

Supplementary Note S2

File identifier: SI-19

Description: Dossier of the open questions. Provided as the PDF file Supplementary_Note_S2.pdf and in the Zenodo record.

Delivery: this PDF file (the Note is a separate file, not part of the Supplementary Information PDF); the same material is in the Zenodo record as SI-19_Supplementary_Note_S2.zip.

Legend

Supplementary Note S2 is the dossier of the open questions that the observations of this work raise and do not answer, with a full card for each of the 12 questions and 2 items reduced to limitation statements. 6 cards correspond to the open questions of the main text, and 6 cards rest mainly on supplementary or limitation observations and appear only in this note. The questions that the main text raises are numbered 1 to 6 in their order of appearance, and the identifiers P01 to P14 are those of this note. Extended Data Figure 9 draws P01 and P02 in panel A, P03 in panel B and P06 in panel C, and it does not draw P05 and P09, which are set out only in this note. Each card gives the observation with its level of support, why it is open, the hypotheses with the predictions that tell them apart, the observation that would refute each hypothesis, the cheapest test with its status, and a figure idea. The same content is given as three tables, with 14 rows for the questions, 127 rows for the observation statements and 33 rows for the hypotheses, together with the column dictionary and the list of references. The count of observations in the questions table covers the observation block of a card only, and the observations table in addition lists the statements on the status of the test and the sentences of literature context. Levels of support are given for every observation statement, and statements that depend on models or definitions built from the cassette itself are marked because they may be circular. Supported means that a rule fixed before the analysis was met, descriptive only that a count or a table is reported without a test, and suggestive, post hoc that the definition or the contrast was chosen after the output was seen. Negative result means that a pre-specified test came out negative, no information that the test could not decide, and limit a statement that travels with the observations and is not itself an observation. The hypotheses are not results, no wet-laboratory result is reported for any of them, and wet-laboratory work on the tests named in the cards is under way. All observations on the cassette itself lie within one genus in the range scanned by the detectors, and several observations rest on models built from the cassette. In this note a pair without further qualification means the -1/+3 gene pair, and two samples are called a sample pair, the focal sample pair being the Canadian and the Chilean sample. The words island, insert, insertion, gain, loss and excision describe a stretch of sequence present at an address in one genome and absent at the same address in a related genome, and they claim no mobility or mechanism. This note does not show that any hypothesis is true, that any test beyond those marked as run has been done, or anything about the function, expression or mobility of the RT or the cassette.

README

Supplementary Note S2 open-question dossier (supplement to the ESOL RT study)

Public edition of the dossier of open questions that the observations of this work raise and do not answer. Tab-separated text files, UTF-8, one header line, no quoting, no empty cell (not applicable marks a cell that does not apply).

Files

File	Content	Rows
SI-19_open_questions_public_en.md	The dossier as text: one card for each open question, with references	
SI-19_questions.tsv	One row for each question: title, placement, number in the main text, figure, levels of support, why it is open, cheapest test, test status, figure idea, limits	14
SI-19_observations.tsv	One row for each observation statement, test-status statement and literature sentence, with its level of support and whether it may depend on a model built from the cassette	127
SI-19_hypotheses.tsv	One row for each hypothesis with its distinguishing prediction and the observation that would refute it	33

File	Content	Rows
SI-19_references.tsv	The works cited in the statements	2
SI-19_COLUMN_DICTIONARY.tsv	Definition, values and blank rule of every column	32
SI-19_legend_public_en.md	Legend of this note	15 sentences

How to read the cards

- Each card has the observation with its level of support, why it is open, the hypotheses with predictions that tell them apart, the observation that would refute each hypothesis, the cheapest test with its status, and a figure idea.
- Six cards correspond to the open questions of the main text, six appear only here, and two items are reduced to limitation statements.
- The questions that the main text raises are numbered 1 to 6 in their order of appearance, and the identifiers P01 to P14 are those of this note.
- Extended Data Figure 9 draws P01 and P02 in panel A, P03 in panel B and P06 in panel C, and it does not draw P05 and P09, which are set out only in this note.
- Hypotheses are not results and no wet-laboratory result is reported: wet-laboratory work on the tests named in the cards is under way.
- All observations on the cassette itself lie within one genus in the range scanned by the detectors, and several observations rest on models or definitions built from the cassette, which the observation table marks as possibly circular.
- Each statement carries a level of support; the questions table counts the observation block of a card only, and the observations table also lists the test-status statements and the literature sentences.
- A pair without further qualification means the $-1/+3$ gene pair, and two samples are called a sample pair, the focal sample pair being the Canadian and the Chilean sample.
- The words island, insert, insertion, gain, loss and excision describe presence or absence of a stretch of sequence at an address relative to related genomes and claim no mobility or mechanism.
- No sequence file is part of this item.

Open questions

Supplementary Note S2. Open-question dossier (P01 to P14)

- This supplementary note holds the full card of each open question that the observations of this work raise and do not answer.
- Six cards (P01, P02, P03, P05, P06 and P09) correspond to the open questions of the main text, six cards (P07, P10, P04, P13, P11 and P08) rest mainly on supplementary or limitation observations and appear only here, and two further items (P12 and P14) are reduced to limitation statements at the end.
- The questions that the main text raises are numbered 1 to 6 in their order of appearance, and the identifiers P01 to P14 are those of this note.
- Extended Data Figure 9 draws P01 and P02 in panel A, P03 in panel B and P06 in panel C, and it does not draw P05 and P09, which are set out only in this note.
- Each card gives, in this order, the observation with the level of support that it carries, why it is open, the hypotheses with the predictions that tell them apart and the observation that would refute each, the cheapest test and its status, and the figure that would show the question.
- Levels of support are stated in plain words: supported means that a rule fixed before the analysis was met, descriptive only means that a count or a table is reported without a test, and suggestive means that the definition or the contrast was chosen after the output was seen.
- A statement that says of itself that its definition or contrast was chosen after the output was seen is marked suggestive.
- A negative result means that a pre-specified test returned a negative outcome, with a null hypothesis rejected or an expectation not met, and no information means that the test could not decide; a limit is a statement that travels with the observations and is not itself an observation.

- Every observation of ours rests on sequence data, the nuclease logic of known DRT systems being literature context, and for the cassette each observation rests on a single genus within the range scanned by the detectors.
- Several observations rest on models and definitions built from the cassette itself and may be circular, and the observation lists mark the cases that depend on such models.
- No wet-laboratory result is reported, and wet-laboratory work on the tests named in the cards is under way.
- The hypotheses are not all mutually exclusive: a card says where two hypotheses exclude each other, and a hypothesis that current data already disfavour is marked as such.
- The ESOL cassette is defined operationally as an RT with an upstream element and neighbouring genes at positions -1 , $+1$, $+2$ and $+3$ counted from the RT, and the name implies no claim about function or mobility.
- The words island, insert, insertion, gain, loss and excision describe a stretch of sequence that is present at an address in one genome and absent at the same address in a related genome, and they claim no mobility, mechanism or direction of change, which only a hypothesis names.
- A pair without further qualification is the $-1/+3$ gene pair, and two samples are always called a sample pair, the focal sample pair being the Canadian and the Chilean sample of P04.

Question	Title	Placement	Figure
P01	An RT with its catalytic box in every evaluable ORF by a pattern widened post hoc, no observed disruption and omega far from neutral that no comparison assigns to a named lineage	main text (open question 1)	Extended Data Figure 9 panel A
P02	The two nearest neighbours of the RT belong to groups for which the literature reports toxicity in cells lacking different host nucleases, and the K family is not monophyletic in the RT-domain tree	main text (open question 2)	Extended Data Figure 9 panel A
P03	Free $-1/+3$ -type gene pairs recur in many genera without an RT and element-bearing K-family RTs occur in further genera, but the combination of cassette parts with the K-lineage RT is found in one genus within the scanned range	main text (open question 3)	Extended Data Figure 9 panel B
P05	A $+3$ protein whose N-terminal domain is predicted metzincin-like lacks the canonical HExxH motif in every ORF examined, while the Peptidase_M54-like gene beside the cassette keeps it, and there is no evidence that the N region is under weaker purifying selection than the rest	main text (open question 4)	not drawn in Extended Data Figure 9, set out in this note
P06	A palindromic motif with a CAAGCTTG core found upstream of free -1 genes in training genera is reproduced in held-out genera by the pre-specified rule, and its frequency varies with the $+3$ context	main text (open question 5)	Extended Data Figure 9 panel C
P09	The 3' end of the element lies in the RT start codon or directly before it in 7 of 16 source windows and within 5 nucleotides of it in 9 of them, both by post hoc definitions, and what kind of element it is remains open	main text (open question 6)	not drawn in Extended Data Figure 9, set out in this note
P07	The island lies at the same address between xseA and recD in related genomes, but how often that address is occupied by a large insert cannot be told from the present scan	supplementary card	none (idea only)
P10	RecBCD genes accompany the RT in the genomes of one family, Chitinophagaceae, and are not detected in the carrier genomes of the other families with several carriers	supplementary card	none (idea only)
P04	The central block of the cassette is almost identical between a Canadian lake sample and a Chilean surface-water sample, while the -1 gene differs	supplementary card	none (idea only)
P13	The cassette MAGs fall in one sub-clade of the genus, the hypothesis that this is a high-GC clade is not met by the pre-specified rule, and cassette-free relatives within it are barely known	supplementary card	none (idea only)
P11	The $+1$ protein resembles a DNA-specificity subunit of type I restriction-modification systems only weakly and sits in genomes where the partner subunits are not detected	supplementary card	none (idea only)
P08	Depth over the island region falls to a fraction of the preceding region in one contig in three samples, but the interpretation is open and the expected island-lacking junction reads are not seen	supplementary card	none (idea only)
P12	Long-read DNA and RNA screens (reduced to a limitation statement)	demoted to a limitation statement	none

Question	Title	Placement	Figure
P14	The pond time series (reduced to a limitation statement)	demoted to a limitation statement	none

P01. An RT with its catalytic box in every evaluable ORF by a pattern widened post hoc, no observed disruption and omega far from neutral that no comparison assigns to a named lineage

Placement: main text (open question 1). Figure: Extended Data Figure 9 panel A.

Observation and level of support

- The aromatic-x-Asp-Asp box is present in 19 of 19 evaluable RT ORFs (17 YADD, 2 FADD), the 3 ORFs without it, among 22 in all, end at a contig edge before the box region, and the search pattern was widened after the pre-specified form missed the FADD copies. [level: suggestive, post hoc]
- No Q.G motif lies in the 59 residues upstream of the box in any of these ORFs (the same holds for the 60 residues named in the pre-specified text), and the same 59 residues carry LPLG in 17 of these 19 ORFs, LPMG in 2 and RGLP in 4, an RGLP window carrying LPLG as well. [level: descriptive only]
- No frameshift, premature stop or truncating deletion is seen in 74 observable copies of the five cassette families, 17 of them RT copies, in the assembled contigs of one genus, and the detector recovers the planted disruptions. [level: negative result]
- The two-sided 95% upper bound on a disruption rate is 0.247 to 0.308 when the copies are collapsed to allele clusters (10 to 13 clusters per module) and 0.9750 when they are counted as the single genus that carries them. [level: negative result]
- Saturation of divergence between the copies inflates the neutral expectation of a disruption and no host pseudogene baseline exists, so a count of zero carries less information than its size suggests, and no neutral probability is quoted here. [limit]
- The substitution-rate ratio omega of the five cassette modules is far from the neutral value, from 0.0896 for the RT, the lowest, to 0.1805 for -1, and negatively selected codons predominate, a direction of constraint of intermediate strength. [level: descriptive only; may depend on a model built from the cassette]
- The classifier DefenseFinder calls a retron, DRT or DRT-like system at 0 of 16 cassette RT loci and at 7 of 39 random bacterial RT loci (Fisher P = 0.0933), but its models are biased toward Proteobacteria and their sensitivity in Bacteroidota is unconfirmed, so the missing call is not evidence about function. [level: negative result]
- The RT alignment is saturated, with a median p-distance of 0.815 between random tip pairs, and the class profiles place the cassette RTs in UG17 for 7, in CRISPR-RT for 5 and in UG3 for 1 of 16 loci (median margin 1.8 bits), an assignment forced to the nearest reference class that weakens when one class is removed. [level: no information; may depend on a model built from the cassette]
- By minimal patristic distance the K family lies nearest to UG17 (1.0993 against 1.1294 for CRISPR-RT, a margin of 0.030; UG17 is nearest in 4 of 6 trees), whereas tree topology places UG3 as the first non-K neighbour (see P02), so the nearest class depends on the method. [level: no information; may depend on a model built from the cassette]

Why it is open

- No comparison assigns the RT to a named lineage, so the first question to ask of it, what it is, has no answer at present: each reading is limited by saturation, by a shared alignment or by classifier models biased toward other phyla, and the honest answer is that it is unassigned.

Hypotheses and the observations that would refute them

- **H1, distance beyond the reach of the classifier profiles.** H1: the cassette RT is a distant member of the UG3 and UG17 group at a distance beyond the reach of the DefenseFinder and PADLOC profiles, so that held-out UG3 and UG17 members matched for distance are missed by both tools as often as the cassette RT.
 - Distinguishing prediction: Distance-matched held-out UG3 and UG17 members are missed at the rate of the cassette RT.
 - Refutation: H1 is refuted if distance-matched held-out UG3 and UG17 members are called by DefenseFinder and PADLOC clearly more often than the cassette RT.
 - Current standing: Tested once with members matched on identity to the closest reference: they were called, the cassette RTs were not; matching on alignment score and coverage was not possible with known members (see the test

status).

- Kind of test: computational (run once)
- **H2, separate branch beside UG3 and UG17.** H2: the cassette RT lies on a branch beside the UG3 and UG17 pair and inside neither, so that matched held-out members are called clearly more often than the cassette RT and AU tests on a saturation-reduced, structure-guided alignment reject nesting within UG3 and within UG17.
 - Distinguishing prediction: Matched held-out members are called more often than the cassette RT, and AU tests reject nesting within UG3 and within UG17.
 - Refutation: H2 is refuted if held-out members are missed as often as the cassette RT or if AU tests on a saturation-reduced, structure-guided alignment fail to reject nesting.
 - Current standing: The detector prediction of H2 was met with members matched on identity only, a comparison that the test status does not count as evidence at the distance of the cassette RT; no AU test on a saturation-reduced, structure-guided alignment is reported here, and the AU placement tests of the K clade are reported in Supplementary Table S2.
 - Kind of test: computational (detector part run once; AU part not reported here)

Cheapest test

- The cheapest computational test of H1 and H2 is the held-out recovery comparison, repeated with members matched on alignment score and coverage as well as on identity, with the matching rule fixed before any call is made, which needs reference members at a greater distance than the present pool offers.
- Membership is addressed by AU tests of nesting within UG3 and within UG17 on a saturation-reduced, structure-guided alignment with additional reference sequences, a computation of moderate cost after which, if saturation remains, the question is reported as unresolved.
- A reference frame for the Bacteroidota-associated subtype of UG17 awaits the release of the reference tree of its source study.

Status of the test

- A held-out recovery test was run under a rule fixed before any result was seen. [level: descriptive only]
- Of 49 known UG17, UG3 and UG8 units matched to the cassette RT on identity to the closest reference, 49 meet the DefenseFinder-style profile threshold and 23 of 49 the PADLOC-style threshold, whereas none of 16 cassette units meets either (Supplementary Table S2, table Q2). [level: descriptive only]
- With the shipped tools, which include these members in training, the same units are called as systems in 3 of 49 ($P = 0.57$ against none of the cassette units), so the counts above refer to profile thresholds and not to system calls (Supplementary Table S2, table Q2). [level: descriptive only]
- The nominal Fisher P of $1.54e-15$ becomes $2.65e-11$ when the known units are collapsed to their 23 sub-lineages and 0.0417 when the cassette is counted as the single genus it comes from, and the test is a reference-pool emulation and not a run of the tools themselves. [level: descriptive only]
- Alignment scores and coverage were not matched: the cassette RT, scored on its 12 full-length ORFs, aligns to its nearest reference with a bit score of at most 79 over 45.2 to 70.7% of its length, whereas the matched members score at least 135 over a median of 98.25% (post hoc; Supplementary Table S2, tables Q1 and Q5). [level: descriptive only]
- So no known member sits as far from its closest reference as the cassette RTs do, the comparison is not evidence in either direction at that distance, and the cassette is a single genus. [level: descriptive only]
- Known UG17 members cut to the aligned core of the cassette RT are not called and most UG3 members are still called, so a missed UG17 call of the cassette RT cannot be read as a property of its lineage (post hoc; UG8 excluded; a single genus). [level: suggestive, post hoc]

Figure idea

- The figure idea is a disagreement panel in which the structure-based, tree-based, class-profile and classifier readings of the RT are set in one line each, with the count of observed disruptions beside them; it is panel A of Extended Data Figure 9.

Limits that travel with this card

- All observations of this card concern the RT of one genus within the range scanned, the K family is a post hoc definition so that several comparisons may be circular, nearest-class readings are directions and not membership, and no wet-

laboratory result is reported.

P02. The two nearest neighbours of the RT belong to groups for which the literature reports toxicity in cells lacking different host nucleases, and the K family is not monophyletic in the RT-domain tree

Placement: main text (open question 2). Figure: Extended Data Figure 9 panel A.

Observation and level of support

- In predicted structures the nearest non-K class of the K family is UG17 (DRT10) for 16 of 16 K units, which is a direction and not membership. [level: suggestive, post hoc; may depend on a model built from the cassette]
- The pre-specified DRT10-specificity rule is met by 0 of 16 units, percentiles against a null of random bacterial RTs run from 0.103 to 0.476, and the Foldseek bit scores of the cassette-lineage models (398 to 495) lie below those of UG17 positive controls (909 to 2161). [level: suggestive, post hoc; may depend on a model built from the cassette]
- In the sequence trees UG3 is the first non-K neighbour in all 7 tree calls, at patristic distances of 1.018 to 1.994 from the K clade, but the calls share one alignment, distances are saturated, and the node support at the junction of the K clade with UG3 ranges from 0.0 to 97.9 among the 6 calls that report it. [level: no information; may depend on a model built from the cassette]
- By minimal patristic distance the nearest class is UG17 (1.0993 against 1.1294 for CRISPR-RT; UG17 is nearest in 4 of 6 trees) although the topology places UG3 beside the K clade, so structure, distance and topology do not agree on one neighbour. [level: no information; may depend on a model built from the cassette]
- The K family is not monophyletic in the RT-domain maximum-likelihood tree, and K is a post hoc definition. [level: suggestive, post hoc; may depend on a model built from the cassette]
- The cassette RT lacks the priming features and the C-terminal tail of DRT10 and no SLATT-family gene occurs in the proteomes of the cassette MAGs, whereas some K-core units are called DRT10-type by several measures (exploratory). [level: suggestive, post hoc; may depend on a model built from the cassette]
- The genes around UG17 windows are neighbours of recombinase-, HsdS-like and other island components and cannot be told from a defense-island context, and carriage of RecBCD, RecJ, ExoVII and SbcB does not differ from the null. [level: suggestive, post hoc; may depend on a model built from the cassette]
- The expectation that carriers of DRT10-type RTs retain SbcB, as would follow if SbcB suppressed DRT10-type toxicity, is not supported as a general rule in the genomes of this study, although the sensitivity of SbcB detection outside Pseudomonadota is uncalibrated, and the pre-specified general test of a RecBCD-axis logic failed. [level: negative result; may depend on a model built from the cassette]
- The literature reports that DRT3 is toxic in cells lacking RecBCD (Wang et al. 2026), and this is literature context and not an observation on the cassette. [literature context]
- A preprint reports that a DRT10 system is toxic in Escherichia coli cells lacking the host exonuclease SbcB (ExoI), where no viable transformants carrying it were recovered, and that SbcB suppresses this toxicity (Sánchez-Nieto et al. 2026), which is literature context and not an observation on the cassette. [literature context]
- In Pollutibacter the RecD-like gene downstream of the RT cannot be told from the host's own RecD, since each of the 8 full-length genomes carries exactly one RecD-like protein, and it is not counted as a cassette module (see P10). [level: suggestive, post hoc; may depend on a model built from the cassette]

Why it is open

- Structure points to UG17, for which the literature reports DRT10 toxicity in cells lacking SbcB (ExoI), and tree topology to UG3, for which it reports DRT3 toxicity in cells lacking RecBCD, so the observations of this work cannot choose between them and it is open whose logic, if any, the cassette follows.

