

Supplementary Note S1

File identifier: SI-24

Description: Predictions and wet-laboratory tests under way (plans, not results). Provided as the PDF file Supplementary_Note_S1.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-24_Supplementary_Note_S1.zip.

Legend

Supplementary Note S1 lists the predictions of this work that observations could show to be wrong, with the observation that would refute each and the status of each test, and it sets out the wet-laboratory tests that are under way, and no wet-laboratory result is reported. The 27 predictions fall into three groups, 3 that were tested once under a rule fixed before the result, 13 that need further analysis or new input, and 11 that are wet-laboratory predictions whose tests are under way. Each prediction has a refutation condition and a condition under which the test returns no information, which is neither support nor refutation. The same content is given as one table of predictions with 27 rows and one table of experiments under way with 7 rows, together with the column dictionary and the list of references. The predictions are hypotheses and not results, the experiments are under way and no wet-laboratory result is reported here, and the experiments depend on strains, vectors, positive-control systems and sample DNA. For the predictions that were tested once, each card points to the place where the outcome is reported, and the outcomes are not repeated in this note. Within the range scanned by the detectors, the cassette itself has been found in one genus only. Several models used in this work were built from the cassette itself. In this note a pair means the $-1/+3$ gene pair, and the pair of samples from Canada and Chile is called the focal sample pair. This note does not show that any prediction is true, that any experiment has been done, or anything about the function or expression of the RT or the cassette, and it contains no sequence file.

README

Supplementary Note S1 predictions and wet-laboratory tests under way (supplement to the ESOL RT study)

Public edition of the supplementary note that lists the predictions of this work that observations could show to be wrong and the wet-laboratory tests under way (no result is reported here). The text files are tab-separated, UTF-8 without a byte-order mark, with one header line, no quoting and no empty cell. The tables contain the minus sign and the prime as Unicode characters, so they must be read as UTF-8 text.

Files

File	Content	Rows x columns	SHA-256
SI-24_plan_public_en.md	The note as text: introduction, a table of the predictions at a glance, one card for each prediction, the experiments under way, the limits and the references	462 lines	fe7e0c5350cba9f019ba7d09a95c00324e7f8d808a9fa0d48923142766255b5b
SI-24_predictions.tsv	One row for each prediction: refutation condition, condition for no information, test status, needs, linked predictions, limits	27 x 14	c56eea605bf87ff48cdf40fa5a8e205bf2360a9dd83c3c0cd2db3e2f656d6ed4
SI-24_experiments.tsv	One row for each experiment under way: design, readout, what it cannot show, needs	7 x 8	b0b44cf13a46107f0fb47ddfffb29598ddda15d1d17de91f13a081609f28e7892
SI-24_COLUMN_DICTIONARY.tsv	Definition, values, unit and blank rule of every column of the three tables	27 x 6	7280e83d8284e32841fffb95a60b71e5da0515d447a0a5f59c286f1e29c5194bf
SI-24_references.tsv	The works cited in the note, in author and year form	2 x 5	e8eba2bc4020332c85c15284fd006ee1c33df2f2a9369f90a575778c898afaab

File	Content	Rows x columns	SHA-256
SI-24_legend_public_en.md	The legend of the note	10 sentences	bcf12d442ee368b555f321de0106f2fda4030967b07e86a6a4f8935336af0a47

No cell is marked NA. The row count of the predictions table is the number of predictions and not a number of loci.

How to read the note

- Every prediction is a hypothesis and not a result, and no wet-laboratory result is reported in this paper.
- A prediction is written with the observation that would refute it and the conditions under which its test returns no information.
- Predictions in group A were tested once under a rule fixed before the result, and this note gives their status and points to the place where the outcome and its limits are reported; a status sentence says what a repeat would need, and what a repeat needs can show which requirement a test did not meet.
- Within the range scanned by the detectors, the cassette itself has been found in one genus only, and several models were built from the cassette itself; not every premise of a prediction concerns the cassette itself, and each card states its own premise.
- The entry on linked predictions of each card names the main predictions that share a premise or a test with it, and the list is not exhaustive.
- The experiments depend on strains, vectors, positive-control systems and sample DNA, and no result of them is part of this paper.
- No sequence file is part of this item.

Plan

Supplementary Note S1. Falsifiable predictions and the wet-laboratory tests under way (no result is reported here)

Tab-separated tables with the same content accompany this text.

This supplementary note lists the predictions of this work that observations could show to be wrong, the observation that would refute each and the status of each test. It also sets out the wet-laboratory tests that are under way, and no wet-laboratory result is reported in this paper. Every prediction is a hypothesis and not a result, and no sentence of this note reports an experimental result.

Each prediction is written with its refutation condition before the test that decides it, and a test that cannot decide returns no information, which is neither support nor refutation. A prediction that is refuted lowers the statement that it extends. The entry on linked predictions of each card names the main predictions that share a premise or a test with it, and the list is not exhaustive.

The 27 predictions, numbered FP01 to FP27, fall into three groups: 3 that were tested once under a rule fixed before the result, 13 that need further analysis or new input, and 11 that need wet-laboratory work. For a prediction that was tested once, this note gives the status and points to the place where the outcome and its limits are reported, and it does not repeat the outcome. A status sentence says what a repeat would need, and what a repeat needs can show which requirement a test did not meet, but the outcome itself is reported only where the card points.

Within the range scanned by the detectors, the cassette itself has been found in one genus only. The premises of the predictions are stated on their cards, and not every premise concerns the cassette itself: some concern free families outside the genus, long-read runs, time series or the order in which parts arose. Several models and definitions were built from the cassette itself and are circular in the sense that they recover what they were built from.

