in the text by their numbers; the Tables and Notes are provided as separate files with the submission and the Data in the Zenodo record only (see the guide below), and the Extended Data Figures follow the references. These section numbers are not the numbers of Supplementary Tables, Data or Notes: Section S1 is the results section of this document, whereas Supplementary Note S1 is a separate file listed in the guide below.

Supplementary items whose numbers do not appear in the guide below are not distributed. Among them, the protocol record (Supplementary Note S4) compiles the rule files of the study; each rule file was sealed by hash before the analysis that it governs (Section S3.14), and the compiled record was itself sealed on 2026-10-13 and is not distributed. So that the sealed file can be recognised if it is disclosed later, its SHA-256 digest is:

b11f5f92c6f6061761b21a2309188928e4855deacc596cf0e89b6a80da19c978

Guide to the supplementary material

This document is accompanied by the items below. The Supplementary Tables are provided as Excel files and the Supplementary Notes as PDF files with the submission, and all of them are also in the Zenodo record (DOI 10.5281/zenodo.23346665); the Supplementary Data are distributed in the Zenodo record only and are not part of the review files.

Item	Contents
Supplementary Table S1 (SI-14)	Catalogue denominators, loci derived from SPIRE and GEM, the cross-catalogue duplicate map, the census funnel (table H), the classes of the candidate groups (table I), the funnels of the later hunts (table J) and the controls of the Pfam hunt (table K). Provided as the Excel file Supplementary_Table_S1.xls (one sheet per table; the Legend sheet comes first) and in the Zenodo record.
Supplementary Table S2 (SI-11)	Comparison ladder for the placement of the RT: classifier, structure, tree and type VI frame comparisons. Provided as the Excel file Supplementary_Table_S2.xls (one sheet per table; the Legend sheet comes first) and in the Zenodo record.
Supplementary Table S3 (SI-04)	Component cards of the arrangement around the ESOL RT: the upstream element and the genes at -2, -1, +1, +2, +3 and +4 and the RecC-like gene. Provided as the Excel file Supplementary_Table_S3.xls (one sheet per table; the Legend sheet comes first) and in the Zenodo record.
Supplementary Table S4 (SI-15)	Co-occurrence, pair-anchor and outgroup tables. Provided as the Excel file Supplementary_Table_S4.xls (one sheet per table; the Legend sheet comes first) and in the Zenodo record.
Supplementary Table S5 (SI-08)	Motif and residue calls with their denominators and controls. Provided as the Excel file Supplementary_Table_S5.xls (one sheet per table; the Legend sheet comes first) and in the Zenodo record.
Supplementary Table S6 (SI-10)	Selection and neutral-expectation tables. Provided as the Excel file Supplementary_Table_S6.xls (one sheet per table; the Legend sheet comes first) and in the Zenodo record.
Supplementary Table S7 (SI-17)	Read-level observations on the cassette in public sequencing data. Provided as the Excel file Supplementary_Table_S7.xls (one sheet per table; the Legend sheet comes first) and in the Zenodo record.
Supplementary Table S8 (SI-29)	Pre-specified result versus post hoc reading. Provided as the Excel file Supplementary_Table_S8.xls (one sheet per table; the Legend sheet comes first) and in the Zenodo record.
Supplementary Table S9 (SI-13)	Source samples, counting units, reasons for empty cells and identifier rules. Provided as the Excel file Supplementary_Table_S9.xls (one sheet per table; the Legend sheet comes first) and in the Zenodo record.
Supplementary Table S10 (SI-16)	Pair-recurrence tables, leave-one-out analyses and the palindrome ladder. Provided as the Excel file Supplementary_Table_S10.xls (one sheet per table; the Legend sheet comes first) and in the Zenodo record.
Supplementary Table S11 (SI-07)	Array coordinates, spacing signature, ORF sizes and the genome-background tables of the cassette. Provided as the Excel file Supplementary_Table_S11.xls (one sheet per table; the Legend sheet comes first) and in the Zenodo record.
Supplementary Table S12 (SI-28)	Tests that did not support a claim or could not test it. Provided as the Excel file Supplementary_Table_S12.xls (one sheet per table; the Legend sheet comes first) and in the Zenodo record.
Supplementary Table S13 (SI-05)	Detectors, profile models and decision thresholds, with the origin of every seed set. Provided as the Excel file Supplementary_Table_S13.xls (one sheet per table; the Legend sheet comes first) and in the Zenodo record.

Item	Contents
Supplementary Table S14 (SI-18)	Long-read screening of public sequencing runs, freshwater enrichment of the K-family, genus habitat records and the pond time series. Provided as the Excel file Supplementary_Table_S14.xls (one sheet per table; the Legend sheet comes first) and in the Zenodo record.
Supplementary Table S15 (SI-20)	Map of literature precedents and how pre-emption was handled. Provided as the Excel file Supplementary_Table_S15.xls (one sheet per table; the Legend sheet comes first) and in the Zenodo record.
Supplementary Table S16 (SI-21)	Statements this work does not make. Provided as the Excel file Supplementary_Table_S16.xls (one sheet per table; the Legend sheet comes first) and in the Zenodo record.
Supplementary Table S17 (SI-27)	The 16 source sample records. Provided as the Excel file Supplementary_Table_S17.xls (one sheet per table; the Legend sheet comes first) and in the Zenodo record.
Supplementary Table S18	Known-type gates applied to the census group of ESOL and their results. Provided as the Excel file Supplementary_Table_S18.xls (one sheet per table; the Legend sheet comes first) and in the Zenodo record.
Supplementary Table S19	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. Provided as the Excel file Supplementary_Table_S19.xls (one sheet per table; the Legend sheet comes first) and in the Zenodo record.
Supplementary Data S1 (SI-09)	Output of the integrity census with derived columns and derived tables. Distributed in the Zenodo record only (DOI 10.5281/zenodo.23346665); not part of the review files.
Supplementary Data S2 (SI-12)	Long-read observation tables. Distributed in the Zenodo record only (DOI 10.5281/zenodo.23346665); not part of the review files.
Supplementary Data S3 (SI-06)	Seed lists by identifier, alignment inventory, model build registry, build commands, tool versions and regeneration checks. Distributed in the Zenodo record only (DOI 10.5281/zenodo.23346665); not part of the review files.
Supplementary Data S4 (SI-02)	Locus atlas: the table of all K-family RT loci, with a column dictionary and blank reasons. Distributed in the Zenodo record only (DOI 10.5281/zenodo.23346665); not part of the review files.
Supplementary Data S5 (SI-03)	Component calls of the K-family RT loci (long form), per-locus summaries, state counts and the annotated loci (GFF3). Distributed in the Zenodo record only (DOI 10.5281/zenodo.23346665); not part of the review files.
Supplementary Note S1 (SI-24)	Predictions and wet-laboratory tests under way (plans, not results). Provided as the PDF file Supplementary_Note_S1.pdf and in the Zenodo record.
Supplementary Note S2 (SI-19)	Dossier of the open questions. Provided as the PDF file Supplementary_Note_S2.pdf and in the Zenodo record.

S1 Supplementary results

S1.1 An orphan by our comparisons: an RT that none of them assigns to a known family

The other 8,178 of the 122,323 RT loci could not be measured against the libraries, so 109,524 named, 4,621 diverged and 8,178 unmeasured loci make up the 122,323 (Supplementary Table S1, table H). Of the 4,621 divergent loci, 871 were not in a candidate group: 525 were singletons and 346 lay in 132 connected sets that were too small to count as candidates (Supplementary Table S1, table H). As

coded in the census partition, in which a group counts as a remote homologue when most of its members had their best, weak, library hit on one published RT family or library entry, the group that became ESOL is one of 7 remote-homologue groups, because 13 of its 14 members had their weak hits on UG17 RTs at the partition's search sensitivity; we do not read this as assignment (Supplementary Table S1, table I). Its census group has 14 loci in 14 genomes and a median RT length of 389 amino acids, and structure prediction followed by a Foldseek search [1] assigns it an RT fold.

Before calling an RT an orphan we had to rule out disguises, so we ran it through gates for DGRs, group II introns, retrons, CRISPR-RT-Cas1 systems and defence-system names, and none assigned it to a known type (Supplementary Table S18; Supplementary Table S2, table N). No DGR-type repeat pair was found (0 of 9 repeat pairs, on the contigs of 12 of the 14 census-group members), although the same detector recovered the BPP-1 positive control. No group II intron maturase domain is present (0 of 17 members). No Cas domain lies among the annotated neighbours of the census-group loci that have them (0 of 8 of the 14 members), and no RT-plus-RecBCD system is defined by the names in DefenseFinder or PADLOC [2,3]. The DGR search window sets an upper bound on the sensitivity of the search. The catalytic aspartates of the YxDD motif are present in 17 of 17 members (the census core and 3 further genome members; these are not independent observations; Supplementary Table S2, table O), so a G2L4-like (group II-like) repair RT can be neither excluded nor confirmed; the remaining motifs cannot be judged. For the remaining RT types the answer is only as good as the libraries available: the retron comparison covers only Bacteroidota, and the check on defence-related names works at the level of names.

Supplementary Table S18. Known-type gates applied to the census group of ESOL and their results.

Gate	Question	How it was applied	Result for ESOL
G1 DGR	Is the RT a diversity-generating retroelement?	bAvd Pfam model; template/variable repeat pairs with adenine-biased mismatches	0 DGR-type pairs among 9 repeat pairs (contigs of 12 of the 14 census-group members)
G2 Group II intron	Is it an intron RT?	Maturase and HNH domains, YADD motif, gene-splitting layout	Maturase domain in 0 of 17 members
G3 Published library	Is it close to a published RT?	DIAMOND against retron, UG / Abi and Toro 2014 RTs	Maximum 34.6%, median 29.3% over the 29 of 30 ESOL queries with a hit (14 census-group members, 3 further genome members, 13 metagenomic RTs); census-group members alone: maximum 31.7%, median 28.5%
G4 Toro 2026 retrons	Does it nest in a Bacteroidota retron group?	Identity and tree against the retron RTs of the survey (Bacteroidota only)	23.2 to 36.5% (24 of the 30 queries with a hit); no comparison places the RT in a retron group
G5 CRISPR-RT-Cas1	Is it fused to or beside Cas1?	Cas1 domain in RT or neighbours	0 of the 8 census-group loci with annotated neighbours (of 14 members) have a Cas domain nearby
G6 Defence names	Does its nearest label name a defence-related system?	Library label against DefenseFinder and PADLOC names (name level)	No RT plus RecBCD system defined

A gate removes a candidate from the unassigned set when it matches a known RT type; none of the six gates assigned the ESOL RT to a known type, and the libraries limit what each gate can say. The 36.5% under G4 is the highest best identity of an ESOL query to the 306 retron-RT candidates assigned to Bacteroidota in the survey, a set that is not part of the library of G3. The values of G3 and G4 are over 30 ESOL queries (14 census-group members, 3 further genome members and 13 metagenomic RTs; Supplementary Table S2, table M), and the counts of G1 and G5 are over the census-group contigs and loci that were available (12 and 8 of the 14 members; Supplementary Table S2, table N). The 17 members of G2 are listed in Supplementary Table S2, table O.

In a post hoc analysis of expanded sequence trees, UG3 was the only sister candidate for the ESOL RT among the RT classes tested, a call that recurs across alignments and tree replicates but reuses the same alignments and so repeats one reading and is not a second line of evidence, whereas predicted-structure distances place the RT nearest UG17; a pre-specified tree test rejected every placement, random ones included, and so was uninformative. None of the sequence-tree or predicted-structure comparisons assigns the RT to a named lineage. In maximum-likelihood trees of the RT domain the K-family RTs are not monophyletic (Figure 3; Extended Data Figure 12). The K family is defined here as a post hoc definition, so this describes the sampled representatives under that definition, not the K core as a whole or every member. The nearest non-K class of the K main clade's predicted structures is UG17 (DRT10) throughout, exceeding a length- and pLDDT-matched random bacterial RT null and holding under the Boltz null, reference removal, and alternative distances; pairwise mean distances also point toward UG17 rather than UG3 in all six measures, though the effect is weaker on specific measures (Figure 3). This is a statement of which of UG3 and UG17 is structurally nearer within the UG3+UG17 supergroup, not of lineage membership: the K main clade is a post hoc definition, and the effect is weaker on specific measures. A second analysis of the same predicted structures asked whether the DRT10-nearest structure is evidence of a DRT10-type RT, and it did not support that reading. On the DRT10-specific scale alone, scores of the cassette clade fall below those of the DRT10 positive control and the same-lineage scores of known lineages, and non-specific proximity does not mean that the ESOL RT is not DRT10-type. Because RT-domain distances are saturated and the relevant joins are weak, we can neither assert that the cassette RT falls inside a UG3/UG17-adjacent lineage nor that it is a relative of ART or UG27, and the honest result of sequence-level placement is that it is unresolved. Of the K-family candidate RTs found in GTDB R232 representative genomes, 32 RTs in 31 genomes have a tip of the main K clade of the post hoc reference tree as their nearest reference tip at an identity of at least 0.30; these genomes belong to 17 genera, 11 orders and 4 phyla, so this nearest-tip class includes RTs from genera beyond the one that carries the cassette (Supplementary Table S2, table P). These are counts of taxa among the carrying genomes and not of independent events (12 of the 17 genera rest on a single genome), and the assignment by nearest-tip identity is a post hoc rule and not a phylogenetic placement; at an identity of at least 0.50 the count falls to 3 phyla and 13 genera, at 0.60 to 2 phyla and 8 genera, and at 0.70 to 7 genera, so the breadth beyond Bacteroidota rests on distant nearest tips.

The upstream element also lies against the RT start with no gene between them, unlike the DRT10 ncRNA placements compared (which vary among subtypes), and UG17 RTs are longer than the cassette RT and carry a C-terminal serine or tyrosine priming signature far more often (a descriptive comparison on complete sequences). Two units resemble the DRT10 lineage by Foldseek and DIAMOND hits, a C-terminal serine with a matching tail and a long thumb region, whereas the

cassette RT lineage lacks these features; tree placement of the two units depends on the method, some neighbouring units show intermediate signal that was not resolved, and the sibling tests point the other way. The upstream element could not be matched to DRT10 ncRNAs, but this test is uninformative because its positive control failed in both directions; the DRT2-type and DRT9-type tests were not supported under the operationalisations we used, and the retron msd-type test (msd is one of the two elements of the retron non-coding RNA) is uninformative because its detector was not validated. In a post hoc comparison on the RT-domain tree, the stem leading to the K-defined RT clade was shorter than the stems of reference RT classes and its long branches lay within the crown, so branch length alone gives no sign that the clade stands apart on a long branch (the percentile shifts with how the random-RT null is defined, and K is itself defined post hoc); a separate post hoc comparison found the K-defined RTs farther than random bacterial RTs, but that null is a common, densely sampled lineage, so neither comparison shows where the clade sits among known RT groups. Post hoc (the K main clade is defined after the fact), the nearest non-K class to the cassette RT in structure-distance trees is UG17 with no UG3 sister, whereas sequence trees show a UG3 sister, so its tree position stays unresolved and nearest is not membership; in the reference-distance structure tree the UG17 monophyly positive control failed, so that tree is uninformative on this point, and the 3Di+amino-acid result is stable only across two core sets. In a UniRef50 search, the closest sequences to the 14 census-group RTs reach a per-query top bit score with a median of 104 and a maximum of 132, 3 clusters score 80 bits or more (one scores 100 bits or more), and the top hit belongs to the sister lineage in *Aquirufa*; this measures distance to database sequences only and does not by itself place the RT inside or outside any known family. In an exploratory, post hoc comparison with reported DRT10 features, the cassette RT's match at the alignment columns that mark UG17 was at the level of non-UG17 RTs, and no SLATT-family protein was detected in the cassette-bearing MAG proteomes; because sensitivity for divergent SLATT-family proteins was not measured and MAG completeness was not controlled, the SLATT result is uninformative, and the comparison leaves a DRT10-like arrangement unresolved. In an exploratory, post hoc comparison of coordinates, loci that carry a SLATT protein place it at or overlapping the RT start with ncRNA-sized room upstream, whereas in the cassette the conserved upstream element sits directly against the RT; because DRT10 subtypes differ in layout, this shows only that the cassette layout does not match the known subtypes.

S1.2 The RT keeps company: an upstream element and four genes at fixed positions

The arrangement of the RT and its downstream +1 and +2 genes is reproduced, in part and not independently, in close relative genomes of one genus (Supplementary Table S11; Extended Data Figure 11). Of the 6 close relative genomes, 4 pass the strict rule, 1 carries the five families in order but fails the strict rule, and 1 is truncated at the contig end, so that its arrangement beyond that point cannot be judged. The reproduction lies within one genus: the independent unit is a single genus, and across the range we swept the independent occurrence of the cassette is a single genus. The test is circular, because the module profiles were built from the cassette MAGs and applied to relatives that are themselves source genomes of members of the RT database, so it re-observes the arrangement within one lineage and does not test it independently. The upstream element was not searched in this test. We therefore claim neither specificity nor evolutionary conservation for the arrangement: seeing the neighbour arrangement again in close same-lineage genomes is expected, and what is unusual is the strict spacing pattern itself. Within the GTDB R232 representative genomes we searched, the near clade defined by RT identity to the ESOL RT (a definition drawn from

the cassette MAGs, so partly circular) lies in the single genus *Pollutibacter*, and formal GTDB-Tk classification places all 16 cassette MAGs, which were selected because they carry the cassette, in that genus; these genomes are close relatives and not independent, K-family RTs and element-RT loci occur in more genera, and this is neither a statement of genus distribution nor a determination of host range.

A consensus structure of the aligned sequences has a 9-bp GC-rich stem, 2 inner hairpins and a 3' U-rich tail, with 12 covarying pairs against a maximum of 3 in column-shuffled controls. The shuffled null does not control for phylogeny. A covariance model for the upstream element [4], built from the same alignment and so circular, and a scan of 5,615 host-lineage genomes found 20 high-scoring hits: 16 in front of an RT (in 16 genomes), none in front of an RT-free repair operon, and 4 that lay inside a gene or too near a contig end to classify (Supplementary Table S4, table A20). A separate scan of *Pollutibacter* genomes found the element in front of an RT in 4 *Pollutibacter* genomes that are not among these 16, which with them gives the 20 genomes, 15 species and 4 genera quoted in the main text. The 794-genome post hoc search quoted in the main text used a different, partly overlapping genome selection; its 19 hits are a separate count, although some genomes occur in both. Sensitivity was measured on seed sequences only, there is no RNA-seq link, and the function of the element is not distinguished. The identity of the element is unresolved and its form description is not diagnostic: our R-scape covariation test found no significant covarying pair among the 23 sequences of the pre-specified alignment and had too little power to test a stem (Extended Data Figure 13), retron positive controls were few, unscanned genomes remain, and a terminator, an attenuator and an ncRNA cannot be told apart. In a post hoc scan, a pair of sigma-A-type promoter-like boxes (TTTG just ahead of the upstream element and TAHHTTTG inside its 5' end) is shared by many cassettes, but a similar shared pair is also found upstream of a non-trivial share of other conserved genes in the genus, the hold-out set declared in advance did not reproduce it at the homologous position, and the RNA-depth rule set in advance was not met, so whether the element carries a promoter remains unresolved (single genus; no transcription-start mapping such as 5' RACE).

Sequence-only re-alignment would place a D instead of a T in the fourth motif column, so the finding depends on the structure mapping; the regex background is not composition-matched, so its p-value is optimistic; and because the family is defined by cassette sequence, the motif fraction is a consequence of that definition. All of this is a motif-level observation within a single genus. Every evaluable ESOL RT ORF carries the aromatic-x-DD catalytic box (Extended Data Figure 2; Supplementary Table S5). The pre-specified regex missed two of them (corrected post hoc), the ORFs lacking a box are contig-end truncations, and the upstream Q.G is absent (the Q position carries a different residue); this is a sequence-level observation within a single genus. The +3 metalloprotease-like ORF lacks the canonical HExxH motif (searched as H[EDQ]xxH) in every ORF, against positive controls (MMP1 and the Peptidase_M54 flank ORF of the same loci, which carry it) and a low regex-shuffle background. This is a statement about the sequence check of the +3 proteins only: it makes no claim about how the absence arose or about enzymatic activity, and no experiment was done. The absence is not specific to the cassette but is a property of the whole +3 family (nearly all free-standing +3-like sequences also lack it, 2 of 127 carrying it, while most Pfam M54 members carry it; Supplementary Table S5, table T10, rows H-8 and H-9), and the independent unit is the genus (closely related ORFs of a single genus, ANI95 units). The low shuffle background supports that the positive-control recovery was not chance; it does not support the significance of the absence, because

composition-matched shuffled sequences would lack the H[EDQ]xxH search pattern in a set as large as our +3 ORF set about half the time (0.44 to 0.51, depending on the region shuffled; Supplementary Table S5, table T10, rows H-8 and H-9). The absence has meaning only if the predicted-structure-based “metalloprotease-like” tag is correct (fold-level evidence of intermediate strength; sequence profile not reproduced). The RecD-like gene at +4 retains its Walker A/B and ATP-contact motifs at the same level of integrity as the host RecD. In the 8 *Pollutibacter* genomes with a complete +4 ORF the island RecD is the only RecD-like protein found in the assembly, so intact motifs show only that it is a normal RecD, not that it is an effector, and the host-versus-island comparison is available in a single genome only. In a post hoc structure annotation, the predicted structure of the –1 protein resembles a von Willebrand A (vWA) fold, whereas the Pfam VWA profiles do not detect the cassette –1 genes, so a sequence-level link to vWA proteins is not shown; all structures are single-sequence predictions. The composition table with the circular and post hoc marks and the full notes is given in Supplementary Table S19. The company is odd: a protein carrying a MIDAS motif, a protein predicted to fold like a metalloprotease, a RecD-like helicase, proteins that match no entry of the standard family libraries and a conserved non-coding element, and the question is whether it truly goes with the RT.

Supplementary Table S19. 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	Evidence level
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)	family descriptive, circular; length descriptive, circular
–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.)	family descriptive; length descriptive, circular; copies negative result, circular; motif descriptive, circular; omega supported, circular, post hoc (partitioned value)
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)	0.088– 0.090 (16 seq.)	family descriptive; length descriptive, circular; copies negative result, circular; motif descriptive, post hoc (pattern widened); omega supported, circular
+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.)	family descriptive; length descriptive, circular; copies negative result, circular; motif no information; omega supported, circular
+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.)	family descriptive; length descriptive, circular; copies negative result, circular; motif no information; omega supported, circular
+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)	family descriptive; length descriptive, circular; copies negative result, circular; motif descriptive, circular; omega supported, circular

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	Evidence level
+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)	family descriptive; length descriptive, circular; copies negative result, circular; omega supported, circular
+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.)	family descriptive; length descriptive, circular; copies negative result, circular; motif descriptive; omega descriptive
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	family not recorded; length not recorded; motif descriptive, circular

The positions are operational (the order of ORFs from the RT, and an upstream element defined by a nucleotide model hit) and assign no activity, no biological role and no transfer of the set between genomes. A term such as vWA-like or metzincin-like is a clue and not an assignment, and the +3 parts are aligned slices of the +3 family. Observable copies and motif counts refer to the assembled loci of one genus; a copy cut by a contig end or lying beyond it is not scored. The columns refer to different sets, so their denominators differ: motif counts are evaluable ORFs, observable copies are observable intact ORFs of the 22 loci, and layer counts are ORFs of the seed contig annotation, as are the length ranges and the element length of the cassette positions; in the RT row the motif denominator, the observable copies and the layer set are different sets. The ω range spans the unpartitioned and the recombination-partitioned codeml M0 estimates on representative sequences of the intact copies; a range end marked post hoc is the partitioned value of an interim recombination model for the –1 slot, which Supplementary Table S6 marks as post hoc. Evidence level gives supported, descriptive, negative result or no information, with circular for quantities measured with sets or models built from the cassette and post hoc for a rule changed after its output was seen. The xseA-like term for slot –3 is a clue and not an assignment: it is an annotation label read from a library model and it was not tested by the layer searches used for the other rows. Because the detectors and family models were built from members of the cassette, detection is circular, and a single genus is the only independent unit.

S1.3 The company is found with the RT inside one genus, and nowhere else we can see

The 27 loci are RT loci and are 19 units, counted as 6 ANI95 clusters plus 13 samples without an ANI computation. This is a census inside one genus: the detector models derive from the 16 seed MAGs, which are not counted as evidence, and the counts test neither function nor origin. In a post hoc

stratification by RT identity, added after the first output was seen, the other 26 of the 27 loci, each with an RT at least 0.70 identical to the cassette RT, show no absence at all (element 20 of 20, -1 17 of 17, +1 25 of 25, +2 18 of 18, +3 17 of 17), so the exception sits at the one locus whose RT is least similar to the cassette RT, within one genus.