Hypotheses and the observations that would refute them

- **H1, ExoI axis (literature-motivated).** Conjecture H1, motivated by the DRT3 and DRT10 literature and not by our data: if the cassette RT matters in a bacterial host it follows DRT10-like logic, so that in E. coli it causes a growth defect only when ExoI (sbcB) is absent, not when recB, recC or recD is absent, and the defect disappears with the YAAD variant.
 - Distinguishing prediction: Growth defect only in the sbcB deletion, abolished by the YAAD variant; no defect in recB, recC or recD deletions.
 - Refutation: H1 is refuted if a defect appears in recB, recC or recD deletions or with Gam or gp5.9, or if validated expression gives no defect in any deletion strain.

- Current standing: Not supported or tested by data of this work; the pre-specified general test of a RecBCD-axis logic failed.
- Kind of test: wet-laboratory test (under way)
- **H2, RecBCD axis (literature-motivated).** Conjecture H2, motivated by the DRT3 and DRT10 literature and not by our data: the cassette RT follows DRT3-like logic, so that the defect appears when RecBCD is absent or inhibited (recB, recC or recD deletion, Gam, gp5.9), is not changed further by an sbcB deletion, and disappears with the YAAD variant.
 - Distinguishing prediction: Growth defect with RecBCD absent or inhibited, unchanged by an sbcB deletion, abolished by the YAAD variant.
 - Refutation: H2 is refuted if no defect appears in recB, recC or recD deletions, with Gam or with gp5.9, or if a defect appears only in the sbcB deletion.
 - Current standing: Not supported or tested by data of this work; H1 and H2 exclude each other in the deletion-strain pattern they predict.
 - Kind of test: wet-laboratory test (under way)
- **H3, mosaic.** Conjecture H3: the K family is a mosaic of lineages with different host-nuclease relationships and the cassette lineage follows neither logic, so that no recBCD or sbcB deletion changes a validated cassette phenotype.
 - Distinguishing prediction: No deletion strain changes a validated cassette phenotype.
 - Refutation: H3 is refuted if a defect that disappears with the YAAD variant appears in one deletion strain.
 - Current standing: Not tested; H3 is the alternative to both H1 and H2 and not exclusive with them across lineages.
 - Kind of test: wet-laboratory test (under way)

Cheapest test

- The cheapest computational step is to classify K-positive genomes by RT lineage type (DRT10-type K core, cassette lineage, other K) and to count sbcB and recBCD carriage with the same caller for each type.
- The caller must be calibrated first, because its RecB and RecD calls are not calibrated and its SbcB sensitivity outside Pseudomonadota is unmeasured.
- The decisive test is expression in Escherichia coli deletion strains with a catalytic-site variant control; it belongs to the wet-laboratory work under way, and no result is reported here.

Status of the test

- No computational step of this test has been run, and the conjectures above rest on the literature and not on our data.

Figure idea

- The figure idea is a schematic with UG3 (first neighbour in sequence trees; literature: DRT3 toxic in cells lacking RecBCD) on one side, UG17 (nearest in structure; literature: DRT10 toxic in cells lacking SbcB, that is ExoI) on the other, and the cassette RT between them.
- Below them stands the RecD-like gene beside the RT, which is not counted as a cassette module, and the schematic is part of panel A of Extended Data Figure 9.

Limits that travel with this card

- Structure-based nearest classes are not specific against a null of random bacterial RTs, tree positions are saturated, K is a post hoc definition, the cassette observations lie within one genus, and no wet-laboratory result is reported.

P03. Free -1/+3-type gene pairs recur in many genera without an RT and element-bearing K-family RTs occur in further genera, but the combination of cassette parts with the K-lineage RT is found in one genus within the scanned range

Placement: main text (open question 3). Figure: Extended Data Figure 9 panel B.

Observation and level of support

- Same-strand -1/+3-type gene pairs without a detected RT occur in 54 genera from 24 families of the confirmatory set (the +3-type gene anchors 348 loci in 149 genera, and 141 of these loci carry such a pair), at a ratio of 132.4 to a positional null (interval 94.4 to 193.4), and across all 155 genera with a +3-type anchor (374 anchor loci), 151 of 157 observed pair loci have no RT of any kind. [level: supported; may depend on a model built from the cassette]
- After the lead genus and the cassette-lineage genera are excluded, 50 genera from 24 families remain at a ratio of 145.0, an analysis that was not in the pre-specified sensitivity list. [level: suggestive, post hoc; may depend on a model built

from the cassette]

- These pair genera are a non-random subset of annotated public Bacteroidota genomes and not Bacteroidota as a whole, the count concerns the repetition of RT-free pairs and not their function or origin, vertical descent and repeated acquisition cannot be told apart without a tree, and the family was defined from cassette sequences so that circularity is not excluded. [limit]
- The number of pair classes is not a number of origins, genera are not independent lineages, and the single-genus limit of the cassette is unchanged by this recurrence. [limit]
- The recurrence of RT-free pairs outside the training genera meets its pre-specified rule, although the null is not calibrated and independence is qualified, and the free pairs and the cassette –1 and +3 genes are distant paralogs with low identity to each other, the direction of any exchange being unresolved. [level: suggestive, post hoc; may depend on a model built from the cassette]
- A vWA-like gene is also found beside a Macro-domain gene outside Bacteroidota, in 1145 of 2275 eligible genera at a ratio of 85.8 (the neighbouring genera are mostly Bacillota), with 29.2% of Macro-domain genes having an adjacent –1-type gene against 0.36% an adjacent generic vWA gene, and in Bacteroidota the corresponding count is 5 of 253 genera, the –1-type gene having more than one partner family. [level: suggestive, post hoc; may depend on a model built from the cassette]
- Outside Pollutibacter, 95 K-family RT loci carry an element at the RT start, in 6 labelled genera and 11 loci without a genus label, and of 1130 observable –1, +1, +2 and +3 slots beside the outside K-family RT loci none holds a cassette part, so the element-plus-RT layer and the cassette-part layer must be counted separately. [level: negative result; may depend on a model built from the cassette]
- The outside RTs are at most 0.33 identical to the cassette RT, so this finding concerns RT lineages far from the cassette lineage, this detector and this range, and the cassette remains a single genus. [level: negative result; may depend on a model built from the cassette]
- In a simulation the part detectors reach a high detection rate only at high identities to the cassette member (the identity needed differs by part), whereas most outside RTs are far from the cassette RT in identity, so that absence outside the genus is uninformative. [level: no information; may depend on a model built from the cassette]
- This reading depends on a simulated evolution model, rests on detectors built from the cassette (circular), has a single genus as its independent unit and has no model-independent verification, and it does not show that the parts are absent outside the genus. [limit]
- Over all Bacteroidota species representatives the strict detectors find same-strand pairs at a short gene-number distance, 23 of them in 14 genera outside the cassette-lineage genera, and the pairs that have an RVT_1 domain between the two genes lie in Pollutibacter at distance 4. [level: descriptive only; may depend on a model built from the cassette]
- These counts belong to this detector, which derives from the cassette, and they are not an independent confirmation of the single-genus result. [limit]

Why it is open

- Parts recur across genera and element-bearing RTs occur in further genera, yet the combination of cassette parts with the K-lineage RT is found in one genus, and whether this reflects sampling and detector reach or a joining event that happened once cannot be separated by the present data.

Hypotheses and the observations that would refute them

- **H1, sampling and detector reach.** H1: the single-genus result is a limit of sampling and detector reach, so that a pre-specified, sensitivity-calibrated scan beyond GTDB representatives finds a –1-type gene next to a K-lineage RT, or an RT-bearing segment between –1- and +3-type genes, in a further genus.
 - Distinguishing prediction: A calibrated scan finds a further locus.
 - Refutation: H1 is refuted if a calibrated scan with adequate reach over enough genomes finds no such locus, and a scan without calibrated sensitivity is uninformative.
 - Current standing: Not tested; a simulation estimates a low detector reach at the identities of the outside RTs, an estimate that depends on a model built from the cassette.
 - Kind of test: computational
- **H2, true confinement.** H2: the combination arose once and stays confined to the Pollutibacter lineage, through the narrow reach of its RT clade or through the joining event, although the parts coexist at genus level elsewhere, so that a calibrated scan finds no further locus.

- Distinguishing prediction: A calibrated scan finds no further locus.
- Refutation: H2 is refuted if a locus whose host genus is confirmed by *rpoB* and *gyrB* is found outside the genus.
- Current standing: Not tested.
- Kind of test: computational
- **H3, pair first (one reading, untested).** H3: the -1 - and $+3$ -type genes were an adjacent pair before the RT-bearing segment lay between them, and the converse, free pairs as remnants of segment-bearing blocks, is equally open.
 - Distinguishing prediction: Outgroup-rooted -1 and $+3$ trees give a polarity of the pair.
 - Refutation: H3 would not be supported if outgroup-rooted trees place the cassette genes in a discordant lineage, if free pairs outside *Pollutibacter* do not lie at the cassette position or if the junction carries no module remnant, and the opposite outcomes would weaken the reverse direction.
 - Current standing: Untestable at present: outgroup-rooted trees are saturated; H3 concerns polarity and is not on the same axis as H1 and H2.
 - Kind of test: computational (untestable at present)

Cheapest test

- The cheapest computational steps are to rebuild the pair model from seeds independent of the cassette and repeat the genome-wide scan, and to check whether free pairs of genera outside *Pollutibacter* lie at the cassette position between *xseA* and *recD*.
- The decisive step is a sensitivity-calibrated scan of catalogues not yet scanned, which would bring new input and is a follow-up that remains open.

Status of the test

- The -1 tree is untestable at a median outgroup-ingroup distance of 6.53 substitutions per site, and the result depends on how the outgroups were chosen. [level: no information]
- The $+3$ tree is untestable at 6.30 substitutions per site, and the AU test of its closest alternative root lies just above the significance threshold, so the polarity of the pair is untested. [level: no information]

Figure idea

- The figure idea is an existence map with three layers of colour on a Bacteroidota genus tree, pair genera, element-bearing K-family RT genera and the cassette genus, with overlaps joined by arcs; it is panel B of Extended Data Figure 9.

Limits that travel with this card

- The cassette observations lie within one genus in the range scanned and the detectors derive from the cassette.

P05. A $+3$ protein whose N-terminal domain is predicted metzincin-like lacks the canonical HExxH motif in every ORF examined, while the Peptidase_M54-like gene beside the cassette keeps it, and there is no evidence that the N region is under weaker purifying selection than the rest

Placement: main text (open question 4). Figure: not drawn in Extended Data Figure 9, set out in this note.

Observation and level of support

- The HExxH-type pattern is found in 0 of 13 $+3$ ORFs (11 complete ORFs, 12 ORFs with a valid test of the N part, 12 ANI95 units), whereas the same pattern is found in 7 of 7 Peptidase_M54-like ORFs beside the cassette and in human MMP1. [level: descriptive only; may depend on a model built from the cassette]
- This is a sequence check of the canonical catalytic motif of the $+3$ N protein that does not establish that the motif was lost, that the protein is inactive or any change in activity, and no wet-laboratory result is reported. [limit]
- The absence is not confined to the cassette: the strict HE..H pattern is found in only 2 of 374 free $+3$ -type sequences, whereas 1565 of 1626 Pfam M54 sequences carry HExxH, and these two counts rest on different definitions. [level: descriptive only; may depend on a model built from the cassette]
- The 13 ORFs fall in 12 ANI95 units of one genus and are not 13 independent observations, and 13 independent absences would not be unlikely by chance. [level: descriptive only; may depend on a model built from the cassette]
- The zinc-compatible part of this analysis rests on a post hoc definition and a null that depends on selection, and it is not carried here. [limit]
- The omega of the $+3$ gene is 0.1125, far from the neutral value, and negatively selected codons predominate in its alignment, a direction of constraint of intermediate strength. [level: descriptive only; may depend on a model built from

the cassette]

- In predicted structures the N region resembles an archaemetzincin-type fold at an intermediate fold-level confidence, a catalytic-loss variant of M54 is not supported by profile searches, and a zinc-compatible geometry may be produced by the sequence prior of the predictor. [level: suggestive, post hoc]
- The C region occurs as a stand-alone family, which may include N-terminally truncated fragments, and as a fusion form extended on the N side that accounts for most public genome hits, and the C family is restricted to Bacteroidota by taxonomy but not by environment (post hoc observations). [level: suggestive, post hoc]
- In the +3 alignment of 11 sequences the fraction of codons that FEL calls negatively selected (P at most 0.1) is 53.5% in the N region (the first 200 columns; 107 of 200) and 46.1% in the rest (71 of 154), with no positively selected codon in either. [level: descriptive only; may depend on a model built from the cassette]
- The test, the interval and the power calculation treat the 354 codons as independent trials, which neighbouring codons need not be, and a circular-shift permutation that keeps the neighbour structure gives $P = 0.47$ with the same negative result, so the smallest difference that this design could detect is at least as large as computed under independence. [limit]
- The difference is 7.4 percentage points with a 95% interval from -3.1 to $+17.6$ and a two-sided Fisher P of 0.1982, so there is no evidence that the N region is under weaker purifying selection than the rest, which is not evidence that the regions are equal. [level: descriptive only; may depend on a model built from the cassette]
- The design has low power for small differences between the N and the C fraction, so the result is not evidence that the two regions are equally constrained. [level: descriptive only; may depend on a model built from the cassette]
- Across an enumerated variant grid, assembled after the primary result was seen and exploratory, the N fraction is at least the C fraction in all but one cell (the partitioned FEL at the strictest threshold, a difference of -0.8 points, $P = 0.90$) and no cell gives significant uncorrected one-sided evidence of a weaker N region. [level: suggestive, post hoc; may depend on a model built from the cassette]
- Two partitioned cell boundaries, at columns 225 and 250, give N minus C differences of $+11.8$ and $+11.6$ points with N higher ($P = 0.036$ and 0.048), also exploratory and uncorrected. [level: suggestive, post hoc; may depend on a model built from the cassette]
- FEL fixes the synonymous rate alpha at 1.000 at 349 of 354 sites, so the contrast rests on an unresolved site-wise synonymous rate, and the calls are departures of the nonsynonymous rate under a fixed synonymous rate. [level: no information; may depend on a model built from the cassette]
- No HExxH-type match occurs in either region of the 11 aligned proteins, so the N/C contrast is not a contrast between the presence and absence of that motif, and the alignment belongs to a single genus. [level: descriptive only]

Why it is open

- A fold that predicts a protease lacks the canonical catalytic signature in every cassette ORF and in nearly all free +3-type sequences (2 of 374 carry it), and there is no evidence that its N region is more relaxed than the rest, so it is open what the N domain is kept for.

Hypotheses and the observations that would refute them

- **H1, catalysis-free scaffold.** H1: the HExxH-lacking N-terminal domain of the +3 protein is kept as a catalysis-free scaffold, for example for contact with the -1 protein, so that the purified domain shows no zinc-dependent proteolysis where an active control of the same fold is active and its removal abolishes -1 binding.
 - Distinguishing prediction: No zinc-dependent proteolysis by the purified N domain; loss of -1 binding without it.
 - Refutation: H1 is refuted if the purified N domain shows proteolysis that EDTA or o-phenanthroline inhibits, or if its removal leaves -1 binding unchanged.
 - Current standing: Not tested; binding between the -1 and +3 proteins has not been tested and a complex prediction carries no information.
 - Kind of test: wet-laboratory test (under way)
- **H2, non-canonical metal site (post hoc).** H2: the N domain keeps a non-HExxH metal site and acts on a substrate other than peptide bonds, as suggested by a His and Cys cluster in some predicted structures, a reading that is post hoc, depends on the predictor and carries no information.
 - Distinguishing prediction: Substituting the His and Cys residues abolishes an activity of the purified domain.
 - Refutation: H2 is refuted if the purified domain has no metal-dependent activity and substituting the His and Cys residues leaves binding and phenotype unchanged.

- Current standing: Not tested; post hoc and predictor-dependent.
- Kind of test: wet-laboratory test (under way)
- **H3, constraint in the C domain.** H3: constraint on the +3 protein is carried by its C-terminal domain and the N domain is a passive fusion partner, so that the N region carries a clearly lower fraction of negatively selected codons than the C region.
 - Distinguishing prediction: A clearly lower N fraction of negatively selected codons at adequate power.
 - Refutation: H3 is refuted if a repeat of the N/C comparison on a larger alignment of free +3-type sequences, with adequate power, shows an N fraction equal to or higher than the C fraction.
 - Current standing: The FEL fractions do not support H3, but the design has low power for small differences and cannot separate selection strength from rate heterogeneity.
 - Kind of test: computational (run once; low power)

Cheapest test

- The cheapest next computation is to repeat the N/C decomposition and its boundary sensitivity on the free +3-type family, which raises power, with the HExxH-bearing M54 flank ORFs as a positive control, and the FEL false-positive calibration and the stratification by information content, not done so far, would then mark the analysis as post hoc.
- The decisive test is a purified N-domain activity assay with an active control and a –1 binding test; it belongs to the wet-laboratory work under way, and no result is reported here.

Status of the test

- The N/C decomposition was run once, on 11 aligned sequences of one genus, as a post hoc analysis that has not been repeated on other sequences.

Figure idea

- The figure idea shows the 354 alignment columns with the FEL negative sites, the N/C boundary at column 200 and the HExxH reference columns, above bars that contrast the free +3 HE..H count with the M54 count; Extended Data Figure 9 does not draw this question.

Limits that travel with this card

- The ORFs are related sequences of one genus, FEL negative calls are not evidence about positive selection, and the family was defined from cassette sequences so that circularity cannot be excluded.

P06. A palindromic motif with a CAAGCTTG core found upstream of free –1 genes in training genera is reproduced in held-out genera by the pre-specified rule, and its frequency varies with the +3 context

Placement: main text (open question 5). Figure: Extended Data Figure 9 panel C.