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. The K family is a set of RTs defined after the fact in RT-domain trees, and the cassette RTs belong to it. The neighbouring families are the -1 , $+1$, $+2$ and $+3$ families, a pair means the $-1/+3$ gene pair, and the pair of samples from Canada and Chile is called the focal sample pair.

The experiments depend on strains, vectors, positive-control systems and sample DNA, and no result of them is part of this paper. No sequence file is part of this note.

Predictions at a glance

ID	Title	Group	Test status
FP01	The missing family assignment of the RT reflects its lineage and not the reach of the classifiers	A	Tested once under a rule fixed before the result
FP02	The RT and its neighbouring families share one history within the genus	A	Tested once under a rule fixed before the result
FP03	The cassette of the Beijing culture series lies on a host genome contig	A	Tested once under a rule fixed before the result
FP04	The RT lies on a branch beside UG3 and UG17	B	Not testable at present
FP05	Disruption stays rare in new copies of the neighbouring families	B	Planned; not run
FP06	The detector of the census recovers lesions in host sequence	B	Planned; not completed
FP07	The constraint on the neighbouring families was exerted inside the genus	B	Planned; not run
FP08	Minor alleles with lesions are rare in the reads of the source runs	B	Planned; not completed
FP09	Cassette reads recur in host-rich long-read runs of further studies	B	Planned; not run
FP10	The same cassette-bearing clone is recovered in more than one year	B	Planned; not run
FP11	The divergence of the cassette tracks the divergence of the host	B	Planned; not run
FP12	The combination of cassette parts and the K-lineage RT recurs in a further genus	B	Planned; not run
FP13	The -1 and +3 genes were an adjacent pair before the RT-bearing segment lay between them	B	Not testable at present
FP14	The N-terminal region of the +3 protein is not under weaker purifying selection than the rest	B	Tested once with low power; repeat planned
FP15	Free -1/+3 pairs outside the genus are maintained as coding units	B	Planned; not run
FP16	Constraint concentrates at the catalytic box and the motif columns of the RT	B	Planned; not run
FP17	Any phenotype of the cassette depends on the catalytic box of the RT	C	Under way; no result reported
FP18	The five gene families act as one unit	C	Under way; no result reported
FP19	ExoI logic: a growth defect only when ExoI is absent	C	Under way; no result reported
FP20	RecBCD logic: a growth defect when RecBCD is absent or inhibited	C	Under way; no result reported
FP21	Mosaic reading: no deletion strain changes a validated cassette phenotype	C	Under way; no result reported
FP22	The N-terminal domain of the +3 protein has no zinc-dependent protease activity	C	Under way; no result reported
FP23	The MIDAS motif of the -1 protein is needed for binding the +3 protein in free pairs	C	Under way; no result reported
FP24	The +3 protein binds the palindromic motif found upstream of free -1 genes	C	Under way; no result reported
FP25	The element is the 5' region of the RT transcript	C	Under way; no result reported
FP26	The element serves as a template or scaffold for RT-made cDNA	C	Under way; no result reported
FP27	The module junctions and the flanking junctions are present in the source DNA	C	Under way; no result reported

Group A. Predictions tested once under a rule fixed before the result

Each prediction of this group was put once to a rule that was fixed before any result was seen, and this note gives only its status and the place where the outcome is reported.

FP01. The missing family assignment of the RT reflects its lineage and not the reach of the classifiers

Family: classification. Test status: Tested once under a rule fixed before the result.

- **Basis.** The classifier DefenseFinder does not call the cassette RTs, its models are biased toward Proteobacteria with unconfirmed sensitivity in Bacteroidota, and the position of the RT among named lineages is unresolved. By structure the nearest named class of the RT is UG17, which is a direction and not a membership.
- **Prediction.** Known UG17 and UG3 members that are held out of the profile-defining sets and matched to the cassette RT in distance from their nearest reference are called by DefenseFinder and PADLOC clearly more often than the

cassette RTs are, so that the missing family assignment reflects the lineage of the RT and not the reach of the classifiers.

- **Refutation.** The prediction is refuted if the matched known members are not called significantly more often than the cassette RTs by the tools as shipped (Fisher exact P at or above 0.05). The prediction is refuted if profiles rebuilt with the members held out do not call the matched known members significantly more often than the cassette RTs (P at or above 0.05).
- **No information.** The test returns no information if the tools cannot be run, fewer than a pre-set minimum number of matched units exist, or only one of the two comparisons, with the cassette RTs and with random RTs, is significant.
- **Status.** This prediction was tested once under a rule fixed before the result, and the outcome and its limits are given in Supplementary Note S2, card P01. Matching used identity alone. A version matched also on alignment score and coverage, a version for the Bacteroidota-associated subtype of UG17 and a run of the shipped tools on held-out members remain open.
- **Linked predictions.** None.
- **Limits.** All cassette RTs come from one genus; the comparison is an emulation inside that genus and not a run of the shipped tools on held-out members; P values computed over units assume independent units.

FP02. The RT and its neighbouring families share one history within the genus

Family: co-inheritance. Test status: Tested once under a rule fixed before the result.