This is a statement bounded to this detector and this scanned range, and is not evidence of absence. The absence therefore concerns RT lineages distant from the cassette lineage: RT identity to the cassette RT is at most 0.33 outside, and in a post hoc identity split no outside locus reaches 0.70, while the 26 loci that do are all *Pollutibacter*. For the -1 part the same detectors also find no outside locus (0 of 445), but they do call the -1-type protein in 8 and 9 of the 1,127 control windows, and the leave-one-out family profile detects every real -1-type pair-family member tested in the identity bands (11 of 11 and 19 of 19; Supplementary Table S4, table R01; Supplementary Table S4, table R02); because those members were found with the same family so the real-member curve is optimistic, the -1 zero is limited to the reach of these detectors. A cassette-like arrangement, a -1-type gene upstream and a +3-type gene downstream of the RT, is found at none of the 322 outside loci with observable windows; a structure-based detector calls the -1-type protein at 1 of 329 evaluable loci (an alignment artefact of a short fragment) and the +3-type protein at none. Of the 445 windows, 103 are complete on both sides, the 1,127 control windows were built from non-K RVT_1 RTs, and every detector family derives from the 16 seed MAGs of one genus, so a variant unlike the seed is not seen; a pipeline test on the 16 seed loci finds the arrangement at 8, the -1-type protein at 15 and the +3-type protein at 9, which is circular and tests the logic only. At 2 of the 445 outside loci, both in *Ferruginibacter* with element-model hits, the profile detectors call a +1-type and a +2-type protein beside the RT and no -1-type or +3-type protein; one of the two genomes was already in the table of +1/+2 pairs of the pair scan (Supplementary Table S4, table P02), so this repeats a known observation and shows that +1/+2-type pairs occur beside an RT outside the genus. Among the labelled outside genera, 6 of 43 with an observable element window carry element hits, and 11 outside loci without a genus label are not counted as genera. The element hits are hits of a model built from the cassette elements, so they are partly circular in the genomes and genera that contributed to it and are bounded by a divergence-dependent detection limit, and the number of genera with hits is descriptive and does not show separate acquisitions of the element. Counted from the other side, 21 of 27 *Pollutibacter* anchors defined by a -1 gene but 0 of 10 anchors outside the genus have a K-family RT beside them, and 5 of 22 dense +1/+2 pairs have a K-lineage RT within the flank window (1 of 4 in *Pollutibacter*, 4 of 18 outside; 3 of 5 pairs with complete windows); these are small descriptive anchor sets that include training genera; the 10 outside anchors are weak -1 hits in 3 genus-level units whose flank windows are mostly cut by contig ends, so 0 of 10 rests on few observable windows (the same holds for the anchors without an RT inside the genus), every outside pair with a K-lineage RT comes from a genome that contributed to a detector family, and the counts are not read as support for or against the existence of parts without an RT outside the genus. In a separate implementation of the cassette-derived part detectors anchored on RT genes that pass a strict cassette-RT model (86 genes in 77 genomes of 53 genera among the processed GTDB representatives), the 3 *Pollutibacter* RT genes have a strict -1 gene in 3 and a strict +3 gene in 2 (both in 2), whereas of the 83 RT genes in 52 other genera none has a -1 gene, one has a +3 gene (genus CAMDMI01) and none has both; the detector models derive from the cassette and contig-end truncation is not distinguished. The genus remains the independent unit: within the range we

searched, and with detectors built from known cassettes, the cassette is seen in one genus, *Pollutibacter*, and near relatives within that genus are not counted separately. Genera with element hits are not thereby genera that carry the cassette.

Within *Pollutibacter* the co-occurrence is asymmetric: of 11 RT-bearing genomes (among 190 examined), 10 carry a strict -1 gene, 10 a strict $+3$ gene and 10 the element somewhere in the genome, while the RT-adjacent windows are filled in 8, 7 and 5 (both) of them. The element behaves differently: all 10 element-bearing genomes carry the RT (10 of 10), a partly circular value because the element model derives from the cassette. The 11 RT-bearing genomes are the same genomes as the GTDB panel of the anchored census, and the genus in which the cassette RT clade was found is the single genus of the range we searched. Within windows a strict -1 or $+3$ gene has an RT of any kind nearby only rarely (30 of 904 and 15 of 374), which is 6.9 and 10.2 times a positional-null expectation, while none of 157 re-detected $-1/+3$ pair loci has a strict cassette-lineage RT (151 of the 157 have no RT of any kind; this is not the count of 151 adjacent RT-free pairs of Supplementary Results S1.8). The genome-level and paired-locus counts are post hoc, whereas the anchor counts of the preceding sentence were pre-specified, all under the strict -1 and $+3$ anchors of the pair-recurrence test, and the asymmetry is one of conditional frequency, not a measure of dependence. So the RT looks like the piece that can be missing: none of 157 re-detected $-1/+3$ pair loci has a strict RT of the cassette RT clade, and 151 of them have no RT of any kind. In a post hoc genome-level count, 2 of the 11 RT-bearing *Pollutibacter* genomes also carry an RT-free adjacent $-1/+3$ pair on another contig, which is genome-level coexistence and not a same-locus relation; the two genomes fall in one ANI95 species cluster, so a statement that pair-bearing and cassette-bearing representatives are disjoint holds only for the GTDB species representatives and not for the whole genus. The next question is how widely the parts occur without it (Question 3).

S1.4 No copy we can see is disrupted, and the codon models are consistent with purifying constraint

As a positive control for the lesion census, synthetic single lesions planted in 12 module ORFs that were observable and intact were called disrupted in 143 of 144 tests; the 12 ORFs were picked by a fixed rule rather than at random. Of the 80 core positions on the 16 seed contigs, 60 carry a called ORF, all on the RT's strand, and each of the other 20 lies where its contig ends, with at most 165 nt left, so no empty position is a gap within an assembled contig; on 6 contigs all five core ORFs are complete. The strict-form rejection is bounded to pseudogene signatures in the assembled contigs of observable module instances: the independent unit is a single genus (a generous allele-cluster upper bound), dispersion saturation inflates the neutral-drift expectation (the expected numbers are an upper-bound artifact), no host pseudogene baseline was available, read support is likewise unmodelled, and identity percentiles do not distinguish non-vertical history. Keeping one MAG locus per study (27 combinations, the 6 public-genome loci always kept) leaves 37 to 46 copies with none disrupted; the public-genome loci carry no study information, and the independent unit remains the genus. The controls of the lesion census and the limits of the exact neutral calculation are in Extended Data Figure 5.

Partition-reflecting and unpartitioned ω are close in every module (largest difference in one module), and both sit far from the neutral value in every module, with FEL negative sites (without synonymous-rate variation) overwhelming and no positive sites (one in -1 unpartitioned only; up to

two per family with synonymous-rate variation; the power of these tests was not resolved) — the same purifying direction as the RT value and the pairwise values reported in the Supplement; the +3, unjudgeable by the saturated cluster-pair route, is judged under an ML tree model instead (Supplementary Table S6; Extended Data Figure 3). By a rule written down before the calculation was run (although exploratory values from a preliminary closed-form calculation had already been seen), an exact calculation on the observed trees rejects the strict neutral-resident codon model for each of the five gene series, which are adjacent genes of one locus and not independent tests: in all 30 decision cells the probability that every leaf is free of premature stop codons is far below the rejection level fixed beforehand ($\log_{10} P$ from -6.57 to -54.82), and the verdict flips only if the time scale is overestimated at least 6.2-fold; the calculation covers substitutions only, rests on stop codons sitting at a composition-neutral equilibrium, and does not say what keeps the copies intact. The calculation counts substitutions only, so frameshifts, truncations and recombination are outside it, it uses no host pseudogene baseline, and a disrupted copy removed or not recognised before the trees were built would bias the comparison toward rejection; the five series are neighbouring genes of one locus in a single genus, so their probabilities are neither independent nor to be multiplied. Recombination attribution is not established: GARD finds breakpoints in every module and topological-mismatch support in most, but the same pipeline gives topological support for all of the few host single-copy control genes, so this is not a module-specific signal. The power simulation for the positive-selection tests could not resolve their power because of its design limits, so the lack of positive sites does not show that positive selection is absent. There is no evidence that the N region of the +3 alignment is under weaker purifying selection than the rest, which is not evidence that the regions are equal, because the design has the stated power only for a difference of about 15.0 percentage points. The numeric FEL decomposition goes to the Supplement, and the counts depend on the P threshold and are not corrected for multiple testing, so they are not a fixed property of the alignment; the alignment belongs to a single genus and the family compared is defined from the cassette, so the comparison is internal to that genus. In a post hoc comparison with the host's own coding genes in each cassette-bearing MAG, the cassette RT's third-position GC (GC3) was not an outlier (median z 0.42, range -0.92 to 1.22 , no z above 2); this range-limited observation, made within a single genus, does not date the RT's presence at the locus or say whether the RT or the module was there first. So no copy we can observe is disrupted, and the codon models are consistent with purifying constraint.

S1.5 Single unassembled long reads carry several modules of the cassette

In the frozen stratum, 10 of the 15 reads cover two or more cassette modules, the class that the screen rule defined before the reads were seen, and a post hoc recount with that rule gives the same 10, together with 4 reads covering four or more modules and one read spanning six, for which no cut-off was fixed beforehand (Supplementary Data S2, OBS_READS, column `n_cassette_modules_covered`, frozen layer). Among reads from samples and studies that contributed no genome to the reference, 36 reads cross a junction between two adjacent cassette modules inside one alignment block (11 in the frozen layer and 25 added post hoc; 8 BioSamples, 4 studies), counted as reads that cross at least one junction (91 crossings in all; the 34 reads that cover two or more modules are all among them), although one river study supplies 32 of the 36 reads, so the effective independent units are the few water systems and not the reads.

The frozen stratum alone gives 10 multi-module reads against 8.60×10^{-5} under M3 (ratio 1.16×10^5 , $\log_{10} P -47.2$). The 34 observed reads over all strata exceed the expectations of 1.95×10^{-6} (M1, run-wise apparent split rate), 1.61×10^{-5} (M2, every cassette read in the pool) and 1.61×10^{-4} (M3, the $p = 1$ ceiling that treats every read as a chimera) by factors of 1.75×10^7 , 2.12×10^6 and 2.11×10^5 (Supplementary Data S2, file SI-12_RANDOM_PAIRING_BY_RUN; Supplementary Data S2, file SI-12_RANDOM_PAIRING_TOTALS). One PCR-selected study supplies 31 of the 34 multi-module reads and 43 of the 49 cassette reads, and the frozen stratum alone gives 10 reads against 8.60×10^{-5} under M3. Of the 49 long reads realigned to the 16 cassette-bearing contigs, 48 fall in one alignment block and 1 in two blocks that the pre-specified rule leaves untestable, so no read is scored as carrying module order or orientation at odds with the reference, although module order was in effect not tested between blocks. Nothing here decides whether an individual read is authentic or what the arrangement does, and the cassette's independent unit is a single genus. So long reads exceed what random pairing of two molecules allows, and the next question is whether any trace of the cassette is found in RNA. This is model-conditional: the expectation is near zero by construction, so the comparison is near-trivial, as declared before the result, and non-random pairing such as template switching between homologous sequences in a PCR-selected library was not tested. Of the 49 long reads realigned to the 16 cassette-bearing contigs, 48 fall in one retained alignment block and 1 falls in two blocks that the pre-specified rule leaves untestable, so 0 reads are scored as carrying module order or orientation at odds with the reference. Of the 34 reads in the multi-module denominator, 33 are a single block and therefore collinear by definition and the remaining 1 is the untestable read, so the zero count is a statement about single-block alignment rather than a passed order test, and no multi-module read had its order and orientation decided between blocks by the pre-specified rule. The single-block count is reference-dependent (5 of the 48 align to two contigs of the same MAG in a reference that holds more contigs than the 16-contig reference), the reads were selected by a prior screen rule, the host genus of the reads themselves was not determined, and the synthetic chimera control builds only cassette-plus-cassette junctions, so its detection power for cassette-plus-non-cassette junctions is unknown (Methods). In a post hoc auxiliary check outside the pre-specified decision rule, 70,978 evaluable long reads aligned to a panel of 142 public reference genomes from 81 genera (Supplementary Data S2, file SI-12_REFERENCE_PANEL_GENOMES) gave 1.49% (95% CI 1.41% to 1.59%) unaligned inside one replicon and 0.85% with two pieces from different genera (Supplementary Data S2, file SI-12_REFERENCE_PANEL_SPLIT_BY_RUN), and the PCR-selected libraries had the lowest unaligned-inside-one-replicon fraction (0.68%, against 3.11% for random and 3.46% for other libraries; Supplementary Data S2, file SI-12_REFERENCE_PANEL_SPLIT_BY_LIBRARY); dividing the different-genus fraction by the mapped fraction (5.01%) gives 16.9%, an inflated ceiling that mixes switching between close relatives with conserved sequence and is not a chimera rate. Contig ends limit what can be observed in the 16 cassette-bearing seed contigs (1,498 to 100,017 bp; 16 sequencing runs, 16 BioSamples, 10 studies): a module reaches the contig end on the downstream side in 8 of 16 contigs and on the upstream side in 0 of 16, the RT itself reaches an end in 4 of 16, and 6 of 16 carry all five families with none cut by a contig end; among the 31 cassette-type loci of a wider catalogue scan the slot upstream cannot be observed at 9 and the third slot downstream at 18, so counts of observable positions rest on these denominators and the lack of a slot is not evidence that the slot is absent.

The focus analysis uses one Canadian LakePulse sample and one Chilean sample; the Chilean bin and the reassembly come from the same sample and the same reads, so they are not independent samples, and the source label of the Canadian sample was checked against ENA records. A Chilean locus has an RT ORF that is identical at the protein level to the ESOL RT.

S1.6 Traces in public RNA and in culture DNA: present, but RNA and DNA are not told apart

A cassette that works should leave traces in RNA, so we looked in public RNA data for sequence that matches it, knowing that public data cannot show transcription. In one public pond RNA run, a post hoc search found reads that match genes of a cassette-bearing contig at near-complete identity, but they came from few molecules, some were not lineage-specific, and none matched the upstream element, so this shows low-level presence of such reads in one environment only and does not show that the cassette, or its element and RT together, are transcribed. In a post hoc reading of public libraries known to be strand-specific, cassette-matching reads from a subset of runs met a pre-set strand rule and lie in the orientation that the library design and the host polarity convention predict (tallies in Supplementary Table S7); this is a conditional polarity cue only, because the strand, reading-frame and paired-DNA checks use the same reads and the same convention and count as one line of evidence, and it does not show an RNA origin.

In a post hoc, assembly-free read check of the same culture series, 0.5% to 1.46% of the cleaned reads of the four conditions place uniquely on the 16 seed MAGs, and for two seed MAGs the depth of the cassette contig divided by the median depth of that genome's contigs of at least 5 kb is 0.7 to 1.13 and 0.82 to 1.08 across the four conditions; this shows depth consistency only and says nothing about which contig or which organism carries the cassette. In a post hoc comparison, 24 of the 93 variant contigs that a sweep assembled from RNA sequencing runs of the culture study match that study's DNA assembly at 95% identity or more over at least 0.90 of the query (16 at 99% or more), and 71 of the 93 have a DNA hit of any quality; this is a sequence-level correspondence between two assemblies within one study and a single genus, counted in contigs rather than independent units, and it is not evidence of transcription or expression.

S1.7 Where in the world: lake, river and pond waters on five continents, and one genus

Each of the 16 samples has its own BioSample record, 15 of 16 carry a coordinate pair and 16 of 16 a collection date. Continent is not a recorded field: it is derived from the country token by a country-to-continent rule table frozen before the fill, which yields 5 named continent labels and leaves 3 units unclassified because the table treats "Korea" as a historically undefined token (Figure 8); mapping Korea to Asia is a post hoc amendment made after the result was seen. Public long reads that match the cassette come from freshwater samples and studies that share neither a sample nor a study with the genome sources: the frozen layer holds 15 reads from 7 samples in 4 studies, 10 of them spanning two or more cassette modules by the rule of the screen (a recount made after the hits gives the same count), and no reverse-sequence decoy read matched. The reads were recognised by their similarity to cassette-derived sequences and so are circular evidence for the cassette itself. Because several units share a submission project and the genomes entered the set by carrying the locus, the country and continent counts describe the span of the source-sample records only and support no prevalence, distribution or independence statement; the independent unit for the cassette is a single genus. The fresh-water labels of Extended Data Figure 7 come from a dictionary reading of BioSample free text

that did not pass its manual pre-check, so every habitat statement there is restricted to the exploratory tier; for the cassette-bearing genomes the genus background is already fresh water, so these data cannot test a habitat effect and the labels are not an environmental measurement. Separately, the genome-level fresh-water excess of K-family RTs in the Genomes from Earth's Microbiomes (GEM) catalogue is carried by one *Chlorobium* OTU.

The 16 source samples, from 9 countries on 5 continents (continents derived from the country text), are all water samples (the genomes entered the set by carrying the cassette, so this describes the records and not where the cassette lives): 6 lakes (among them Lake Kivu in Rwanda and Lake Lanier in the USA), 3 samples of one river, the Chaobai River in China, 2 ponds in Israel and in Maryland, USA, and 5 described only as fresh, surface or aquatic water (Figure 8; Supplementary Table S17). Separately from these source records, in the catalogues we scanned (SPIRE, the Searchable planetary-scale microbiome REsource, and part of GEM), loci with a -1 gene beside a K main-clade RT came from freshwater lake samples only, and most were repeats of one lineage nearly identical to the known *Pollutibacter* cassette RT, within a partial sample and with this detector. In a post hoc keyword reading of the environmental-description text of the 253 distinct core units of the habitat analysis, 32 carry no water-body noun and 5 carry explicitly non-aquatic text (soil, decomposing wood in a forest biome, wetland soil, a bioreactor and contaminated soil), so a filled field does not by itself mean that its text names a water body. In a post hoc reading of two pre-specified database tests, hits to the $+3$ C-domain family fall mostly in Bacteroidota and show no freshwater enrichment: freshwater samples are a smaller share of its environmental hits than of random background sequences, although the primary set only marginally met the pre-specified rule for a restricted habitat and some comparison families of unannotated genes behaved similarly; these counts are sequence-weighted, rest on incomplete taxonomy and were not tested against a genus-level null, so the water sources of the cassette-bearing genomes should not be read as a habitat property of this gene family.

The locus atlas table (Supplementary Data S4) is an accounting ledger of 4186 rows over three source tables, 489 of them nucleotide-window rows, and not a detection. Rows are not loci, because the same locus can appear in more than one catalogue, so row count is not locus count. The evidence behind the K-positive call is not uniform across the atlas: in the source table, 182 of the 1399 SPIRE-derived rows have a profile score of zero and were called K-positive, while the other 4004 atlas rows carry a positive profile score. This is a property of the source table, made visible by a post hoc per-row flag, and not a new detection; the flag leaves the source K-positive calls unchanged and neither shows the calls to be wrong nor confirms them. The genus-unit column is filled only for the 489 nucleotide-window rows and holds 48 distinct values, of which 45 are labelled genera — the rest are family-level or unknown placeholders, so a genus count over this column is 45 and not 48.

S1.8 Two pieces that meet in one genus: the $-1/+3$ pair recurs without the RT, and the element is found with it

Dense same-strand $+1/+2$ pairs in cassette order also occur in several Bacteroidota genera other than *Pollutibacter*, so a second pair of parts, not only the $-1/+3$ pair, turns up outside *Pollutibacter*. The pair-family count is not a count of independent origins, the genera are not phylogenetically independent, the family defined from cassette sequences may be circular, and the single-genus bound of the cassette itself (*Pollutibacter* across the range swept) is not replaced by this pair recurrence. In a post hoc check of the 54 pair genera among 149 anchor genera in the confirmatory population (ratio

132.4 to a positional null), dropping any single order leaves at least 40 of them, the minimum being reached when Chitinophagales is dropped (40 pair genera among 124 anchor genera, in 21 families; ratio 126.2), and dropping any single family leaves at least 44. These removals are re-counts of the same 54-genus observation, defined after the primary result; they do not re-run the pre-specified rule, no P value was recomputed after a removal, and they are not an independent line of evidence. The ratios are taken against a uniform positional null that does not model local gene order, so their absolute size may be optimistic.

A recurrent palindromic core (CAAGCTTG) sits upstream of the start of the $-1/+3$ operational units, discovered in the training genera and verified in the held-out genera, which meets the pre-specified rule as written. The strict site (FIMO p at most 10^{-4}) is present in 78 of 103 processed genera (held-out 51 of 66, training 27 of 37), the exact CAAGCTTG core in 47 of the 103 genera, and 25 genera carry no strict site — so this is not conservation across every processed genus or of the full palindrome. The genera are not independent lineages, co-inheritance of a mobile locus is not excluded, whether the +3 protein binds this palindrome has never been tested, and the “non-coding” placement is an exploratory analysis, so we do not read the palindrome as a regulatory motif or an operator.

In a post hoc tally over the 372 genomes that contribute only K-family RT loci outside *Pollutibacter*, this detector set labels 139 ORFs in 126 genomes as -1 family and 174 ORFs in 111 genomes as $+3$ family, whereas its $+1$ and $+2$ models label none; this describes this detector set, this ORF calling and this genome set only, the models derive from the seed loci, and a count of genomes is not a count of independent units. In a post hoc leave-one-out check of the 26 pairs found by the cassette-derived strict detectors, dropping any single order leaves at least 18 pairs in at least 12 genera and dropping any single genus leaves at least 20, so no single order or genus accounts for all of them; the 3 pairs with an RVT_1 gene between the two genes all lie in the cassette lineage and so vanish when its order or genus is dropped, as follows from how they were selected. In a post hoc control, vWA-like genes outside the cassette-derived -1 profile showed no excess adjacency to $+3$ -type genes over a positional null, but so few were expected that the control has little power, and both the -1 and $+3$ calls derive from cassette sequence. In a post hoc held-out analysis, recurrence of the RT-free cassette-part pairs was also observed in genera and families left out of model training, meeting the pre-set recurrence rule for presence only; because the training pairs were found with the cassette profile models, recall is circular, the held-out genera share orders and families with the training genera, and the rotation null over-predicts ordinary adjacency.

Counted again from the genomes, the strong $-1/+3$ -type pairs number 178: the 158 rows of the pair table of the first count were all reproduced (152 of them with no RT of any kind in the window) and 20 further pairs were found, all without an RT of any kind in the window; of the 172 pairs without an RT, 151 have the two genes adjacent (131 of the first-count rows and the 20 further pairs), and 9 of these overlap by exactly 4 nt (6 and 3) (Supplementary Table S10, sheet F2). At the 151 adjacent pair junctions, direct repeats sit at the nominal boundary (8-mer: 35 observed against 27.9 expected, one-sided $P = 0.050$) and inverted repeats show no excess (7-mer: 27 against 23.0, $P = 0.18$) over 48,186 null regions, and the two cassette-specific motifs of the -1 to RT region hit 1 and 0 of the 151

junctions against 26 and 33 of 3,020 length-matched random regions; this gives no clear sign that the RT, +1 and +2 were inserted between the pair genes, and it does not exclude an insertion because the junctions are short and the power is low.