Observation and level of support

- A palindromic motif with the core CAAGCTTG lies upstream of the –1 gene of free –1/+3 pairs as a strict site in 78 of 103 genera (51 of 66 held-out genera against 5.8% of null windows, and 27 of 37 training genera), ending a median of 61 nucleotides before the start codon (10th percentile 49.7, 90th percentile 113.4). [level: supported; may depend on a model built from the cassette]
- The exact CAAGCTTG core occurs in 47 of 103 genera against 16 of 4049 null windows, the 14-bp extended palindrome in 16 of 103 against 0 of 4049, and 25 genera have no strict site, so the motif is not present in every genus. [level: supported; may depend on a model built from the cassette]
- The support is that of a rule fixed before the analysis being met, the genera are phylogenetically related and are not independent observations (an order-level test would be more honest because many held-out genera belong to one order), and co-inheritance of a mobile locus is not excluded. [limit]
- Binding of the +3 protein to the palindrome was never tested and the motif is not read as regulating anything, and its position in non-coding sequence is an exploratory analysis. [limit]
- At the cassette, the exact core occurs at 11 of 14 upstream windows of the –1 gene, at 55 to 67 nucleotides from its start (post hoc), and at 0 of the 16 cassette element windows. [level: suggestive, post hoc; may depend on a model built from the cassette]
- The fraction of –1 upstream windows with a strict site is 0.877 beside the +3 gene, 0.576 where a +3 gene lies a few genes away (post hoc), 0.333 where the +3 gene comes first, 0.286 elsewhere in the genome and 0.049 where it is absent,

in classes of 46, 23, 12, 35 and 168 genera, and the comparison is descriptive. [level: suggestive, post hoc; may depend on a model built from the cassette]

- This contrast mixes the presence of the +3 gene with the lineage of the –1 gene, because –1 genes without a +3 gene belong to a different –1 lineage, so the gradient does not show that the +3 protein binds the motif. [limit]
- The site covaries with the sequence of the N domain of the +3 protein, used as a negative control, more strongly ($r = 0.296$; $P = 0.0014$) than with that of its C domain ($r = 0.194$; $P = 0.0239$). [level: no information; may depend on a model built from the cassette]
- A WYL-family neighbour lies near the –1 gene in 17 of 258 units, so the cis test of a host regulator has little power and a regulator acting in trans is untested. [level: no information; may depend on a model built from the cassette]
- The C-terminal domain of free +3 proteins is predicted winged-helix-like (a predictor-dependent reading that is weaker for the cassette +3), but the pre-specified structure and sequence arms did not support the resemblance and basic-residue enrichment is not specific to contact positions, so the prediction is not evidence of DNA binding. [level: suggestive, post hoc; may depend on a model built from the cassette]
- WYL-domain ORFs lie beside the cassette more often than in random windows of the same genomes, but their orientation and distance do not match known regulator arrangements. [level: suggestive, post hoc; may depend on a model built from the cassette]

Why it is open

- A palindromic site recurs upstream of free –1 genes at a similar distance and its frequency follows the +3 context, but whether the +3 protein, a host regulator or the mobility of the pair locus explains it cannot be told from the present data.

Hypotheses and the observations that would refute them

- **H1, candidate recognition site of the +3 protein (untested).** H1: the +3 protein, whose C-terminal domain is predicted winged-helix-like (a predictor-dependent reading), recognises the palindrome, so that purified +3 protein binds the core-bearing oligonucleotide and not a core-scrambled one, and +3-free members of the same –1 lineage lose the site faster than neutral spacing.
 - Distinguishing prediction: Core-dependent binding of the +3 protein; loss of the site faster than neutral spacing in +3-free members of the same –1 lineage.
 - Refutation: H1 is refuted by an EMSA with a validated positive control that shows no core-dependent binding, and it is weakened if +3-free members of the same –1 lineage keep the site.
 - Current standing: Untested; the pre-specified structure and sequence arms did not support the winged-helix resemblance.
 - Kind of test: wet-laboratory test (under way); computational
- **H2, host regulator in trans.** H2: a host regulator, for example a WYL-domain protein, reads the site, so that the site persists in +3-free members of the same –1 lineage and recurs at other loci regulated by the same factor.
 - Distinguishing prediction: The site persists without the +3 gene within one –1 lineage and recurs beside a common neighbouring factor.
 - Refutation: H2 is refuted if the genome-wide copies of the 14-bp site (22 outside the focal window against 7.4 expected from composition) share no neighbouring factor, or if members without the +3 gene lose the site within one –1 lineage, which the present contrast, mixing +3 presence with –1 lineage, cannot decide.
 - Current standing: Untested; the cis test is nearly uninformative.
 - Kind of test: computational
- **H3, mobility marker.** H3: the palindrome marks the mobile pair locus, as an integration or recombination site, and tracks +3-bearing lineages because the pair travels as a unit, so that copies lie at pair boundaries beside integrase-type genes.
 - Distinguishing prediction: Copies of the site at pair boundaries beside integrase-type genes.
 - Refutation: H3 is refuted if the genome-wide copies outside the focal window show no neighbouring integrase or recombination function and the site does not co-travel with the pair locus; the similar distance of the site from the –1 start does not separate H3 from H1 and H2.
 - Current standing: Not tested; co-inheritance of a mobile locus is not excluded.
 - Kind of test: computational

Cheapest test

- The cheapest computations are, first, to enlarge the catalogue so that more members of the same –1 lineage with and without a +3 neighbour can be compared, which the present contrast could not do because few close –1 members lack a +3 neighbour; second, to find the common neighbouring families of the genome-wide 14-bp copies; and third, to look for copies of the site outside pair boundaries.
- The decisive test is an EMSA of purified +3 protein, its C and N domains, with wild-type and core-scrambled oligonucleotides and a validated positive control; it belongs to the wet-laboratory work under way, and no result is reported here.

Status of the test

- None of these computations has been run beyond the analyses described above.

Figure idea

- The figure idea combines sequence logos of the site by genus, the distribution of the distance to the –1 start, the bars of the +3 context gradient and the cells for the cassette –1 genes and the element windows.
- Panel C of Extended Data Figure 9 draws a logo of the site and the distribution of the distance to the –1 start, and the bars of the +3 context gradient and the cells for the cassette –1 genes and the element windows are not drawn in Extended Data Figure 9.

Limits that travel with this card

- The motif analysis concerns one protein family, the genera are related, and the cassette observations lie within one genus within the range scanned.

P09. The 3' end of the element lies in the RT start codon or directly before it in 7 of 16 source windows and within 5 nucleotides of it in 9 of them, both by post hoc definitions, and what kind of element it is remains open

Placement: main text (open question 6). Figure: not drawn in Extended Data Figure 9, set out in this note.

Observation and level of support

- Relative to the RT start annotated in each of 16 source windows, the 3' end of the element lies in the start codon or directly before it in 7 windows and within 5 nucleotides of it in 9 windows, both windows being post hoc and the narrower one chosen to reproduce a count of 7 windows, so that it has no definition independent of that count. [level: suggestive, post hoc; may depend on a model built from the cassette]
- The RT call extends 95 nucleotides upstream of the element end in 3 of the 16 windows. [level: descriptive only; may depend on a model built from the cassette]
- The element models derive from the cassette and may be circular, so the abutment partly reflects how the models were made, and the coordinate calculation is the same observation and not a separate one. [limit]
- Among loci with an element hit within a short window of the re-called RT start, the hit ends within a few nucleotides of the RT start at 15 seed windows, 21 other *Pollutibacter* loci and 95 outside K-family RT loci, and the non-coding space upstream of the RT cannot hold an ncRNA of the size of a DRT2 ncRNA. [level: descriptive only; may depend on a model built from the cassette]
- Extending the windows of cassettes in known bins confirms in some cases an element 3' end that overlaps the first RT start codon, but these are sister copies of known cassette bins whose elements are identical to a known element of the same bin, so they are not independent evidence, and the other windows failed for lack of reads, which is not evidence of absence. [level: suggestive, post hoc; may depend on a model built from the cassette]
- Sequence constraint of the element across genomes of the same genus is circular, because the anchor interval continues from the element 3' end and so coincides with the element-model boundary, but the contrast between its 5' part and the part near the RT start suggests that the constraint is not explained by the coding overlap with the RT start alone. [level: suggestive, post hoc; may depend on a model built from the cassette]
- R-scape finds 0 significant covariation pairs with a passing positive control, which is not informative about a long stem at this power, msd-, DRT2- and DRT9-type templates are not supported by this operationalisation, and the comparison with DRT10 ncRNA covariance models failed its positive controls, so a structured RNA is neither supported nor excluded. [level: no information; may depend on a model built from the cassette]
- An exploratory clue is that element loci without an RT exist, which an intron-like derivative from which the RT was deleted could explain. [level: suggestive, post hoc; may depend on a model built from the cassette]

- A sigma-A-type promoter-like pair of boxes, TTTG before the element start and TAHHTTTG near its 5' end, is shared by many cassettes (post hoc; the putative transcription start lies inside the element), but it is not unusual against gene-upstream base rates and was not reproduced at homologous positions in held-out sequences. [level: suggestive, post hoc; may depend on a model built from the cassette]
- An element-first search of 794 genomes finds 19 hits, 18 with the RT ahead, 1 with the RT side cut by a contig end and 0 element-only, after a post hoc rule that treats the cut case as unobservable (1 element-only case before the rule), within a population restricted to element-bearing genera plus *Pollutibacter* and with an element model derived from the cassette. [level: suggestive, post hoc; may depend on a model built from the cassette]

Why it is open

- The end of a non-coding stretch that looks constrained lies at the RT start codon in some windows, and the present tests neither support nor exclude a structured RNA, a retron-type template or a DRT10-type ncRNA for it, so a separate RNA, the 5' region of the RT transcript and a remnant of an intron-like ancestor all remain open.

Hypotheses and the observations that would refute them

- **H1, 5' region of the RT transcript.** H1: the element is the 5' untranslated region, or an N-terminal coding extension, of the RT transcript with no separate RNA product, so that 5' RACE places the transcription start at or inside the 5' end of the element and no stable element-length RNA accumulates apart from the RT mRNA.
 - Distinguishing prediction: A transcription start at or within the 5' end of the element, and no stable element-length RNA apart from the RT mRNA.
 - Refutation: H1 is refuted if a stable element-length RNA with its own 5' and 3' ends accumulates apart from the RT mRNA.
 - Current standing: Not tested; the promoter-like pair predicts a start inside the element (post hoc).
 - Kind of test: wet-laboratory test (under way)
- **H2, separate template or scaffold RNA.** H2: the element is a structured RNA used by the RT as template or scaffold, so that cassettes with an active RT, but not the YAAD variant, yield element-derived cDNA in a validated expression system.
 - Distinguishing prediction: Element-derived cDNA from the active RT and not from the YAAD variant.
 - Refutation: H2 is refuted if the active RT does not make more element-derived cDNA than the YAAD variant in a validated system.
 - Current standing: Not tested; structure-based tests were uninformative.
 - Kind of test: wet-laboratory test (under way)
- **H3, intron-like vestige.** H3: the element is a degenerate remnant of an intron-like ancestor whose constraint is a by-product of its overlap with the RT start, so that outside the overlap element loci without an RT are no more constrained than neutral spacing.
 - Distinguishing prediction: Element loci without an RT are no more constrained than neutral spacing once the overlap is masked.
 - Refutation: H3 is refuted if element loci without an RT stay clearly more constrained than neutral spacing once the overlap is masked, although the constraint analysis is circular through its anchor.
 - Current standing: Cannot be tested at present because the element-first search found no element-only locus after the post hoc rule.
 - Kind of test: computational (not testable at present)

Cheapest test

- The cheapest computation is to compare the constraint of RT-less element loci with that of elements beside an RT with the overlap masked, by the same method.
- The decisive tests are 5' RACE and an RNA size fractionation; they belong to the wet-laboratory work under way, and no result is reported here.

Status of the test

- Neither computation has been run.

Figure idea

- The figure idea sets the coordinates of the element 3' end against the RT start codon in the source windows beside a constraint percentile profile of the element region across genomes of the genus, in one column; Extended Data Figure 9 does not draw this question.

Limits that travel with this card

- The element models derive from the cassette and the constraint analysis is anchored on it, so the observations of this card may be circular, and they lie within one genus within the range scanned.

P07. The island lies at the same address between xseA and recD in related genomes, but how often that address is occupied by a large insert cannot be told from the present scan

Placement: supplementary card. Figure: none (idea only).

Observation and level of support

- In related genomes the cassette locus lies at the same address, between xseA/xseB and recD, recC and recB, and island-free references carry the same arrangement without a gap, but the genomes are not independent, the observation concerns one genus and the polarity is unknown. [level: suggestive, post hoc; may depend on a model built from the cassette]
- Among 3 non-carrier *Pollutibacter* species in which the address can be judged, 1 has an empty address and 2 are rearranged, which gives a 95% Wilson interval of 0.061 to 0.792 for the empty fraction, and rearranged means only that xseA and recD are not adjacent, usually on different contigs, and is not active evidence of an occupied address. [level: negative result; may depend on a model built from the cassette]
- In a scan of GTDB representatives, of which 69.7% were processed (139,400 of 199,923), a strict class of large gap between xseA and recD is found in 3 of 106,040 xseA-bearing genomes: the cassette island genome, one genome of the CAZUSO01 lineage whose insert carries outer-membrane- and porin-like ORFs, and one genome in which the gap holds RecB-like, RecC-domain and signalling genes, which a post hoc reading of the scan output treats as a RecBCD arrangement and not as an insertion. [level: no information; may depend on a model built from the cassette]
- The strict class is a post hoc definition whose sensitivity was not measured and the scan covered only part of the representatives, so the count is not an occupancy frequency, and the insert of the CAZUSO01 genome is not a cargo family shared with the island. [limit]
- Tests of a hotspot are closed: too few evaluable junctions reached the pre-declared minimum (no information), enrichment of large gaps upstream of recD outside the lineage failed its positive control (no information), the enrichment upstream of recD within the cassette genus is post hoc and partly circular, and junction repeats and target-site duplications failed the boundary positive control (no information). [level: no information; may depend on a model built from the cassette]

Why it is open

- The same island sits at the same address, yet the address is empty or rearranged in the non-carrier species that can be judged and a large insert is found in very few genomes, so it is open whether the address is selected or the island is one old insertion kept by descent, because an occupancy frequency cannot be measured with the present scan.

Hypotheses and the observations that would refute them

- **H1, DNA-end processing neighbourhood as a preferred landing place.** H1: the xseA-recD interval is a preferred landing place because it lies among DNA-end processing genes (xseA/xseB, recBCD, sbcCD), so that large inserts occur there more often than at other positions and equivalent end-processing clusters elsewhere (recJ, sbcCD, recFOR) are occupied at comparable rates.
 - Distinguishing prediction: Occupancy by large inserts is raised at xseA-recD and at equivalent end-processing clusters, above a same-genome null.
 - Refutation: H1 is refuted if the occupancy of recJ, sbcCD and recFOR neighbourhoods by large inserts is clearly lower than at xseA-recD or does not exceed a same-genome null.
 - Current standing: Not tested; the scan covers only part of the representatives and the strict class is post hoc.
 - Kind of test: computational
- **H2, recognised integration site.** H2: a short sequence at the address is recognised by an integration or recombination function of the island or the host, so that occupied junctions share a boundary repeat or target-site duplication and

empty references carry one copy.

- Distinguishing prediction: Occupied junctions share a boundary repeat or target-site duplication beyond a shuffled control.
- Refutation: H2 is refuted if the boundary repeats or target-site duplications of the occupied junctions are not shared beyond a shuffled control and the cargo carries no integrase, on the premise that the boundary positive control passes, which it did not in the present analysis.
- Current standing: Not testable now: the boundary positive control failed (no information).
- Kind of test: computational (not testable now)
- **H3, one old insertion.** H3: the island entered once, long ago, and persists within one subclade, so that the other large inserts are unrelated events and the shared address is incidental.
 - Distinguishing prediction: Island-carrying genomes lie within one subclade with a shared junction, and other cargo families do not repeat the address.
 - Refutation: H3 is refuted if island-carrying genomes occur outside the subclade with independent junctions or if the same cargo family repeatedly occupies this address.
 - Current standing: Not tested.
 - Kind of test: computational

Cheapest test

- The cheapest computation is to process the 30.3% of the representatives that the scan has not yet covered and to apply the same strict criterion to the recJ, sbcCD and recFOR neighbourhoods, comparing occupancy with a same-genome null.
- A search for junction repeats gives information only after the boundary positive control, the point at which the present analysis failed, has been passed.

Status of the test

- None of these computations has been run.

Figure idea

- The figure idea is a mosaic gene map of the address in which empty, rearranged, island-carrying, CAZUSO01 insertion and RecBCD-arrangement genomes are aligned in one column, with a box for the count of 3 of 106,040.

Limits that travel with this card

- The same address is compatible with vertical descent and with site-specific insertion, the strict class is a post hoc definition, and the observation concerns one genus within the range scanned.

P10. RecBCD genes accompany the RT in the genomes of one family, Chitinophagaceae, and are not detected in the carrier genomes of the other families with several carriers

Placement: supplementary card. Figure: none (idea only).

Observation and level of support

- Among 27 K main-clade units that carry the RT, 13 have RecBCD (Wilson interval 0.31 to 0.66) against 52 of 943 background units, and in an exact test stratified by genus, in 7 informative genera with 14 carriers, 13 are RecBCD-positive against 2.09 expected ($P = 1.5e-12$), which is companionship and not a requirement, rests on few confirmatory units and has an optimistic P value. [level: suggestive, post hoc; may depend on a model built from the cassette]
- The statement that RecBCD companionship is a property of the whole K main clade is partly rejected, because a substantial share of the high-quality carrier units lack RecBCD, while the narrow cassette clade is not refuted. [level: negative result; may depend on a model built from the cassette]
- By family, the fractions of high-quality carrier genomes with RecBCD are 13 of 13 in Chitinophagaceae, 0 of 6 in Saprospiraceae, 0 of 6 in Spirosomataceae and 0 of 5 in CAIMTB01, and 3 of the 7 families represented by a single genome also carry it. [level: descriptive only; may depend on a model built from the cassette]
- These are genome counts and differ from the unit counts above, and the family fractions rest on few genomes. [limit]
- In the genus BJGN01, a genus of Saprospiraceae that carries the element plus RT unit, no RecC or RecD call is made, and the RecB and RecD callers are not calibrated. [level: descriptive only; may depend on a model built from the cassette]

- With a definition of K_E units (element plus K-family RT) that does not depend on recD, the K_E units of Chitinophagaceae are RecC-positive in 8 of 8 against 51 of 437 K_0 units, whereas in BJGN01 the figures are 0 of 4 and 0 of 5, and the comparison is uninformative because too few genera are informative. [level: no information; may depend on a model built from the cassette]
- In Pollutibacter the RecD-like gene downstream of the RT cannot be told from the host's own RecD, since each of the 8 genomes with a full-length gene carries exactly one RecD-like protein, and it is not counted as a cassette module. [level: suggestive, post hoc; may depend on a model built from the cassette]
- Of 16 island RecD-like sequences from 3 genera, 15 are evaluable and all keep their Walker motifs and ATP-contact columns, which shows only that these are ordinary RecD-like proteins and does not show that any of them acts as an effector. [level: descriptive only]
- In the Ferruginibacter island, additional RecD-, RecC- and RecB-like genes form a full-length set with intact motifs, but this is one clone and one independent unit, the association of RecD with the island is supported in one independent unit only, and complete non-carrier genomes of the same species keep RecBCD, which favours a lineage-background reading. [level: suggestive, post hoc; may depend on a model built from the cassette]
- The literature reports that DRT3 is toxic in cells lacking RecBCD (Wang et al. 2026), which is context and not an observation on the cassette. [literature context]
- The pre-specified general test of a RecBCD-axis logic failed, and the caller for alternative end-processing enzymes is not calibrated, so the alternatives are uninformative. [level: no information; may depend on a model built from the cassette]

Why it is open

- RecBCD companionship is complete in the genomes of one family, not detected in several other families, among them Saprospiraceae with the genus BJGN01 that carries the element plus RT, and present in a few single-genome families, so whether the RT unit needs, tolerates or merely lands beside RecBCD genes is not resolved and an association is not evidence of need.