- **Basis.** The cassette RT clade is confined to one genus in GTDB release R232, and the loci analysed come from closely related genomes of that genus, which are not independent samples.
- **Prediction.** At the same loci, the gene trees of the RT and of each neighbouring family are more congruent with one another than with a null in which locus labels are permuted, and no supported topological conflict separates the RT tree from a neighbour tree.
- **Refutation.** The prediction is refuted if a supported branch of the RT tree conflicts with a supported branch of a neighbour tree, support meaning that the branch meets a pre-set branch-support rule, and the refutation then holds for that neighbouring family.
- **No information.** The test returns no information if fewer than a pre-set minimum number of units are shared between two trees, the trees are too poorly resolved to beat the null of label permutations at P at or below 0.05, or the positive control of the rule fails at the threshold used.
- **Status.** This prediction was tested once under a rule fixed before the result, and the outcome and its limits are given in the main text and in Extended Data Figure 13, panel C. A repeat needs a rule whose positive control passes at its threshold, a nucleotide-level screen for recombination breakpoints and additional close relatives of the loci.
- **Linked predictions.** None.
- **Limits.** The five trees re-use the same loci and are not independent samples; the effective independent unit is one genus.

FP03. The cassette of the Beijing culture series lies on a host genome contig

Family: culture. Test status: Tested once under a rule fixed before the result.

- **Basis.** Reads that match the cassette windows occur in a laboratory culture series of one study from Beijing (BioProject PRJNA1313158), which is a single study and a single culture lineage.
- **Prediction.** Reads of the DNA of the culture series assemble into a contig that holds the cassette modules together with single-copy host genes of a *Pollutibacter*-like bacterium.
- **Refutation.** The prediction is refuted if the cassette contig carries no host single-copy gene and carries only plasmid-type or phage-type genes, while the host marker genes assemble on other contigs.
- **No information.** The test returns no information if no cassette-positive contig reaches the pre-set contig-length floor, a cassette contig carries too few host genes, or it carries neither host genes nor plasmid-type or phage-type genes.
- **Status.** This prediction was tested once under a rule fixed before the result, and the outcome belongs to the results on the culture series and is not repeated in this note. A decisive version needs cassette contigs that reach the length floor of the rule, for which an isolate of the culture or a longer-read data set of the same series would serve.
- **Linked predictions.** None.
- **Limits.** The prediction concerns one culture series of one study, and the independent unit is one genus.

Group B. Predictions that need further analysis or new input

The predictions of this group are computational and have not been decided, and some need input data or a method that is not yet in hand.

FP04. The RT lies on a branch beside UG3 and UG17

Family: classification. Test status: Not testable at present.

- **Basis.** The position of the RT relative to UG3 and UG17 is unresolved because distances are saturated and the support for the junctions is weak, and a rejection of nesting within UG17 does not reproduce on the post hoc K family.
- **Prediction.** The cassette RT is not nested within UG17, including its Bacteroidota-associated subtype, or within UG3, but lies on a branch beside the pair, so that approximately unbiased (AU) tests on a saturation-reduced, structure-guided alignment with UG17 and UG3 references reject nesting within each group.
- **Refutation.** The prediction is refuted if the AU tests do not reject nesting within UG17 or within UG3 at P below 0.05, or maximum likelihood places the cassette RT inside the Bacteroidota-associated UG17 subtype with ultrafast bootstrap support at or above a pre-set level.
- **No information.** The test returns no information if the alignment remains saturated after reduction, or the labelled references of the UG17 subtype are not available.
- **Status.** Not testable at present, because the labelled references of the UG17 subtype are not among the inputs and the alignment is saturated. The related comparison of distance-matched held-out members was run once and is reported in Supplementary Note S2, card P01.
- **Linked predictions.** None.
- **Limits.** The prediction concerns the RT of one genus, and the K family is a post hoc definition.

FP05. Disruption stays rare in new copies of the neighbouring families

Family: lesions. Test status: Planned; not run.

- **Basis.** None of the 57 observable copies of the neighbouring families, which were located through the RT and not through their own sequence, carries a frameshift, premature stop or truncating deletion in the assembled contigs of 22 loci of one genus. With none of 45 allele clusters disrupted the two-sided 95% upper limit on the disruption rate is 7.9 per cent, and with none of 57 copies disrupted it is 6.3 per cent. Divergence saturation inflates the neutral expectation of a disruption, no host pseudogene baseline exists in these contigs, and the constraint is of intermediate strength.
- **Prediction.** Among new observable copies of the neighbouring families in cassette-bearing *Pollutibacter* genomes outside these loci, the fraction disrupted by a frameshift, premature stop or truncating deletion stays below 7.9 per cent.
- **Refutation.** The prediction is refuted if enough new observable copies are disrupted for an exact one-sided test to reject the unrounded upper limit of 7.9 per cent at a pre-set level.
- **No information.** The test returns no information if the host baseline of the same detector is also zero (FP06), because the census then cannot tell the absence of lesions from a floor of the detector.
- **Status.** Planned and not run, because the detector has not yet been applied to cassette-bearing genomes outside the first census, which are the first source of new copies.
- **Linked predictions.** FP06 tests the host baseline of the same detector, and a host baseline of zero is the condition under which this prediction returns no information.
- **Limits.** The copies belong to one genus, divergence saturation inflates the neutral expectation, the host baseline is not yet available, and the exact rule has low power for modest true rates.

FP06. The detector of the census recovers lesions in host sequence

Family: lesions. Test status: Planned; not completed.

- **Basis.** The detector of the census recovered the disruptions that were planted as a positive control, and no host pseudogene baseline was measured with it in the assembled contigs.
- **Prediction.** The same detector, applied to pseudogene-prone host sequence of the same assemblies, recovers at least one lesion, so that the lower bound of the 95% Clopper-Pearson interval of the lesion rate is above zero and a count of zero for the cassette is not a floor of the detector.
- **Refutation.** The prediction is refuted if the host baseline recovers no lesion, so that the lower bound of the interval is not above zero.
- **No information.** The test returns no information if too few host gene families are evaluable in the assemblies for a rate to be estimated.
- **Status.** Planned; the test has not been completed, and no output of it is used in this work.
- **Linked predictions.** FP05 returns no information if the host baseline of the same detector is also zero, which is the refutation condition of this prediction. FP08 also takes pseudogene-prone host sequence of the same assemblies as its

baseline, so that a host baseline that cannot be interpreted here bears on it as well.