A second rooting attempt, with every strong –1-type and +3-type hit of the census, a closer outgroup, three alignments and a profile-mixture model, met the saturation criteria in all 30 tree variants (Supplementary Table S4, table O08) (median outgroup-to-ingroup distances of 4.29 to 6.82 substitutions per site for the –1-type tree and 4.24 to 9.09 for the +3-type tree, against a limit fixed beforehand), so it was untestable again; this is a limit of the data and not a result about direction, and the cassette sequences come mostly from one genus. Trees built from predicted structures did not settle the direction either: in none of the five structure-based variants of either gene tree (ten trees in all) did the direction of the RT meet the pre-specified support rule (bootstrap support could not be computed, so only SH-aLRT was used, and the cassette leaves did not form one clade in any tree), so the direction of the RT remains unresolved and neither a gain nor a loss is supported. In the structure-based distances none of the saturation criteria was met, but this gives no evidence that the saturation problem has been overcome: the 3Di substitution units are not comparable with the amino-acid limit, the amino-acid tree guided by the same structural alignment exceeds the limit for both distances in both families, and in the +3-type set the outgroup-to-ingroup template-modelling score (TM-score; 0.734) is higher than the cassette-to-free (0.684) and free-to-free (0.669) scores, so this score does not rank evolutionary distance. Post hoc, in the FastTree variants the nearest RT-free leaf of almost every cassette leaf is a sequence at a contig end whose RT window cannot be read to the end (25 of 25 in 5 of the 7 –1-type variants and 24 of 25 in the other two; 18 of 18 for the +3-type; Supplementary Table S4, table O09), and only a few free sequences serve as nearest leaf for all cassette leaves (at most 2 distinct for the –1-type and at most 4 for the +3-type), so these nearest free leaves cannot be counted as genes without an RT and may belong to fragmented cassette loci; the median distance from a cassette leaf to its nearest free leaf across the analysed variants is 0.63 to 1.76 (–1-type) and 0.31 to 0.76 (+3-type) substitutions per site, and 2.93 to 4.36 and 1.53 to 4.84 after the window-unobservable free sequences are removed, so the post hoc ranges reach into the saturated range. Post hoc, in the seven FastTree variants of the second attempt, the nearest free sequence of the 25 –1-type cassette leaves is one and the same *Pollutibacter* sequence at a contig end for all 25 in two variants and for 24 in the other five, in which the one other nearest free sequence, from a different genus, lies at 2.56 to 3.58 substitutions per site; for the +3-type the nearest free sequence of each of the 18 cassette leaves is a *Pollutibacter* genome sequence or an unlabelled sequence from a metagenome-assembled genome of the seed runs (2 to 4 distinct sequences), so the closest relatives of the cassette genes in the genomes searched are same-genus or seed-run fragments at contig ends, and no –1-type gene of another genus lies closer than about 2.56 substitutions per site in any of the seven variants (Supplementary Table S4, table O09; Supplementary Table S4, table O14). In the FastTree variants of the second attempt, the 22 RT-free adjacent-pair members of *Pollutibacter* lie far from the cassette leaves: the distance to the nearest cassette leaf has a minimum of 3.15 to 4.58 (–1-type) and 2.38 to 4.48 (+3-type) and a median of 3.34 to 5.04 and 2.65 to 5.05 substitutions per site; the –1-type distances are in the saturated range in all of these variants, whereas for the +3-type the medians of only 3 of the 7 FastTree variants reach the saturated range, so the trees show that pair members and cassette genes are far apart, and the distances do not measure how long ago they separated. With a closer Bacteroidota outgroup of 60 sequences, the outgroup is monophyletic in 11 of 11 –1-type trees

and 8 of 11 +3-type trees, whereas in the first attempt the -1-type outgroup was not monophyletic in any variant and the +3-type outgroup was monophyletic in all variants; for the +3-type the weak profile scores make the homology of the outgroup uncertain, and a monophyletic outgroup does not make rooting testable when the distances are saturated. In approximately-unbiased root tests on two maximum-likelihood trees, the root candidates that place the outgroup cluster inside the cassette cluster were all rejected for the -1-type tree (the maximum-likelihood root has p_AU 0.998 among 52 candidates and the best cassette-side root 0.00407), but not for the +3-type tree (maximum-likelihood root 0.93 among 40 candidates, best cassette-side root 0.289, 11 candidates with p_AU of 0.05 or more; Supplementary Table S4, table O10); the two families differ, and neither result changes the finding that the trees are saturated. Descriptive classifications of where the cassette leaves sit relative to the outgroup-rooted trees depend on the leaf set and were not used for the decision: the pre-specified rule with all cassette leaves gave a paraphyletic placement in 30 of 30 variants, a post hoc version without the window-unobservable free sequences gave a nested placement in 24, a basal placement in 1 and an undefined root in 5, and a post hoc version with only the cassette core gave a nested placement in 30; all 30 variants were untestable, so none of these labels describes the direction of the RT (Supplementary Table S4, table O08, for the 30 variants and the pre-specified class; Supplementary Table S8, row P06, for the two post hoc counts). In stochastic mapping on the trees that have a defined outgroup root (5 trees of the -1-type, 7 of the +3-type), the mean number of RT-in-window gains and losses was 4.07 to 29.42 gains and 1.04 to 15.8 losses (-1-type) and 8.91 to 17.93 gains and 4.79 to 14.46 losses (+3-type), while for the adjacent partner gene the mean losses exceeded the gains (-1-type 17.25 to 63.84 losses against 11.81 to 20.22 gains; +3-type 29.96 to 50.73 against 12.83 to 17.51); these counts are conditional on the root, are descriptive, and carry no weight for inference because all variants were untestable (Supplementary Table S4, table O11). Searching the genomes with detector hits in the census together with the *Pollutibacter* genomes (not an exhaustive census) with the same detectors, the position rule alone gave 32 -1-type and 20 +3-type cassette leaves, more than half of them in the *Pollutibacter* lineage and the rest in other genera or unclassified genomes; the position rule marks lineage membership and not identity of composition. Gene trees of the RT-free pair members agree with the host tree much better than a permutation null (patristic-distance correlation $r = 0.51$ and 0.66 ; permutation $P < 0.0001$; genus monophyly 11 of 26 = 0.42 against a null of 0.002 and 0.001; Supplementary Table S4, table O12), but fall short of the pre-specified gate, and the median between-genus distance of the +3-type gene tree, 3.16 substitutions per site, is in the saturated range. Within *Pollutibacter*, on a tree built from core single-copy proteins, the RT-free pair state needs 12 changes against a permutation mean of 13.1 ($P = 0.23$) and the cassette state needs 4 against 3.95 ($P = 1.00$), so no clustering of either state is detected at the level of these units; the power is low, and the result neither supports nor excludes a shared history of the two states within *Pollutibacter*.

S1.9 Is this orphan unusual, and how far can the search be trusted?

To ask whether this RT is unusual or one of many unassigned lineages, we ran untargeted searches that did not use it as the target and measured how often the pipeline would have flagged a lineage that is known (Extended Data Figure 15; Extended Data Figure 14). The hunt scored every protein (144,625 windows, 1,797,083 proteins) and clustered them into 1,735 groups, which gave 329 significant groups; of the top-ranked clusters 15 had no name-level marker of a known RT type and became candidates (Supplementary Table S1, table J). Only 8 members per cluster were checked for

DGR repeat pairs, group II intron features and neighbouring genes (the nearest identities were computed for every member), the Toro control was tested only in Bacteroidota, the 9 clusters beside retron lineages were placed by identity, without a retron tree or msr/msd analysis, and the identities come from alignments of different coverage (median query coverage 52.5% for the ESOL queries and 92.6% for the members of the candidate clusters). An untargeted Pfam hunt of enriched neighbour domains over the same windows also recovered the ESOL cassette, although it recovered the positive controls only in part. The Pfam hunt scanned all 30,134 Pfam-A models (Pfam 38.2) [5] and recovered 2 of 4 testable UG34 NERD clusters, 2 of 3 UG35 UvrD-like clusters, 114 of 125 CRISPR-RT-Cas1 clusters and 151 of 216 DGR clusters (Supplementary Table S1, table K). Its most prominent unassigned candidate, a short RT with an adjacent HTH protein found mostly in Patescibacteriota (33 of its 36 RTs; 36 RTs in 34 genera in all), matches the retron type VI layout and is classed as close to known retrons; its effector, ncRNA and function are unknown. Among 52,515 metagenome-assembled genomes [6], only 12 of the RT groups seen in 2 or more samples are as isolated as the ESOL RT, with a nearest identity of 35% or lower, and the comparison is approximate. In those genomes 23,072 RTs formed 3,968 groups at 50% identity, 497 of which had no neighbour at or above 50% identity in GTDB or in a library of 6,039 published RTs (86 of them seen in 2 or more samples; Supplementary Table S1, table J). Most groups come from a single sample, part of the metagenome-only class reflects sparse sampling, the comparison with the ESOL RT is approximate because its value refers to the published library only, and none of the groups has structure or tree follow-up, so we call them candidates and nothing more. From the Logan resource [7], which holds assembled unitigs and contigs for the Sequence Read Archive, we clustered 952,088 RT sequences without a close match in the published library with GTDB, library and ESOL RTs, and recovering the neighbourhoods of 24 metagenomic clusters found none of the 24 to be a close relative of the ESOL RT; the result is preliminary. We drew an RT pool from the Logan resource, which holds assembled unitigs and contigs for the Sequence Read Archive. From the RT hits we took those with no close match in the published library and kept sequences of plausible RT length without internal stops and with a YxDD-like motif, which left 952,088 sequences. The scans were incomplete, nearest identities came from DIAMOND in a lower-sensitivity mode than planned (re-searching 40,000 sampled queries below 40% identity or without a hit in the planned mode raised 8 of them to 40% or more; Supplementary Table S1, table J), the unitigs are short, the selection favours widespread sequences, and the clusters are candidates only.

The neighbour-domain axes did better, recovering 9 of 25 lineages with a testable cluster for the DefensePredictor axis (36%) and 33 of 37 for the metagenome axis (89%; Supplementary Table S12, file SI-28_removed_lineage_recall), and the Logan axis could not be measured.

S1.10 Two pre-set predictions: a contrast at matched identity, and a tree test that told us nothing

We tested two predictions that were written down before the results were seen. The first was that profiles rebuilt from a reference pool, which stands in for the shipped training set of existing detection tools, call known members of related lineages at the same amino-acid identity as the ESOL RT but do not call the ESOL RT; the second was that the gene trees of the neighbouring cassette families agree with the ESOL RT tree. In a post hoc control declared after the main results, removing only a group's own profile from the shipped DefenseFinder library left 1 of 336 UG17, 0 of 484 UG3 and 0 of 512 UG8 members called by any other reverse transcriptase profile, so the 0 of 16 cassette RT

units called with the unchanged tools cannot be told apart from the outcome for members of these three groups when their own profile is absent; only these three groups were measured, the count uses DefenseFinder component hits only (PADLOC was not counted), and the cassette RTs come from one genus (Supplementary Table S2, table Q4). A single profile built from the union of the three known groups (UG17, UG3 and UG8; 1332 members) called none of the identity-matched known units at any of four thresholds (0 of 13, 0 of 49, 0 of 234 and 0 of 300), so its 0 of 16 for the cassette RTs carries little information; in a post hoc variant, 4 of 13, 18 of 49 and 35 of 234 matched known units were called at the three highest thresholds against 0 of 16 cassette RTs (nominal two-sided Fisher P 0.03 and 0.0031 at the two highest thresholds and larger at the third) and 0 of 300 at the lowest (P 1), with the 16 cassette RTs counted as units although they come from one genus (Supplementary Table S2, table Q2). At the lowest of the four thresholds the per-group rebuild of the emulation of existing detection tools could not be run for UG17, so there that emulation covers 246 of the 300 distance-matched units, whereas the single union profile covers all 300; at the three highest thresholds UG17 units are inside that emulation (2 of the 49 matched known units at the primary threshold). In a post hoc follow-up run after the main results, PADLOC 2.0.0 applied to the single-protein inputs of the detector test gave a component hit for 1184 of 1332 known UG17, UG3 and UG8 members and for 0 of 16 cassette RTs, and replacing the emulated component flag by the native one changed the any-tool call of no protein and left the pre-specified verdict category unchanged at all four thresholds; these are raw counts that are not matched for distance from known members, the known members lie inside the profiles' own training sets, and the check replaces only the PADLOC component flag, not the reference-pool rebuild used to emulate the held-out profiles (Supplementary Table S2, table Q3). Rescaling the domain E-values of the emulated profile hits to a genome-sized file of 5000 sequences (a conversion of E-values and not a rerun of the classifier; post hoc) gave a hit for 1166 known members against 1170 in the emulation, and the cassette RT count stayed 0 over a wide range of search-file sizes, so over this range the zero did not depend on search-file size (Supplementary Table S2, table Q3). The cassette side has a single independent unit, the genus, and the two-sided Fisher P of 0.0417 collapsed to that genus is nominal, and the result does not speak to what the shipped tools would themselves detect. Under a rule sealed before any result was seen, known UG17, UG3 and UG8 members whose closest reference sits at the same DIAMOND amino-acid identity as the cassette RT (24.8 to 33.3 per cent, median 28.35, for the 12 full-length cassette RTs) were scored against profiles rebuilt from the reference pool after each member's own subclade was removed. At the primary threshold the 50 matched members were called by the unchanged tool thresholds in 49 of 49 units, while the cassette RTs were called in 0 of 16 units and context-free random RTs in 2 of 10 genome units; the two-sided Fisher P is nominal because it assumes independent units — 1.54×10^{-15} at 49 units, 2.65×10^{-11} collapsed to 23 subclades, and 0.0417 collapsed to the single genus the cassette comes from, which is the 1 independent unit this branch has (Supplementary Table S2, table Q1). Matching was on identity alone, and this is a reference-pool rebuild emulation inside one genus — the reference pool stands in for the shipped training set, which the tools do not distribute — rather than a holdout run of the tools themselves, so this result does not speak to what the shipped tools would themselves detect. The direction is the same at the other thresholds we ran: the other identity cut-offs (13 of 13, 234 of 234 and 246 of 246 units), which rely on MMseqs2 clustering at sensitivity 7.5, a setting absent from the pre-specified rule file (Supplementary Table S2, table Q2).

The positive control pre-specified for the same test also failed: the *rpoB* and *gyrB* trees of the same genomes carry a supported split conflict (normalised Robinson-Foulds distance 0.556), and the nominal permutation P value of 0.0054 for that pair is uncorrected for multiple comparisons and measures agreement against a null and not the size of the conflict, so at the pre-specified threshold the test does not separate shared descent from tree-estimation noise. The whole comparison lies within one genus: the same 22 loci, collapsed to 18 ANI95 units, are re-used across all five trees, so the 4 families are not independent samples and the positive control lies in the same genus as the test; a result with no conflict would likewise not be read as support for co-inheritance.

S1.11 Hypotheses the data leave open, and the observations that would decide them

The work ends in six open questions, each with the hypotheses we hold and the observation that would settle it.

In structure-based comparison, the post hoc K-family models lie closer to UG17 than to UG3 within the UG3+UG17 supergroup, a direction and not a membership, and the pre-set test of DRT10 proximity against a random bacterial-RT null was not supported. Two readings remain open: the RT is a divergent member of a UG3/UG17-adjacent lineage that denser reference sampling would resolve, or it is a separate lineage that sequence trees cannot separate from its neighbours.

Post hoc, within species that carry the cassette-lineage RT, complete genomes lacking the RT retain RecBCD, so a lineage-background reading of the RecBCD pairing is favoured over an island-specific one, although the sister-pair test has too little power to settle it. In an analysis of K-family RT island loci, a member of the end-processing or recombination gene category lay within the neighbour window at only about one in five evaluable loci, and the partner types seen, tabulated post hoc, were mixed (RecD/RecBCD-type, SbcCD-type, UPF1/AAA15-type), so these data do not point to a single recurring partner; the result rests on annotation, and many neighbouring ORFs carry none. The same category was found more often beside the RT than in length-matched windows elsewhere in the same genomes, but only at the immediate neighbours, with a borderline contrast against non-K RTs, and the category was partly derived from a neighbour-nuclease enrichment search, so the two readings are not independent. In *Sediminibacterium*, a genus separate from the one carrying the cassette RT, every genome with a K-family RT has *recD*, *recC* and *recB* inside the RT unit, but in this thin, post hoc comparison few non-carrier empty sites could be judged, a second locus had no empty-site non-carrier, and most anchor pairs reflect rearrangement rather than insertion, so it leaves open whether the RecBCD block was inserted. In a post hoc analysis, a species-tree model in which the gain rate of K-family RTs differs by RecBCD status fitted much better than a single-rate model and held under a parametric bootstrap, genus stratification, a slowly rotating K null and a coding variant; this is co-occurrence only, every K-bearing species in the model's set carries RecBCD (so the effect size sits at a parameter boundary), site availability beside *recD* is not separated from it, and under the broader K main-clade definition about half of the high-quality carrier units lack RecBCD. In this post hoc classification, the +4 protein beside the island was called a stand-alone RecD2-type helicase in the cases assessed, with no RecC marker immediately downstream and a length in the RecD range; this is a post hoc reading and does not exclude a RecBCD reading. In *Pollutibacter*, every evaluable island RecD keeps its Walker A/B and ATP-contact motifs, but the island RecD is the only RecD-like protein

found in the assembly of the 8 *Pollutibacter* genomes with a complete +4 ORF, so preserved motifs show a normal RecD and not an effector: they cannot separate an island-specific effector or decoy from the host's own RecD, and a decoy reading cannot be tested with RecD data.

A second rooting attempt and trees built from predicted structures leave the direction unresolved. Only one genome offers an empty insertion site by gene order, it is not a close relative of any carrier, and it lies inside the radiation of the carrying lineage, so loss after insertion cannot be told from repeated insertion, and rooting the -1 and $+3$ gene trees with outgroups was untestable under saturation rules fixed beforehand. On the single *Ferruginibacter* island (one clone, one visible contig-end boundary), mobile-gene HMMs, CONJscan, ISfinder and ARAGORN detected no mobility marker under rules fixed before the analysis, and the scanner did detect its one *Bacteroides* ICE positive control; this is a negative for those four classes only, with no search for target-site duplications, att sites or new integrases.

Structure prediction of the free-standing $+3$ protein gives an archaemetzincin-like N-terminal domain with a His+Cys zinc-site pattern and, in a post hoc reading, a C-terminal domain that resembles OmpR/PhoB-type winged-helix domains; both readings are prediction-only, weaker for the cassette's own $+3$, and not evidence of DNA binding, but they make the palindrome-binding question testable, and the pre-specified structure and sequence tests had not supported the winged-helix resemblance, with the validity check of the structure test failing. The palindrome found upstream of the free $-1/+3$ units is absent from the cassette's upstream element itself (exact core, dyad and extended form), so the element is not a copy of that motif.

The six questions meet at three joints: where the parts come from, what the RT is, and how the cassette might be read. The informative next observations are a further genus that carries the combination, outgroup-rooted trees for which polarity can be tested, a 5'-end map of the RT transcript and an activity assay of the $+3$ N-terminal domain. A wet-laboratory programme to run these tests is under way; until it reports, the paper offers a defined association, a mapped list of alternatives and the observations that would choose between them.

S2 Supplementary discussion and limitations

S2.1 What we found, and what it might mean

The downstream arrangement of the RT with its $+1$ and $+2$ neighbours (same-strand proximity of that pair only; the upstream element was not searched) recurs in some close-relative genomes of that genus, not all (Figure 1; Figure 5). This is a non-independent re-observation made with detectors built from the cassette assemblies on genomes that already contributed to our RT reference database, and it claims neither specificity nor evolutionary conservation of the arrangement. Within the assembled contigs of one genus, no observable copy of any cassette module shows a frameshift, premature stop or truncating deletion, far fewer than neutral drift predicts, so the strict neutral-resident model is rejected under a pre-specified rule; separately, the codon-model ω of every module lies far from the neutral value in the purifying direction (Figure 6). Both rest on one genus, the only independent unit; the neutral expectation is inflated by variance saturation, there is no host pseudogene baseline, the constraint is moderate, the ω null is nearly self-evident, the direction depends on the BUSTED-S choice, and function is not decided. Public long reads that span two or

more cassette-specific modules number 34; even if every cassette read in the pool were treated as a chimera, random pairing of two molecules would be expected to give 1.61×10^{-4} such reads (Figure 7), and 24 of the 34 lie in strata added after the rule file was sealed. The comparison is near-trivial because the expectation is near zero by construction, it holds under random pairing only (non-random pairing such as template switching in a PCR-selected library was not tested), 31 of the 34 multi-module reads come from one PCR-selected study, the effective independent units are 4 water systems while the cassette's independent unit is one genus, and nothing here decides whether an individual read is authentic or what the arrangement does.

Four preprints from three author teams (Sánchez-Nieto et al.; Tang et al. and Ramirez et al.; Chen et al.) describe DRT10 (UG17) as a tripartite system of a structured ncRNA, a SLATT (SMODS and SLOG-associating two-transmembrane) effector protein and an RT. Three of them report that the DRT10 RT synthesises tandem-repeat DNA from an ncRNA template, primed from a C-terminal serine or tyrosine of the RT itself, depending on the homolog; these are unreviewed results from *E. coli* model systems that differ between homologs [9–11]. The fourth divides UG17 into three subtypes (UG17-A, UG17-B and UG17-C) by RT phylogeny, gene architecture, taxonomic distribution and ncRNA features, associates UG17-B with both Pseudomonadota and Bacteroidota, and finds that in a tree restricted to the conserved RT catalytic core the three subtypes are not recovered as strictly monophyletic clades [12]; we do not assign the cassette RT to any subtype. In the ESOL cassette (as seen in *Pollutibacter*) the RT is accompanied by an upstream non-coding element of unknown identity, by a –1 gene family, defined from cassette sequences, whose evaluable members all carry a MIDAS motif in predicted structures, and by a +3 gene family with a predicted metalloprotease-like fold (a structure-based label of intermediate confidence); we compare layout only and make no claim that the cassette acts as, or belongs to, a DRT10 system. In every +3 ORF examined the H[EDQ]xxH motif is absent, a sequence-motif check that is not specific to the cassette (nearly all free-standing +3-like sequences outside it also lack the motif; 2 of 127 carry it; Supplementary Table S5, table T10, rows H-8 and H-9), is weakly informative unless the metzincin-like label holds, covers closely related ORFs from one genus rather than independent cases, and does not show that activity is lost; no experiment tests it. Wang et al. (Cell 2026, author manuscript) [13] report that DRT3 and DRT9 are synthetically lethal with loss of RecBCD or ExoI, respectively. The cassette sits beside a RecD-like gene (+4) that, in the 8 *Pollutibacter* genomes with a complete +4 ORF, is the only RecD-like protein found in the assembly and so cannot be told apart from the host's own RecD; by analogy to these two systems, dependence on a host nuclease is a question worth testing (Question 2), but the analogy is not support, no experiment has tested it, and our genome-association tests of general RecBCD- and SbcB-dependence rules were negative or uninformative (Limitations and the Supplementary Information). The candidate msr/msd-like ncRNA motifs that Toro (mSystems 2026) [14] predicts computationally upstream of the RT in subsets of loci of two candidate type XI-like lineages are literature context for an RT beside an upstream non-coding element, not evidence about our element: we assert no identity between that element and ours (Question 6). In the range we searched we found precedents for the individual components of this composition (a RecD-like helicase, an upstream non-coding element beside an RT, a MIDAS-type protein, and a vWA-type partner family at retron-clade RT loci that is not RT-adjacent [15]) but no description of the composition itself. Searching PDB and the AlphaFold Swiss-Prot set with a local structure search, the top hits of RT structures (pLDDT 75 to 83) are reverse-transcriptase structures (including PDB 9c0j and 6ar1 and a retron RT; lowest E

per query 7×10^{-10} to 7×10^{-8}); the +1-type protein has only 2 structures with pLDDT of 70 or more, and its neighbours are weak hits to type I restriction-enzyme specificity subunits ($E 4 \times 10^{-6}$ to 1.3×10^{-4} ; sequence identity 0.11 to 0.16, a partial fold resemblance), and the +2-type structures (pLDDT 33 to 58) were not searched. For the +1 protein, several weak signals point towards a distant resemblance to HsdS, the specificity subunit of type I restriction-modification enzymes, but none reaches its pre-set threshold, the signals are not independent of one another, they rest on near-clonal members of one genus, direct alignment and Pfam-specific profiles do not show homology, and a pre-specified structure-mapping test of the idea was not supported, so the +1 protein remains an unassigned candidate lead.

By a post hoc nearest-tip rule, K main-clade RTs sit in 17 genera, 11 orders and 4 phyla (12 genera rest on one genome; at identity 0.50, 13 genera in 3 phyla), so the RT's wider family is not confined to the cassette's genus.

By gene order, in one *Pollutibacter* reference genome without the island that is not a close relative of any carrier, the flanking sequence runs on uninterrupted where carrier genomes have that sequence plus the island; with no polarity-resolved tree, insertion and loss both remain open. A post hoc comparison, added after the primary results had been seen, suggests that, in related *Pollutibacter* genomes, the pairs and the cassettes sit at different sites (flanking sequence next to RT-free pairs matched flanking sequence next to cassettes in 0 of 27 comparisons, whereas the same method matched cassette flanks in 22 of 22 cassette-to-cassette comparisons), but this adds no direction for the RT, and reconstruction of the state of the *Pollutibacter* ancestor depends on the model. One partner looks intact: in every evaluable copy the island RecD keeps its Walker A/B and ATP-contact residues at the level of the host's own RecD, although in the 8 *Pollutibacter* genomes with a complete +4 ORF it is the only RecD-like protein found in the assembly, so intactness cannot separate an island-specific effector from the host's RecD.