Hypotheses and the observations that would refute them

- **H1, RecBCD axis (literature-motivated).** Conjecture H1, motivated by the DRT literature and not by our data: the cassette RT product intersects RecBCD-dependent processing of DNA ends, so that *E. coli* lacking recB, recC or recD, or expressing Gam or gp5.9, shows a growth defect absent in the wild type and in an sbcB deletion and abolished by the YAAD variant.
 - Distinguishing prediction: Growth defect with RecBCD absent or inhibited, absent in the wild type and in an sbcB deletion, abolished by the YAAD variant.
 - Refutation: H1 is refuted if validated expression shows no such defect, and an effect only in the sbcB deletion points to the ExoI axis (see P02).
 - Current standing: Not supported or tested by data of this work; the pre-specified general test of this logic failed.
 - Kind of test: wet-laboratory test (under way)
- **H2, site availability and lineage background.** H2: the companionship reflects insertion-site availability next to recD and the RecBCD background of Chitinophagaceae and not a requirement of the RT, so that K_E units are not enriched for RecC over same-genus K_0 units in informative genera outside Chitinophagaceae.
 - Distinguishing prediction: No enrichment of RecC in K_E units over same-genus K_0 units outside Chitinophagaceae.
 - Refutation: H2 is refuted if K_E units, defined without recD, are RecC-positive significantly more often than same-genus K_0 units in a pre-set number of informative genera outside Chitinophagaceae.
 - Current standing: Not tested; only 4 genera are informative so far.
 - Kind of test: computational
- **H3, end-processing set in cis in one island lineage.** H3: in one island lineage (Ferruginibacter; one independent unit) the RT-bearing island carries its own end-processing set of RecB-, RecC- and RecD-like genes, so that the companionship travels with the RT unit there and the island-borne set is full length with intact motifs and not sister to the host copy.
 - Distinguishing prediction: The island-borne set is full length with intact motifs and not sister to the host copy.
 - Refutation: H3 is refuted if the island-borne set is a recent duplicate and sister of the host copy or if its motifs are destroyed.

- Current standing: Supported in one independent unit only; H1 and H3 can both hold in part.
- Kind of test: computational

Cheapest test

- The cheapest computation is to widen the K_E comparison to catalogues not yet scanned and to more genera, so that the number of informative genera rises to the pre-set minimum, with the same caller and pre-specified rules.
- The RecD and RecB calls must be calibrated first.
- The decisive test is a growth test in *Escherichia coli* deletion strains; it belongs to the wet-laboratory work under way, and no result is reported here.

Status of the test

- None of these computations has been run.

Figure idea

- The figure idea is a dot plot of the RecBCD-positive fraction per genus unit, K carriers against background, by family and with BJGN01 marked, beside a box that states the literature context: DRT3 toxic in cells lacking RecBCD and DRT10 toxic in cells lacking ExoI.

Limits that travel with this card

- Insertion-site availability is not separated from the association, the independent units are few genera, and the cassette observations lie within one genus within the range scanned.

P04. The central block of the cassette is almost identical between a Canadian lake sample and a Chilean surface-water sample, while the -1 gene differs

Placement: supplementary card. Figure: none (idea only).

Note: Pairs in this card are sample pairs, the focal sample pair being the Canadian and the Chilean sample, and the -1/+3 gene pair of the other cards is not meant.

Observation and level of support

- The focal sample pair is a Canadian LakePulse sample (run ERR4869792; latitude 49.01 degrees north) and a Chilean surface-water sample, and the RT of the Chilean contig is identical at protein level to the RT of the Chilean MAG cassette, so that it is neither a variant nor an independent unit. [level: descriptive only; may depend on a model built from the cassette]
- Protein identity between the two samples is 1.0000 for +2 (152 residues), 0.9962 for +1, 0.9894 for the RT and 0.9818 for +3 (a partial 220 residues in the Chilean reassembly), against 0.8967 for -1. [level: suggestive, post hoc]
- Of the mismatches of the -1 protein, 10 of 22 fall in a single window of 30 residues (permutation $P = 0.0057$; post hoc). [level: suggestive, post hoc]
- Against 2339 host orthologs of the focal sample, which have a median amino-acid identity of 0.9296, the observed synonymous distance is lower than that of 0.9957 of the length-matched host orthologs for the RT, 0.9991 for +1 and 0.9926 for +2. [level: suggestive, post hoc]
- Counted against all 2262 host orthologs that have a synonymous distance, only 11, 2 and 7 are at or below the value observed for the RT, +1 and +2, and the amino-acid identity of -1 exceeds that of 0.3923 of the length-matched host orthologs. [level: suggestive, post hoc]
- Among 66 sample pairs, 1 (the focal sample pair, selected post hoc) has an RT amino-acid identity percentile of 0.95 or more, the next highest being 0.731, and the other sample pairs follow the divergence of their hosts (Spearman $\rho = 0.929$). [level: suggestive, post hoc]
- The covariation of the distances of -1 and the RT is a cassette-wide effect and not specific to the focal sample pair, and the charge complementarity of -1 and the RT is not more retained than in charge- and identity-matched host genes. [level: negative result; may depend on a model built from the cassette]
- The focal sample pair was selected post hoc, the Chilean MAG and its reassembly share the same reads, the Chilean contig ends without the downstream junction so that it cannot be judged whether the block moved as a unit, and continent is confounded with clade, so this sample pair is not read as one moving unit. [limit]

Why it is open

- Four modules are almost identical between samples from two continents while -1 differs, and only the focal sample pair among 66 related sample pairs stands out in RT identity, so the history of the central block relative to -1 is open.

Hypotheses and the observations that would refute them

- **H1, central block shared more recently than -1.** H1: the identity of RT, +1, +2 and +3 between the Canadian and the Chilean sample reflects a more recent shared history of the central block than of -1, by recent exchange or recent common descent, which the Chilean contig, ending without the downstream junction, cannot tell apart, so that synonymous and intergenic positions inside the block are as identical as coding ones with a sharp identity break at the -1 edge.
 - Distinguishing prediction: Synonymous and non-coding identity inside the block as high as coding identity, with an identity break at the -1 edge.
 - Refutation: H1 is refuted if synonymous and non-coding identity inside the block lies at the host level or if no identity break lies at the -1 edge.
 - Current standing: Not tested; needs a nucleotide-level alignment.
 - Kind of test: computational
- **H2, slow block and fast -1.** H2: RT, +1 and +2 are unusually constrained while -1 evolves under relaxed constraint, so that the Canadian and the Chilean sample carry old alleles and synonymous divergence inside the block lies at host levels.
 - Distinguishing prediction: Synonymous divergence inside the block at host levels.
 - Refutation: H2 is refuted if the observation that the synonymous distance of the RT, +1 and +2 lies at the extreme of host orthologs reproduces in a nucleotide-level alignment.
 - Current standing: Disfavoured by the present synonymous-distance percentiles; the nucleotide-level alignment is pending.
 - Kind of test: computational
- **H3, assembly artefact.** H3: one member carries a chimeric or cross-sample block, so that read-pair or long-read support across the block-to-host junction fails in at least one sample.
 - Distinguishing prediction: Junction support fails in at least one sample.
 - Refutation: H3 is refuted if junction support is confirmed in both samples.
 - Current standing: Not tested.
 - Kind of test: computational

Cheapest test

- The cheapest computation needs little CPU: align the two contigs at nucleotide level and read the identity of coding and non-coding positions in windows fixed in advance at gene boundaries, which avoids the post hoc window centre used above, and count the junction support separately from the read pairs of the two samples; the host ortholog baseline exists.

Status of the test

- This computation has not been run.

Figure idea

- The figure idea is a map of the two sites with bars of the module identities, the median identity of host orthologs as a dotted line, and a scatter of the related sample pairs of cassette RT divergence against host divergence.

Limits that travel with this card

- The focal sample pair is a post hoc selection, and the observation lies within one genus within the range scanned.

P13. The cassette MAGs fall in one sub-clade of the genus, the hypothesis that this is a high-GC clade is not met by the pre-specified rule, and cassette-free relatives within it are barely known

Placement: supplementary card. Figure: none (idea only).

Observation and level of support

- All 16 cassette MAGs are classified as Pollutibacter by GTDB-Tk and fall into one sub-clade. [level: supported; may depend on a model built from the cassette]
- The mean GC of the MAGs is 54.06% against 43.32% for the genus representatives (95th percentile of the null 45.19%), but the pre-specified hypothesis of a high-GC sub-clade is not met as a whole, only its clumping component being supported, and the GC excess is borderline and confounded with a single clade. [level: no information; may depend on a model built from the cassette]
- The 7 carrier genomes on the GTDB tree (6 species; carriers are called by DIAMOND similarity to the cassette RT) form neither a single clone nor a pattern of repeated gain and loss, but cluster in one sub-clade of the genus. [level: no information; may depend on a model built from the cassette]
- The Fitch score is 2 against a permutation null mean of 5.84, with no permutation reaching it, a result that carries no information because the carriers were selected for carrying. [level: no information; may depend on a model built from the cassette]
- Within the sub-clade only 2 non-carrier species are known, 1 of them of high quality, so the pre-specified sibling criterion fails and carries no information. [level: no information; may depend on a model built from the cassette]
- Whether the island was gained once and kept by descent or gained several times cannot be decided, because the sign of the polarity flips when an outgroup is included. [level: no information; may depend on a model built from the cassette]
- The MAGs are a sample selected for carrying and cannot be used to measure the carrier proportion, the null of the clustering test is almost trivial, and seed circularity remains in the carrier call. [level: no information; may depend on a model built from the cassette]

Why it is open

- Within one genus the cassette clumps in one sub-clade and the cassette-free relatives inside that sub-clade are barely known, so it cannot be told whether an ancestor gained the island or whether some members lost it, which repeats the single-genus bottleneck inside the genus.

Hypotheses and the observations that would refute them

- **H1, single acquisition with losses.** H1: the island was acquired once by the ancestor of the sub-clade and persisted by descent with a few losses, so that cassette-free sisters inside the sub-clade lack the island but keep an empty or rearranged xseA-recD address and the island gene trees match the core-genome tree.
 - Distinguishing prediction: Island gene trees match the core tree; cassette-free sisters keep an empty or rearranged address.
 - Refutation: H1 is refuted if the island gene trees disagree with the core tree beyond saturation noise or if a sister inside the sub-clade carries a different cargo at the same address.
 - Current standing: Not tested; the present data cannot separate H1 from H2.
 - Kind of test: computational
- **H2, repeated acquisition into a permissive sub-clade.** H2: the genomic context of the sub-clade, for example high GC and a RecBCD arrangement, permits the island, which entered several times from outside the genus, so that junction sequences and island gene trees differ among carriers beyond host divergence.
 - Distinguishing prediction: Junction sequences and island gene trees differ among carriers beyond host divergence.
 - Refutation: H2 is refuted if the junction sequences are identical among carriers and the island tree matches the species tree.
 - Current standing: Not tested; the junctions of the cassette MAGs can be compared at once.
 - Kind of test: computational
- **H3, detection decay with distance.** H3: the clumping reflects declining detection sensitivity with distance from the seed lineage and not biology, so that detectors built from seeds of a different lineage do not show it.
 - Distinguishing prediction: Detectors built from other seed lineages do not show the clumping.
 - Refutation: H3 is refuted if the same clumping persists with detectors built from other seed lineages.
 - Current standing: Not tested; the carrier call retains seed circularity.
 - Kind of test: computational

Cheapest test

- The cheapest computation compares the junction sequences of the cassette MAGs with one another, which is possible now, and the next step is to obtain more high-quality non-carrier sister species inside the sub-clade by widening the catalogues, an input that is new, and to test the concordance of island gene trees with the core tree by AU tests with an outgroup, where polarity may remain unresolved.

Status of the test

- None of these computations has been run.

Figure idea

- The figure idea is a genus core tree painted with island presence and GC, and the GC distribution of the genus with the mean of the cassette MAGs marked, on one panel.

Limits that travel with this card

- The independent unit is one genus, carriers are defined by similarity to the cassette RT, and no wet-laboratory result is reported.

P11. The +1 protein resembles a DNA-specificity subunit of type I restriction-modification systems only weakly and sits in genomes where the partner subunits are not detected

Placement: supplementary card. Figure: none (idea only).

Note: The +2 protein has no card of its own and is covered by one limitation sentence, the last sentence of the observation list.

Observation and level of support

- At profile level +1 resembles HsdS specificity subunits at a distance: a loose UniRef50 hit, a predicted-structure hit and a relaxed-threshold DefenseFinder profile point to the same family, but each is below its rule and they are not independent. [level: suggestive, post hoc; may depend on a model built from the cassette]
- The resemblance is reproduced only with profiles derived from REBASE, and it is a weak signal that rests on one genus. [level: suggestive, post hoc; may depend on a model built from the cassette]
- No HsdM, HsdR or HsdS is detected in the cassette-bearing genomes, but non-carrier genomes of the same genus also mostly lack them, so a carrier effect cannot be separated from lineage background. [level: no information; may depend on a model built from the cassette]
- The genomes that carry +1-like ORFs have short contigs, so a test of a type I neighbourhood was not possible, and none of HsdM, HsdR and HsdS is found in those genomes. [level: no information]
- Permissive restriction-modification calls in K-positive genus units reject the absence of type I systems, including outside Bacteroidota, and the PF12705 slot meets its rule but is not specific to +1. [level: negative result; may depend on a model built from the cassette]
- Among the typed units of the main -1 cluster of the module references, +2 and +3 each fall in one reference cluster and +1 in several, a post hoc contrast between reference clusters and not between alleles, and +1 is 0.9962 identical between the two samples of the focal sample pair and has an omega of 0.1563. [level: no information; may depend on a model built from the cassette]
- The module references were built from the same MAGs, so the contrast may be circular and the larger +1 reference set may inflate it, and because the allele clusters counted for P01 number 10 to 13 per module, the contrast does not show that +2 and +3 are invariant between alleles. [limit]
- The +2 protein gives no functional clue in sequence, profile, structure or embedding layers (an exploratory extended MSA model gives one macrodomain-like hit), but dark ORFs of matched length are common, so being dark is not unusual, and +2 has an omega of 0.1111 and is 1.0000 identical between the two samples of the focal sample pair. [level: no information]

Why it is open

- The +1 protein looks like a specificity subunit at a distance but sits in genomes in which no partner subunit is detected, so it is open whether it is an adaptor for a nuclease-dependent system or an unrelated protein that resembles one through the profiles.

Hypotheses and the observations that would refute them

- **H1, targeting adaptor.** H1: +1 is an HsdS-type DNA-recognition adaptor that directs an effector to target sequences, so that its sequence variation, which gives several module reference clusters where +2 and +3 give one each (post hoc), concentrates in recognition-like regions on both sides of a centre that varies little, as in domain shuffling of specificity subunits.
 - Distinguishing prediction: Variation among +1 alleles concentrates in recognition-like regions on both sides of a centre that varies little.
 - Refutation: H1 is refuted if variation in the +1 allele alignment does not concentrate in recognition-like regions on both sides of the centre but is uniform or lies elsewhere.
 - Current standing: Not tested; the contrast of module reference clusters on which it rests is post hoc.
 - Kind of test: computational
- **H2, convergent resemblance.** H2: the HsdS resemblance of +1 is a profile artefact of a single protein of the size of a recognition domain matched through REBASE-derived profiles only, and +1 has an unrelated role, so that profiles built without REBASE sequences do not retrieve +1 and +1 lacks a paired recognition-domain architecture.
 - Distinguishing prediction: Profiles built without REBASE sequences do not retrieve +1; no paired recognition-domain architecture.
 - Refutation: H2 is refuted if HsdS profiles built from non-REBASE sources retrieve +1 and a paired recognition-domain architecture is seen.
 - Current standing: Not tested.
 - Kind of test: computational

Cheapest test

- The cheapest computation needs little CPU: compare the distribution of variation along the +1 allele alignment with the positions of the recognition domains in an HsdS reference alignment, and repeat the post hoc contrast of module reference clusters between +1, +2 and +3 by a rule fixed in advance.
- A second step is to build HsdS profiles from non-REBASE sources and repeat the retrieval of +1, and for +2 no further information is expected without a new input.

Status of the test

- Neither computation has been run.

Figure idea

- The figure idea overlays the HsdS reference recognition-domain intervals on the curve of variation along the +1 allele alignment, with bars for the identity and omega of each module on the same panel.

Limits that travel with this card

- The observations concern one genus, being dark is a result of the annotation layers searched, and the HsdS resemblance of +1 is below its pre-specified threshold.

P08. Depth over the island region falls to a fraction of the preceding region in one contig in three samples, but the interpretation is open and the expected island-lacking junction reads are not seen

Placement: supplementary card. Figure: none (idea only).

Observation and level of support

- In one contig of the *Ferruginibacter* block the depth over the island region is 0.392 of the preceding region (95% interval 0.340 to 0.450; change point at 13000 bp; depth 14.846 before and 5.823 in the island), the step is seen in three samples, and the pairing of read pairs across it is normal. [level: suggestive, post hoc]
- The pre-specified competitive mapping ratio of 2.406 exceeds the criterion for no step, so the step is not supported by that rule, and a re-measure in one study alone is post hoc and not independent. [level: suggestive, post hoc]
- The cause is unresolved: occupancy polymorphism, low copy number of the island, a heterogeneous contig, cross-mapping and misassignment of the contig to the bin all remain possible, and the window was chosen after the step was seen. [level: no information]

- The expected number of island-lacking junction reads is 4.51 or 2.33 at alternative fractions of 0.58 and 0.30, against 1 clipped read and 2 of 65 abnormal reads left of the step (expected 7.83 and 4.05), and small fractions cannot be excluded. [level: suggestive, post hoc]
- In two closed *Pollutibacter* MAGs no read pair supports a deletion (0 against 38.0 preserved, and 0 against 19.0), and the depth ratios of island to flank are 1.018 and 1.171, a one-time-point upper bound under assumptions and without a spike-in test. [level: no information; may depend on a model built from the cassette]
- Long-read occupancy is a conditional clue that rests on few samples and is not an independent unit, and closure of the address carries no information. [level: no information; may depend on a model built from the cassette]

Why it is open

- The depth step falls to a fraction of its neighbours and does not appear in the closed MAGs, while the expected island-lacking junction reads are not seen, so whether the step reflects the island being absent from part of the population or a mapping artefact is open.

Hypotheses and the observations that would refute them

- **H1, excisable island and mixed population.** H1: cells with and without the island coexist because the island can excise or be lost, so that island-lacking junction reads appear at the rate predicted by the depth step once mapping bias is removed.
 - Distinguishing prediction: Island-lacking junction reads at the rate predicted by the depth step.
 - Refutation: H1 is refuted if the step persists in k-mer depth although, at a depth where the expected read count is sufficient, no island-lacking junction reads appear, not even a fraction of the number expected.
 - Current standing: Not tested; the observed junction reads fall short of the expectation without excluding small fractions.
 - Kind of test: computational
- **H2, mapping artefact or copy number.** H2: the step comes from multi-copy or mis-assembled front sequence and cross-mapping while the island is single-copy in every cell, so that island-unique and flank-unique k-mers show equal depth.
 - Distinguishing prediction: Island-unique and flank-unique k-mers show equal depth.
 - Refutation: H2 is refuted if the depth ratio of island-unique to host-unique k-mers on both sides of the junction is near the step ratio.
 - Current standing: Not tested.
 - Kind of test: computational

Cheapest test

- The cheapest computation uses the reads of the same samples: remeasure the step by k-mer depth without a reference, comparing island-unique k-mers with the host-unique k-mers on both sides of the junction.
- The expected number of junction-spanning reads is fixed by a rule before counting, which avoids the design flaw of the competitive mapping, and the denominator of island-free reads is widened with the long reads at the recD boundary already held.

Status of the test

- None of these computations has been run.

Figure idea

- The figure idea shows the read-depth profile around the junction in the samples with the step marked, the expected and observed junction reads as bars, and the island-to-flank depth ratios of the closed MAGs on a separate panel.

Limits that travel with this card

- The window was chosen after the step was seen, the samples may share one structural artefact, a block contig must not be read as one uniform genome fragment, and the observation concerns one genus within the range scanned.

P12. Long-read DNA and RNA screens (reduced to a limitation statement)

Placement: demoted to a limitation statement. Figure: none.

Limitation statement

- Long-read DNA screens recover cassette reads (15 reads from 7 samples of 4 studies, in 265 runs and 202.9 Gbp screened; existence only). [level: supported; may depend on a model built from the cassette]
- The long-read RNA screens are uninformative for the cassette lineage because almost no run carries host-lineage signal, and K-family-like reads in runs labelled as RNA cannot be assigned to the cassette lineage or to transcription. [level: no information; may depend on a model built from the cassette]

Why this is not an open question

- This item was reduced to a limitation statement because its negative result carries no information and its interpretation would read as absence or expression, which the data do not support.