- **Limits.** The baseline concerns assembled contigs of one genus and has no read-level support, and a single recovered lesion meets the rule without showing how large the host rate is.

FP07. The constraint on the neighbouring families was exerted inside the genus

Family: lesions. Test status: Planned; not run.

- **Basis.** The substitution-rate ratio omega of the cassette modules is far from the neutral value and higher than that of single-copy host orthologs, which is a constraint of intermediate strength.
- **Prediction.** If the constraint on the neighbouring families was exerted inside the genus, the divergence of these families between genomes of one close group follows the divergence of single-copy host genes of the same genomes.
- **Refutation.** The prediction is refuted if the divergence of the neighbouring families between genomes of one close group exceeds that of single-copy host genes of the same genomes, as expected if alleles that were already selected had been introduced from a donor population.
- **No information.** The test returns no information if divergence between the genomes is saturated, or the close groups hold too few cassette-bearing genomes for a comparison.
- **Status.** Planned and not run; it needs cassette-bearing genomes grouped by close relationship and a divergence measure that is not saturated.
- **Linked predictions.** None.
- **Limits.** Cassette divergence and host divergence both come from genomes of one genus.

FP08. Minor alleles with lesions are rare in the reads of the source runs

Family: lesions. Test status: Planned; not completed.

- **Prediction.** In a read-level census of the 16 source runs of the cassette MAGs, minor alleles that carry a frameshift or a premature stop in the neighbouring families are rarer than in pseudogene-prone host sequence of the same assemblies.
- **Refutation.** The prediction is refuted if the fraction of lesion-carrying minor alleles in the neighbouring families is equal to or higher than in the host baseline.
- **No information.** The test returns no information if the read depth at the module positions stays below a pre-set floor, or the numbers of lesion-carrying minor alleles are too small for a comparison.
- **Status.** Planned; the test has not been completed, and the reads are taken from the public archive.
- **Linked predictions.** FP06 also takes pseudogene-prone host sequence of the same assemblies as its baseline, so that a host baseline that cannot be interpreted there bears on this prediction as well.
- **Limits.** Each run is one sample of one genus, and the census depends on the read depth of the cassette contigs.

FP09. Cassette reads recur in host-rich long-read runs of further studies

Family: reads. Test status: Planned; not run.

- **Basis.** Of the 265 runs of the long-read screen, which come from 61 studies, 44 are non-amplicon runs with at least the minimum number of host reads, from 17 studies, and 7 of these 44 runs, from 4 studies, carry at least one strict cassette read. A single study, the Texas river study (BioProject PRJNA1126728), supplies 4 of the 7 positive runs and 9 of the 44 host-rich runs, and the host-read threshold was set after the counts were known.
- **Prediction.** In host-rich non-amplicon freshwater long-read runs of studies other than the Texas river study that have not yet been screened, the per-run probability of at least one strict cassette read is not below 0.066, the lower 95% bound of 7 positive runs among 44.
- **Refutation.** The prediction is refuted if 44 consecutive host-rich runs from several studies other than the Texas river study carry no strict cassette read, which would reject a per-run probability of 0.066 or more at the 5% level.
- **No information.** The test returns no information if the runs are amplicon-like, hold fewer than the minimum number of host reads, come from the Texas river study, or come from studies that were already screened.
- **Status.** Planned and not run; it needs new run files from the public archive.
- **Linked predictions.** None.
- **Limits.** The bound is a lower limit for a per-run probability among runs selected after the fact and is not a prevalence; it treats the runs as independent although 4 of the 7 positive runs come from one study, and the same bound without that study is 0.018, whose refutation at the same level would take 165 consecutive runs; reads do not show the host genus; PacBio and sediment runs are not informative.

FP10. The same cassette-bearing clone is recovered in more than one year

Family: persistence. Test status: Planned; not run.

- **Basis.** A time series of one pond gave no information, because it holds few time points, a coarse spacing and a carrier host that is barely detected at the read cap of each run, so that no statement about persistence is made from it.
- **Prediction.** In multi-year series from one water body, the same cassette-bearing ANI95 clone is recovered in more than one collection year.
- **Refutation.** The prediction is not supported if cassette-bearing clones occur only at single time points although the clone itself is recovered at several time points.
- **No information.** The test returns no information if no series records the same clone in more than one year, or collection dates are missing or too coarse to separate years.
- **Status.** Planned and not run; it needs the collection date and the clone assignment of each cassette-bearing genome from multi-year series, and the dates of the source samples are compiled in Supplementary Table S9.
- **Linked predictions.** None.
- **Limits.** Collection dates are recorded to different precision, and cassette-bearing genomes are a sample selected for carrying the cassette.

FP11. The divergence of the cassette tracks the divergence of the host

Family: divergence. Test status: Planned; not run.