In post hoc recounts of the same 54 RT-free pair genera, which are not an independent line of evidence, dropping any single order or family still leaves at least 40 and 44 of them, against a positional null that may be optimistic, with the order minimum reached only when Chitinophagales is dropped; and a palindrome found in training genera meets its pre-specified rule as written in held-out genera, of which most from orders absent in training are Chitinophagales (17 of 26) and which carry no test of whether the +3 protein binds the palindrome. Genera in both checks are not independent units and neither check adds a unit beyond the single genus; both bear on the RT-free pair recurrence and on a motif in free $-1/+3$ units, not on the close-relative recurrence above, which stays non-independent, and the $-1/+3$ families and the training sample were defined from cassette sequences, so some circularity remains.

The work also bears on how to search for RT lineages outside the known types. In our own hunt, scored with DefensePredictor [8], 6 of 15 unassigned candidate clusters reduced to retrons, CRISPR-RT-Cas1 or UG19 and the other 9 lay beside retron lineages (only 8 members per cluster checked; no retron tree or msr/msd analysis). Because 6 of the 15 clusters reduced to known types, absence of a name-level marker did not show that a cluster was outside the known types, and such absence is not evidence of novelty.

Either answer matters: if the company works with the RT, the RT is one part of a multi-gene unit that profile-based detection could miss (in an emulation at matched amino-acid identity, known relatives were called in 49 of 49 units and the ESOL RTs in 0 of 16, a contrast that is not evidence about the shipped tools because the match was not on alignment coverage), and if the RT joined an older module, the cassette shows how a reverse transcriptase can be recruited into a gene set that exists without it. How widely the cassette is distributed is a separate question from what it does: a system found in one genus can still be characterised and tested in another host, and for ESOL no activity has been tested yet. Within the same genus, carriers of the K main clade have RecBCD more often than non-carriers (post hoc; extremely wide confidence interval; the carrying units had been seen before), and in Chitinophagaceae all high-quality carrying species units are RecBCD-positive, so the association is local and not a rule.

S2.2 Limitations

The part detectors were built from the cassette-bearing MAGs, so a search that re-observes those MAGs is circular for detection claims; hits in genomes that were not used to build them are new observations, but sensitivity cannot be judged with the same models. The cassette-bearing MAGs were selected because they carry the locus, so they support no estimate of prevalence. The 16 cassette-bearing MAGs and their ANI-level clusters are grains inside one genus and add no independent unit; the cluster count shifts with the clustering tool and rule, and counts at different grains are not added, so a test run at the MAG or cluster level is not a test across independent units.

Nearly all of the 445 K-family RT loci outside the genus have an RT much less identical to the cassette RT. In real MAGs, core-gene loss at the same CheckM2 completeness was at least twice that expected under a random contig-removal model (post hoc), and random breaks at arbitrary positions did not reproduce it, so the false-negative rate of a single genome's absence call lies between the random-model floor and this real-MAG anchor, which we read as an upper bound because core genes are biased markers in MAGs. For most candidate public long-read runs, a zero cassette result means that the cassette may not have been visible at that depth and not that it is absent; the run-level detection limits come from a calculator checked only for self-consistency with its own model, and two later revisions of it were made after the results were seen, so the stated limits should be read as optimistic. The emulation of existing tools has a similar limit: although UG17 is the nearest reference lineage of 11 of the 12 full-length cassette RTs, the identity-matched known set holds only 2 UG17 units among its 49 at the primary threshold, so a UG17-only reading has no power there and the contrast rests mostly on UG3 and UG8 units.

Our tests of a general SbcB-dependence rule did not support one either: UG17-carrying genomes do not carry SbcB-class nucleases less often than non-carriers of the same genus, and carriers in the K main clade carry an alternative nuclease less often than non-carriers of the same genus. The tests that did not support an idea, and the post hoc readings that sit beside the pre-specified results, are tabulated side by side in Supplementary Table S12 and Supplementary Table S8. Each limit points to a test: detectors that do not derive from the cassette (profile-profile and structure-based) for reach, further genera for independence, and activity assays for function.

S3 Supplementary methods

This section gives the core protocol of the study: the definitions of the objects and classes that the main text uses, the data sources and denominators, the design of each analysis with its tools, versions and key parameters, the counting conventions, and the software. This section describes the data sources and denominators, the design of each analysis with its public tools and versions, the counting conventions and the software; the detector models, thresholds, seed lists and build commands are given in Supplementary Table S13 and Supplementary Data S3.

S3.1 Census data, RT detection and clustering

The first five blocks of this section describe the census stage of the study, which led to the ESOL RT group, and the blocks that follow describe the later analyses of the cassette and of the locus atlas of K-family RTs, which use their own definitions. 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. The census stage called members of the ESOL RT group with the ESOL RT profile. The later stage calls K-family RT candidates and detects the cassette modules. In both stages the independent unit of the cassette observations is the genus, and the census group and the K-family loci are related only by identifier correspondence, for which no sequence was aligned.

Both the census and the later hunt used species-representative genomes of GTDB R232, in nested sets. The census covered the 189,630 bacterial representatives that have an NCBI assembly (archaea were not included, and 171 further bacterial representatives have no retrievable assembly). The later hunt read the GTDB tarball of predicted proteins for all species representatives, 199,923 genomes (189,801 bacterial and 10,122 archaeal), which contains every census genome. The release label is the same for both runs: the taxonomy files they used are byte-identical. The GEM catalogue was used as the published protein tarball with Prodigal coordinates. Logan unitigs were searched and are described as preliminary.

RTs were called with the Pfam RVT_1 model at the gathering threshold (HMMER [16], Pfam [5]) with a protein length filter. A trailing stop symbol was removed before downstream tools. Genes were predicted with Prodigal [17]. Domain assignment used the Pfam-A gathering cutoff. In the census stage, DefenseFinder [2] and PADLOC [3] system names were used at name level only; the later comparison with defense-system classifiers runs both tools.

Clusters were formed at 50% identity with 80% coverage (DIAMOND-based clustering for the GTDB hunt and MMseqs2 [18] for GEM). Library comparisons used DIAMOND [19] in more-sensitive mode. The published-library reference combined the RT sets of retrons and UG/Abi systems [20,21] and the bacterial RT classification of Toro and Nisa-Martínez [22]. A separate library held the Bacteroidota retron-RT candidates of the metagenome survey [14]; this set contains sequences reconstructed from contigs, in which a few RTs may have been assigned to the wrong paralogue.

S3.2 Known-type gates, statistics and phylogeny

Every candidate cluster was checked against the following gates and the result was recorded as a column in the candidate table. G1 (DGR) looked for the bAvd Pfam model [5] and for template/variable repeat pairs with adenine-biased mismatches [23]. G2 (group II intron) looked for maturase and HNH domains, the YADD motif and a gene-splitting arrangement. G3 (published library) took the nearest identity to the library. G4 (Toro retrons) repeated this against the 306 Bacteroidota retron-RT candidates and is limited to that phylum. G5 (CRISPR–Cas1) looked for a Cas1 domain fused to the RT or among its neighbours. G6 (defence-system names) matched the nearest library label against the names in DefenseFinder and PADLOC and works at the level of names only. The untargeted Pfam hunt used an additional gate for retro-element domains (G7). G7 is a census-level filter for retro-element domains that was applied before the candidate groups were formed.

The independent unit was the genus (or the sample for metagenomes), because related genomes are not independent. Enrichment used one-sided hypergeometric or Fisher tests with Benjamini–Hochberg correction. Permutation tests resampled units within the phylum. The smallest p-value a permutation test can return (its floor) is m divided by (N plus one), where N is the number of permutations and m the number of tested hypotheses. If that floor reaches the significance level, the test cannot detect an effect, and we label the result "not measurable" and not "not significant". In the Pfam hunt we used $N = 20,000$ permutations (floor 5×10^{-5}), so the floor was below the level; for other permutation tests we did not record N in the sources used for this paper. A non-significant result for a test whose floor is at or above the level was never treated as a negative.

Sequences were aligned with MAFFT [24] and trees inferred with IQ-TREE [25] under the model that ModelFinder chose for each alignment (for example LG+G4, LG+R4 or Q.pfam+F+R4), with branch support by ultrafast bootstrap or SH-aLRT as stated for each tree. Alternative placements were tested with the AU test. Profile–profile comparison used HMM-to-HMM scoring [26] on non-redundant reference groups.

S3.3 Upstream RNA element, defence and mobile-element tools

Consensus structures used RNAalifold from ViennaRNA [27]. 15 Rfam models [28] were searched.

DefensePredictor [8] was run on RT-bearing contigs at a fixed score threshold. geNomad [29] was run with default thresholds. For the untargeted hunt, DefensePredictor was run on windows that lack nucleotide sequence, which lowers scores slightly, so we read the result at a fixed score threshold as primary.

S3.4 Metagenome, Logan and additional public-source searches

GEM RTs were extracted with the same RT call, clustered at 50% identity with 80% coverage, and compared with the GTDB and library RTs with MMseqs2 at a query coverage of at least 0.5. Gates G1–G6 were applied to groups. For the Logan pool hunt, RT hits were pooled, collapsed by exact sequence and filtered as described in the Results. Filtered sequences were searched against the combined reference (GTDB RTs, library RTs, ESOL RTs) with DIAMOND [19] (version 2.2.8; one hit per query; sensitive mode in sharded searches, whose effect is described in the Results) and clustered with MMseqs2 [18] (version 18-8cc5c; 50% and 80% identity, 80% coverage of both sequences,

bidirectional coverage). Run metadata came from the ENA portal API. Contigs of cluster members were streamed from the public Logan bucket and cut at the target header; genes were called with pyrodigal [17] in metagenome mode (version 3.7.1), the RT located by exact protein match, and neighbouring proteins searched against Pfam-A at the gathering cutoff with HMMER (version 3.4). Gates G1, G2, G3 and G5 were applied as defined above. The software versions named for the Logan pool hunt are the versions installed when the numbers were recomputed, and the run logs do not record them.

Changes made after results were seen are marked as such.

S3.5 Sensitivity analysis and software

The sensitivity analysis was pre-specified in writing before its results were computed, with the SHA-256 digest of the rule file recorded; changes made after results were seen are listed at the end of the rule file and are reported in the Results. The sensitivity analysis tests the pipeline on published lineages that were removed from the reference libraries. For the DefensePredictor and GEM axes the same removal is applied to the known-type tags and to the metagenome reference, respectively; lineages with no cluster are reported as untestable and are not counted as negative.

Versions recorded in run logs or reports for the analyses described above: geNomad 1.12.0 (database 1.9), Infernal 1.1.5 and ViennaRNA 2.7.2. Software cited in this section: skani [30], codeml of PAML 4 [31], HyPhy [32] with FEL [33] and GARD [34], minimap2 [35], BLAST+ [36] and GTDB-Tk [37]. The versions of DIAMOND, MMseqs2, MAFFT, IQ-TREE and DefensePredictor were not recorded for some of the analyses described above and are then reported as not recorded; the versions that the logs do record, among them HMMER 3.4 and pyrodigal 3.7.1, are given in Section S3.16, in Supplementary Table S13 and in Supplementary Data S3. For the Logan pool hunt we give the installed versions at the time the numbers were recomputed (DIAMOND 2.2.8, MMseqs2 18-8cc5c, HMMER 3.4, pyrodigal 3.7.1); the run logs do not record them.

S3.6 Data sources, denominators and sample provenance

We scanned public genome catalogs for K-family reverse transcriptase (RT) loci, and we report coverage as genomes processed and never as completeness. We call a locus cassette-positive when a -1 gene lies on the strand of a K-family RT, which is a post hoc operational class and not a clade, and which does not require the upstream element or the +1, +2 and +3 genes. The Zenodo global surface-water catalog (record 18977176) holds 20,017 species-level representative MAGs from 4,128 rivers and lakes, and its record gives the taxonomy of each MAG. The LIMNOS lake catalog (Zenodo record 10533926) holds 76,918 MAGs, and we selected its Bacteroidota MAGs by the taxonomy in LIMNOS_mag_metadata.csv. The GEM catalog holds 52,515 genomes, and we scanned the Bacteroidota genomes among the 23,626 that the first protein scan had not covered, with the taxonomy of genome_metadata.tsv. The GTDB K-positive panel is a list of GTDB genomes chosen because they carry a K-family RT, and we downloaded the genome files of its entries from NCBI.

We processed 3,791 Zenodo Bacteroidota representatives and 4,687 Zenodo non-Bacteroidota representatives, and the second pass is partial because we stopped it when the scan ran out of time. We processed 9,998 LIMNOS Bacteroidota MAGs, 4,049 GEM remainder Bacteroidota genomes and 237 genomes of the GTDB panel. The five sets add up to 22,762 genomes processed. The second

Zenodo pass listed 16,040 non-Bacteroidota bacterial representatives and 186 archaeal representatives, and it processed 4,687 of the 16,040 and none of the archaeal representatives. The list of LIMNOS Bacteroidota MAGs had 10,037 lines that name 9,998 distinct MAGs, and we processed all 9,998. We looked up the 262 entries of the GTDB panel in the NCBI GenBank and RefSeq assembly summary tables, directly or through the paired accession, and we processed the 237 entries that have a genome path in these tables and not the 25 that have none. We took each processed count from the table of genomes actually processed per catalog piece and not from the candidate list.

We scanned the GEM remainder from protein FASTA files only, so these genomes have no nucleotide context and no element or array call. For this scan we read the per-genome protein files of the GEM protein archive with the gene calls that they hold, and we applied the same K-family RT rule and the same windows to these proteins. For the other catalogs we called genes with pyrodigal from the nucleotide sequences, trained on each genome with translation table 11 and in the metagenomic mode for genomes below 100 kb in total, which provides the contig context around each locus. A locus reported as not detected in these genomes means only that this detector did not find it in the genomes we processed, and it is not evidence of absence.

A first protein-level scan covered SPIRE gene-call files, GEM proteins and two MGnify marine protein catalogs, and its extension applied the same scanner and rules to further SPIRE studies and to the rest of GEM. Loci from these protein scans, with their own denominators and the duplicate map between catalogs, are listed in Supplementary Table S1 and are not added to the genome-level counts. We did not scan the MGnify soil and human-gut protein catalogs, OceanDNA, ENA assemblies or the LakePulse MAG collection, and the genome-level scan left out the non-Bacteroidota remainder of LIMNOS and GEM. Together the two scans covered 12,047 SPIRE samples, which is 0.2122 of the SPIRE gene-call archive by size with part of that size estimated, and all 52,515 GEM genomes, 28,889 in the first scan and 23,626 in the extension. A run of the scanner on the 16 seed MAGs recovered a K-family RT in 16 and a cassette-positive locus in 14 of them. This control is a plumbing check, because we built the detector from the same seed MAGs, and it does not measure sensitivity for diverged cassettes of other genera. The recomputations that checked the extended scan started from the stored tables and not from the original catalogs, so they are not independent of the pipeline. The sensitivity of the K-positive rule is circular, and its recalls are upper bounds.

We classified the 16 seed MAGs and one control MAG with GTDB-Tk 2.7.2 (command “classify_wf”) and the GTDB R232 reference package, at the default minimum alignment fraction of 0.5. We computed ANI with skani 0.3.2 and joined two genomes into one ANI95 unit when their ANI was at least 95 per cent and the larger of their two alignment fractions was at least 50 per cent, by single linkage. We ran skani in its default learned-ANI mode and repeated the clustering with --no-learned-ani as a sensitivity setting. We ran CheckM2, in the version that its run log records, with the general model in low-memory mode on the 16 seed MAGs and one control MAG. We took the quality values of public genomes from the GTDB metadata, and we call a genome Q90 when completeness is at least 90 per cent and contamination is below 5 per cent. The GEM MAGs and four GTDB-panel genomes with a cassette-positive locus and a blank taxonomy received a genus from rpoB and gyrB amino-acid identity to the seed MAGs and to GTDB genomes, a post hoc step because no GTDB-Tk run was made for them.

We count a set of genomes and their source samples at up to seven rungs, the genus, the ANI95 unit, the BioSample, the BioProject, the coordinate cluster, the country and the continent, and we never add counts of different rungs. For the cassette the independent unit is the genus, and the counts at the other rungs describe source records of the same genomes. Genus counts use labelled genera only. Of the 16 seed MAGs, 8 meet Q90, and all 16 are separate ANI95 units. The number of ANI95 units depends on the skani mode and the alignment-fraction rule, and the 16 seed MAGs are samples selected because they carry the locus, so they must not be used to estimate carriage.

The remaining long-read RNA runs, which are the runs outside the freshwater-keyword runs of the first two long-read screens, form a universe of 4,263 runs from 3,827 BioSamples and 151 BioProjects, with 2,151.8 Gbp (Supplementary Table S14, table B3). We excluded runs of library strategy AMPLICON, runs with zero bases or zero reads, and runs that were not available from AWS Open Data, and we do not use these exclusions as evidence of absence. The design exclusions removed 1,015 amplicon runs and 57 runs with zero bases. We assigned each remaining run to one of ten habitat classes with keyword rules applied in a fixed priority order to the run table and to the sample records, with no manual edits. A run took the class of the first rule that matched, tested in the order host-associated by host fields or scientific name, wastewater, estuary, sediment, wetland, biofilm, freshwater, soil, marine, host-associated by words in the free text, and other. The freshwater rule applied only when the environment fields and the scientific name held no marine word. The classes freshwater, sediment, wetland, soil, biofilm, estuary and wastewater form the near set (NEAR), and the classes marine, host-associated and other form the far set (FAR).

We processed every eligible NEAR run, taking the first 15 Gbp of each run at the nucleotide stage or the whole run when it was smaller. For FAR we drew runs at random with a fixed seed from strata of habitat and BioProject, taking runs of at most 3 Gbp, or the smallest run of a stratum that held none, and we processed the first 200,000 spots of each run. We raised the cap of FAR runs per stratum, with the same seed and the same rule, so that each sample contains the smaller one. We screened 257 of the 258 NEAR runs, which cover 0.842 of the NEAR bases at the nucleotide stage, and 739 of the 1,096 FAR runs sampled, 156.4 Gbp in all. We did not validate the accuracy of the habitat classes independently, and the FAR runs were screened at shallow depth, so the far habitats have no power to show absence for a whole habitat.

S3.7 Locus definition, cassette-type call and atlas states

The K family is a post hoc definition. The K-family RT set is not monophyletic in the RT-domain maximum-likelihood tree. The number of independent units is far smaller than the number of tree tips, and the deep placement of the K family cannot be relied on.

Slots are counted in the orientation of the RT, with negative slots upstream and positive slots downstream, so that slot k is ORF $i+k$ for a plus-strand RT at ORF i and ORF $i-k$ for a minus-strand RT. For genome and MAG files we called genes with pyrodigal, trained on the genome itself. We numbered the ORFs of each contig from 1 in the order of the gene calls. The genome-level scan applied the same frozen detection rules as the first SPIRE/GEM-derived scan to nucleotide genome and MAG files. We scanned the GEM remainder as protein sets, so its rows have no upstream nucleotide window.

A cassette-positive call is a post hoc operational class and not a clade. We annotated the window ORFs with Pfam profiles at their gathering thresholds, NCBIfam and TIGRFAM profiles at their trusted cutoffs. Calls inside the seed MAGs are circular, and we reported them separately. Throughout the Methods the parts of the cassette are the upstream element and the genes at -1 , $+1$, $+2$ and $+3$, the RT is the anchor of each locus, and the RecD-like $+4$ gene is an annotation slot that we do not count as a part.

RecD-like and RecC-like genes are kept as annotation columns that we do not count among the parts of the cassette, because they cannot be told from the host copies. A cassette-type call was assessable, and window ORFs missing because of contig ends were recorded as unobserved and not as absent.

Each component of a locus is one cell with one of four states: present, absent but observable, not observable, or not applicable (NA). A cell is PRESENT when an ORF in a search slot of the part, on the strand of the RT, carries a label of the part. PRESENT counts whatever the observability of the cell, and absence is counted only in observable cells. A cell without a label is called absent but observable when the window is long enough to judge the part, and not observable when it does not. We report "not detected" only for cells that are absent but observable, and it means not found by this detector inside the observable window.

We searched this window with the nucleotide model of the element. The element cell is NA for protein-only windows, because no nucleotide window was scanned. The element boundaries that set these positions depend on the element model.

The element calls depend on the seed, because the element model was built from the single genus of the cassette, and they cover only the range that the scans reached.

The locus atlas (Supplementary Data S4) is an accounting ledger and not a detection: it holds 4,186 rows, one for each K-positive RT row of the three source tables (nucleotide-window catalogue 489, SPIRE/GEM-derived catalogue 1,399, extended SPIRE/GEM catalogue 2,298). Each K-positive row of the three source tables (489, 1,399 and 2,298 rows) is listed exactly once in the atlas, and the reference tables of the locus data deposit hold the same rows by source. Rows are not loci: the same locus can occur in more than one catalogue, so the row count is a table figure, and the duplicate-link columns, not this number, bear on how many distinct loci there are. Within the scanned range and for this detector the cassette-type arrangement is described from a single genus, which counts as one independent unit, and no row, genome or sample count in the atlas is a count of independent units.

S3.8 Detection models and re-calling of the component states

We use cassette-positive for a post hoc operational class and not for a clade: a locus in which a K-family RT is preceded on the same strand by a -1 gene. We read a hit of these profiles as similarity to the seed sequences and a missing hit as the reach of a narrow profile, not as absence.

We re-called the component states of cassette-type loci from genome coordinates with the detectors that had produced the recorded states, and we tuned none of them. Fifty LIMNOS genomes that were missing from the stored copies were extracted afterwards from the public catalogue archive, and the 59 loci touched by such post hoc input revisions are marked. We called ORFs in every genome with pyrodigal under translation table 11, in meta mode for genomes of less than 100 kb and with training on all contigs of the genome otherwise, and numbered the ORFs along each contig.

S3.9 Motifs, structure prediction and structure-based placement

A cassette-positive locus is a post hoc operational class and not a clade: a K-family RT locus whose –1 gene lies on the RT strand upstream of the RT. The motif scans in this section used the 22 loci of the intactness census (Section S3.11): 16 MAG loci and 6 GTDB Pollutibacter carrier genomes. Module labels –1, +1, +2, +3 and +4 give positions relative to the ESOL RT. An ORF cut by a contig end was excluded from motif calls and counted as not evaluable, never as absent. Genus was the independent unit, and because all 22 loci belong to the genus Pollutibacter, every cassette-module statistic rests on one genus unit; ANI95 clusters were reported in addition.

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. We superposed each ESMFold model of a –1 ORF on chain A of PDB 1T0P with TM-align and read the residue at each reference position from the structural alignment. The –1 models were four-recycle ESMFold models. A model was evaluable when its TM-score normalised by the reference was at least 0.5 and all five reference positions were aligned within 4 Å between Cα atoms after superposition. 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. –1 ORFs without a model were read at the alignment columns defined by the structure-mapped members. The MIDAS readout rested on 14 ESMFold models mapped to 1T0P by TM-align and on 4 ORFs read by alignment-column transfer. A sequence-only screen, the pattern D.[ST].[ST], was reported separately and could not override the structural call.

We predicted monomer structures with ESMFold v1. A panel of ten K-family RT cores, taken at a fixed stride from the 145 K-family sequences and cut to at most 380 residues, was submitted for ESMFold prediction, and nine models were returned (the request for the tenth failed). We took mean pLDDT as the mean Cα B-factor and treated a model as confident at a mean pLDDT of at least 70; models below 70 were listed and read as no information, never as evidence for or against. Every ESMFold model is a single-sequence prediction, so a confident model does not measure accuracy and a structural similarity is not evidence of function, DNA binding, interaction or homology. Boltz-2 served as a secondary predictor for pair complexes and OmegaFold for selected +3 C-domain models, and because their training data overlap with that of ESMFold we count their agreement as partly independent.

To place the ESOL RT by structure we built a panel of RT-domain windows, each the envelope of the best RVT_1 (PF00078) domain plus 40 residues on each side, from reference retron, UG/Abi, DRT, group II intron, DGR, CRISPR-RT and outgroup RTs, from SPIRE type XI-like, VI and XIII RTs, and from experimental RT structures. We parsed Supplementary Table S2 of Toro 2026 from its PDF text by code and checked the per-group counts against the totals of the published Table 1. The table lists identifiers and not sequences, so we restored each sequence by fetching the contig named in its identifier from NCBI nuccore, re-calling genes with pyrodigal in metagenomic mode and scanning with the RVT_1 profile at the gathering cutoff. ESMFold models were graded ok at a mean pLDDT of at least 70, low between 50 and 70 and fail below 50. The 13 ESOL RT windows were treated as one genus lineage and not as 13 independent samples.