P14. The pond time series (reduced to a limitation statement)

Placement: demoted to a limitation statement. Figure: none.

Limitation statement

- The pond time series of 14 time points cannot test host-virus covariation, because the carrier is detected at only 1 time point at the read cap and host and virus abundance share a winter minimum, so the series carries no information about covariation. [level: no information]
- The Beaver Lake monthly series is a single independent unit with a post hoc pattern of presence. [level: suggestive, post hoc]

Why this is not an open question

- This item was reduced to a limitation statement because its observations concern a single unit and carry no information about covariation, so it cannot be offered as an open question.

Summary and what this note does not show

- Taken together, the open questions meet at three joints: where the parts come from (P03, P07 and P13), what the RT is (P01, P02 and P10), and how the cassette might be read (P05, P06, P09 and P11), while P04 and P08 ask what a sample pair and a depth profile say about the history of the island.
- Each hypothesis comes with the observation that would end it, no wet-laboratory result is reported for any of them, and the informative next observations are a further genus that carries the combination, outgroup-rooted –1 and +3 trees (polarity may stay unresolved), a 5' end map of the RT transcript and an activity assay of the +3 N-terminal domain.
- The RecD-like gene next to the RT is not distinguishable from the host RecD and is not counted as a cassette module.
- This note does not show that any hypothesis is true, that any test beyond those marked as run has been done, or anything about the function, expression or mobility of the RT or the cassette.

References

Titles are quoted from the cited works.

- Wang et al. 2026: Wang M, Yoneyama K, Žedaveinytė R, Ishikawa J, Tang S, Le HC, Wiegand T, Ramirez JL, Nagahata N, Ma Y, Zhang DJ, Helmeczi E, Berisa M, Jovanovic M, Hiraizumi M, Yamashita K, Nishimasu H, Sternberg SH. Coordinated RNA- and protein-templated synthesis of double-stranded DNA by a dual reverse transcriptase immune system. *Cell*, 189: 4997-5011.e11, 2026.
- Sánchez-Nieto et al. 2026: Sánchez-Nieto E, Wilkinson ME, Millán V, Wenardy C, Cabrerizo D, Martínez-Abarca F, Toro N. Host exonuclease SbcB and a phage-encoded SSB-like protein control activation of the DRT10 reverse transcriptase defense system. *bioRxiv* (preprint), 2026.

Table SI-19_COLUMN_DICTIONARY.tsv (33 rows x 5 columns)

table	column	definition	values_or_format	blank_rule
SI-19_questions.tsv	question_id	Identifier of the open question (P01 to P14).	P and two digits; P12 and P14 are reduced to limitation statements.	never blank
SI-19_questions.tsv	title	Title of the question as a statement of the observation that is open.	text	never blank
SI-19_questions.tsv	placement	Where the question is carried in this work.	main text (open question 1 to 6); supplementary card; demoted to a limitation statement	never blank
SI-19_questions.tsv	main_text_question	Number under which the main text and the legend of Extended Data Figure 9 call the question: the questions that the main text raises are numbered 1 to 6 in their order of appearance, and the identifiers P01 to P14 are those of this note.	integer 1 to 6 for a question of the main text	not applicable = the question is not raised in the main text (a supplementary card or an item reduced to a limitation statement)
SI-19_questions.tsv	figure	Panel of Extended Data Figure 9 that draws the question, or the statement that no panel does.	Extended Data Figure 9, panel A to C; not drawn in Extended Data Figure 9, set out in this note (a question of the main text that this figure does not draw); none (idea only) when only a figure idea exists; none when the item is reduced to a limitation statement	never blank
SI-19_questions.tsv	n_observations	Number of observation statements in the observation block of the card, limit statements included. The sentences of literature context and the statements of the test-status block are not counted here; they are listed in SI-19_observations.tsv (column where_stated) and the literature sentences carry the level literature context, so the sum over the questions is smaller than the number of rows of that table.	integer	never blank
SI-19_questions.tsv	levels_of_support	Counts of the observation statements counted in n_observations by level of support; the levels are those defined for SI-19_observations.tsv (level_of_support).	each level name followed by its count, the pairs separated by semicolons (for example: descriptive only 3; limit 1)	not applicable = the card holds no observation statement
SI-19_questions.tsv	why_it_is_open	Why the observation is open, in one or two sentences.	text	not applicable = the item is reduced to a limitation statement and is not an open question
SI-19_questions.tsv	n_hypotheses	Number of hypotheses of the card.	integer; 0 for limitation statements	never blank
SI-19_questions.tsv	cheapest_test	The cheapest test that bears on the hypotheses; wet-laboratory tests are under way, and no result is reported here. When a cell holds more than one sentence the sentences are numbered (1), (2) and so on.	text	not applicable = the item is reduced to a limitation statement and no test is proposed
SI-19_questions.tsv	test_status	Whether the test has been run and what the run showed. When a cell holds more than one sentence the sentences are numbered (1), (2) and so on.	text	not applicable = the item is reduced to a limitation statement and has no test
SI-19_questions.tsv	figure_idea	The figure that would show the question.	text	not applicable = the item is reduced to a limitation statement and has no figure idea
SI-19_questions.tsv	limits_carried	Limits that travel with the card.	text	never blank
SI-19_observations.tsv	question_id	Question of the statement.	P and two digits	never blank
SI-19_observations.tsv	observation_no	Running number of the statement within its question, in the order of the card (observation block first, then the test-status block).	integer	never blank
SI-19_observations.tsv	statement	The observation, test-status or literature sentence, as written in the card.	text	never blank
SI-19_observations.tsv	level_of_support	Level of support of the statement; a statement that says of itself that its definition or contrast was chosen after the output was seen is marked suggestive, post hoc. Definitions: supported = a rule fixed before the analysis was met; descriptive only = a count or a table is reported without a test; suggestive, post hoc = the definition or the contrast was chosen after the output was seen; negative result = a pre-specified test came out negative; no information = the test could not decide; limit = a statement that travels with the observations and is not itself an observation; literature context = a sentence that reports the literature and is not an observation on the cassette.	supported; descriptive only; suggestive, post hoc; negative result; no information; limit; literature context	never blank

table	column	definition	values_or_format	blank_rule
SI-19_observations.tsv	may_depend_on_cassette_derived_model	Whether the statement depends on a model or definition built from the cassette itself, so that it may be circular; limit statements and literature sentences are not observations and are marked not applicable.	yes; no; not applicable (limit statement); not applicable (literature context)	never blank
SI-19_observations.tsv	where_stated	Block of the card in which the statement stands; the why-it-is-open and limits blocks are not listed in this table but stand in SI-19_questions.tsv.	observation; test status; limitation statement	never blank
SI-19_hypotheses.tsv	question_id	Question of the hypothesis.	P and two digits	never blank
SI-19_hypotheses.tsv	hypothesis_id	Identifier within the question.	H1 to H3	never blank
SI-19_hypotheses.tsv	short_name	Short name of the hypothesis.	text	never blank
SI-19_hypotheses.tsv	statement	The hypothesis, stated with the prediction that tells it apart from the others.	text	never blank
SI-19_hypotheses.tsv	distinguishing_prediction	What this hypothesis predicts that the others do not.	text	never blank
SI-19_hypotheses.tsv	refuted_if	The observation that would show the hypothesis wrong.	text	never blank
SI-19_hypotheses.tsv	current_standing	Standing of the hypothesis on the present data; a hypothesis is not a result.	text	never blank
SI-19_hypotheses.tsv	test_type	Kind of test that decides the hypothesis.	computational; wet-laboratory test (under way; no result is reported); with the state of the computation in brackets	never blank
SI-19_references.tsv	citation_in_text	The citation as it stands in the statements (author and year).	text	never blank
SI-19_references.tsv	authors	Authors of the cited work.	text	never blank
SI-19_references.tsv	year	Year of the cited work.	four digits	never blank
SI-19_references.tsv	title	Title of the cited work, quoted from it.	text	never blank
SI-19_references.tsv	venue	Journal or preprint server, with volume and pages where they exist.	text	never blank

Table SI-19_hypotheses.tsv (34 rows x 8 columns)

question_id	hypothesis_id	short_name	statement	distinguishing_prediction	refuted_if	current_standing	test_type
P01	H1	distance beyond the reach of the classifier profiles	H1: the cassette RT is a distant member of the UG3 and UG17 group at a distance beyond the reach of the DefenseFinder and PADLOC profiles, so that held-out UG3 and UG17 members matched for distance are missed by both tools as often as the cassette RT.	Distance-matched held-out UG3 and UG17 members are missed at the rate of the cassette RT.	H1 is refuted if distance-matched held-out UG3 and UG17 members are called by DefenseFinder and PADLOC clearly more often than the cassette RT.	Tested once with members matched on identity to the closest reference: they were called, the cassette RTs were not; matching on alignment score and coverage was not possible with known members (see the test status).	computational (run once)
P01	H2	separate branch beside UG3 and UG17	H2: the cassette RT lies on a branch beside the UG3 and UG17 pair and inside neither, so that matched held-out members are called clearly more often than the cassette RT and AU tests on a saturation-reduced, structure-guided alignment reject nesting within UG3 and within UG17.	Matched held-out members are called more often than the cassette RT, and AU tests reject nesting within UG3 and within UG17.	H2 is refuted if held-out members are missed as often as the cassette RT or if AU tests on a saturation-reduced, structure-guided alignment fail to reject nesting.	The detector prediction of H2 was met with members matched on identity only, a comparison that the test status does not count as evidence at the distance of the cassette RT; no AU test on a saturation-reduced, structure-guided alignment is reported here, and the AU placement tests of the K clade are reported in SI-11.	computational (detector part run once; AU part not reported here)
P02	H1	ExoI axis (literature-motivated)	Conjecture H1, motivated by the DRT3 and DRT10 literature and not by our data: if the cassette RT matters in a bacterial host it follows DRT10-like logic, so that in E. coli it causes a growth defect only when ExoI (sbcB) is absent, not when recB, recC or recD is absent, and the defect disappears with the YAAD variant.	Growth defect only in the sbcB deletion, abolished by the YAAD variant; no defect in recB, recC or recD deletions.	H1 is refuted if a defect appears in recB, recC or recD deletions or with Gam or gp5.9, or if validated expression gives no defect in any deletion strain.	Not supported or tested by data of this work; the pre-specified general test of a RecBCD-axis logic failed.	wet-laboratory test (under way)

Supplementary Note S2: Dossier of the open questions

question_id	hypothesis_id	short_name	statement	distinguishing_prediction	refuted_if	current_standing	test_type
P02	H2	RecBCD axis (literature-motivated)	Conjecture H2, motivated by the DRT3 and DRT10 literature and not by our data: the cassette RT follows DRT3-like logic, so that the defect appears when RecBCD is absent or inhibited (recB, recC or recD deletion, Gam, gp5.9), is not changed further by an sbcB deletion, and disappears with the YAAD variant.	Growth defect with RecBCD absent or inhibited, unchanged by an sbcB deletion, abolished by the YAAD variant.	H2 is refuted if no defect appears in recB, recC or recD deletions, with Gam or with gp5.9, or if a defect appears only in the sbcB deletion.	Not supported or tested by data of this work; H1 and H2 exclude each other in the deletion-strain pattern they predict.	wet-laboratory test (under way)
P02	H3	mosaic	Conjecture H3: the K family is a mosaic of lineages with different host-nuclease relationships and the cassette lineage follows neither logic, so that no recBCD or sbcB deletion changes a validated cassette phenotype.	No deletion strain changes a validated cassette phenotype.	H3 is refuted if a defect that disappears with the YAAD variant appears in one deletion strain.	Not tested; H3 is the alternative to both H1 and H2 and not exclusive with them across lineages.	wet-laboratory test (under way)
P03	H1	sampling and detector reach	H1: the single-genus result is a limit of sampling and detector reach, so that a pre-specified, sensitivity-calibrated scan beyond GTDB representatives finds a -1-type gene next to a K-lineage RT, or an RT-bearing segment between -1- and +3-type genes, in a further genus.	A calibrated scan finds a further locus.	H1 is refuted if a calibrated scan with adequate reach over enough genomes finds no such locus, and a scan without calibrated sensitivity is uninformative.	Not tested; a simulation estimates a low detector reach at the identities of the outside RTs, an estimate that depends on a model built from the cassette.	computational
P03	H2	true confinement	H2: the combination arose once and stays confined to the Pollutibacter lineage, through the narrow reach of its RT clade or through the joining event, although the parts coexist at genus level elsewhere, so that a calibrated scan finds no further locus.	A calibrated scan finds no further locus.	H2 is refuted if a locus whose host genus is confirmed by rpoB and gyrB is found outside the genus.	Not tested.	computational
P03	H3	pair first (one reading, untested)	H3: the -1- and +3-type genes were an adjacent pair before the RT-bearing segment lay between them, and the converse, free pairs as remnants of segment-bearing blocks, is equally open.	Outgroup-rooted -1 and +3 trees give a polarity of the pair.	H3 would not be supported if outgroup-rooted trees place the cassette genes in a discordant lineage, if free pairs outside Pollutibacter do not lie at the cassette position or if the junction carries no module remnant, and the opposite outcomes would weaken the reverse direction.	Unstable at present: outgroup-rooted trees are saturated; H3 concerns polarity and is not on the same axis as H1 and H2.	computational (unstable at present)
P05	H1	catalysis-free scaffold	H1: the HExxH-lacking N-terminal domain of the +3 protein is kept as a catalysis-free scaffold, for example for contact with the -1 protein, so that the purified domain shows no zinc-dependent proteolysis where an active control of the same fold is active and its removal abolishes -1 binding.	No zinc-dependent proteolysis by the purified N domain; loss of -1 binding without it.	H1 is refuted if the purified N domain shows proteolysis that EDTA or o-phenanthroline inhibits, or if its removal leaves -1 binding unchanged.	Not tested; binding between the -1 and +3 proteins has not been tested and a complex prediction carries no information.	wet-laboratory test (under way)
P05	H2	non-canonical metal site (post hoc)	H2: the N domain keeps a non-HExxH metal site and acts on a substrate other than peptide bonds, as suggested by a His and Cys cluster in some predicted structures, a reading that is post hoc, depends on the predictor and carries no information.	Substituting the His and Cys residues abolishes an activity of the purified domain.	H2 is refuted if the purified domain has no metal-dependent activity and substituting the His and Cys residues leaves binding and phenotype unchanged.	Not tested; post hoc and predictor-dependent.	wet-laboratory test (under way)
P05	H3	constraint in the C domain	H3: constraint on the +3 protein is carried by its C-terminal domain and the N domain is a passive fusion partner, so that the N region carries a clearly lower fraction of negatively selected codons than the C region.	A clearly lower N fraction of negatively selected codons at adequate power.	H3 is refuted if a repeat of the N/C comparison on a larger alignment of free +3-type sequences, with adequate power, shows an N fraction equal to or higher than the C fraction.	The FEL fractions do not support H3, but the design has low power for small differences and cannot separate selection strength from rate heterogeneity.	computational (run once; low power)
P06	H1	candidate recognition site of the +3 protein (untested)	H1: the +3 protein, whose C-terminal domain is predicted winged-helix-like (a predictor-dependent reading), recognises the palindrome, so that purified +3 protein binds the core-bearing oligonucleotide and not a core-scrambled one, and +3-free members of the same -1 lineage lose the site faster than neutral spacing.	Core-dependent binding of the +3 protein; loss of the site faster than neutral spacing in +3-free members of the same -1 lineage.	H1 is refuted by an EMSA with a validated positive control that shows no core-dependent binding, and it is weakened if +3-free members of the same -1 lineage keep the site.	Untested; the pre-specified structure and sequence arms did not support the winged-helix resemblance.	wet-laboratory test (under way); computational
P06	H2	host regulator in trans	H2: a host regulator, for example a WYL-domain protein, reads the site, so that the site persists in +3-free members of the same -1 lineage and recurs at other loci regulated by the same factor.	The site persists without the +3 gene within one -1 lineage and recurs beside a common neighbouring factor.	H2 is refuted if the genome-wide copies of the 14-bp site (22 outside the focal window against 7.4 expected from composition) share no neighbouring factor, or if members without the +3 gene lose the site within one -1 lineage, which the present contrast, mixing +3 presence with -1 lineage, cannot decide.	Untested; the cis test is nearly uninformative.	computational
P06	H3	mobility marker	H3: the palindrome marks the mobile pair locus, as an integration or recombination site, and tracks +3-bearing lineages because the pair travels as a unit, so that copies lie at pair boundaries beside integrase-type genes.	Copies of the site at pair boundaries beside integrase-type genes.	H3 is refuted if the genome-wide copies outside the focal window show no neighbouring integrase or recombination function and the site does not co-travel with the pair locus; the similar distance of the site from the -1 start does not separate H3 from H1 and H2.	Not tested; co-inheritance of a mobile locus is not excluded.	computational