- **Basis.** Across 66 pairs of cassette-bearing genomes, the divergence of the RT is rank-correlated with the divergence of host markers (Spearman 0.929; bootstrap interval 0.696 to 0.977). Of these 66 pairs, only 1, the focal sample pair from Canada and Chile, stands out against the identities of host orthologs, with an RT identity of 0.9894 against a median amino-acid identity of 0.9296 among its host orthologs, and it was chosen after inspection.
- **Prediction.** In further pairs of cassette-bearing genomes, the divergence of the cassette tracks the divergence of host markers irrespective of the continent of the samples, with the focal sample pair as the one known deviation.
- **Refutation.** The prediction is refuted if a pre-set number of further pairs have an RT identity of at least 98 per cent while their host-marker identity lies below a threshold that is fixed before any new pair is scored.
- **No information.** The test returns no information if too few further pairs with host-marker orthologs exist, or continent and clade stay confounded in the new pairs.
- **Status.** Planned and not run on further pairs; the threshold for the host-marker identity is fixed before new pairs are scored, with the focal pair as the reference case.
- **Linked predictions.** None.
- **Limits.** The rank correlation and the pair divergence come from the same sequences, partial ORF pairs break the rule fixed for the comparison, and continent is confounded with clade, so the correlation is not independent evidence.

FP12. The combination of cassette parts and the K-lineage RT recurs in a further genus

Family: arrangement. Test status: Planned; not run.

- **Basis.** Within the range scanned the combination of the cassette parts with an RT of the K lineage is found in one genus, and the detectors were built from the cassette, so their sensitivity for diverged loci is unproven.
- **Prediction.** A pre-specified, calibrated scan of public catalogues beyond the GTDB representatives finds in a genus other than *Pollutibacter* either a -1-type gene next to an RT of the K lineage or an RT-bearing segment between a -1-type and a +3-type gene, confirmed by *rpoB* and *gyrB* identity.
- **Refutation.** The prediction is refuted if a scan that meets its pre-set genome number and sensitivity thresholds finds no such locus. The opposite reading, that the combination arose once and stays within the genus, is refuted if such a locus is confirmed in another genus.
- **No information.** The test returns no information if the sensitivity of the scan at the identities of the target loci is not shown.
- **Status.** Planned and not run, because the simulated reach of the present detectors is low at the identities of the RTs outside the genus, so that a scan with them would be uninformative (Supplementary Note S2, card P03).
- **Linked predictions.** None.
- **Limits.** The detectors derive from the cassette, and genera that carry only one of the parts are not evidence about the combination.

FP13. The -1 and +3 genes were an adjacent pair before the RT-bearing segment lay between them

Family: arrangement. Test status: Not testable at present.

- **Basis.** The order in which the pair and the RT-bearing segment arose is unresolved, because the trees of the free families are unrooted and without support values and the test of RT degeneration has no positive control with a distant RT.
- **Prediction.** If the -1- and +3-type genes were an adjacent pair before the RT-bearing segment lay between them, the cassette -1 and +3 genes fall within the diversity of the free genes in outgroup-rooted trees with support values. In that case the pair also lies at the cassette insertion site in cassette-free *Pollutibacter* genomes, and remnants of the segment are not enriched at the junction beyond a control.
- **Refutation.** The prediction is refuted if the cassette -1 and +3 genes fall into different free lineages with support, or the pairs of cassette-free *Pollutibacter* genomes lie only elsewhere. The reading is weakened if remnants of the segment are enriched at the junction above the control (genus-level permutation P below a pre-set level).
- **No information.** The test returns no information if outgroup-rooted trees stay saturated, so that the root cannot be placed.
- **Status.** Not testable at present for the tree part, because outgroup-rooted trees of the -1 and +3 families are saturated (Supplementary Note S2, card P03), while the junction parts are planned. The reading assumes that the cassette -1 and +3 genes belong to the free pair families, an assumption that is nearly circular.
- **Linked predictions.** FP15 shares the assumption that the cassette -1 and +3 genes belong to the free pair families.
- **Limits.** The cassette -1 and +3 genes are assumed to belong to the free pair families, an assumption that is nearly circular.

FP14. The N-terminal region of the +3 protein is not under weaker purifying selection than the rest

Family: selection. Test status: Tested once with low power; repeat planned.

- **Basis.** The direction of selection on the +3 family is purifying, as for the other families, and the function of the protein is not determined.
- **Prediction.** In a repeat of the comparison on a larger alignment of free +3-type sequences, the N-terminal region does not carry a lower fraction of negatively selected codons than the C-terminal region. A decision rule written before the pre-specified re-analyses were run, with a margin set post hoc, decides whether the prediction is treated as not contradicted.
- **Refutation.** The prediction is refuted if the fraction of negatively selected codons is lower in the N-terminal region than in the C-terminal region with a two-sided Fisher exact P at or below 0.05, as the same decision rule requires.
- **No information.** The test returns no information if the sample is too small to tell a weaker N-terminal region from none, or selection strength and rate heterogeneity cannot be separated.
- **Status.** The comparison was run once on the alignment of the cassette +3 sequences, with low power for small differences, and it is described in Supplementary Note S2, card P05; a repeat needs a larger alignment of free +3-type sequences.
- **Linked predictions.** FP22 concerns the same N-terminal region, and a refutation of this prediction, which would show weaker purifying selection on that region, bears on the statement that the region is maintained.
- **Limits.** The comparison is post hoc, the boundary between the regions is arbitrary, and a negatively selected codon may reflect a structural constraint; the first design had low power for small differences.

FP15. Free -1/+3 pairs outside the genus are maintained as coding units

Family: selection. Test status: Planned; not run.