Contig re-calling covered 204 identifiers, with no failed fetch. The median TM-score between ESMFold models of the 26 experimental RT windows and the experimental structures was 0.9212, and the pre-declared pass criterion was met. Unrelated non-RT models scored a mean TM-score of 0.2658 and a maximum of 0.3605 against the panel. The best TM-score of an ESOL RT window to any non-K tip had a median of 0.8235, so all RT windows share the RT fold and class discrimination rests on small distance differences. SPIRE sequences are restored by contig re-calling and not taken from the authors of the table.

To test whether the ESOL RT shares features specific to DRT10, we fixed rules before any alignment of the ESOL RT to a DRT10 structure was run or read. The DRT10 references were PDB 30QM (Eco1), 30QU and 30QX (Eco3, the latter with dNTPs), released by the PDB on 7 October 2026, and we counted two independent subtypes, Eco1 and Eco3. Regions were assigned by RVT_1 HMM coordinates (PF00078.33), an operational approximation of fingers, palm and thumb with the C-terminal extension. Similarity was TM_avg, the mean of the TM-scores normalised by the reference and by the query from order-dependent US-align, computed on region-cut structures and, separately, on regions within full-length alignments. An effect had to exceed a level fixed beforehand in length-matched fragments of unrelated non-RT models. We computed TM-scores with US-align version 20260920. The Foldseek PDB database snapshot did not contain the DRT10 entries, so DRT10 structures entered the comparisons as coordinate files downloaded from the RCSB. We used pydssp and we ran pyhmmer with its heuristic filters off to find the RVT_1 envelopes.

The K-family tree analysis used 435 RT-domain sequences, each the RVT_1 (PF00078) envelope plus 40 residues on each side, comprising 51 K-family tips and reference classes such as retron types, UG/Abi groups, group II introns, DGR, CRISPR-RT, LINE, LTR, HBV, TERT and GTDB background RTs. References were drawn by farthest-point sampling in 3-mer space, so each reference class is a sample of its diversity and not a census. Alignments were MAFFT L-INS-i, an hmalign alignment of the PF00078 match columns, and MAFFT E-INS-i, each with columns of gap fraction above 0.5 removed. ModelFinder, with the candidates LG, WAG, JTT, Q.pfam and rtREV and up to four rate categories, chose Q.pfam+F+R4 by BIC, but the default IQ-TREE search with 1000 ultrafast bootstrap replicates did not fit the time limit, so the main tree used the fixed model Q.pfam+F+R4 with a reduced search and 1000 SH-aLRT replicates and no ultrafast bootstrap. Robustness trees used LG+F+G4 on the hmalign, E-INS-i and L-INS-i alignments and on the L-INS-i alignment with the fastest 25% of sites removed, each with a reduced search. A six-group Dayhoff recoding of the L-INS-i alignment was analysed in IQ-TREE as morphological data under MK+G4, the fallback model named in the pre-specified rule file, with the default search.

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. Position calls for the K unit on sequence trees used a backbone alignment that several analyses share: the L-INS-i alignment of 493 sequences, from which columns with gap fraction above 0.5 were removed, leaving 258 columns. New sequences were added to this alignment with MAFFT --add --keeplength. The relation of the K unit to a target group was called NESTED, SISTER, NEAR or INDEPENDENT by fixed rules. Alternative positions were tested with the approximately unbiased test on regrafts of the K unit and of the K clade onto target stems and onto 30 random internal

branches. Because p-AU values below 0.001 differed between two runs on the same trees, the log-likelihood difference and its percentile among the random placements were the primary effect sizes and p-AU was shown only for orientation.

Each placement was read by edge category and likelihood weight ratio. The edge categories were inside the K unit, the stem edge of the K unit, the other edges inside the K clade, the stem edge of the K clade and all other edges, which we split further by the class composition of the clade below the edge.

S3.10 Defense-system classifier comparison and the gap test

In this section a cassette-positive locus is a post hoc operational class of loci in which a -1 gene lies upstream of a K+ RT, and it is not a clade. We asked whether two published classifiers of RT-containing systems assign any of the 16 ESOL RTs to a named class. We used DefenseFinder 3.0.0 [2] with MacSyFinder 2.1.4 and defense-finder-models 3.1.0, and PADLOC 2.0.0 [3] with padloc-db 2.0.0. DefenseFinder ran on the 16 MAG bins that carry the cassette in gembase mode, with each contig as a replicon and Prodigal gene calls, the anti-defense models added, the default gathering thresholds, a profile coverage of 0.4 and the default system-completeness rules. PADLOC was run on the 16 bins and on six near-relative genomes, and it was not run on the control genomes of the host-range analysis or on the positive-control genomes. An ESOL RT counted as called when either tool placed it in a Retron, DRT or Abi (AbiA, AbiK, AbiP2) system.

As a positive control, we ran the same DefenseFinder pipeline on seven public complete genomes that carry known retron RTs and counted the genomes in which the known RT was called as a member of a Retron system. The known retrons of the positive control are proteobacterial retrons of the kind on which the DefenseFinder models were built, and the public genomes carry RefSeq gene calls whereas the bins carry Prodigal gene calls. The shipped model libraries are weighted toward Proteobacteria and we did not measure the sensitivity of either classifier for Bacteroidota RTs, so a missing call is a model-limited tool output and not evidence for or against any function of the RT. As a baseline for how often the same DefenseFinder pipeline calls an RT that is not an ESOL RT, we counted how many of 39 RT loci in 10 random-RT genomes were called as RT-containing systems in their genome context. The baseline rests on DefenseFinder alone. We compared the proportion of RT loci called between the baseline and the 16 ESOL RTs by a two-sided Fisher exact test. The 39 baseline loci come from 10 genomes, so loci within a genome are not independent units and the P value of this comparison is nominal.

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. Before sealing we had seen the classifier counts of the run on the 16 ESOL RTs (this section), the nearest-reference identities of the ESOL RTs and the metadata headers of the shipped profiles, and the design was shaped by that prior knowledge. The reference pool holds the reference RT sequences of the nearest-class analysis of this section, taken from published UG/Abi and retron RT sets [20,21] and other RTs, with the K lineage and the GTDB background RTs excluded. The known members are all pool sequences of the classes UG17, UG3 and UG8. The Bacteroidota-associated subtype of UG17 could not be matched, because its marker list is not in the permanent copy of the inputs, so the known members are not restricted by

subtype. The ESOL RTs are the 16 RT ORFs of the cassette-carrying MAG bins, of which 12 are full-length and 4 end at a contig end. The control sets are the RT ORFs of 39 loci in 10 random-RT genomes and seven known retron RTs of the classifier analysis of this section as a positive control. The ESOL RTs all come from one genus, *Pollutibacter*, so a test at genus level has one unit and no power.

The test therefore has two layers: layer 1 ran the tools unchanged, with the known members presumed to lie inside the training sets, and layer 2 rebuilt the profiles from the reference pool with one subclade withheld. The shipped profiles for UG17, UG3 and UG8 record training-sequence counts of 335, 482 and 510 in the DefenseFinder models (UG17, drt3a and drt3b) and of 335, 482 and 512 in padloc-db (drt-UG17, drt-UG3 and drt-UG8), against reference-pool group sizes of 336, 484 and 512. The counts differ by 0 to 2 sequences, so the profiles appear to have been built from nearly all of the published group members, but the shipped models carry only the counts and not the sequence lists, so the reference pool serves as a stand-in for the training set and every layer-2 number rests on that stand-in. A layer-1 call rate of the known members is read only as an upper bound of their holdout call rate, on the assumption that removing training members cannot raise a call, and layer 1 alone does not declare support.

Layer 1 ran both tools unchanged on single-protein inputs, under the same condition for the known members, the ESOL RTs, the random RTs and the positive controls. DefenseFinder 3.0.0 ran in gembase mode with the anti-defense models added and the default thresholds and completeness rules, and PADLOC 2.0.0 ran on the protein and GFF files of each chunk with --fix-prodigal and its default thresholds. A system call means that the protein is a member of a DefenseFinder system of type Retron, DRT, DRT-like UG, AbiA, AbiK or AbiP2, or of a PADLOC system named DRT, retron, AbiP, AbiK or AbiA. A component call means a hit of an RT-type profile in the DefenseFinder hmmer table at the model threshold, and because DRT_3 is a two-gene system a lone UG3 or UG8 protein can be called only at this component level, so the primary layer-1 call is a system call or a component call by either tool.

Layer 2 rebuilt the profiles of UG17, UG3 and UG8 from the reference pool with one subclade withheld and applied the shipped tool thresholds to the rebuilt profiles; at the lowest of the four thresholds of S1.10 the per-group rebuild could not be run for UG17, so UG17 units are absent from the layer-2 emulation there. For each group the pool members were aligned with MAFFT [24] and the full-group profile was rebuilt from this alignment. Profiles were built and searched with HMMER 3.4. A holdout profile was built from the same alignment with one subclade withheld. All profiles, rebuilt and shipped, were searched against the combined target file with hmmsearch.

The verdict of a closed branch is the literal outcome of its rule and is not a statement of support or refutation.

S3.11 Intactness census and selection analyses

In this section a cassette-positive locus is a post hoc operational class of loci in which a -1 gene lies upstream of a K^+ RT on the same strand, and it is not a clade. 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). The cassette has five counted families, the RT and the four neighbouring families -1, +1, +2 and +3, and we tally the RecD-like +4 gene separately because it cannot be told from the host RecD. We defined the loci through the RT, and for the MAG loci through the adjacent element, and not through the neighbouring families, so a disrupted neighbour would be scored as disrupted, whereas the census of the RT is conditional on the recognition of the RT.

An INTACT call is a classification by coverage and not a proof that the ORF carries no lesion.

Of the 132 slots of the 22 loci (six slots per locus), 82 are INTACT, 13 are cut by a contig end and 37 are not observable, so none of the 132 slots is scored disrupted or absent. None of the 24 not-observable slots of the five families is a hidden-deletion candidate, and it is not evidence about the randomness of the not-observable slots. The census concerns assembled contigs of recognised loci in one genus and has neither read support nor a host pseudogene baseline. We did not examine reads for assembly errors, because no reads were downloaded for this analysis.

The census classified all 12 edited contigs as disrupted. In this first control no lesion was placed in +1, +2 or +4 ORFs and none near an ORF end, so it does not show that the census recognises lesions in these families or lesions that leave one stop-free segment.

We bounded the disruption rate of a module by the upper limit of the two-sided exact (Clopper-Pearson) 95% interval, computed from the beta quantile, which equals $1 - 0.025^{(1/n)}$ when none of n observable copies is disrupted. Copies at one locus or in one genome are not independent observations, so the limit by copy is arithmetic, and the genus is the independent unit of the cassette observations. We also counted the observable copies by ANI95 unit, taking the cluster label of the genome from the ANI95 clustering table, and by allele cluster. An allele cluster is a single-linkage cluster of the full-length copies of a module at an amino-acid identity of at least 95% over at least 50 aligned positions in the family alignment. The pre-specification said that near-identical members count once, but the allele clusters were computed after the census had been run, so we treat them as a post hoc accounting. The unit ladder in Supplementary Data S1 counts the same observable neighbour copies at nine units (the copy, the locus, the genome, the ANI95 unit in two labellings, the allele cluster, the study or public genome, the country and the genus), which are different ways of counting the same copies and are not added across layers. The two ANI95 labellings, one from the co-occurrence census table and one from the module table of the census, give 17 and 18 units among the 22 loci and differ for one pair of genomes. Only the copy, the ANI95 unit of the census table and the genus are pre-specified units, and the other units are post hoc.

We therefore state no neutral expectation and no rejection of neutrality in this section.

We estimated one global ω per alignment with codeml of PAML 4.9 under the one-ratio model M0, with CodonFreq F3x4, a free κ , a free ω and cleandata 0, on the FastTree topology. We tested $\omega = 1$ by a likelihood-ratio test of the free fit against a fit with ω fixed at 1, with one degree of freedom for an unpartitioned alignment. For a partitioned analysis we ran M0 on each segment with the FastTree tree of that segment, summed the log-likelihoods, took the ω weighted by codon number and tested

$\omega = 1$ with the summed statistic on as many degrees of freedom as there are segments. We report ω as one value for all sites and all branches of an alignment, which is a mean and does not describe single sites, and the estimate carries no information on the heterogeneity of ω among sites. We checked saturation from the tree length and from the mean branch dS of codeml, and we marked an alignment as saturated when the mean dS per branch was too high to be informative. We repeated the unpartitioned fit on the set of all members, on the untrimmed alignment and on the columns without gaps, to see how far ω depends on the choice of members and columns.

S3.12 Long-read screen, chimera baseline and read-level controls

In this section a cassette-positive locus is a post hoc operational class and not a clade. 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. In the read screens we represented the cassette by 16 cassette-bearing contigs of assembled genomes, and a read counted as a cassette read only through a rule-based similarity to these references. 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.

We drew the candidate runs from the run table of the European Nucleotide Archive, queried through its portal interface for runs with library source METAGENOMIC on the Oxford Nanopore or PacBio platform. A run was a candidate when its library strategy was WGS and the archive held FASTQ files for it. We assigned a run to the freshwater stratum S1 when the text fields of its metadata matched one of the words freshwater, lake, pond, river, stream, creek, reservoir, groundwater, aquifer, wetland, marsh, limnic, oxbow, brook or paddy, and to the sediment stratum S2 when they matched sediment and none of these words. We excluded a run when its metadata matched an exclusion term, for example for marine, saline, sewage, sludge, bioreactor, drinking-water, hot-spring, glacier or human or animal gut samples, or when its host field was filled. We kept the class and the reason for the exclusion of every run of the archive table, and we used no source that required a login or an acceptance of terms. The selection left 1,064 candidate runs, 927 in the freshwater stratum and 137 in the sediment stratum.

We processed the runs of the freshwater stratum before those of the sediment stratum. For each run we processed at most the first 2.0 Gbp of its FASTQ file in file order. We fixed a clock time after which no new run was started, and we recorded the runs that had not been reached as not reached. We stopped before the fixed stop time and froze the sample at that point, which was after the first high-confidence hits had been seen, and the amendments record both facts. 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. The processed runs are not a clean prefix of the pre-specified order, because 251 of the first 263 runs of the freshwater stratum in that order were processed and the processed runs reach order index 314. We did not use the runs that were not reached, and the failures and the runs not examined, as evidence of absence. The deeper protein pass did not finish for some runs, and we did not repeat it for them.

We aligned the reads with minimap2 2.27 using the preset map-ont and the options -c and --secondary=no, and we used the preset map-pb only for the oldest PacBio models (RS and Sequel I). The identity of an alignment was the number of matching bases divided by the length of the alignment block, columns 10 and 11 of the PAF line.

The screen carried positive controls in the same runs: 16S rRNA reads, reads of the host genomes and single-copy genes. The screen detected 1,018,899 reads of small-subunit ribosomal RNA in 198 runs and 15,456 host-genome reads in 132 runs, and no read hit a reversed DECOY reference. We recovered the sequence of each hit read and examined it with three tools that differ from the screen: BLASTN against the 16 cassette contigs, minimap2 with the preset asm20 against the same contigs, and DIAMOND blastx against the cassette-specific module proteins. As a negative control we shuffled the nucleotides of every hit read and put the shuffled reads through the same three tools. Of the 49 hit reads of all layers, BLASTN confirmed 48, minimap2 with asm20 confirmed 47 and strict protein blastx confirmed 48; every read was confirmed by at least one tool, 46 by all three, and none of the 245 shuffled reads gave a hit with any tool. The reversed DECOY and the shuffled reads are trivial nulls that cannot produce homology hits, so a count of zero in them is not evidence of specificity against unrelated RT or vWA-family homologs. A read that covers a single module cannot be shown not to come from a native host homolog.

We counted independence at the levels of the read, the BioSample, the study and the water system, we counted one study as one water system, and we counted the reads of one sample as one unit. We marked whether the sample or the study of a read was also the source of one of the genomes of the reference, because the same sample is not independent evidence. After high-confidence hits appeared in two runs of one river study, we switched from the sediment stratum to a rescreen of these two runs at full depth, first with a cap of 40 Gbp and after a restart with a cap of 12 Gbp, and we kept this stratum apart from the statistics of the sample. After cassette-hit reads appeared in a lake study and in a second lake study, we screened 14 runs of these two studies, 12 of them not yet processed and 2 already in the frozen sample, with the same pipeline and the same caps, and we reported them apart. After the sample was frozen, three runs rich in hits were each screened to full depth with a cap of 30 Gbp, and we added a run that was in progress at the time of the freeze as a further post hoc layer. We call the layers chosen after hits had been seen post hoc layers, and we do not pool them with the frozen layer. The frozen layer holds 15 cassette-hit reads from 7 runs of 7 samples and 4 studies, and with the post hoc layers the 49 reads come from 9 runs of 9 samples and 4 studies. One study of one river supplies 43 of the 49 reads and used PCR-selected libraries, so the effective independent units are the four water systems and not the nine samples.

We reported the exact two-sided 95 per cent Clopper-Pearson interval of the share of synthetic reads with a block in both parts that the audit flagged as discordant. Of the 200 synthetic chimeras, 169 showed a block in both parts, 25 showed one part and 6 showed none, and of the 169 the audit flagged 138 as discordant, a share of 0.817 with an exact interval of 0.750 to 0.872. The remaining 31 of the 169 were not separated, because the two parts joined in the same order and orientation, so this audit cannot tell such a join from a single molecule. All 49 source reads are cassette reads, so the control covers joins of a cassette read to a cassette read only and says nothing about a join between a cassette read and a non-cassette molecule, which would present as a single block. The value 0.817 is therefore a bound on what the audit can separate and not a chimera rate of the real reads, which we

did not measure, and the 25 reads with one part and the 6 reads without a block lie outside its denominator. The reads and the reference lie inside the set of genomes of a single genus, which stays the independent unit, and the host genus of the reads was not determined.

To ask whether the reads that cover two or more cassette-specific modules exceed the number that a random joining of two molecules during library preparation would give, we designed an empirical baseline for chimeric nanopore reads and sealed its pre-specified rule file before the first download finished, the first alignment and the first aggregation. We took the 9 runs that held cassette-hit reads, from the Lake Malawi, Lavaca River, Umiam Lake and Pampa River studies, and for each of these studies runs that the screen had processed and that had no cassette-hit read. From each run we took a subsample from the start of the FASTQ file, and this subsample is the start of the sequencer output and not a random sample of the run. We aligned the subsample with minimap2 2.27, preset map-ont, to a host reference without the cassette contigs, so that cassette sequence could not mix into the host alignment. We called a read split when two or more pieces remained after merging, and we classified it by its strongest piece pair as same contig on opposite strands, same contig on one strand with inconsistent order or gap, a different contig of the same genome, or a different genome. The share of split reads among the reads above a minimum length that map to the host reference is an apparent split rate and not a chimera rate, because a chimera is seen only when both of its pieces align to the reference and fragmented assemblies add splits.

S3.13 Co-occurrence census, pair-anchor scan, part recurrence, palindromes and outgroup trees

In this section a cassette-positive locus is a post hoc operational class of loci in which a -1 gene lies upstream of a K^+ RT, and it is not a clade. 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. Counts under the frozen rules are reported as counts only, the crossings and category calls first defined in this analysis are post hoc aggregates, and values that come from the 16 seed MAGs carry a circularity mark. An absence always means that this detector did not find the part in this range and in the cells where it could have been seen.

We counted a locus as *Pollutibacter* when its genus label was *Pollutibacter* or SXYR01, or when it came from one of five genomes without a label. Layer B held all other K^+ loci with their genus label as given, a locus without a genus label being labelled by its family and never counted as a genus. Layer C was anchored on the 158 same-strand adjacent pair rows of -1 and $+3$ genes, with their RT status recorded, and layer D was anchored on the -1 gene, the $+3$ gene, $+1/+2$ pairs or the element. The units of counting were the locus (genome, contig and RT index, counted once across catalogues), the genome, the genus (GTDB R232 label), the ANI95 cluster and the sample, and we counted loci and genera separately and never added them. For ANI95 units we used the skani clusters for GTDB genomes and computed fastANI clusters for genomes without a label, and a locus in neither set was counted by its sample as ANI95 not computed. fastANI ran with fragment length 3,000 and minimum fraction 0.1, and genomes were joined by single linkage when ANI was at least 95 and at

least 50 fragments (about 150 kb) matched. Because the LIMNOS catalogue is about 54 GB, we did not download its genomes, so LIMNOS, Zenodo and GEM loci were counted as sample units and the independent units of the first counting layer are a mixture of ANI95 clusters and samples.

For the one seed cell that was observed absent and was checkable, a genome-wide search with the element model found the element at $E 5 \times 10^{-18}$, -112 nt from the RT start, so the window had missed it because the RT start call cut the window, whereas the 2 observed-absent cells of the non-seed census were not checkable for lack of nucleotide sequence. This case is a window cut at one seed locus and not an absence of the element, and the value belongs to a seed MAG, which carries a circularity mark because the part and element models were built from the seed MAGs and is not counted as evidence. Within the scanned range and for this detector the loci of the ESOL cassette come from a single genus, *Pollutibacter*, which counts as one independent unit, and the loci, genomes and samples counted in this census are not further independent units.

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 scan tests neither function, mobility nor origin, the genus is the unit of every inferential statement, and pairs, genomes and loci are descriptive. We sealed the pre-specification on 2026-10-11 by recording its SHA-256 hash and time, and we froze the detectors by hash before the full scan was launched. Before sealing we had seen the pair counts made before sealing and the head of the pair table, so the recomputation of those numbers and the replicate inside the new scan were declared recomputations, and everything computed with the detectors on the full universe was declared new and judged by the rules written beforehand.

The universe was the 23,901 GTDB R232 species representatives of the phylum Bacteroidota, taken from the R232 taxonomy table. Genes came from the GTDB R232 representative proteomes (Prodigal ORFs), streamed from public GTDB mirrors, with the gene index set to the order of the ORF along its contig and contig boundaries never crossed. The table of the count made before sealing, which used NCBI annotation where available and prodigal otherwise, served only for an annotation-source sensitivity analysis. Genomes that failed to parse or had too few ORFs were counted in the summary and never read as absence, and the subsets Q90 (CheckM2 completeness of at least 90 and contamination below 5) and a subset with more ORFs served as sensitivity sets. Unprocessed genomes and failed downloads were recorded and never read as absence.

A pair was a -1-type gene and a +3-type gene on the same contig and strand at gene-number distance d , the absolute difference of their indices, with $d = 1$ for adjacent genes and $d - 1$ genes between the members. The primary range was the range of the count made before sealing and a literal reading of that range was also reported. A pair was called -1-upstream, the cassette order, when the -1 gene preceded the +3 gene in the direction of transcription, and +3-upstream otherwise. An RT lay between the members when at least one RT gene had an index strictly between them on the same contig, on any strand in the primary analysis. For every pair we also recorded whether an RT gene lay before the upstream member, after the downstream member or nowhere, and the reverse arrangement with an RT before the -1 gene. The cassette geometry flag required -1-upstream order, $d = 4$ and an RT gene on the same strand exactly one gene downstream of the -1 gene, which is the

arrangement of the *Pollutibacter* cassette with -1 , RT and $+3$ at positions -1 , 0 and $+3$. A genome counted as a both genome when it had at least one -1 -type and one $+3$ -type gene at the given threshold level, and the null and its denominators used these genomes.

The rotation null was checked against the controls only at short gene-number distances, one of the controls went in the opposite direction, and its extension to larger distances is weakly supported, so we report ratios with bootstrap intervals and do not use its exact upper-tail probability as evidence. The pre-specified rule still contained that probability, so its label above chance is the literal outcome of the rule and not the result of a P-value based test. A post hoc check outside the pre-specification restricted the comparison to pairs in which no member had an adjacent partner of the other class, and its values were not recomputed. For each pair we also rotated the RT gene set and took the share of offsets that put at least one RT gene strictly between the members, and the sum over pairs gave the expected number of pairs with an RT between, which serves as the power statement for an absence. If this expected number outside the cassette lineage was very small, an observed absence there was the chance outcome and said nothing about an RT association.