Supplementary Note S2: Dossier of the open questions

question_id	hypothesis_id	short_name	statement	distinguishing_prediction	refuted_if	current_standing	test_type
P09	H1	5' region of the RT transcript	H1: the element is the 5' untranslated region, or an N-terminal coding extension, of the RT transcript with no separate RNA product, so that 5' RACE places the transcription start at or inside the 5' end of the element and no stable element-length RNA accumulates apart from the RT mRNA.	A transcription start at or within the 5' end of the element, and no stable element-length RNA apart from the RT mRNA.	H1 is refuted if a stable element-length RNA with its own 5' and 3' ends accumulates apart from the RT mRNA.	Not tested; the promoter-like pair predicts a start inside the element (post hoc).	wet-laboratory test (under way)
P09	H2	separate template or scaffold RNA	H2: the element is a structured RNA used by the RT as template or scaffold, so that cassettes with an active RT, but not the YAAD variant, yield element-derived cDNA in a validated expression system.	Element-derived cDNA from the active RT and not from the YAAD variant.	H2 is refuted if the active RT does not make more element-derived cDNA than the YAAD variant in a validated system.	Not tested; structure-based tests were uninformative.	wet-laboratory test (under way)
P09	H3	intron-like vestige	H3: the element is a degenerate remnant of an intron-like ancestor whose constraint is a by-product of its overlap with the RT start, so that outside the overlap element loci without an RT are no more constrained than neutral spacing.	Element loci without an RT are no more constrained than neutral spacing once the overlap is masked.	H3 is refuted if element loci without an RT stay clearly more constrained than neutral spacing once the overlap is masked, although the constraint analysis is circular through its anchor.	Cannot be tested at present because the element-first search found no element-only locus after the post hoc rule.	computational (not testable at present)
P07	H1	DNA-end processing neighbourhood as a preferred landing place	H1: the xseA-recD interval is a preferred landing place because it lies among DNA-end processing genes (xseA/xseB, recBCD, sbcCD), so that large inserts occur there more often than at other positions and equivalent end-processing clusters elsewhere (recJ, sbcCD, recFOR) are occupied at comparable rates.	Occupancy by large inserts is raised at xseA-recD and at equivalent end-processing clusters, above a same-genome null.	H1 is refuted if the occupancy of recJ, sbcCD and recFOR neighbourhoods by large inserts is clearly lower than at xseA-recD or does not exceed a same-genome null.	Not tested; the scan covers only part of the representatives and the strict class is post hoc.	computational
P07	H2	recognised integration site	H2: a short sequence at the address is recognised by an integration or recombination function of the island or the host, so that occupied junctions share a boundary repeat or target-site duplication and empty references carry one copy.	Occupied junctions share a boundary repeat or target-site duplication beyond a shuffled control.	H2 is refuted if the boundary repeats or target-site duplications of the occupied junctions are not shared beyond a shuffled control and the cargo carries no integrase, on the premise that the boundary positive control passes, which it did not in the present analysis.	Not testable now: the boundary positive control failed (no information).	computational (not testable now)
P07	H3	one old insertion	H3: the island entered once, long ago, and persists within one subclade, so that the other large inserts are unrelated events and the shared address is incidental.	Island-carrying genomes lie within one subclade with a shared junction, and other cargo families do not repeat the address.	H3 is refuted if island-carrying genomes occur outside the subclade with independent junctions or if the same cargo family repeatedly occupies this address.	Not tested.	computational
P10	H1	RecBCD axis (literature-motivated)	Conjecture H1, motivated by the DRT literature and not by our data: the cassette RT product intersects RecBCD-dependent processing of DNA ends, so that E. coli lacking recB, recC or recD, or expressing Gam or gp5.9, shows a growth defect absent in the wild type and in an sbcB deletion and abolished by the YAAD variant.	Growth defect with RecBCD absent or inhibited, absent in the wild type and in an sbcB deletion, abolished by the YAAD variant.	H1 is refuted if validated expression shows no such defect, and an effect only in the sbcB deletion points to the ExoI axis (see P02).	Not supported or tested by data of this work; the pre-specified general test of this logic failed.	wet-laboratory test (under way)
P10	H2	site availability and lineage background	H2: the companionship reflects insertion-site availability next to recD and the RecBCD background of Chitinophagaceae and not a requirement of the RT, so that K_E units are not enriched for RecC over same-genus K_0 units in informative genera outside Chitinophagaceae.	No enrichment of RecC in K_E units over same-genus K_0 units outside Chitinophagaceae.	H2 is refuted if K_E units, defined without recD, are RecC-positive significantly more often than same-genus K_0 units in a pre-set number of informative genera outside Chitinophagaceae.	Not tested; only 4 genera are informative so far.	computational
P10	H3	end-processing set in cis in one island lineage	H3: in one island lineage (Ferruginibacter; one independent unit) the RT-bearing island carries its own end-processing set of RecB-, RecC- and RecD-like genes, so that the companionship travels with the RT unit there and the island-borne set is full length with intact motifs and not sister to the host copy.	The island-borne set is full length with intact motifs and not sister to the host copy.	H3 is refuted if the island-borne set is a recent duplicate and sister of the host copy or if its motifs are destroyed.	Supported in one independent unit only; H1 and H3 can both hold in part.	computational
P04	H1	central block shared more recently than -1	H1: the identity of RT, +1, +2 and +3 between the Canadian and the Chilean sample reflects a more recent shared history of the central block than of -1, by recent exchange or recent common descent, which the Chilean config, ending without the downstream junction, cannot tell apart, so that synonymous and intergenic positions inside the block are as identical as coding ones with a sharp identity break at the -1 edge.	Synonymous and non-coding identity inside the block as high as coding identity, with an identity break at the -1 edge.	H1 is refuted if synonymous and non-coding identity inside the block lies at the host level or if no identity break lies at the -1 edge.	Not tested; needs a nucleotide-level alignment.	computational
P04	H2	slow block and fast -1	H2: RT, +1 and +2 are unusually constrained while -1 evolves under relaxed constraint, so that the Canadian and the Chilean sample carry old alleles and synonymous divergence inside the block lies at host levels.	Synonymous divergence inside the block at host levels.	H2 is refuted if the observation that the synonymous distance of the RT, +1 and +2 lies at the extreme of host orthologs reproduces in a nucleotide-level alignment.	Disfavoured by the present synonymous-distance percentiles; the nucleotide-level alignment is pending.	computational
P04	H3	assembly artefact	H3: one member carries a chimeric or cross-sample block, so that read-pair or long-read support across the block-to-host junction fails in at least one sample.	Junction support fails in at least one sample.	H3 is refuted if junction support is confirmed in both samples.	Not tested.	computational

Supplementary Note S2: Dossier of the open questions

question_id	hypothesis_id	short_name	statement	distinguishing_prediction	refuted_if	current_standing	test_type
P13	H1	single acquisition with losses	H1: the island was acquired once by the ancestor of the sub-clade and persisted by descent with a few losses, so that cassette-free sisters inside the sub-clade lack the island but keep an empty or rearranged xseA-recD address and the island gene trees match the core-genome tree.	Island gene trees match the core tree; cassette-free sisters keep an empty or rearranged address.	H1 is refuted if the island gene trees disagree with the core tree beyond saturation noise or if a sister inside the sub-clade carries a different cargo at the same address.	Not tested; the present data cannot separate H1 from H2.	computational
P13	H2	repeated acquisition into a permissive sub-clade	H2: the genomic context of the sub-clade, for example high GC and a RecBCD arrangement, permits the island, which entered several times from outside the genus, so that junction sequences and island gene trees differ among carriers beyond host divergence.	Junction sequences and island gene trees differ among carriers beyond host divergence.	H2 is refuted if the junction sequences are identical among carriers and the island tree matches the species tree.	Not tested; the junctions of the cassette MAGs can be compared at once.	computational
P13	H3	detection decay with distance	H3: the clumping reflects declining detection sensitivity with distance from the seed lineage and not biology, so that detectors built from seeds of a different lineage do not show it.	Detectors built from other seed lineages do not show the clumping.	H3 is refuted if the same clumping persists with detectors built from other seed lineages.	Not tested; the carrier call retains seed circularity.	computational
P11	H1	targeting adaptor	H1: +1 is an HsdS-type DNA-recognition adaptor that directs an effector to target sequences, so that its sequence variation, which gives several module reference clusters where +2 and +3 give one each (post hoc), concentrates in recognition-like regions on both sides of a centre that varies little, as in domain shuffling of specificity subunits.	Variation among +1 alleles concentrates in recognition-like regions on both sides of a centre that varies little.	H1 is refuted if variation in the +1 allele alignment does not concentrate in recognition-like regions on both sides of the centre but is uniform or lies elsewhere.	Not tested; the contrast of module reference clusters on which it rests is post hoc.	computational
P11	H2	convergent resemblance	H2: the HsdS resemblance of +1 is a profile artefact of a single protein of the size of a recognition domain matched through REBASE-derived profiles only, and +1 has an unrelated role, so that profiles built without REBASE sequences do not retrieve +1 and +1 lacks a paired recognition-domain architecture.	Profiles built without REBASE sequences do not retrieve +1; no paired recognition-domain architecture.	H2 is refuted if HsdS profiles built from non-REBASE sources retrieve +1 and a paired recognition-domain architecture is seen.	Not tested.	computational
P08	H1	excisable island and mixed population	H1: cells with and without the island coexist because the island can excise or be lost, so that island-lacking junction reads appear at the rate predicted by the depth step once mapping bias is removed.	Island-lacking junction reads at the rate predicted by the depth step.	H1 is refuted if the step persists in k-mer depth although, at a depth where the expected read count is sufficient, no island-lacking junction reads appear, not even a fraction of the number expected.	Not tested; the observed junction reads fall short of the expectation without excluding small fractions.	computational
P08	H2	mapping artefact or copy number	H2: the step comes from multi-copy or mis-assembled front sequence and cross-mapping while the island is single-copy in every cell, so that island-unique and flank-unique k-mers show equal depth.	Island-unique and flank-unique k-mers show equal depth.	H2 is refuted if the depth ratio of island-unique to host-unique k-mers on both sides of the junction is near the step ratio.	Not tested.	computational

Table SI-19_observations.tsv (128 rows x 6 columns)

question_id	observation_no	statement	level_of_support	may_depend_on_cassette_derived_model	where_stated
P01	1	The aromatic-x-Asp-Asp box is present in 19 of 19 evaluable RT ORFs (17 YADD, 2 FADD), the 3 ORFs without it, among 22 in all, end at a contig edge before the box region, and the search pattern was widened after the pre-specified form missed the FADD copies.	suggestive, post hoc	no	observation
P01	2	No Q.G motif lies in the 59 residues upstream of the box in any of these ORFs (the same holds for the 60 residues named in the pre-specified text), and the same 59 residues carry LPLG in 17 of these 19 ORFs, LPMG in 2 and RGLP in 4, an RGLP window carrying LPLG as well.	descriptive only	no	observation
P01	3	No frameshift, premature stop or truncating deletion is seen in 74 observable copies of the five cassette families, 17 of them RT copies, in the assembled contigs of one genus, and the detector recovers the planted disruptions.	negative result	no	observation
P01	4	The two-sided 95% upper bound on a disruption rate is 0.247 to 0.308 when the copies are collapsed to allele clusters (10 to 13 clusters per module) and 0.9750 when they are counted as the single genus that carries them.	negative result	no	observation
P01	5	Saturation of divergence between the copies inflates the neutral expectation of a disruption and no host pseudogene baseline exists, so a count of zero carries less information than its size suggests, and no neutral probability is quoted here.	limit	not applicable (limit statement)	observation
P01	6	The substitution-rate ratio omega of the five cassette modules is far from the neutral value, from 0.0896 for the RT, the lowest, to 0.1805 for -1, and negatively selected codons predominate, a direction of constraint of intermediate strength.	descriptive only	yes	observation
P01	7	The classifier DefenseFinder calls a retron, DRT or DRT-like system at 0 of 16 cassette RT loci and at 7 of 39 random bacterial RT loci (Fisher P = 0.0933), but its models are biased toward Proteobacteria and their sensitivity in Bacteroidota is unconfirmed, so the missing call is not evidence about function.	negative result	no	observation
P01	8	The RT alignment is saturated, with a median p-distance of 0.815 between random tip pairs, and the class profiles place the cassette RTs in UG17 for 7, in CRISPR-RT for 5 and in UG3 for 1 of 16 loci (median margin 1.8 bits), an assignment forced to the nearest reference class that weakens when one class is removed.	no information	yes	observation

question_id	observation_no	statement	level_of_support	may_depend_on_cassette_derived_model	where_stated
P01	9	By minimal patristic distance the K family lies nearest to UG17 (1.0993 against 1.1294 for CRISPR-RT, a margin of 0.030; UG17 is nearest in 4 of 6 trees), whereas tree topology places UG3 as the first non-K neighbour (see P02), so the nearest class depends on the method.	no information	yes	observation
P01	10	A held-out recovery test was run under a rule fixed before any result was seen.	descriptive only	no	test status
P01	11	Of 49 known UG17, UG3 and UG8 units matched to the cassette RT on identity to the closest reference, 49 meet the DefenseFinder-style profile threshold and 23 of 49 the PADLOC-style threshold, whereas none of 16 cassette units meets either (Supplementary Table S2, table Q2).	descriptive only	no	test status
P01	12	With the shipped tools, which include these members in training, the same units are called as systems in 3 of 49 (P = 0.57 against none of the cassette units), so the counts above refer to profile thresholds and not to system calls (Supplementary Table S2, table Q2).	descriptive only	no	test status
P01	13	The nominal Fisher P of 1.54e-15 becomes 2.65e-11 when the known units are collapsed to their 23 sub-lineages and 0.0417 when the cassette is counted as the single genus it comes from, and the test is a reference-pool emulation and not a run of the tools themselves.	descriptive only	no	test status
P01	14	Alignment scores and coverage were not matched: the cassette RT, scored on its 12 full-length ORFs, aligns to its nearest reference with a bit score of at most 79 over 45.2 to 70.7% of its length, whereas the matched members score at least 135 over a median of 98.25% (post hoc; Supplementary Table S2, tables Q1 and Q5).	descriptive only	no	test status
P01	15	So no known member sits as far from its closest reference as the cassette RTs do, the comparison is not evidence in either direction at that distance, and the cassette is a single genus.	descriptive only	no	test status
P01	16	Known UG17 members cut to the aligned core of the cassette RT are not called and most UG3 members are still called, so a missed UG17 call of the cassette RT cannot be read as a property of its lineage (post hoc; UG8 excluded; a single genus).	suggestive, post hoc	no	test status
P02	1	In predicted structures the nearest non-K class of the K family is UG17 (DRT10) for 16 of 16 K units, which is a direction and not membership.	suggestive, post hoc	yes	observation
P02	2	The pre-specified DRT10-specificity rule is met by 0 of 16 units, percentiles against a null of random bacterial RTs run from 0.103 to 0.476, and the Foldseek bit scores of the cassette-lineage models (398 to 495) lie below those of UG17 positive controls (909 to 2161).	suggestive, post hoc	yes	observation
P02	3	In the sequence trees UG3 is the first non-K neighbour in all 7 tree calls, at patristic distances of 1.018 to 1.994 from the K clade, but the calls share one alignment, distances are saturated, and the node support at the junction of the K clade with UG3 ranges from 0.0 to 97.9 among the 6 calls that report it.	no information	yes	observation
P02	4	By minimal patristic distance the nearest class is UG17 (1.0993 against 1.1294 for CRISPR-RT; UG17 is nearest in 4 of 6 trees) although the topology places UG3 beside the K clade, so structure, distance and topology do not agree on one neighbour.	no information	yes	observation
P02	5	The K family is not monophyletic in the RT-domain maximum-likelihood tree, and K is a post hoc definition.	suggestive, post hoc	yes	observation
P02	6	The cassette RT lacks the priming features and the C-terminal tail of DRT10 and no SLATT-family gene occurs in the proteomes of the cassette MAGs, whereas some K-core units are called DRT10-type by several measures (exploratory).	suggestive, post hoc	yes	observation
P02	7	The genes around UG17 windows are neighbours of recombinase-, HsdS-like and other island components and cannot be told from a defense-island context, and carriage of RecBCD, RecJ, ExoVII and SbcB does not differ from the null.	suggestive, post hoc	yes	observation
P02	8	The expectation that carriers of DRT10-type RTs retain SbcB, as would follow if SbcB suppressed DRT10-type toxicity, is not supported as a general rule in the genomes of this study, although the sensitivity of SbcB detection outside Pseudomonadota is uncalibrated, and the pre-specified general test of a RecBCD-axis logic failed.	negative result	yes	observation
P02	9	The literature reports that DRT3 is toxic in cells lacking RecBCD (Wang et al. 2026), and this is literature context and not an observation on the cassette.	literature context	not applicable (literature context)	observation
P02	10	A preprint reports that a DRT10 system is toxic in Escherichia coli cells lacking the host exonuclease SbcB (ExoI), where no viable transformants carrying it were recovered, and that SbcB suppresses this toxicity (Sánchez-Nieto et al. 2026), which is literature context and not an observation on the cassette.	literature context	not applicable (literature context)	observation
P02	11	In Pollutibacter the RecD-like gene downstream of the RT cannot be told from the host's own RecD, since each of the 8 full-length genomes carries exactly one RecD-like protein, and it is not counted as a cassette module (see P10).	suggestive, post hoc	yes	observation
P03	1	Same-strand -1/+3-type gene pairs without a detected RT occur in 54 genera from 24 families of the confirmatory set (the +3-type gene anchors 348 loci in 149 genera, and 141 of these loci carry such a pair), at a ratio of 132.4 to a positional null (interval 94.4 to 193.4), and across all 155 genera with a +3-type anchor (374 anchor loci), 151 of 157 observed pair loci have no RT of any kind.	supported	yes	observation
P03	2	After the lead genus and the cassette-lineage genera are excluded, 50 genera from 24 families remain at a ratio of 145.0, an analysis that was not in the pre-specified sensitivity list.	suggestive, post hoc	yes	observation
P03	3	These pair genera are a non-random subset of annotated public Bacteroidota genomes and not Bacteroidota as a whole, the count concerns the repetition of RT-free pairs and not their function or origin, vertical descent and repeated acquisition cannot be told apart without a tree, and the family was defined from cassette sequences so that circularity is not excluded.	limit	not applicable (limit statement)	observation
P03	4	The number of pair classes is not a number of origins, genera are not independent lineages, and the single-genus limit of the cassette is unchanged by this recurrence.	limit	not applicable (limit statement)	observation
P03	5	The recurrence of RT-free pairs outside the training genera meets its pre-specified rule, although the null is not calibrated and independence is qualified, and the free pairs and the cassette -1 and +3 genes are distant paralogs with low identity to each other, the direction of any exchange being unresolved.	suggestive, post hoc	yes	observation
P03	6	A vWA-like gene is also found beside a Macro-domain gene outside Bacteroidota, in 1145 of 2275 eligible genera at a ratio of 85.8 (the neighbouring genera are mostly Bacillota), with 29.2% of Macro-domain genes having an adjacent -1-type gene against 0.36% an adjacent generic vWA gene, and in Bacteroidota the corresponding count is 5 of 253 genera, the -1-type gene having more than one partner family.	suggestive, post hoc	yes	observation

question_id	observation_no	statement	level_of_support	may_depend_on_cassette_derived_model	where_stated
P03	7	Outside Pollutibacter, 95 K-family RT loci carry an element at the RT start, in 6 labelled genera and 11 loci without a genus label, and of 1130 observable -1, +1, +2 and +3 slots beside the outside K-family RT loci none holds a cassette part, so the element-plus-RT layer and the cassette-part layer must be counted separately.	negative result	yes	observation
P03	8	The outside RTs are at most 0.33 identical to the cassette RT, so this finding concerns RT lineages far from the cassette lineage, this detector and this range, and the cassette remains a single genus.	negative result	yes	observation
P03	9	In a simulation the part detectors reach a high detection rate only at high identities to the cassette member (the identity needed differs by part), whereas most outside RTs are far from the cassette RT in identity, so that absence outside the genus is uninformative.	no information	yes	observation
P03	10	This reading depends on a simulated evolution model, rests on detectors built from the cassette (circular), has a single genus as its independent unit and has no model-independent verification, and it does not show that the parts are absent outside the genus.	limit	not applicable (limit statement)	observation
P03	11	Over all Bacteroidota species representatives the strict detectors find same-strand pairs at a short gene-number distance, 23 of them in 14 genera outside the cassette-lineage genera, and the pairs that have an RVT_1 domain between the two genes lie in Pollutibacter at distance 4.	descriptive only	yes	observation
P03	12	These counts belong to this detector, which derives from the cassette, and they are not an independent confirmation of the single-genus result.	limit	not applicable (limit statement)	observation
P03	13	The -1 tree is untestable at a median outgroup-ingroup distance of 6.53 substitutions per site, and the result depends on how the outgroups were chosen.	no information	no	test status
P03	14	The +3 tree is untestable at 6.30 substitutions per site, and the AU test of its closest alternative root lies just above the significance threshold, so the polarity of the pair is untested.	no information	no	test status
P05	1	The HExxH-type pattern is found in 0 of 13 +3 ORFs (11 complete ORFs, 12 ORFs with a valid test of the N part, 12 ANI95 units), whereas the same pattern is found in 7 of 7 Peptidase_M54-like ORFs beside the cassette and in human MMP1.	descriptive only	yes	observation
P05	2	This is a sequence check of the canonical catalytic motif of the +3 N protein that does not establish that the motif was lost, that the protein is inactive or any change in activity, and no wet-laboratory result is reported.	limit	not applicable (limit statement)	observation
P05	3	The absence is not confined to the cassette: the strict HE..H pattern is found in only 2 of 374 free +3-type sequences, whereas 1565 of 1626 Pfam M54 sequences carry HExxH, and these two counts rest on different definitions.	descriptive only	yes	observation
P05	4	The 13 ORFs fall in 12 ANI95 units of one genus and are not 13 independent observations, and 13 independent absences would not be unlikely by chance.	descriptive only	yes	observation
P05	5	The zinc-compatible part of this analysis rests on a post hoc definition and a null that depends on selection, and it is not carried here.	limit	not applicable (limit statement)	observation
P05	6	The omega of the +3 gene is 0.1125, far from the neutral value, and negatively selected codons predominate in its alignment, a direction of constraint of intermediate strength.	descriptive only	yes	observation
P05	7	In predicted structures the N region resembles an archaeometzincin-type fold at an intermediate fold-level confidence, a catalytic-loss variant of M54 is not supported by profile searches, and a zinc-compatible geometry may be produced by the sequence prior of the predictor.	suggestive, post hoc	no	observation
P05	8	The C region occurs as a stand-alone family, which may include N-terminally truncated fragments, and as a fusion form extended on the N side that accounts for most public genome hits, and the C family is restricted to Bacteroidota by taxonomy but not by environment (post hoc observations).	suggestive, post hoc	no	observation
P05	9	In the +3 alignment of 11 sequences the fraction of codons that FEL calls negatively selected (P at most 0.1) is 53.5% in the N region (the first 200 columns; 107 of 200) and 46.1% in the rest (71 of 154), with no positively selected codon in either.	descriptive only	yes	observation
P05	10	The test, the interval and the power calculation treat the 354 codons as independent trials, which neighbouring codons need not be, and a circular-shift permutation that keeps the neighbour structure gives P = 0.47 with the same negative result, so the smallest difference that this design could detect is at least as large as computed under independence.	limit	not applicable (limit statement)	observation
P05	11	The difference is 7.4 percentage points with a 95% interval from -3.1 to +17.6 and a two-sided Fisher P of 0.1982, so there is no evidence that the N region is under weaker purifying selection than the rest, which is not evidence that the regions are equal.	descriptive only	yes	observation
P05	12	The design has low power for small differences between the N and the C fraction, so the result is not evidence that the two regions are equally constrained.	descriptive only	yes	observation
P05	13	Across an enumerated variant grid, assembled after the primary result was seen and exploratory, the N fraction is at least the C fraction in all but one cell (the partitioned FEL at the strictest threshold, a difference of -0.8 points, P = 0.90) and no cell gives significant uncorrected one-sided evidence of a weaker N region.	suggestive, post hoc	yes	observation
P05	14	Two partitioned cell boundaries, at columns 225 and 250, give N minus C differences of +11.8 and +11.6 points with N higher (P = 0.036 and 0.048), also exploratory and uncorrected.	suggestive, post hoc	yes	observation
P05	15	FEL fixes the synonymous rate alpha at 1.000 at 349 of 354 sites, so the contrast rests on an unresolved site-wise synonymous rate, and the calls are departures of the nonsynonymous rate under a fixed synonymous rate.	no information	yes	observation
P05	16	No HExxH-type match occurs in either region of the 11 aligned proteins, so the N/C contrast is not a contrast between the presence and absence of that motif, and the alignment belongs to a single genus.	descriptive only	no	observation
P06	1	A palindromic motif with the core CAAGCTTG lies upstream of the -1 gene of free -1/+3 pairs as a strict site in 78 of 103 genera (51 of 66 held-out genera against 5.8% of null windows, and 27 of 37 training genera), ending a median of 61 nucleotides before the start codon (10th percentile 49.7, 90th percentile 113.4).	supported	yes	observation
P06	2	The exact CAAGCTTG core occurs in 47 of 103 genera against 16 of 4049 null windows, the 14-bp extended palindrome in 16 of 103 against 0 of 4049, and 25 genera have no strict site, so the motif is not present in every genus.	supported	yes	observation