- **Basis.** The recurrence of RT-free -1 / +3 gene pairs is found in held-out genera of GTDB representatives as an existence statement, and it is circular because the pairs were first found with models built from the cassette.
- **Prediction.** In genera outside the cassette lineage that carry RT-free -1 / +3 pairs, the pairs are maintained as coding units and not carried as decaying remnants. Accordingly the other member of a pair stays adjacent more often than a null of adjacency loss along the genus tree allows, and the disrupted fraction of the pair lies below that of insertion-sequence ORFs in the same genomes. Both genes of the pair have an omega below a pre-set limit in the codon alignment of each genus, as a necessary condition only.
- **Refutation.** The prediction is refuted if in a genus the adjacency of the pair is not above the null of adjacency loss, or the disrupted fraction of the pair is at or above that of insertion-sequence ORFs, or omega is at or above a pre-set higher limit in a pre-set number of genera, and the unit reading is then dropped for that genus.

- **No information.** The test returns no information if the pair definition cannot be rebuilt from seeds independent of the cassette, or too few genera carry both an informative pair and a codon alignment with enough sequences.
- **Status.** Planned and not run; the thresholds are fixed before the result, and the same test is repeated after the pair definition is rebuilt from seeds independent of the cassette.
- **Linked predictions.** FP13 shares the assumption that the cassette -1 and +3 genes belong to the free pair families.
- **Limits.** The pair models derive from the cassette, the genera are not phylogenetically independent, and an omega below the threshold is met by most host core genes, so that it is only a necessary condition.

FP16. Constraint concentrates at the catalytic box and the motif columns of the RT

Family: selection. Test status: Planned; not run.

- **Basis.** The RT keeps its aromatic-x-Asp-Asp catalytic box in every evaluable ORF, but a retained box can also reflect a retained fold or retained binding, so that the observation alone does not separate these readings.
- **Prediction.** In the codon alignment of the RT, the fraction of negatively selected codons at pre-defined motif columns, the YxDD box and the RVT_1 motifs, is higher than at the remaining columns of the same alignment, judged against random column sets of the same size (P below 0.05, with an odds ratio above one).
- **Refutation.** The prediction is refuted if the fraction at the motif columns is equal to or lower than at the remaining columns, and the reading of constraint concentrated at the motifs is then dropped.
- **No information.** The test returns no information if too few variable motif columns exist for the permutation test to have power.
- **Status.** Planned and not run; it needs the set of motif columns to be defined before the test.
- **Linked predictions.** None.
- **Limits.** The test is a necessary condition for selection on catalysis and does not show catalytic activity, and the alignment comes from one genus.

Group C. Wet-laboratory predictions (tests under way; no result is reported)

The predictions of this group can be decided only by wet-laboratory work, which is under way, and no result of it is reported here.

FP17. Any phenotype of the cassette depends on the catalytic box of the RT

Family: wet-laboratory. Test status: Under way; no result reported.

- **Basis.** The RT keeps its aromatic-x-Asp-Asp catalytic box in every evaluable ORF, and the substitution-rate ratio of the RT lies far from the neutral value.
- **Prediction.** In a validated expression system, every phenotype that the wild-type cassette produces disappears when the catalytic box of the RT is altered from YADD to YAAD, so that the phenotype depends on RT catalytic activity.
- **Refutation.** The prediction is refuted if the box-altered variant keeps a phenotype of the wild-type cassette in a validated system, and the refutation then holds for that phenotype.
- **No information.** The test returns no information if the wild-type cassette produces no phenotype in any tested condition, or the expression system is not validated.
- **Status.** Under way; no wet-laboratory result is reported in this paper.
- **Linked predictions.** FP19, FP20, FP21 and FP26 assume that the phenotype or the product is lost with the YAAD variant, so that a refutation or a null result for this prediction bears on them. FP18, FP19, FP20 and FP21 use the same expression system as EX01. The reaction of the purified RT (EX02) is a separate test, and a negative result there is not a refutation condition of this prediction.
- **Limits.** A null result in a heterologous host does not show that the cassette has no function in its natural host.

FP18. The five gene families act as one unit

Family: wet-laboratory. Test status: Under way; no result reported.

- **Basis.** Where a contig spans the whole array, all families of the array are present and intact and no family is missing between observed neighbours, which does not show cooperation.
- **Prediction.** An in-frame deletion of any one of the -1, +1, +2 or +3 genes changes the phenotype of the cassette in a readout that is defined before the experiment.

- **Refutation.** The prediction is refuted if all deletion strains equal the wild type in the readout, and the unit reading is then dropped for that readout.
- **No information.** The test returns no information if the wild-type cassette gives no phenotype in the readout, or the deletions are not confirmed to be in frame.
- **Status.** Under way; no wet-laboratory result is reported in this paper.
- **Linked predictions.** FP17, FP19, FP20 and FP21 use the same expression system as EX01, so that a failed validation returns no information for all of them.
- **Limits.** A null result in a heterologous host does not show that the families do not cooperate in the natural host.

FP19. *ExoI* logic: a growth defect only when *ExoI* is absent

Family: wet-laboratory. Test status: Under way; no result reported.