To ask whether rooting with outgroups can place the cassette -1 and $+3$ genes inside or at the base of the free-gene diversity, we built outgroup-containing gene trees under rules written before any outgroup sequence, tree or intervening-gene annotation of this analysis had been seen. We sealed the rules on 2026-10-11 by recording the file hash and time. The ingroup was the -1 family and the $+3$ family with their cassette and free sequences. The -1 outgroups came from the Pfam seed alignments of VWA (PF00092), VWA_2 (PF13519), VWA_3 (PF13768) and VWA_CoxE (PF05762), and the $+3$ outgroups from the Peptidase_M54 (PF07998) seeds, the reviewed UniProtKB entries with Pfam PF07998 and the first 300 unreviewed Bacteroidota entries with PF07998 by accession. The selection kept sequences above a minimum length and with few ambiguous residues. It then clustered each family at 50% identity and 80% coverage with MMseqs2 easy-cluster. If the counts or the composition after selection differed from the rule, the rule was not changed and the selection was used as it came out.

S3.14 Pre-specification and time stamps

No wet-laboratory result is reported for this work: its results come from computation and from the reading of the literature. Analyses were specified in rule files before they were run, and a rule file was sealed by recording its SHA-256 hash and the time; changes made after results were seen are marked as post hoc. The date of sealing and the SHA-256 digest of the protocol record are given in the guide to the supplementary material at the start of this document.

A stamp is an internal record: there is no third-party clock behind any stamp or amendment. The time of a stamp is the time that the analyst's own system gave. We cannot create third-party evidence afterwards for work that is already done.

S3.15 Reading the deposited tables

In this section a cassette-positive locus is a post hoc operational class of loci in which a -1 gene lies on the strand of a K^+ RT, and it is not a clade. An atlas row is one K -family RT gene in one catalog table and not a locus, and duplicate groups across catalogs are marked so that counts of loci are taken from primary rows only.

Within the scanned range, which consists of the Zenodo freshwater, LIMNOS lake, GEM and GTDB K-positive collections, and for this detector the cassette-type arrangement is described from a single genus, *Pollutibacter*, which counts as one independent unit, and the counts of rows, genomes or samples are not independent units. The unit counts at the levels of genus, sample and cassette genus are not summed. Models and the definition of the K family were derived in part from the 16 seed MAGs, so values inside the seed set are not independent evidence, and the seed-derived columns are flagged. In every table, not detected means that this detector did not find the element in the range processed and inside the observable window of the contig, and it is not evidence of absence. Nothing in the deposited tables is evidence of function, expression or transcription.

The locus atlas, the component tables, the GFF3 files, the recorded trees and the sequence files are deposited in a public archive (Data availability). No wet-laboratory result is reported.

S3.16 Software versions

We took tool versions from the version banners and package records of the runs. Within the logs we scanned, HMMER appears only as version 3.4, Infernal only as 1.1.5 and pyrodigal only as 3.7.1. For MAFFT, MMseqs2 and DIAMOND the logs of different analyses record different versions, so we state no single version for these tools, and the version of an analysis is the one in its own run log.

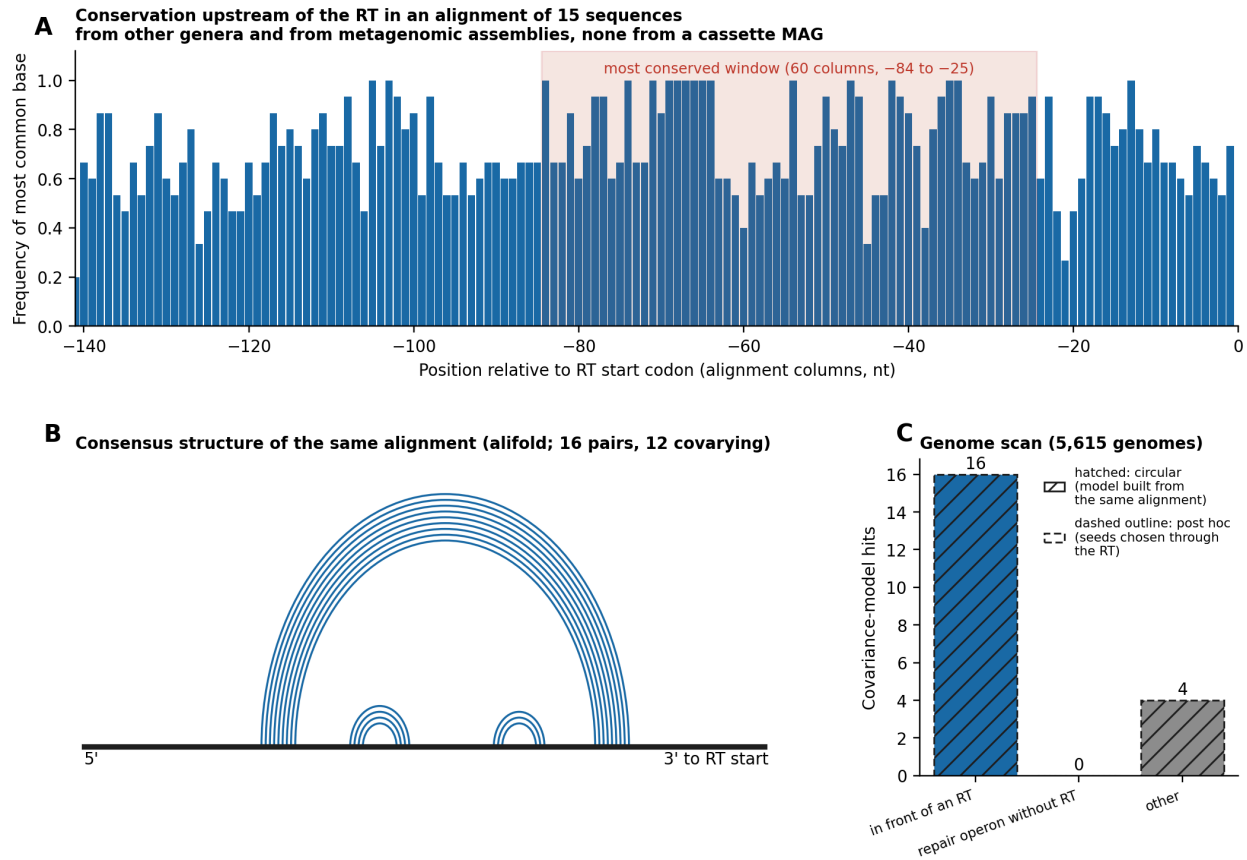
References

1. 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>
2. Tesson F, Hervé A, Mordret E, Touchon M, d'Humières C, Cury J, et al. Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nat Commun* 2022;13:2561. <https://doi.org/10.1038/s41467-022-30269-9>
3. Payne LJ, Todeschini TC, Wu Y, Perry BJ, Ronson CW, Fineran PC, et al. Identification and classification of antiviral defence systems in bacteria and archaea with PADLOC reveals new system types. *Nucleic Acids Res* 2021;49:10868-10878. <https://doi.org/10.1093/nar/gkab883>
4. 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>
5. 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>
6. Nayfach S, Roux S, Seshadri R, Udwy D, Varghese N, Schulz F, et al. A genomic catalog of Earth's microbiomes. *Nat Biotechnol* 2021;39:499-509. <https://doi.org/10.1038/s41587-020-0718-6>
7. Chikhi R, Lemane T, Loll-Krippelbein R, Montoliu-Nerin M, Raffestin B, Camargo AP, et al. Logan: Planetary-Scale Genome Assembly Surveys Life's Diversity. *bioRxiv preprint (version 2)* 2025. <https://doi.org/10.1101/2024.07.30.605881>
8. 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>
9. Tang S, Ramirez JL, Mestre MR, Zhang DJ, Wang M, Wiegand T, et al. Antiviral reverse transcriptases reveal the evolutionary origin of telomerase. *bioRxiv (preprint)* 2025. <https://doi.org/10.1101/2025.10.16.682844>

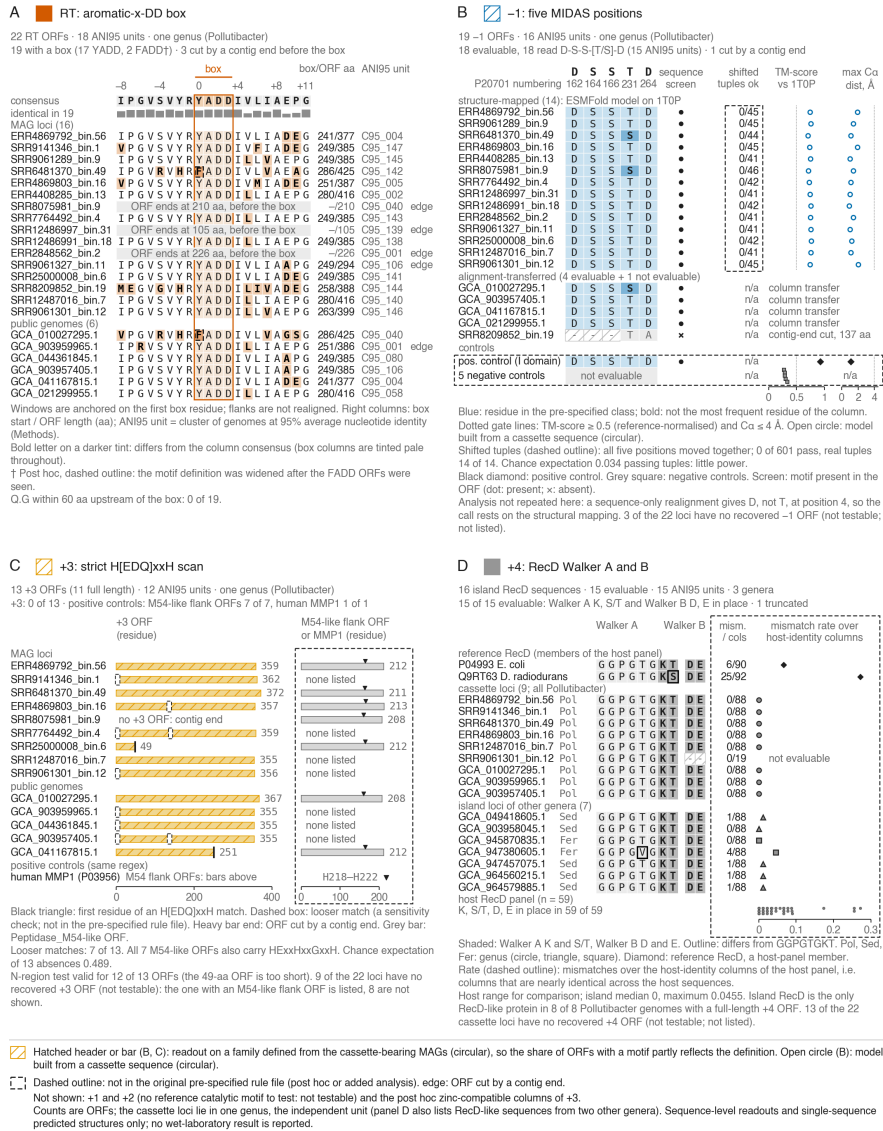
10. Ramirez JL, He Q, Wiegand T, Lampe GD, Wang J, Tang S, et al. Mechanism of tandem-repeat DNA synthesis by an antiviral reverse transcriptase. *bioRxiv* (preprint) 2026. <https://doi.org/10.64898/2026.06.29.735271>
11. Chen S, Zhang S, Zhang Y, Fu P, Mao T, Zhan Y, et al. Defense-associated reverse transcriptases synthesize DNA in a telomerase-like manner. *bioRxiv* (preprint) 2026. <https://doi.org/10.64898/2026.10.08.757509>
12. Sánchez-Nieto E, Wilkinson ME, Millán V, Wenardy C, Cabrerizo D, Martínez-Abarca F, et al. Host exonuclease SbcB and a phage-encoded SSB-like protein control activation of the DRT10 reverse transcriptase defense system. *bioRxiv* (preprint) 2026. <https://doi.org/10.64898/2026.04.29.721553>
13. Wang M, Yoneyama K, Žedaveinytė R, Ishikawa J, Tang S, Le HC, et al. Coordinated RNA- and protein-templated synthesis of double-stranded DNA by a dual reverse transcriptase immune system. *Cell* 2026;189:4997-5011.e11. <https://doi.org/10.1016/j.cell.2026.07.012>
14. 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>
15. 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>
16. Eddy SR. Accelerated Profile HMM Searches. *PLoS Comput Biol* 2011;7:e1002195. <https://doi.org/10.1371/journal.pcbi.1002195>
17. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 2010;11:119. <https://doi.org/10.1186/1471-2105-11-119>
18. 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>
19. Buchfink B, Reuter K, Drost HG. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods* 2021;18:366-368. <https://doi.org/10.1038/s41592-021-01101-x>
20. Mestre MR, Gao LA, Shah SA, López-Beltrán A, González-Delgado A, Martínez-Abarca F, et al. UG/Abi: a highly diverse family of prokaryotic reverse transcriptases associated with defense functions. *Nucleic Acids Res* 2022;50:6084-6101. <https://doi.org/10.1093/nar/gkac467>
21. Mestre MR, González-Delgado A, Gutiérrez-Rus LI, Martínez-Abarca F, Toro N. Systematic prediction of genes functionally associated with bacterial retrons and classification of the encoded tripartite systems. *Nucleic Acids Res* 2020;48:12632-12647. <https://doi.org/10.1093/nar/gkaa1149>
22. Toro N, Nisa-Martínez R. Comprehensive phylogenetic analysis of bacterial reverse transcriptases. *PLoS One* 2014;9:e114083. <https://doi.org/10.1371/journal.pone.0114083>
23. Wu L, Gingery M, Abebe M, Arambula D, Czornyj E, Handa S, et al. Diversity-generating retroelements: natural variation, classification and evolution inferred from a large-scale genomic survey. *Nucleic Acids Res* 2018;46:11-24. <https://doi.org/10.1093/nar/gkx1150>
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. Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, Söding J. HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* 2019;20:473. <https://doi.org/10.1186/s12859-019-3019-7>
27. 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>
28. 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>
29. 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>
30. Shaw J, Yu YW. Fast and robust metagenomic sequence comparison through sparse chaining with skani. *Nat Methods* 2023;20:1661-1665. <https://doi.org/10.1038/s41592-023-02018-3>
31. Yang Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol Biol Evol* 2007;24:1586-1591. <https://doi.org/10.1093/molbev/msm088>
32. Kosakovsky Pond SL, Poon AFY, Velazquez R, Weaver S, Hepler NL, Murrell B, et al. HyPhy 2.5—A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies. *Mol Biol Evol* 2020;37:295-299. <https://doi.org/10.1093/molbev/msz197>
33. Kosakovsky Pond SL, Frost SDW. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol Biol Evol* 2005;22:1208-1222. <https://doi.org/10.1093/molbev/msi105>
34. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW. Automated phylogenetic detection of recombination using a genetic algorithm. *Mol Biol Evol* 2006;23:1891-1901. <https://doi.org/10.1093/molbev/msl051>
35. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;34:3094-3100. <https://doi.org/10.1093/bioinformatics/bty191>
36. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. *BMC Bioinformatics* 2009;10:421. <https://doi.org/10.1186/1471-2105-10-421>
37. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* 2022;38:5315-5316. <https://doi.org/10.1093/bioinformatics/btac672>

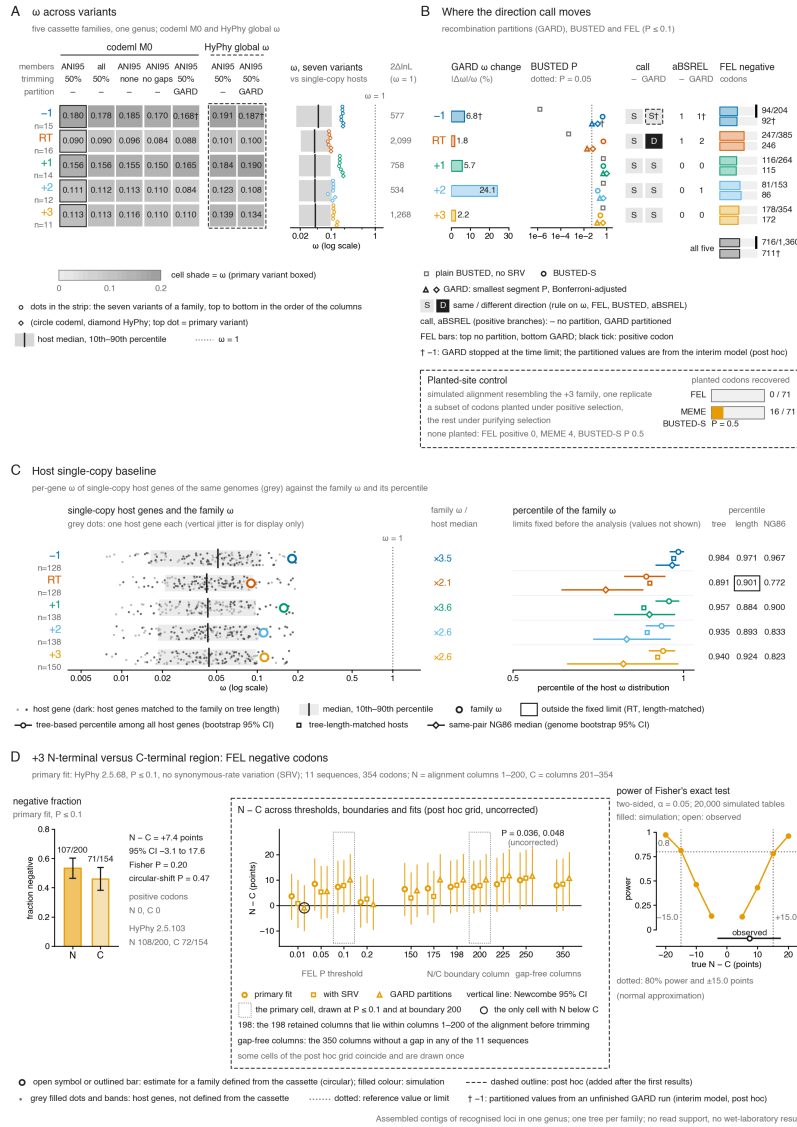
Extended Data Figures



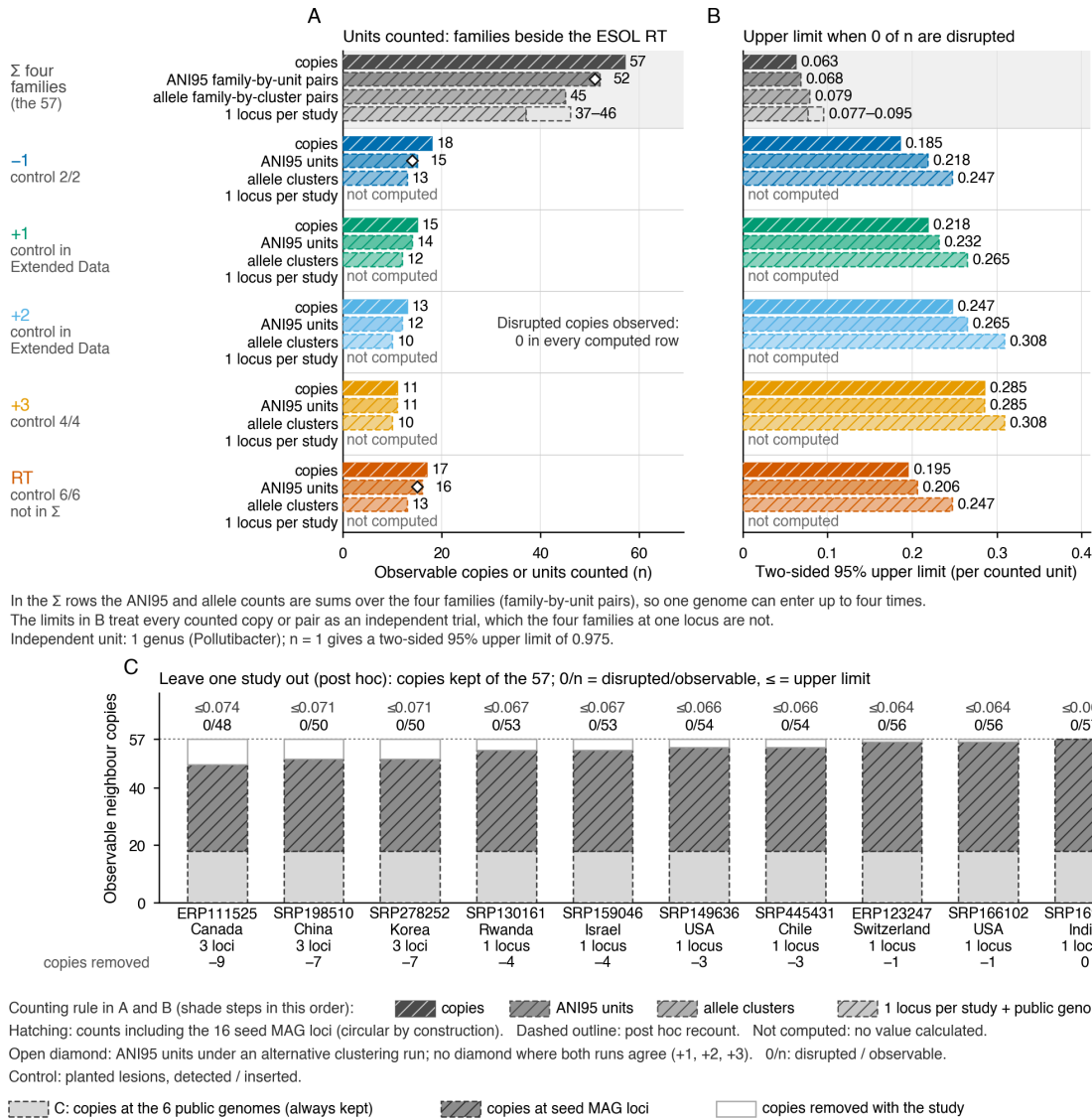
Extended Data Figure 1. Upstream non-coding element. (A) Conservation of the aligned region in front of the RT start in an alignment of 15 upstream sequences from genera other than *Pollutibacter* and from metagenomic assemblies, none of them from a cassette MAG; the shaded window is the stretch of 60 consecutive columns with the highest mean conservation, found by computation (positions -84 to -25 relative to the RT start). (B) Consensus secondary structure of the same aligned sequences with the pairs predicted by alifold (arcs); 12 of 16 pairs covary, against at most 3 in column-shuffled controls (no phylogenetic control). (C) The 20 high-scoring hits of a covariance model built from the same alignment (circular) and from seeds chosen through the RT after it was observed (post hoc; Supplementary Table S13, detector D23), in a scan of 5,615 genomes of the host lineages, classified by what lies downstream ('in front of an RT': an RT among the nearby downstream genes; 'repair operon without RT': nearby repair-type genes and no RT; 'other': inside a gene or too near a contig end to classify).



Extended Data Figure 2. Residue-level readouts of the RT box and of motifs in the -1 and +3 proteins of the ESOL cassette and in the RecD-like +4 protein found at the same loci, shown with the number of open reading frames (ORFs) behind each call. Hatched headers and bars and open circles mark readouts made on families or models that were defined from the cassette-bearing MAGs, so the share of ORFs carrying a motif partly reflects that definition, and dashed outlines mark items that are not in the pre-specified rule file. Panel A aligns the 4-residue box and eight residues on each side for the RT ORFs of the 22 loci, anchoring each window on the first box residue without realigning the flanks. Panel B reads the residue at each of the five MIDAS positions of the reference integrin I domain (D162, S164, S166, T231 and D264 of UniProt P20701) in the 19 -1 ORFs. Panel C scans the +3 ORFs for the strict pattern H[EDQ]xxH and marks the first residue of each match on a bar that shows the length of the ORF. Panel D lists the Walker A 8-mer and the Walker B Asp and Glu of the 16 RecD-like island sequences next to E. coli RecD (P04993) and D. radiodurans RecD2 (Q9RT63), with nine sequences from the cassette loci (all Pollutibacter) and seven from island loci of two other genera.