question_id	observation_no	statement	level_of_support	may_depend_on_cassette_derived_model	where_stated
P06	3	The support is that of a rule fixed before the analysis being met, the genera are phylogenetically related and are not independent observations (an order-level test would be more honest because many held-out genera belong to one order), and co-inheritance of a mobile locus is not excluded.	limit	not applicable (limit statement)	observation
P06	4	Binding of the +3 protein to the palindrome was never tested and the motif is not read as regulating anything, and its position in non-coding sequence is an exploratory analysis.	limit	not applicable (limit statement)	observation
P06	5	At the cassette, the exact core occurs at 11 of 14 upstream windows of the -1 gene, at 55 to 67 nucleotides from its start (post hoc), and at 0 of the 16 cassette element windows.	suggestive, post hoc	yes	observation
P06	6	The fraction of -1 upstream windows with a strict site is 0.877 beside the +3 gene, 0.576 where a +3 gene lies a few genes away (post hoc), 0.333 where the +3 gene comes first, 0.286 elsewhere in the genome and 0.049 where it is absent, in classes of 46, 23, 12, 35 and 168 genera, and the comparison is descriptive.	suggestive, post hoc	yes	observation
P06	7	This contrast mixes the presence of the +3 gene with the lineage of the -1 gene, because -1 genes without a +3 gene belong to a different -1 lineage, so the gradient does not show that the +3 protein binds the motif.	limit	not applicable (limit statement)	observation
P06	8	The site covaries with the sequence of the N domain of the +3 protein, used as a negative control, more strongly ($r = 0.296$; $P = 0.0014$) than with that of its C domain ($r = 0.194$; $P = 0.0239$).	no information	yes	observation
P06	9	A WYL-family neighbour lies near the -1 gene in 17 of 258 units, so the cis test of a host regulator has little power and a regulator acting in trans is untested.	no information	yes	observation
P06	10	The C-terminal domain of free +3 proteins is predicted winged-helix-like (a predictor-dependent reading that is weaker for the cassette +3), but the pre-specified structure and sequence arms did not support the resemblance and basic-residue enrichment is not specific to contact positions, so the prediction is not evidence of DNA binding.	suggestive, post hoc	yes	observation
P06	11	WYL-domain ORFs lie beside the cassette more often than in random windows of the same genomes, but their orientation and distance do not match known regulator arrangements.	suggestive, post hoc	yes	observation
P09	1	Relative to the RT start annotated in each of 16 source windows, the 3' end of the element lies in the start codon or directly before it in 7 windows and within 5 nucleotides of it in 9 windows, both windows being post hoc and the narrower one chosen to reproduce a count of 7 windows, so that it has no definition independent of that count.	suggestive, post hoc	yes	observation
P09	2	The RT call extends 95 nucleotides upstream of the element end in 3 of the 16 windows.	descriptive only	yes	observation
P09	3	The element models derive from the cassette and may be circular, so the abutment partly reflects how the models were made, and the coordinate calculation is the same observation and not a separate one.	limit	not applicable (limit statement)	observation
P09	4	Among loci with an element hit within a short window of the re-called RT start, the hit ends within a few nucleotides of the RT start at 15 seed windows, 21 other Pollutibacter loci and 95 outside K-family RT loci, and the non-coding space upstream of the RT cannot hold an ncRNA of the size of a DRT2 ncRNA.	descriptive only	yes	observation
P09	5	Extending the windows of cassettes in known bins confirms in some cases an element 3' end that overlaps the first RT start codon, but these are sister copies of known cassette bins whose elements are identical to a known element of the same bin, so they are not independent evidence, and the other windows failed for lack of reads, which is not evidence of absence.	suggestive, post hoc	yes	observation
P09	6	Sequence constraint of the element across genomes of the same genus is circular, because the anchor interval continues from the element 3' end and so coincides with the element-model boundary, but the contrast between its 5' part and the part near the RT start suggests that the constraint is not explained by the coding overlap with the RT start alone.	suggestive, post hoc	yes	observation
P09	7	R-scape finds 0 significant covariation pairs with a passing positive control, which is not informative about a long stem at this power, msd-, DRT2- and DRT9-type templates are not supported by this operationalisation, and the comparison with DRT10 ncRNA covariance models failed its positive controls, so a structured RNA is neither supported nor excluded.	no information	yes	observation
P09	8	An exploratory clue is that element loci without an RT exist, which an intron-like derivative from which the RT was deleted could explain.	suggestive, post hoc	yes	observation
P09	9	A sigma-A-type promoter-like pair of boxes, TTTG before the element start and TAHHTTTG near its 5' end, is shared by many cassettes (post hoc; the putative transcription start lies inside the element), but it is not unusual against gene-upstream base rates and was not reproduced at homologous positions in held-out sequences.	suggestive, post hoc	yes	observation
P09	10	An element-first search of 794 genomes finds 19 hits, 18 with the RT ahead, 1 with the RT side cut by a contig end and 0 element-only, after a post hoc rule that treats the cut case as unobservable (1 element-only case before the rule), within a population restricted to element-bearing genera plus Pollutibacter and with an element model derived from the cassette.	suggestive, post hoc	yes	observation
P07	1	In related genomes the cassette locus lies at the same address, between xseA/xseB and recD, recC and recB, and island-free references carry the same arrangement without a gap, but the genomes are not independent, the observation concerns one genus and the polarity is unknown.	suggestive, post hoc	yes	observation
P07	2	Among 3 non-carrier Pollutibacter species in which the address can be judged, 1 has an empty address and 2 are rearranged, which gives a 95% Wilson interval of 0.061 to 0.792 for the empty fraction, and rearranged means only that xseA and recD are not adjacent, usually on different contigs, and is not active evidence of an occupied address.	negative result	yes	observation
P07	3	In a scan of GTDB representatives, of which 69.7% were processed (139,400 of 199,923), a strict class of large gap between xseA and recD is found in 3 of 106,040 xseA-bearing genomes: the cassette island genome, one genome of the CAZUSO01 lineage whose insert carries outer-membrane- and porin-like ORFs, and one genome in which the gap holds RecB-like, RecC-domain and signalling genes, which a post hoc reading of the scan output treats as a RecBCD arrangement and not as an insertion.	no information	yes	observation
P07	4	The strict class is a post hoc definition whose sensitivity was not measured and the scan covered only part of the representatives, so the count is not an occupancy frequency, and the insert of the CAZUSO01 genome is not a cargo family shared with the island.	limit	not applicable (limit statement)	observation

question_id	observation_no	statement	level_of_support	may_depend_on_cassette_derived_model	where_stated
P07	5	Tests of a hotspot are closed: too few evaluable junctions reached the pre-declared minimum (no information), enrichment of large gaps upstream of recD outside the lineage failed its positive control (no information), the enrichment upstream of recD within the cassette genus is post hoc and partly circular, and junction repeats and target-site duplications failed the boundary positive control (no information).	no information	yes	observation
P10	1	Among 27 K main-clade units that carry the RT, 13 have RecBCD (Wilson interval 0.31 to 0.66) against 52 of 943 background units, and in an exact test stratified by genus, in 7 informative genera with 14 carriers, 13 are RecBCD-positive against 2.09 expected ($P = 1.5e-12$), which is companionship and not a requirement, rests on few confirmatory units and has an optimistic P value.	suggestive, post hoc	yes	observation
P10	2	The statement that RecBCD companionship is a property of the whole K main clade is partly rejected, because a substantial share of the high-quality carrier units lack RecBCD, while the narrow cassette clade is not refuted.	negative result	yes	observation
P10	3	By family, the fractions of high-quality carrier genomes with RecBCD are 13 of 13 in Chitinophagaceae, 0 of 6 in Saprospiraceae, 0 of 6 in Spirosomataceae and 0 of 5 in CAIMTB01, and 3 of the 7 families represented by a single genome also carry it.	descriptive only	yes	observation
P10	4	These are genome counts and differ from the unit counts above, and the family fractions rest on few genomes.	limit	not applicable (limit statement)	observation
P10	5	In the genus BJGN01, a genus of Saprospiraceae that carries the element plus RT unit, no RecC or RecD call is made, and the RecB and RecD callers are not calibrated.	descriptive only	yes	observation
P10	6	With a definition of K_E units (element plus K-family RT) that does not depend on recD, the K_E units of Chitinophagaceae are RecC-positive in 8 of 8 against 51 of 437 K_0 units, whereas in BJGN01 the figures are 0 of 4 and 0 of 5, and the comparison is uninformative because too few genera are informative.	no information	yes	observation
P10	7	In Pollutibacter the RecD-like gene downstream of the RT cannot be told from the host's own RecD, since each of the 8 genomes with a full-length gene carries exactly one RecD-like protein, and it is not counted as a cassette module.	suggestive, post hoc	yes	observation
P10	8	Of 16 island RecD-like sequences from 3 genera, 15 are evaluable and all keep their Walker motifs and ATP-contact columns, which shows only that these are ordinary RecD-like proteins and does not show that any of them acts as an effector.	descriptive only	no	observation
P10	9	In the Ferruginibacter island, additional RecD-, RecC- and RecB-like genes form a full-length set with intact motifs, but this is one clone and one independent unit, the association of RecD with the island is supported in one independent unit only, and complete non-carrier genomes of the same species keep RecBCD, which favours a lineage-background reading.	suggestive, post hoc	yes	observation
P10	10	The literature reports that DRT3 is toxic in cells lacking RecBCD (Wang et al. 2026), which is context and not an observation on the cassette.	literature context	not applicable (literature context)	observation
P10	11	The pre-specified general test of a RecBCD-axis logic failed, and the caller for alternative end-processing enzymes is not calibrated, so the alternatives are uninformative.	no information	yes	observation
P04	1	The focal sample pair is a Canadian LakePulse sample (run ERR4869792; latitude 49.01 degrees north) and a Chilean surface-water sample, and the RT of the Chilean contig is identical at protein level to the RT of the Chilean MAG cassette, so that it is neither a variant nor an independent unit.	descriptive only	yes	observation
P04	2	Protein identity between the two samples is 1.0000 for +2 (152 residues), 0.9962 for +1, 0.9894 for the RT and 0.9818 for +3 (a partial 220 residues in the Chilean reassembly), against 0.8967 for -1.	suggestive, post hoc	no	observation
P04	3	Of the mismatches of the -1 protein, 10 of 22 fall in a single window of 30 residues (permutation $P = 0.0057$; post hoc).	suggestive, post hoc	no	observation
P04	4	Against 2339 host orthologs of the focal sample, which have a median amino-acid identity of 0.9296, the observed synonymous distance is lower than that of 0.9957 of the length-matched host orthologs for the RT, 0.9991 for +1 and 0.9926 for +2.	suggestive, post hoc	no	observation
P04	5	Counted against all 2262 host orthologs that have a synonymous distance, only 11, 2 and 7 are at or below the value observed for the RT, +1 and +2, and the amino-acid identity of -1 exceeds that of 0.3923 of the length-matched host orthologs.	suggestive, post hoc	no	observation
P04	6	Among 66 sample pairs, 1 (the focal sample pair, selected post hoc) has an RT amino-acid identity percentile of 0.95 or more, the next highest being 0.731, and the other sample pairs follow the divergence of their hosts (Spearman $\rho = 0.929$).	suggestive, post hoc	no	observation
P04	7	The covariation of the distances of -1 and the RT is a cassette-wide effect and not specific to the focal sample pair, and the charge complementarity of -1 and the RT is not more retained than in charge- and identity-matched host genes.	negative result	yes	observation
P04	8	The focal sample pair was selected post hoc, the Chilean MAG and its reassembly share the same reads, the Chilean contig ends without the downstream junction so that it cannot be judged whether the block moved as a unit, and continent is confounded with clade, so this sample pair is not read as one moving unit.	limit	not applicable (limit statement)	observation
P13	1	All 16 cassette MAGs are classified as Pollutibacter by GTDB-Tk and fall into one sub-clade.	supported	yes	observation
P13	2	The mean GC of the MAGs is 54.06% against 43.32% for the genus representatives (95th percentile of the null 45.19%), but the pre-specified hypothesis of a high-GC sub-clade is not met as a whole, only its clumping component being supported, and the GC excess is borderline and confounded with a single clade.	no information	yes	observation
P13	3	The 7 carrier genomes on the GTDB tree (6 species; carriers are called by DIAMOND similarity to the cassette RT) form neither a single clone nor a pattern of repeated gain and loss, but cluster in one sub-clade of the genus.	no information	yes	observation
P13	4	The Fitch score is 2 against a permutation null mean of 5.84, with no permutation reaching it, a result that carries no information because the carriers were selected for carrying.	no information	yes	observation
P13	5	Within the sub-clade only 2 non-carrier species are known, 1 of them of high quality, so the pre-specified sibling criterion fails and carries no information.	no information	yes	observation

question_id	observation_no	statement	level_of_support	may_depend_on_cassette_derived_model	where_stated
P13	6	Whether the island was gained once and kept by descent or gained several times cannot be decided, because the sign of the polarity flips when an outgroup is included.	no information	yes	observation
P13	7	The MAGs are a sample selected for carrying and cannot be used to measure the carrier proportion, the null of the clustering test is almost trivial, and seed circularity remains in the carrier call.	no information	yes	observation
P11	1	At profile level +1 resembles HsdS specificity subunits at a distance: a loose UniRef50 hit, a predicted-structure hit and a relaxed-threshold DefenseFinder profile point to the same family, but each is below its rule and they are not independent.	suggestive, post hoc	yes	observation
P11	2	The resemblance is reproduced only with profiles derived from REBASE, and it is a weak signal that rests on one genus.	suggestive, post hoc	yes	observation
P11	3	No HsdM, HsdR or HsdS is detected in the cassette-bearing genomes, but non-carrier genomes of the same genus also mostly lack them, so a carrier effect cannot be separated from lineage background.	no information	yes	observation
P11	4	The genomes that carry +1-like ORFs have short contigs, so a test of a type I neighbourhood was not possible, and none of HsdM, HsdR and HsdS is found in those genomes.	no information	no	observation
P11	5	Permissive restriction-modification calls in K-positive genus units reject the absence of type I systems, including outside Bacteroidota, and the PF12705 slot meets its rule but is not specific to +1.	negative result	yes	observation
P11	6	Among the typed units of the main -1 cluster of the module references, +2 and +3 each fall in one reference cluster and +1 in several, a post hoc contrast between reference clusters and not between alleles, and +1 is 0.9962 identical between the two samples of the focal sample pair and has an omega of 0.1563.	no information	yes	observation
P11	7	The module references were built from the same MAGs, so the contrast may be circular and the larger +1 reference set may inflate it, and because the allele clusters counted for P01 number 10 to 13 per module, the contrast does not show that +2 and +3 are invariant between alleles.	limit	not applicable (limit statement)	observation
P11	8	The +2 protein gives no functional clue in sequence, profile, structure or embedding layers (an exploratory extended MSA model gives one macrodomain-like hit), but dark ORFs of matched length are common, so being dark is not unusual, and +2 has an omega of 0.1111 and is 1.0000 identical between the two samples of the focal sample pair.	no information	no	observation
P08	1	In one contig of the Ferruginibacter block the depth over the island region is 0.392 of the preceding region (95% interval 0.340 to 0.450; change point at 13000 bp; depth 14.846 before and 5.823 in the island), the step is seen in three samples, and the pairing of read pairs across it is normal.	suggestive, post hoc	no	observation
P08	2	The pre-specified competitive mapping ratio of 2.406 exceeds the criterion for no step, so the step is not supported by that rule, and a re-measure in one study alone is post hoc and not independent.	suggestive, post hoc	no	observation
P08	3	The cause is unresolved: occupancy polymorphism, low copy number of the island, a heterogeneous contig, cross-mapping and misassignment of the contig to the bin all remain possible, and the window was chosen after the step was seen.	no information	no	observation
P08	4	The expected number of island-lacking junction reads is 4.51 or 2.33 at alternative fractions of 0.58 and 0.30, against 1 clipped read and 2 of 65 abnormal reads left of the step (expected 7.83 and 4.05), and small fractions cannot be excluded.	suggestive, post hoc	no	observation
P08	5	In two closed Pollutibacter MAGs no read pair supports a deletion (0 against 38.0 preserved, and 0 against 19.0), and the depth ratios of island to flank are 1.018 and 1.171, a one-time-point upper bound under assumptions and without a spike-in test.	no information	yes	observation
P08	6	Long-read occupancy is a conditional clue that rests on few samples and is not an independent unit, and closure of the address carries no information.	no information	yes	observation
P12	1	Long-read DNA screens recover cassette reads (15 reads from 7 samples of 4 studies, in 265 runs and 202.9 Gbp screened; existence only).	supported	yes	limitation statement
P12	2	The long-read RNA screens are uninformative for the cassette lineage because almost no run carries host-lineage signal, and K-family-like reads in runs labelled as RNA cannot be assigned to the cassette lineage or to transcription.	no information	yes	limitation statement
P14	1	The pond time series of 14 time points cannot test host-virus covariation, because the carrier is detected at only 1 time point at the read cap and host and virus abundance share a winter minimum, so the series carries no information about covariation.	no information	no	limitation statement
P14	2	The Beaver Lake monthly series is a single independent unit with a post hoc pattern of presence.	suggestive, post hoc	no	limitation statement