- **Basis.** The literature on DRT systems links DRT3 to RecBCD, with toxicity in cells lacking RecBCD (Wang et al. 2026), and the pre-specified general test of a RecBCD-surveillance logic did not meet its rule on our data, so that the logics below are conjectures motivated by the literature and not by our data. For the *ExoI* logic, the cited preprint of Sánchez-Nieto et al. 2026 reports that no viable transformants carrying the DRT10 system were recovered in an *Escherichia coli* strain lacking *sbcB*, which describes the cited system and not the cassette.
- **Prediction.** Conjecture, motivated by the DRT literature (Sánchez-Nieto et al. 2026) and not by our data: the cassette RT follows an *ExoI* logic, so that in *Escherichia coli* the cassette causes a growth defect only when *ExoI* (*sbcB*) is absent and not when *recB*, *recC* or *recD* is absent, and the defect disappears with the YAAD variant.
- **Refutation.** The prediction is refuted if a defect appears in *recB*, *recC* or *recD* deletions, with Gam or gp5.9, or in the wild-type host in which *ExoI* is present, or validated expression gives no defect in any deletion strain.
- **No information.** The test returns no information if the expression system is not validated, or the deletion strains and the inhibitors are not confirmed to show the expected phenotypes.
- **Status.** Under way; no wet-laboratory result is reported in this paper.
- **Linked predictions.** FP17 bears on this prediction, because the defect is expected to be lost with the YAAD variant. FP18, FP20 and FP21 use the same strains as EX01, and FP20 and FP21 are the other readings that the same deletion strains distinguish.
- **Limits.** The test concerns *Escherichia coli* and not the natural host.

FP20. *RecBCD* logic: a growth defect when *RecBCD* is absent or inhibited

Family: wet-laboratory. Test status: Under way; no result reported.

- **Prediction.** Conjecture, motivated by the DRT literature (Wang et al. 2026) and not by our data: the cassette RT follows a *RecBCD* logic, with a growth defect when *RecBCD* is absent or inhibited (deletion of *recB*, *recC* or *recD*, or expression of Gam or gp5.9), not changed further by an *sbcB* deletion, and lost with the YAAD variant.
- **Refutation.** The prediction is refuted if no defect appears in *recB*, *recC* or *recD* deletions, with Gam or with gp5.9, or a defect appears only in the *sbcB* deletion.
- **No information.** The test returns no information if the expression system is not validated, or Gam and gp5.9 are not confirmed to inhibit *RecBCD* in the strains used.
- **Status.** Under way; no wet-laboratory result is reported in this paper.
- **Linked predictions.** FP17 bears on this prediction, because the defect is expected to be lost with the YAAD variant. FP18, FP19 and FP21 use the same strains as EX01, and FP19 and FP21 are the other readings that the same deletion strains distinguish.
- **Limits.** The test concerns *Escherichia coli* and not the natural host, and the *RecD*-like gene beside the cassette cannot be told from the host *RecD*.

FP21. *Mosaic* reading: no deletion strain changes a validated cassette phenotype

Family: wet-laboratory. Test status: Under way; no result reported.

- **Prediction.** Conjecture: the cassette lineage follows neither logic, so that no deletion strain changes a validated cassette phenotype.
- **Refutation.** The prediction is refuted if a defect that disappears with the YAAD variant appears in one deletion strain.
- **No information.** The test returns no information if the wild-type cassette gives no validated phenotype in any strain.
- **Status.** Under way; no wet-laboratory result is reported in this paper.

- **Linked predictions.** FP17 bears on this prediction, because a defect that is lost with the YAAD variant is part of its refutation condition. FP18, FP19 and FP20 use the same strains as EX01, and FP19 and FP20 are the two logics to which this reading is the alternative.
- **Limits.** The mosaic reading is the alternative to both logics and is not exclusive with them.

FP22. The N-terminal domain of the +3 protein has no zinc-dependent protease activity

Family: wet-laboratory. Test status: Under way; no result reported.

- **Prediction.** The purified N-terminal domain of the +3 protein shows no zinc-dependent proteolytic activity under conditions in which an active control of the same fold is active.
- **Refutation.** The prediction is refuted if the purified domain shows proteolysis that EDTA or o-phenanthroline inhibits, under conditions in which the active control is active.
- **No information.** The test returns no information if the active control is inactive or the domain cannot be purified in a folded state.
- **Status.** Under way; no wet-laboratory result is reported in this paper. The observations behind this prediction are given in Supplementary Note S2, card P05.
- **Linked predictions.** FP14 tests whether the same N-terminal region is under weaker purifying selection, and a refutation there bears on the statement that the region is maintained.
- **Limits.** A purified domain tested in vitro does not show what the domain does in the cell.

FP23. The MIDAS motif of the -1 protein is needed for binding the +3 protein in free pairs

Family: wet-laboratory. Test status: Under way; no result reported.

- **Basis.** Every evaluable -1 ORF carries a MIDAS motif by structure mapping, in a family defined from the cassette itself, and no metal binding is shown.
- **Prediction.** In RT-free adjacent pairs of a -1-type and a +3-type gene from several orders, the wild-type pair binds in a bacterial two-hybrid assay and in a pull-down that both pass their positive controls, a MIDAS-null -1 protein binds the +3 protein less, and non-cognate combinations bind less than cognate pairs.
- **Refutation.** The prediction is refuted if an expressed and folded MIDAS-null -1 protein binds the +3 protein as well as the wild-type -1 protein in both assays, the 90% interval of the signal ratio lying within a pre-set margin. The premise of binding is rejected in this system if a wild-type pair does not bind under positive controls that pass, which is not a conclusion about the natural host.
- **No information.** The test returns no information if the proteins are not expressed or folded, or the positive controls of the assays fail.
- **Status.** Under way; no wet-laboratory result is reported in this paper.
- **Linked predictions.** None.
- **Limits.** The prediction concerns adjacent free pairs and does not claim that +3 proteins require -1 proteins at the family level, and carrying it over to the cassette genes, which are separated by the RT and two further genes, is a separate step.

FP24. The +3 protein binds the palindromic motif found upstream of free -1 genes

Family: wet-laboratory. Test status: Under way; no result reported.