Extended Data Figure 3. Across the five codeml variants, ω of each of the five cassette families lies between 0.084 and 0.185, far below $\omega = 1$ and above the median ω of single-copy host genes, and among the four families whose GARD run finished the direction call changes only for the RT when recombination partitions are used. A, Codon-model ω of each family (rows) in seven variants (columns): codeml M0 on ANI95 representatives with 50% occupancy trimming (primary, boxed), on all members, without trimming, on gap-free columns and with GARD recombination partitions, followed by the HyPhy MG94xREV global ω without and with partitions (dashed, a cross-check added after the first results). B, Partitioning the alignments at the GARD breakpoints changes ω by 6.8% or less in four families (1.8% for the RT, 6.8%† for –1, 5.7% for +1 and 2.2% for +3) and by 24% in +2. C, Left, ω of single-copy host genes of the same genomes (128–150 per family; grey dots, dark where the host genes are matched to the family on tree length) against the ω of each family (open circle), with the host median and the 10th–90th percentile. C, Right, ω of each family is 2.1–3.6 times the host median and lies at the 0.89–0.98 percentile of the host distribution by tree-based ω (bootstrap 95% CI), at 0.88–0.97 when host genes are matched on tree length and at 0.77–0.97 by same-pair NG86 medians (genome bootstrap 95% CI). D, In the +3 alignment, the fraction of codons with FEL-negative selection in the N-terminal and the C-terminal part, the difference between them across a post hoc grid of FEL thresholds, boundaries and fits, and the power of Fisher's exact test for that difference.



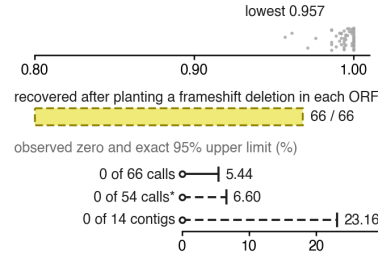
Extended Data Figure 4. Recounting the lesion census by ANI95 unit, allele cluster and study leaves the observed number of disrupted copies at zero and changes only the denominator and the upper limit. (A) At 22 assembled loci of one genus (16 seed MAG loci and 6 public genomes), the four gene families beside the ESOL RT have 57 observable copies (-1, 18; +1, 15; +2, 13; +3, 11). Hatching marks counts that include the 16 seed MAG loci, which are circular by construction, and dashed outlines mark post hoc recounts that change the denominator and not the zero. The control labels in (A) give the planted lesions detected out of those inserted for the -1, RT and +3 families, and the planted-lesion control of the +1 and +2 families is in Extended Data Figure 5. (B) With 0 disrupted copies, the two-sided 95% upper limit on the disruption rate is 0.063 for 57 copies, 0.068 for 52 family-by-unit pairs and 0.079 for 45 family-by-cluster pairs. (C) Dropping any one of the 10 studies that supplied the 16 seed MAG loci leaves 48 to 57 of the 57 observable neighbour copies, with 0 disrupted among those that remain, so the denominator changes and the zero does not.

Lesion-census controls

host-ORF baseline, planted lesions and read depth for the census of the cassette-family copies; the baseline and the read census carry no information

A Host-ORF baseline

13 families of host ORFs, same census procedure
142 ORFs, 77 clusters, 25 families shared across genomes,
13 of them with larger membership, 66 calls on 14 contigs
no internal stop in any of the 142 ORFs: zero is by design
model coverage of the 66 calls



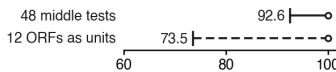
* host ORFs and host-ORF calls (not defined from the cassette) planted lesions (simulation)
■ histogram bars: read depth —○ observed value and exact two-sided 95% limit
---- dashed: post hoc (the deletion control was added after the run; * and ORF-unit limits were computed afterwards)
* the 54 calls from the larger families (12 calls from the smaller families left out)

B Planted lesions

synthetic single lesions planted in +1 and +2 ORFs

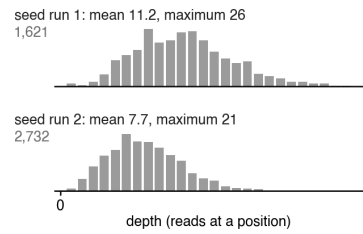
	position in the coding sequence		
	early	middle	late
frameshift deletion	12/12	12/12	12/12
frameshift deletion	12/12	12/12	12/12
frameshift insertion	12/12	12/12	11/12
premature stop	12/12	12/12	12/12
all four kinds	48/48	48/48	47/48

137 frameshift or stop, 6 truncated, 1 intact calls
one ORF holds all 6 truncated calls and the miss
observed 100% and exact 95% lower limit (%)



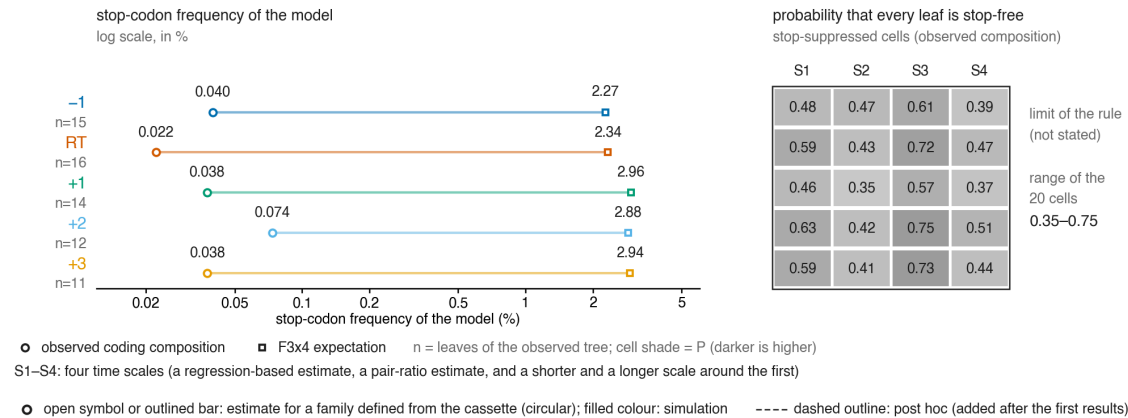
C Read depth

2 of the seed runs, aligned to a seed contig
positions per depth (reads)



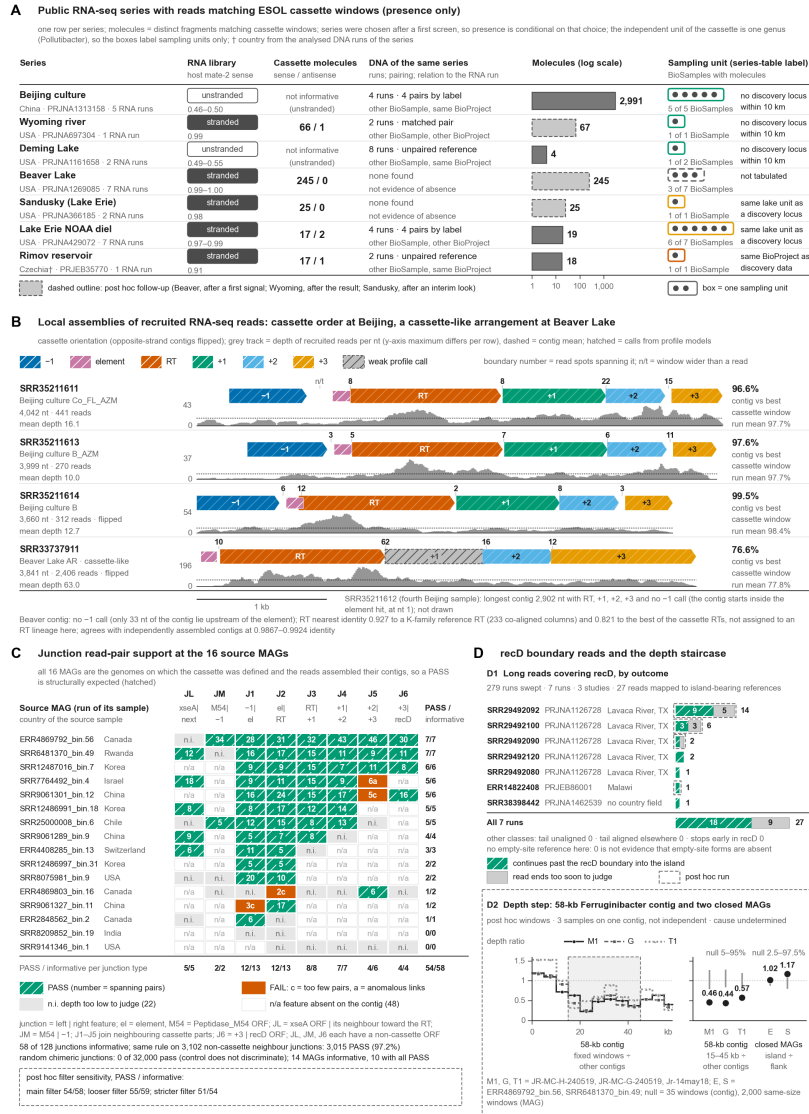
D Stop-suppressed model: exact probability that every leaf is free of premature stops

exact pruning on the observed trees; substitutions only; the five series are neighbouring genes of one locus and are not combined

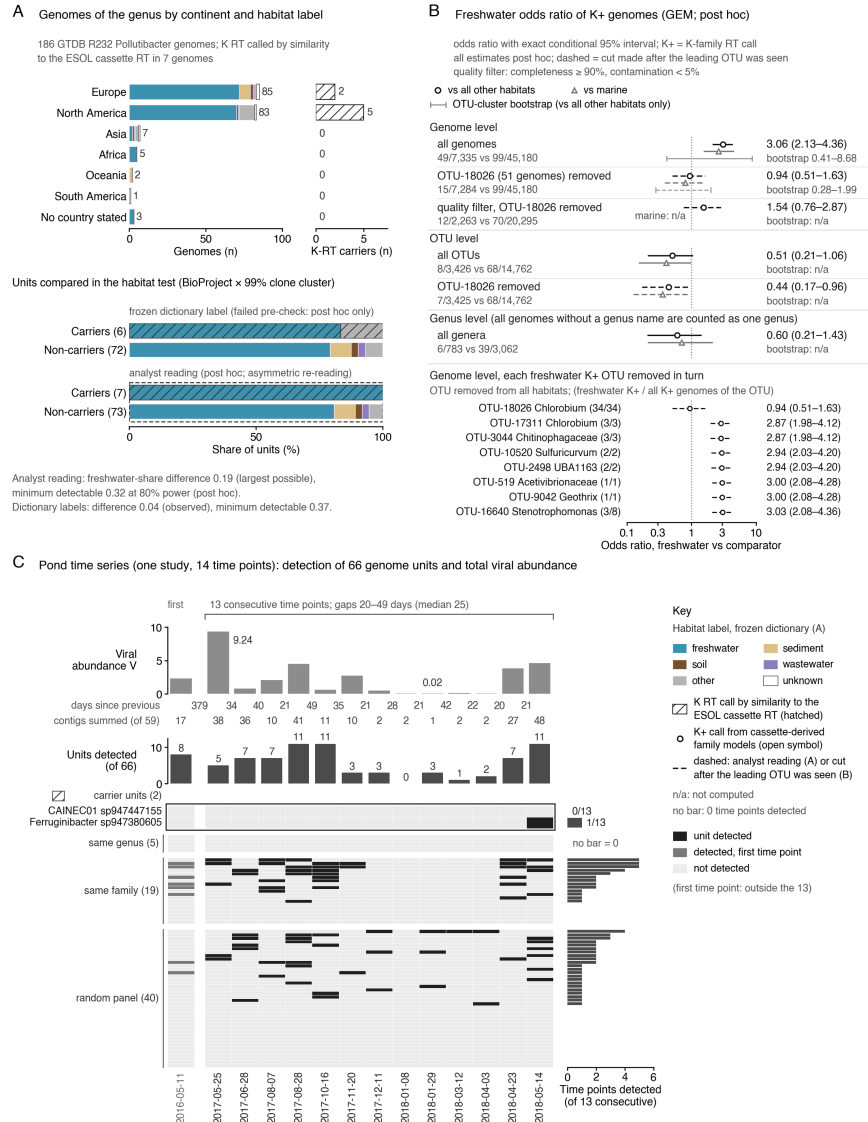


Assembled contigs of recognised loci in one genus; one tree per family; the read census (panel C) is too small to inform; no wet-laboratory result.

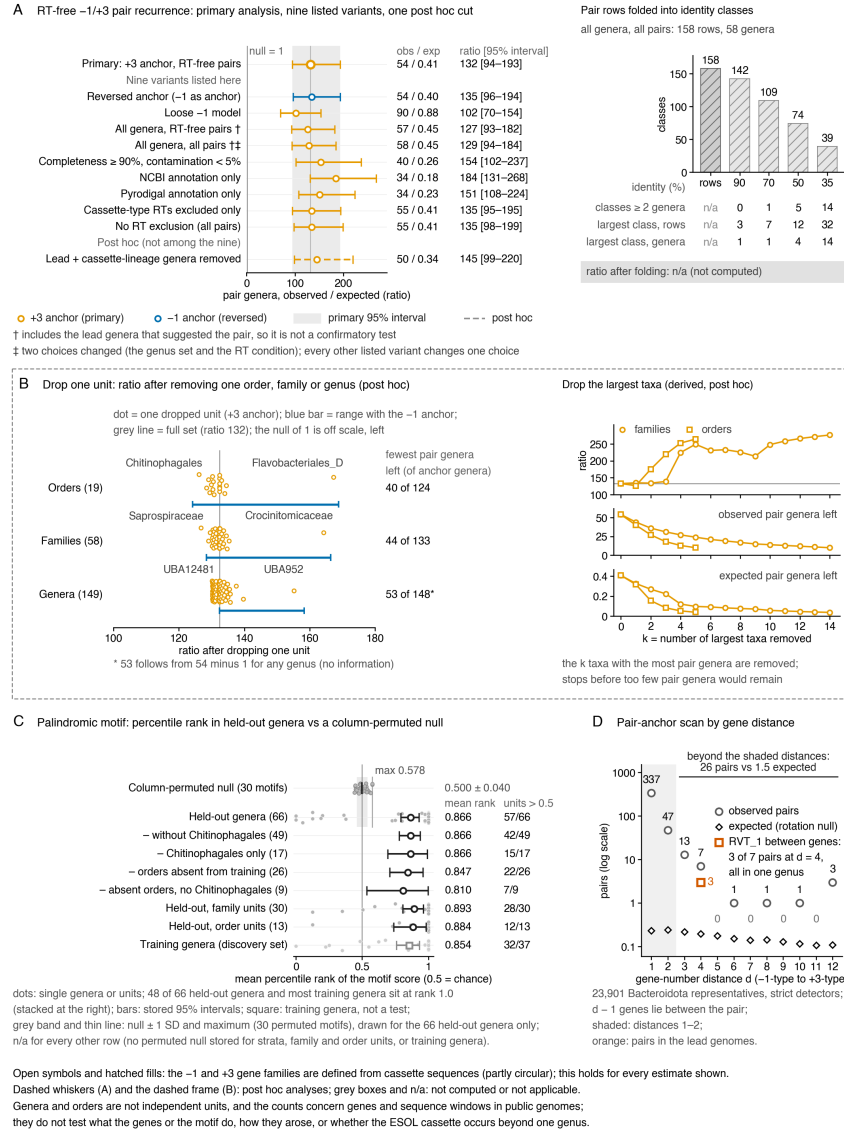
Extended Data Figure 5. This figure gives the controls of the lesion census of the cassette-family copies, namely a host-ORF baseline (A), planted lesions (B) and a read-level census (C), and the stop-suppressed cells of the exact neutral calculation (D). A, the census procedure used for the cassette families was applied to 13 families of host ORFs that flank the module block, and all 66 calls on 14 contigs were intact (0 of 66 disrupted). B, synthetic single lesions of four kinds (two frameshift deletions, a frameshift insertion and a premature stop codon) were planted at an early, a middle and a late position of the coding sequence in intact ORFs of the +1 and +2 families (144 tests), and 143 of them were called disrupted. C, reads of 2 of the seed runs were aligned to a seed cassette contig, and the mean depth was 11.2 and 7.7 over 16,225 and 20,689 pileup positions (maximum 26 and 21). D, the stop-codon frequency of the model for each family under the observed coding composition and under the F3x4 expectation (log scale, in per cent), and the exact probability that every leaf of the observed tree is free of premature stop codons in the stop-suppressed cells (observed composition) at four time scales S1 to S4, which ranges from 0.35 to 0.75 in the 20 cells, the limit of the rule being not stated. Panels A to C do not show a natural lesion rate of host genes, a floor of the census, a sensitivity beyond the planted single lesions or a lesion frequency at read level, because the host families are defined from called ORFs and only 2 of the seed runs were read.



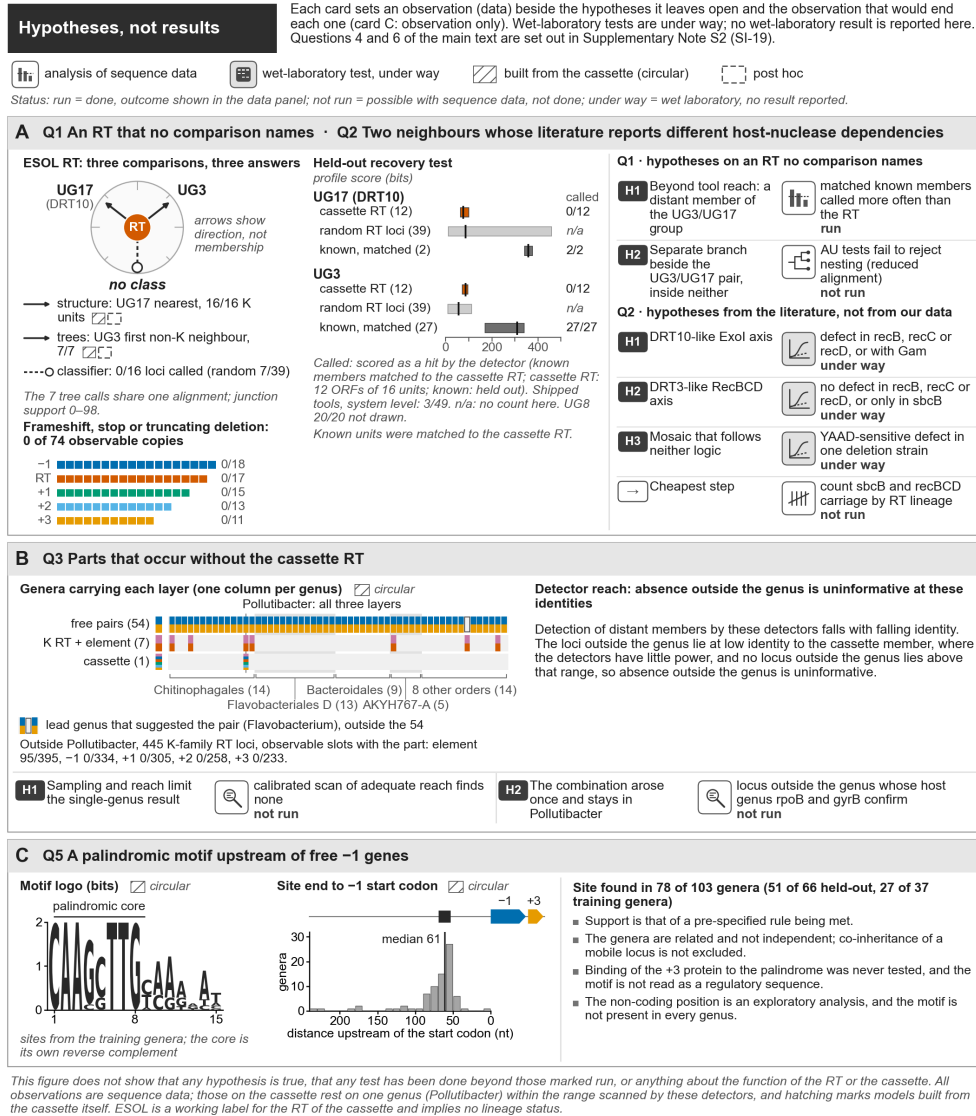
Extended Data Figure 6. This figure collects read-level observations on the ESOL cassette from public sequencing data, limited to presence, and marks for each row how few sampling units stand behind it, the independent unit of the cassette being one genus. (A) Public RNA-seq libraries of 7 series contain reads that match cassette windows: Beijing culture, Wyoming river, Deming Lake, Beaver Lake, Sandusky, Lake Erie and Rimov reservoir. A dagger marks a country taken from the analysed DNA runs of the series because the Rimov RNA run record has no country. Dashed outlines mark Beaver Lake, which was followed up after a first signal, Wyoming, which passed after the scan was extended, and Sandusky, which was chosen after an interim look (Supplementary Table S8, row P03). (B) Local assembly of the recruited reads gives contigs of 4,042, 3,999 and 3,660 nt for 3 of the 4 Beijing RNA-seq samples assembled, each carrying the -1, RT, +1, +2 and +3 modules in cassette order with the element between -1 and RT, at a mean depth of the recruited reads of 16.1, 10.0 and 12.7; a fourth contig, from a Beaver Lake sample, is cassette-like (3,841 nt, 2,406 reads, mean depth 63.0). (C) Of 128 junction cells (8 junction types at 16 source MAGs) 58 can be judged by read pairs, 54 of these pass and 4 fail, while 22 cells are too shallow to judge and 48 do not exist on the contig. (D) Of 279 long-read runs swept, 7 runs from 3 studies carry 27 reads that cover recD, and 18 of them continue past the recD boundary into the island while 9 end too soon to judge. The controls for whether such reads could be chimeras are in Figure 7D.



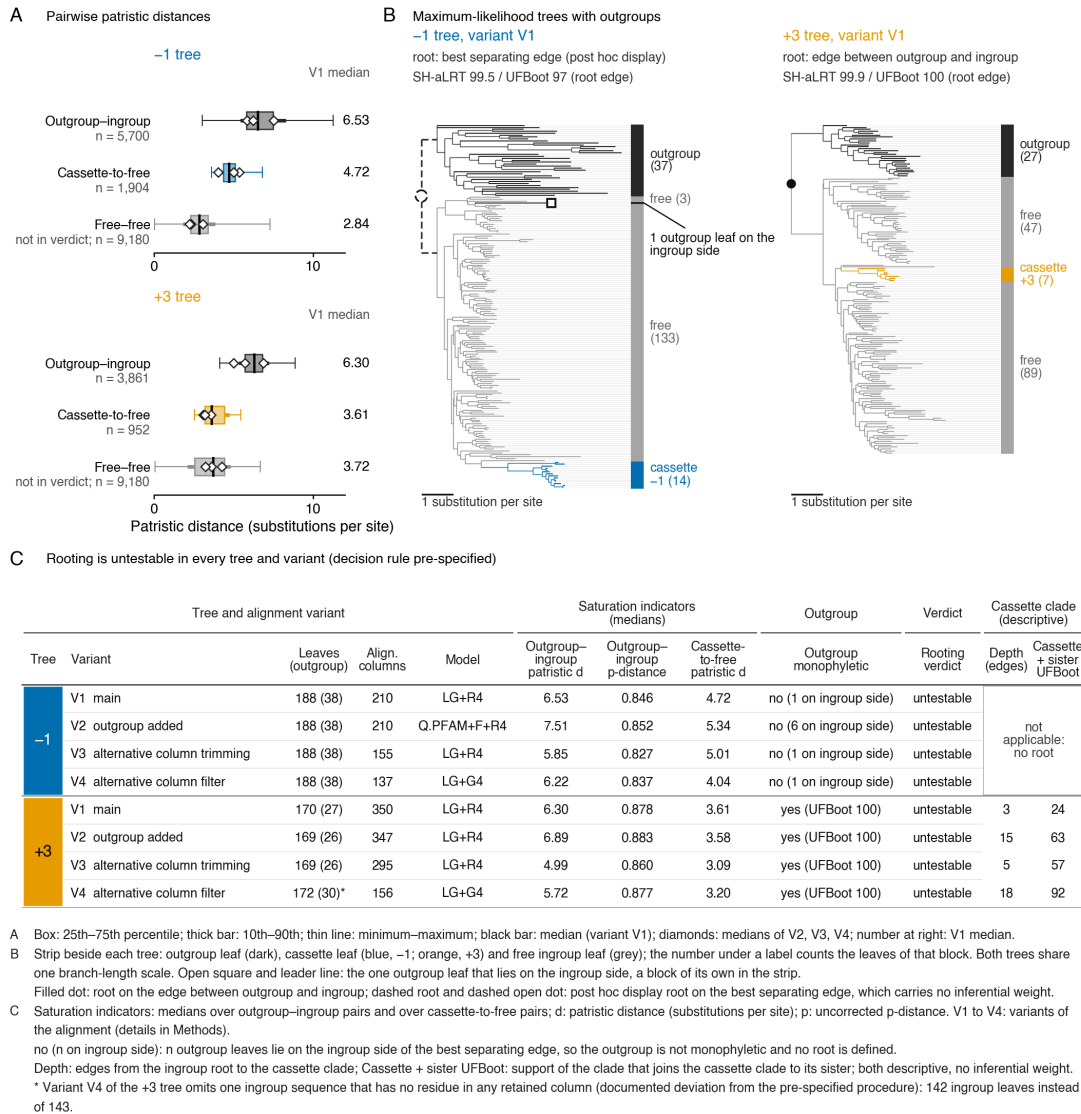
Extended Data Figure 7. Habitat labels of the 186 genomes of the genus *Pollutibacter* and a pond series of 14 time points are shown as records that carry no information on a habitat or host-dynamics statement, and a freshwater odds ratio for K-family RTs is shown to rest on one OTU. A, The 186 GTDB R232 genomes of *Pollutibacter* are counted by continent, which is derived from the country text of the source record with a fixed country-to-continent table and is not a recorded field, and by the habitat label of the frozen dictionary (152 freshwater). Seven genomes of six species carry a K RT called by similarity to the cassette RT proteins (hatched), five from North America and two from Europe. B, Odds ratios with exact conditional 95% intervals compare K+ genomes of freshwater with all other habitats (circles) and with marine habitats (triangles) in the Genomes from Earth's Microbiomes (GEM) catalogue of 52,515 genomes, at genome, OTU and genus level. C, The one pond series has 14 time points from one study (PRJNA429145), 13 of them consecutive with gaps of 20–49 days (median 25) and the first 379 days before the second.