Table SI-19_questions.tsv (15 rows x 13 columns)

question_id	title	placement	main_text_question	figure	n_observations	levels_of_support	why_it_is_open	n_hypotheses	cheapest_test	test_status	figure_idea	limits_carried
P01	An RT with its catalytic box in every evaluable ORF by a pattern widened post hoc, no observed disruption and omega far from neutral that no comparison assigns to a named lineage	main text (open question 1)	1	Extended Data Figure 9, panel A	9	negative result 3; descriptive only 2; no information 2; limit 1; suggestive, post hoc 1	No comparison assigns the RT to a named lineage, so the first question to ask of it, what it is, has no answer at present: each reading is limited by saturation, by a shared alignment or by classifier models biased toward other phyla, and the honest answer is that it is unassigned.	2	(1) The cheapest computational test of H1 and H2 is the held-out recovery comparison, repeated with members matched on alignment score and coverage as well as on identity, with the matching rule fixed before any call is made, which needs reference members at a greater distance than the present pool offers. (2) Membership is addressed by AU tests of nesting within UG3 and within UG17 on a saturation-reduced, structure-guided alignment with additional reference sequences, a computation of moderate cost after which, if saturation remains, the question is reported as unresolved. (3) A reference frame for the Bacteroidota-associated subtype of UG17 awaits the release of the reference tree of its source study.	(1) A held-out recovery test was run under a rule fixed before any result was seen. (2) Of 49 known UG17, UG3 and UG8 units matched to the cassette RT on identity to the closest reference, 49 meet the DefenseFinder-style profile threshold and 23 of 49 the PADLOC-style threshold, whereas none of 16 cassette units meets either (Supplementary Table S2, table Q2). (3) With the shipped tools, which include these members in training, the same units are called as systems in 3 of 49 ($P = 0.57$ against none of the cassette units), so the counts above refer to profile thresholds and not to system calls (Supplementary Table S2, table Q2). (4) The nominal Fisher P of 1.54e-15 becomes 2.65e-11 when the known units are collapsed to their 23 sub-lineages and 0.0417 when the cassette is counted as the single genus it comes from, and the test is a reference-pool emulation and not a run of the tools themselves. (5) Alignment scores and coverage were not matched: the cassette RT, scored on its 12 full-length ORFs, aligns to its nearest reference with a bit score of at most 79 over 45.2 to 70.7% of its length, whereas the matched members score at least 135 over a median of 98.25% (post hoc; Supplementary Table S2, tables Q1 and Q5). (6) So no known member sits as far from its closest reference as the cassette RTs do, the comparison is not evidence in either direction at that distance, and the cassette is a single genus. (7) Known UG17 members cut to the aligned core of the cassette RT are not called and most UG3 members are still called, so a missed UG17 call of the cassette RT cannot be read as a property of its lineage (post hoc; UG8 excluded; a single genus).	The figure idea is a disagreement panel in which the structure-based, tree-based, class-profile and classifier readings of the RT are set in one line each, with the count of observed disruptions beside them; it is panel A of Extended Data Figure 9.	All observations of this card concern the RT of one genus within the range scanned, the K family is a post hoc definition so that several comparisons may be circular, nearest-class readings are directions and not membership, and no wet-laboratory result is reported.
P02	The two nearest neighbours of the RT belong to groups for which the literature reports toxicity in cells lacking different host nucleases, and the K family is not monophyletic in the RT-domain tree	main text (open question 2)	2	Extended Data Figure 9, panel A	9	suggestive, post hoc 6; no information 2; negative result 1	Structure points to UG17, for which the literature reports DRT10 toxicity in cells lacking SbcB (Exol), and tree topology to UG3, for which it reports DRT3 toxicity in cells lacking RecBCD, so the observations of this work cannot choose between them and it is open whose logic, if any, the cassette follows.	3	(1) The cheapest computational step is to classify K-positive genomes by RT lineage type (DRT10-type K core, cassette lineage, other K) and to count sbcB and recBCD carriage with the same caller for each type. (2) The caller must be calibrated first, because its RecB and RecD calls are not calibrated and its SbcB sensitivity outside Pseudomonadota is unmeasured. (3) The decisive test is expression in Escherichia coli deletion strains with a catalytic-site variant control; it belongs to the wet-laboratory work under way, and no result is reported here.	No computational step of this test has been run, and the conjectures above rest on the literature and not on our data.	(1) The figure idea is a schematic with UG3 (first neighbour in sequence trees; literature: DRT3 toxic in cells lacking RecBCD) on one side, UG17 (nearest in structure; literature: DRT10 toxic in cells lacking SbcB, that is Exol) on the other, and the cassette RT between them. (2) Below them stands the RecD-like gene beside the RT, which is not counted as a cassette module, and the schematic is part of panel A of Extended Data Figure 9.	Structure-based nearest classes are not specific against a null of random bacterial RTs, tree positions are saturated, K is a post hoc definition, the cassette observations lie within one genus, and no wet-laboratory result is reported.

Supplementary Note S2: Dossier of the open questions

question_id	title	placement	main_text_question	figure	n_observations	levels_of_support	why_it_is_open	n_hypotheses	cheapest_test	test_status	figure_idea	limits_carried
P03	Free -1/+3-type gene pairs recur in many genera without an RT and element-bearing K-family RTs occur in further genera, but the combination of cassette parts with the K-lineage RT is found in one genus within the scanned range	main text (open question 3)	3	Extended Data Figure 9, panel B	12	limit 4; suggestive, post hoc 3; negative result 2; descriptive only 1; no information 1; supported 1	Parts recur across genera and element-bearing RTs occur in further genera, yet the combination of cassette parts with the K-lineage RT is found in one genus, and whether this reflects sampling and detector reach or a joining event that happened once cannot be separated by the present data.	3	(1) The cheapest computational steps are to rebuild the pair model from seeds independent of the cassette and repeat the genome-wide scan, and to check whether free pairs of genera outside Pollutibacter lie at the cassette position between xseA and recD. (2) The decisive step is a sensitivity-calibrated scan of catalogues not yet scanned, which would bring new input and is a follow-up that remains open.	(1) The -1 tree is untestable at a median outgroup-ingroup distance of 6.53 substitutions per site, and the result depends on how the outgroups were chosen. (2) The +3 tree is untestable at 6.30 substitutions per site, and the AU test of its closest alternative root lies just above the significance threshold, so the polarity of the pair is untested.	The figure idea is an existence map with three layers of colour on a Bacteroidota genus tree, pair genera, element-bearing K-family RT genera and the cassette genus, with overlaps joined by arcs; it is panel B of Extended Data Figure 9.	The cassette observations lie within one genus in the range scanned and the detectors derive from the cassette.
P05	A +3 protein whose N-terminal domain is predicted metzincin-like lacks the canonical HExxH motif in every ORF examined, while the Peptidase_M54-like gene beside the cassette keeps it, and there is no evidence that the N region is under weaker purifying selection than the rest	main text (open question 4)	4	not drawn in Extended Data Figure 9, set out in this note	16	descriptive only 8; suggestive, post hoc 4; limit 3; no information 1	A fold that predicts a protease lacks the canonical catalytic signature in every cassette ORF and in nearly all free +3-type sequences (2 of 374 carry it), and there is no evidence that its N region is more relaxed than the rest, so it is open what the N domain is kept for.	3	(1) The cheapest next computation is to repeat the N/C decomposition and its boundary sensitivity on the free +3-type family, which raises power, with the HExxH-bearing M54 flank ORFs as a positive control, and the FEL false-positive calibration and the stratification by information content, not done so far, would then mark the analysis as post hoc. (2) The decisive test is a purified N-domain activity assay with an active control and a -1 binding test; it belongs to the wet-laboratory work under way, and no result is reported here.	The N/C decomposition was run once, on 11 aligned sequences of one genus, as a post hoc analysis that has not been repeated on other sequences.	The figure idea shows the 354 alignment columns with the FEL negative sites, the N/C boundary at column 200 and the HExxH reference columns, above bars that contrast the free +3 HE...H count with the M54 count; Extended Data Figure 9 does not draw this question.	The ORFs are related sequences of one genus, FEL negative calls are not evidence about positive selection, and the family was defined from cassette sequences so that circularity cannot be excluded.
P06	A palindromic motif with a CAAGCTTG core found upstream of free -1 genes in training genera is reproduced in held-out genera by the pre-specified rule, and its frequency varies with the +3 context	main text (open question 5)	5	Extended Data Figure 9, panel C	11	suggestive, post hoc 4; limit 3; no information 2; supported 2	A palindromic site recurs upstream of free -1 genes at a similar distance and its frequency follows the +3 context, but whether the +3 protein, a host regulator or the mobility of the pair locus explains it cannot be told from the present data.	3	(1) The cheapest computations are, first, to enlarge the catalogue so that more members of the same -1 lineage with and without a +3 neighbour can be compared, which the present contrast could not do because few close -1 members lack a +3 neighbour; second, to find the common neighbouring families of the genome-wide 14-bp copies; and third, to look for copies of the site outside pair boundaries. (2) The decisive test is an EMSA of purified +3 protein, its C and N domains, with wild-type and core-scrambled oligonucleotides and a validated positive control; it belongs to the wet-laboratory work under way, and no result is reported here.	None of these computations has been run beyond the analyses described above.	(1) The figure idea combines sequence logos of the site by genus, the distribution of the distance to the -1 start, the bars of the +3 context gradient and the cells for the cassette -1 genes and the element windows. (2) Panel C of Extended Data Figure 9 draws a logo of the site and the distribution of the distance to the -1 start, and the bars of the +3 context gradient and the cells for the cassette -1 genes and the element windows are not drawn in Extended Data Figure 9.	The motif analysis concerns one protein family, the genera are related, and the cassette observations lie within one genus within the range scanned.

Supplementary Note S2: Dossier of the open questions

question_id	title	placement	main_text_question	figure	n_observations	levels_of_support	why_it_is_open	n_hypotheses	cheapest_test	test_status	figure_idea	limits_carried
P09	The 3' end of the element lies in the RT start codon or directly before it in 7 of 16 source windows and within 5 nucleotides of it in 9 of them, both by post hoc definitions, and what kind of element it is remains open	main text (open question 6)	6	not drawn in Extended Data Figure 9, set out in this note	10	suggestive, post hoc 6; descriptive only 2; limit 1; no information 1	The end of a non-coding stretch that looks constrained lies at the RT start codon in some windows, and the present tests neither support nor exclude a structured RNA, a retror-type template or a DRT10-type ncRNA for it, so a separate RNA, the 5' region of the RT transcript and a remnant of an intron-like ancestor all remain open.	3	(1) The cheapest computation is to compare the constraint of RT-less element loci with that of elements beside an RT with the overlap masked, by the same method. (2) The decisive tests are 5' RACE and an RNA size fractionation; they belong to the wet-laboratory work under way, and no result is reported here.	Neither computation has been run.	The figure idea sets the coordinates of the element 3' end against the RT start codon in the source windows beside a constraint percentile profile of the element region across genomes of the genus, in one column; Extended Data Figure 9 does not draw this question.	The element models derive from the cassette and the constraint analysis is anchored on it, so the observations of this card may be circular, and they lie within one genus within the range scanned.
P07	The island lies at the same address between xseA and recD in related genomes, but how often that address is occupied by a large insert cannot be told from the present scan	supplementary card	not applicable	none (idea only)	5	no information 2; limit 1; negative result 1; suggestive, post hoc 1	The same island sits at the same address, yet the address is empty or rearranged in the non-carrier species that can be judged and a large insert is found in very few genomes, so it is open whether the address is selected or the island is one old insertion kept by descent, because an occupancy frequency cannot be measured with the present scan.	3	(1) The cheapest computation is to process the 30.3% of the representatives that the scan has not yet covered and to apply the same strict criterion to the recJ, sbcCD and recFOR neighbourhoods, comparing occupancy with a same-genome null. (2) A search for junction repeats gives information only after the boundary positive control, the point at which the present analysis failed, has been passed.	None of these computations has been run.	The figure idea is a mosaic gene map of the address in which empty, rearranged, island-carrying, CAZUSO01 insertion and RecBCD-arrangement genomes are aligned in one column, with a box for the count of 3 of 106,040.	The same address is compatible with vertical descent and with site-specific insertion, the strict class is a post hoc definition, and the observation concerns one genus within the range scanned.
P10	RecBCD genes accompany the RT in the genomes of one family, Chitinophagaceae, and are not detected in the carrier genomes of the other families with several carriers	supplementary card	not applicable	none (idea only)	10	descriptive only 3; suggestive, post hoc 3; no information 2; limit 1; negative result 1	RecBCD companionship is complete in the genomes of one family, not detected in several other families, among them Saprospiraceae with the genus BJGN01 that carries the element plus RT, and present in a few single-genome families, so whether the RT unit needs, tolerates or merely lands beside RecBCD genes is not resolved and an association is not evidence of need.	3	(1) The cheapest computation is to widen the K_E comparison to catalogues not yet scanned and to more genera, so that the number of informative genera rises to the pre-set minimum, with the same caller and pre-specified rules. (2) The RecD and RecB calls must be calibrated first. (3) The decisive test is a growth test in Escherichia coli deletion strains; it belongs to the wet-laboratory work under way, and no result is reported here.	None of these computations has been run.	The figure idea is a dot plot of the RecBCD-positive fraction per genus unit, K carriers against background, by family and with BJGN01 marked, beside a box that states the literature context: DRT3 toxic in cells lacking RecBCD and DRT10 toxic in cells lacking ExoI.	Insertion-site availability is not separated from the association, the independent units are few genera, and the cassette observations lie within one genus within the range scanned.
P04	The central block of the cassette is almost identical between a Canadian lake sample and a Chilean surface-water sample, while the -1 gene differs	supplementary card	not applicable	none (idea only)	8	suggestive, post hoc 5; descriptive only 1; limit 1; negative result 1	Four modules are almost identical between samples from two continents while -1 differs, and only the focal sample pair among 66 related sample pairs stands out in RT identity, so the history of the central block relative to -1 is open.	3	The cheapest computation needs little CPU: align the two contigs at nucleotide level and read the identity of coding and non-coding positions in windows fixed in advance at gene boundaries, which avoids the post hoc window centre used above, and count the junction support separately from the read pairs of the two samples; the host ortholog baseline exists.	This computation has not been run.	The figure idea is a map of the two sites with bars of the module identities, the median identity of host orthologs as a dotted line, and a scatter of the related sample pairs of cassette RT divergence against host divergence.	The focal sample pair is a post hoc selection, and the observation lies within one genus within the range scanned.

Supplementary Note S2: Dossier of the open questions

question_id	title	placement	main_text_question	figure	n_observations	levels_of_support	why_it_is_open	n_hypotheses	cheapest_test	test_status	figure_idea	limits_carried
P13	The cassette MAGs fall in one sub-clade of the genus, the hypothesis that this is a high-GC clade is not met by the pre-specified rule, and cassette-free relatives within it are barely known	supplementary card	not applicable	none (idea only)	7	no information 6; supported 1	Within one genus the cassette clumps in one sub-clade and the cassette-free relatives inside that sub-clade are barely known, so it cannot be told whether an ancestor gained the island or whether some members lost it, which repeats the single-genus bottleneck inside the genus.	3	The cheapest computation compares the junction sequences of the cassette MAGs with one another, which is possible now, and the next step is to obtain more high-quality non-carrier sister species inside the sub-clade by widening the catalogues, an input that is new, and to test the concordance of island gene trees with the core tree by AU tests with an outgroup, where polarity may remain unresolved.	None of these computations has been run.	The figure idea is a genus core tree painted with island presence and GC, and the GC distribution of the genus with the mean of the cassette MAGs marked, on one panel.	The independent unit is one genus, carriers are defined by similarity to the cassette RT, and no wet-laboratory result is reported.
P11	The +1 protein resembles a DNA-specificity subunit of type I restriction-modification systems only weakly and sits in genomes where the partner subunits are not detected	supplementary card	not applicable	none (idea only)	8	no information 4; suggestive, post hoc 2; limit 1; negative result 1	The +1 protein looks like a specificity subunit at a distance but sits in genomes in which no partner subunit is detected, so it is open whether it is an adaptor for a nuclease-dependent system or an unrelated protein that resembles one through the profiles.	2	(1) The cheapest computation needs little CPU: compare the distribution of variation along the +1 allele alignment with the positions of the recognition domains in an HsdS reference alignment, and repeat the post hoc contrast of module reference clusters between +1, +2 and +3 by a rule fixed in advance. (2) A second step is to build HsdS profiles from non-REBASE sources and repeat the retrieval of +1, and for +2 no further information is expected without a new input.	Neither computation has been run.	The figure idea overlays the HsdS reference recognition-domain intervals on the curve of variation along the +1 allele alignment, with bars for the identity and omega of each module on the same panel.	The observations concern one genus, being dark is a result of the annotation layers searched, and the HsdS resemblance of +1 is below its pre-specified threshold.
P08	Depth over the island region falls to a fraction of the preceding region in one contig in three samples, but the interpretation is open and the expected island-lacking junction reads are not seen	supplementary card	not applicable	none (idea only)	6	no information 3; suggestive, post hoc 3	The depth step falls to a fraction of its neighbours and does not appear in the closed MAGs, while the expected island-lacking junction reads are not seen, so whether the step reflects the island being absent from part of the population or a mapping artefact is open.	2	(1) The cheapest computation uses the reads of the same samples: remeasure the step by k-mer depth without a reference, comparing island-unique k-mers with the host-unique k-mers on both sides of the junction. (2) The expected number of junction-spanning reads is fixed by a rule before counting, which avoids the design flaw of the competitive mapping, and the denominator of island-free reads is widened with the long reads at the recD boundary already held.	None of these computations has been run.	The figure idea shows the read-depth profile around the junction in the samples with the step marked, the expected and observed junction reads as bars, and the island-to-flank depth ratios of the closed MAGs on a separate panel.	The window was chosen after the step was seen, the samples may share one structural artefact, a block contig must not be read as one uniform genome fragment, and the observation concerns one genus within the range scanned.
P12	Long-read DNA and RNA screens (reduced to a limitation statement)	demoted to a limitation statement	not applicable	none	2	no information 1; supported 1	not applicable	0	not applicable	not applicable	not applicable	This item was reduced to a limitation statement because its negative result carries no information and its interpretation would read as absence or expression, which the data do not support.

question_id	title	placement	main_text_question	figure	n_observations	levels_of_support	why_it_is_open	n_hypotheses	cheapest_test	test_status	figure_idea	limits_carried
P14	The pond time series (reduced to a limitation statement)	demoted to a limitation statement	not applicable	none	2	no information 1; suggestive, post hoc 1	not applicable	0	not applicable	not applicable	not applicable	This item was reduced to a limitation statement because its observations concern a single unit and carry no information about covariation, so it cannot be offered as an open question.

Table SI-19_references.tsv (3 rows x 5 columns)

citation_in_text	authors	year	title	venue
Wang et al. 2026	Wang M, Yoneyama K, Žedaveinytė R, Ishikawa J, Tang S, Le HC, Wiegand T, Ramirez JL, Nagahata N, Ma Y, Zhang DJ, Helmeczi E, Berisa M, Jovanovic M, Hiraizumi M, Yamashita K, Nishimasu H, Sternberg SH.	2026	Coordinated RNA- and protein-templated synthesis of double-stranded DNA by a dual reverse transcriptase immune system.	Cell, 189: 4997-5011.e11
Sánchez-Nieto et al. 2026	Sánchez-Nieto E, Wilkinson ME, Millán V, Wenardy C, Cabrerizo D, Martínez-Abarca F, Toro N.	2026	Host exonuclease SbcB and a phage-encoded SSB-like protein control activation of the DRT10 reverse transcriptase defense system.	bioRxiv (preprint)