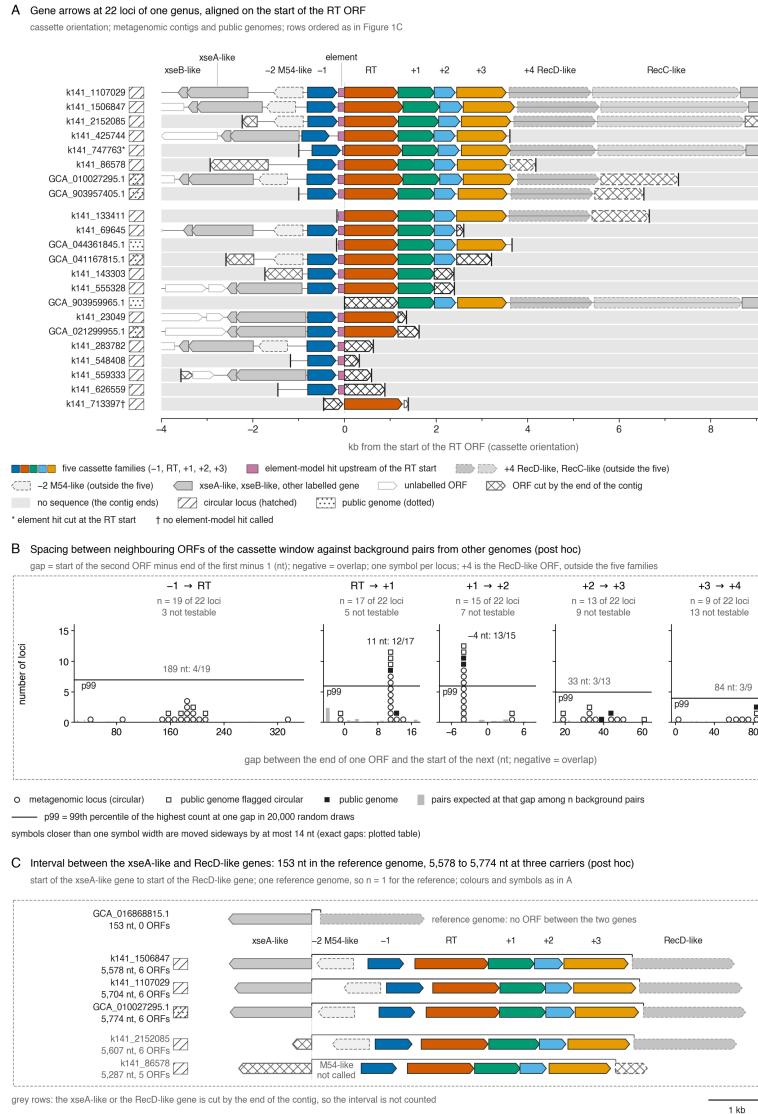
Extended Data Figure 8. Recurrence of RT-free -1/+3 gene pairs in public Bacteroidota genomes is shown across the nine variants of the primary analysis listed here (A), after dropping orders, families, genera and the largest taxa (B), for a palindromic motif scored in held-out genera (C) and by gene distance in a scan of 23,901 representatives (D), with the genus as the counted unit. A, Among the 149 genera of the confirmatory set that carry a strict +3-type gene (the lead genera that suggested the pair are excluded), 54 genera in 24 families have a same-strand -1-type gene and no RT in the window, against 0.41 genera expected under a positional null (ratio 132, 95% genus-bootstrap interval 94–193). A (right), The 158 pair rows of all genera (156 genomes, 58 genera) fall into 142 classes when both members are folded at 90% identity, the largest class holding 3 rows of one genus, and into 109, 74 and 39 classes at 70%, 50% and 35%. B (left), Each dot is the ratio recomputed after dropping one order (19 orders), one family (58) or one genus (149) with all its loci, the grey line is the ratio of the full set (132) and the blue bar is the range of the same drops when the -1 gene is the anchor (22 orders, 86 families, 248 genera). B (right), The k orders or families with the most pair genera are removed in turn (ties broken by the larger expected count) and removal stops before too few pair genera would remain.



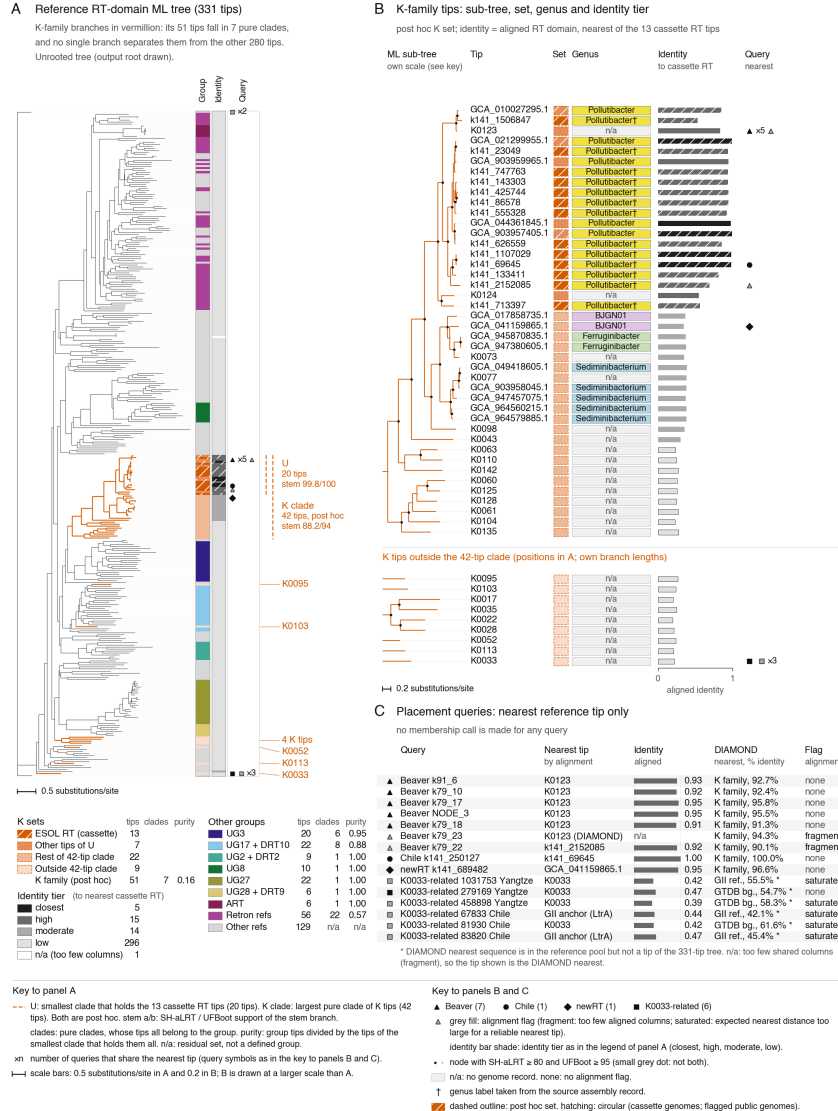
Extended Data Figure 9. Open questions that the observations raise and do not answer, with the hypotheses that questions 1 to 3 leave open and, for each, the observation that would end it; the pair-first hypothesis of question 3, which concerns polarity and is untestable at present, is listed in Supplementary Note S2 and not drawn. Panel C gives the observation on question 5 alone, and questions 4 and 6 of the main text are set out in Supplementary Note S2. Hatching denotes models built from the cassette itself (circular), dashed outlines denote post hoc definitions, and a grey-filled icon denotes a wet-laboratory test that is under way. A, Predicted structures rank UG17 (DRT10) first and sequence trees place UG3 as the first non-K neighbour, but both are directions and not membership. B, Free -1/+3-type gene pairs with no RT detected by HMM in the search window occur in 54 genera from 24 families; the text beside the plot states that detection of distant members by these detectors falls with falling identity, so that absence outside the genus is uninformative. C, A palindromic motif with the core CAAGCTTG lies upstream of free -1 genes as a strict site in 78 of 103 genera (51 of 66 held-out and 27 of 37 training genera), ending a median of 61 nucleotides before the start codon (10th percentile 49.7 and 90th percentile 113.4).



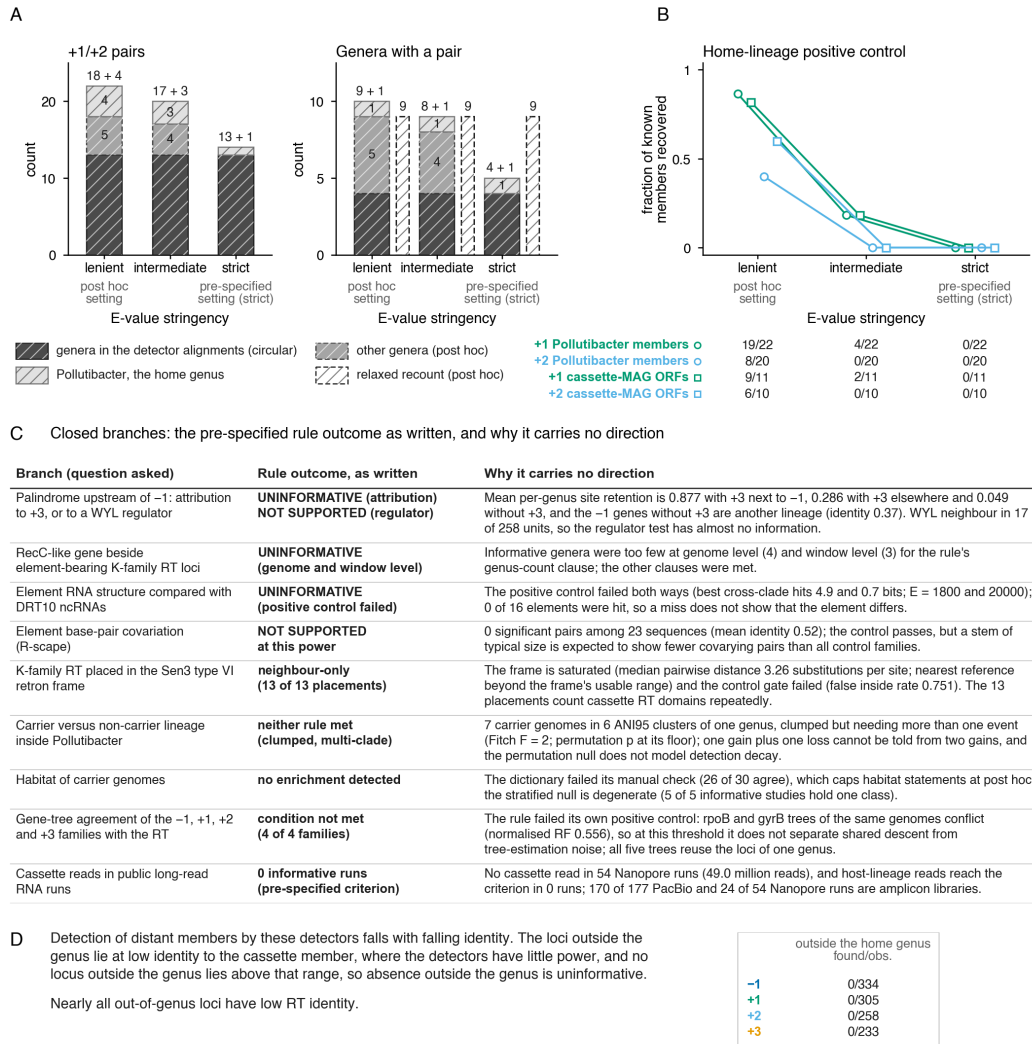
Extended Data Figure 10. Rooting of the -1 gene tree with outgroups is untestable under the pre-specified saturation rules in 4 of 4 alignment variants, and this verdict depends on the outgroups chosen, because none of the 38 selected outgroup sequences has a hit to the -1 profile. (A) In variant V1 the median outgroup–ingroup patristic distance is 6.53 substitutions per site in the -1 tree and 6.30 in the +3 tree, and the median patristic distance of cassette-to-free pairs (a cassette leaf and a non-cassette ingroup leaf) is 4.72 and 3.61. (B) The -1 tree in variant V1 has 38 outgroup, 14 cassette and 136 free (non-cassette ingroup) leaves, and no edge separates all outgroup leaves from the rest, because 1 outgroup leaf lies on the ingroup side of the best separating edge (SH-aLRT 99.5, UFBoot 97). The root drawn on this tree is a post hoc display root on that edge, shown dashed, and carries no inferential weight, while the terminal branch of that outgroup leaf is 2.39 substitutions per site. (C) The four variants are V1, the main alignment, V2, in which the outgroup is added to the fixed ingroup alignment, V3, an alternative column trimming, and V4, an alternative column filter.



Extended Data Figure 11. Array coordinates, spacing signature and the xseA-like to RecD-like interval of the cassette. A, Gene arrows at 22 loci of one genus, drawn in the cassette orientation with positions in kb from the start of the RT ORF. The five cassette families are -1, RT, +1, +2 and +3, and the +4 RecD-like ORF, the RecC-like ORF and the -2 M54-like ORF are outside the five and are drawn with dashed outlines. Hatching marks a locus flagged circular in Supplementary Data S1 (the 16 metagenomic loci and 4 of the 6 public genomes), and a dotted fill marks a public genome; in B an open symbol marks a circular locus. B, For each pair of neighbouring ORFs in the cassette window (-1 and RT, RT and +1, +1 and +2, +2 and +3, and +3 and +4) the gap from the end of the first ORF to the start of the second is drawn as one symbol per locus, with negative values for overlaps. C, The interval from the start of the xseA-like gene to the start of the RecD-like gene is 153 nt in the reference genome, which has no ORF between the two genes.



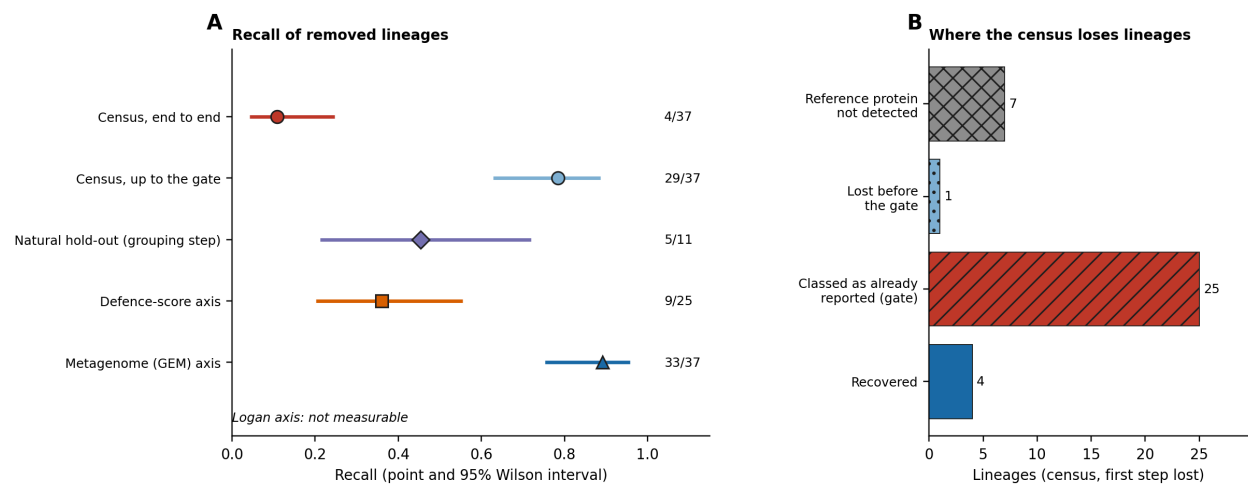
Extended Data Figure 12. The post hoc K-family set is split into several clades in the reference RT-domain maximum-likelihood tree, and its tips are shown with their aligned identity to the cassette RT and the genus of their source genome, with placement queries drawn at their nearest tip. Panel A draws the reference maximum-likelihood tree of the 331 tips of the reduced set (Supplementary Table S2, table I1) without an inferred root, with the group II intron reference tips at the position of the output root, and adds a group band, an identity band and a query column in which a count after a symbol gives the number of queries that share the nearest tip; the 435-tip main tree of Supplementary Table S2 (table G1), which gives the 42-tip K-family clade a stem of SH-aLRT 83.7, is a different run. The cassette RT tips are the cassette-bearing reference itself, so these tips and their identity values are hatched as circular, as are the tips of the public genomes flagged circular in Supplementary Data S1 in panels A and B, and dashed outlines mark post hoc sets. Panel B enlarges the K-family tips as a sub-tree of the 42-tip clade and of the K-family tips outside it, with the set, the genus of the source genome and the aligned identity of each tip. Panel B is drawn at a larger scale than panel A, with scale bars of 0.2 and 0.5 substitutions per site. Panel C lists the placement queries (Beaver, Chile, newRT and K0033-related) with the reference tip nearest by aligned identity and by DIAMOND, and it makes no call on clade membership for any query.



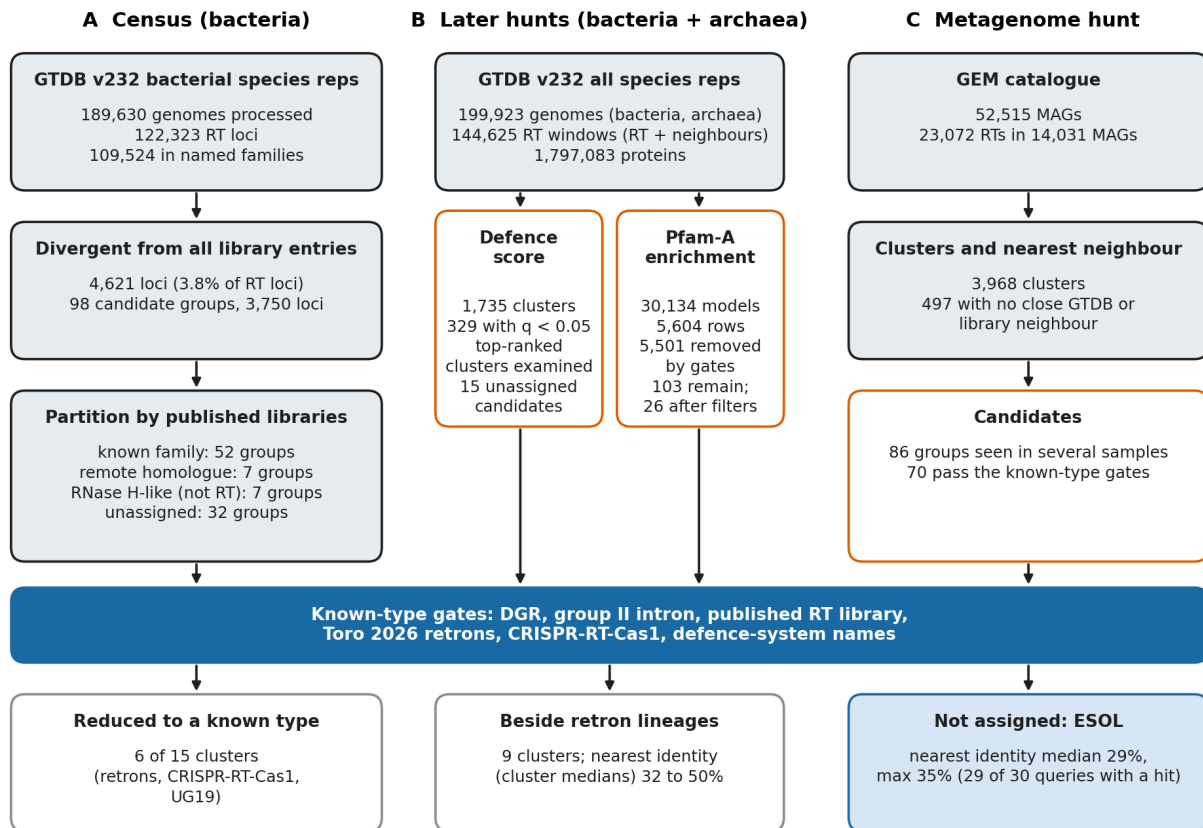
A: bar labels give the pairs or genera outside Pollutibacter + those in Pollutibacter. One setting was chosen post hoc after a calibration scan; the strict setting was pre-specified in the rule. Hatching: the models or control sets derive from the cassette lineage (circular); dashed outline: post hoc split or recount; open markers: circular control. The pair-module rule outcome changed from reclassified to intermediate as stringency increased. A and B are counts reported by the scan at each threshold; C quotes each rule outcome as written and is not read as support, refutation, absence or difference, and its element rows use models built from our own cassette data.

Extended Data Figure 13. This figure reports how the E-value threshold changes the +1/+2 pair scan (A) and the recovery of known members by the same detectors (B), the pre-specified rule outcomes of the branches that closed without a direction (C), and the counts of parts outside the genus, where absence is uninformative (D). (A) The scan of 199,923 GTDB R232 representative proteomes finds 22 tight cassette-order pairs of a +1 and a +2 gene at the lenient (post hoc) setting, 20 at the intermediate and 14 at the pre-specified strict setting. Hatching marks counts and control sets that derive from the cassette lineage, dashed outlines mark post hoc splits and recounts, the recount bars carry both marks, and the bar labels in A give the pairs or genera outside Pollutibacter plus those in Pollutibacter. The dark segment counts pairs in genera that contribute sequences to the alignments of the +1 and +2 detectors (Supplementary Data S3); this circular category is not the four-genus flag of Supplementary Table S4. (B) At the lenient E-value setting the detectors recover 19 of 22 known +1 members and 8 of 20 known +2 members that an earlier cassette-seeded search found in Pollutibacter genomes, at an intermediate setting they recover 4 of 22 and 0 of 20, and at the strict setting they recover 0 of 22 and 0 of 20. (C) Each row quotes the outcome of a pre-specified rule as written, and none of these outcomes is read as support, refutation or absence. (D) Detection of distant members by these detectors falls with falling identity; the loci

outside the genus lie at low identity to the cassette member, where the detectors have little power, and none of the observable slots outside the genus carries a part (−1 0 of 334, +1 0 of 305, +2 0 of 258, +3 0 of 233), so absence outside the genus is uninformative.



Extended Data Figure 14. Sensitivity of the search pipeline. (A) Recall of published lineages that were removed from the reference libraries, with Wilson confidence intervals: the census pipeline end to end and up to the published-sequence gate, a natural hold-out of 11 UG lineages absent from the census library (grouping step), the DefensePredictor defence-score axis (25 lineages with a testable cluster) and the GEM metagenome axis (37 testable lineages). The Logan axis could not be measured. (B) Step at which each of the 37 census lineages was lost. Markers and hatching distinguish the groups without colour.



Extended Data Figure 15. Search routes and known-type gates. Candidate RT lineages come from a census of GTDB representative genomes (A), neighbourhood-enrichment hunts that did not use ESOL as the target, over RT windows (B) and a metagenome hunt (C). Panel A covers the bacterial species representatives of GTDB R232 (189,630 genomes); panel B covers the later hunt over all species representatives of that release (199,923 genomes, bacteria and archaea), which includes the census genomes. Every candidate is checked against the known-type gates before it is described. The Pfam-A rows are cluster-by-domain enrichment rows. Our hunt, which scored every RT window with DefensePredictor [8], gave 15 unassigned candidate clusters, of which 6 reduced to known types and 9 lay beside retron lineages (nearest identity to the library, cluster medians, 32 to 50% for the 9); the ESOL RT is not assigned to a known type by these gates, a result for these libraries and not a novelty claim.