- **Prediction.** The +3 protein binds a DNA fragment that carries the core of the palindromic motif found upstream of free -1 genes, and the binding depends on the core sequence.
- **Refutation.** The prediction is refuted if an electrophoretic mobility shift assay with a validated positive control shows no core-dependent binding of the +3 protein. The prediction is weakened if members of the same -1 lineage that lack the +3 gene keep the site.
- **No information.** The test returns no information if the +3 protein cannot be purified in a folded state or the positive control does not bind.
- **Status.** Under way; no wet-laboratory result is reported in this paper. The observations behind this prediction are given in Supplementary Note S2, card P06.
- **Linked predictions.** None.
- **Limits.** The motif is a candidate recognition site that has not been tested, and binding in vitro would not show a function in the cell.

FP25. The element is the 5' region of the RT transcript*Family: wet-laboratory. Test status: Under way; no result reported.*

- **Basis.** The 3' end of the element abuts the RT start codon in a minority of the source windows, and the kind of element it is remains open.
- **Prediction.** The upstream element is the 5' region of the RT transcript: a transcription start lies at or within the 5' end of the element, and no stable element-length RNA with its own ends accumulates apart from the RT mRNA.
- **Refutation.** The prediction is refuted if a stable element-length RNA with its own 5' and 3' ends accumulates apart from the RT mRNA, as shown by 5' RACE and RNA size fractionation.
- **No information.** The test returns no information if the RNA of the construct is not detected, or the promoter of the vector alone accounts for the start sites in the host used.
- **Status.** Under way; no wet-laboratory result is reported in this paper.
- **Linked predictions.** FP26 states the competing reading of the element, a separate RNA used as template or scaffold, so that a result for one of the two predictions bears on how the other is read.
- **Limits.** A start site found in a heterologous host need not be the start site in the natural host.

FP26. The element serves as a template or scaffold for RT-made cDNA*Family: wet-laboratory. Test status: Under way; no result reported.*

- **Prediction.** The element is a separate RNA used as template or scaffold: the active RT, and not the YAAD variant, makes element-derived cDNA in a validated system.
- **Refutation.** The prediction is refuted if the active RT does not make more element-derived cDNA than the YAAD variant in a validated system.
- **No information.** The test returns no information if the system is not validated or the positive-control RT is inactive under the conditions used.
- **Status.** Under way; no wet-laboratory result is reported in this paper. The observations behind this prediction are given in Supplementary Note S2, card P09.
- **Linked predictions.** FP17 bears on this prediction, because both tests use the YAAD variant as the box-altered control. FP25 states the competing reading of the element, the 5' region of the RT transcript with no stable element-length RNA apart from the RT mRNA, so that a result for one of the two predictions bears on how the other is read.
- **Limits.** In vitro synthesis shows what the RT can do with an RNA and not what it does in the cell.

FP27. The module junctions and the flanking junctions are present in the source DNA*Family: wet-laboratory. Test status: Under way; no result reported.*

- **Prediction.** On DNA from the habitats of the source samples, PCR with primers that span the module junctions and the junction between the cassette and its flanking host genes gives amplicons of the sizes predicted from the assemblies, and their sequences match the assemblies.
- **Refutation.** The prediction is refuted if a junction gives no product of the predicted size, or a product whose sequence differs from the assembly at the junction, with DNA of the source habitat in a reaction whose synthetic positive control amplifies.
- **No information.** The test returns no information if sample DNA is not available, the primers do not match the sequences of the population because the flanks diverge, or the positive control fails.
- **Status.** Under way; no wet-laboratory result is reported in this paper. The observations behind this prediction are given in Supplementary Note S2, cards P07 and P08.
- **Linked predictions.** None.
- **Limits.** The assemblies come from metagenomes, and the sample DNA must come from the same habitat.

Experiments under way

- The experiments under way are described here at the level of design, and they depend on strains, vectors, positive-control systems, sample DNA and purified proteins; no result of them is part of this paper.

ID	Experiment	Bears on
EX01	Expression of the cassette in Escherichia coli deletion strains with a catalytic-box variant of the RT	FP17, FP18, FP19, FP20, FP21

ID	Experiment	Bears on
EX02	In vitro reaction of the purified RT with element RNA and analysis of the cDNA	FP26
EX03	5' RACE, read-through RT-PCR and RNA size fractionation of the RT transcript	FP25
EX04	Electrophoretic mobility shift assay of the +3 protein with the palindromic motif	FP24
EX05	Activity assay of the purified +3 N-terminal domain	FP22
EX06	Junction PCR and amplicon sequencing on DNA of the source habitats	FP27
EX07	Bacterial two-hybrid and pull-down assays of free -1/+3 pairs with a MIDAS variant	FP23

EX01. Expression of the cassette in *Escherichia coli* deletion strains with a catalytic-box variant of the RT

Bears on: FP17, FP18, FP19, FP20, FP21. Status: under way; no result reported.

- **Design.** The cassette core, from the -1 gene to the +3 gene and including the element, is expressed from an arabinose-inducible plasmid in the wild-type strain BW25113 and in strains deleted for recB, recC, recD or sbcB, with and without Gam or gp5.9. Each condition is run with the wild-type RT and with the YAAD variant and against an empty vector. The design uses two independently assembled plasmid clones, several independent transformants or cultures, and two alleles from different source genomes of the genus. In-frame deletions of the -1, +1, +2 or +3 gene are tested one at a time in the conditions that give a phenotype.
- **Readout.** Growth is followed as the area under the growth curve over a fixed time window and as colony-forming units, against the empty vector under the same induction.
- **What it cannot show.** The experiment does not show whether the cassette has a function in its natural host, or the identity of an effector.
- **Needs.** The experiment needs synthesis of the cassette core and its variants from sequence files, strains with the deletions, vectors and positive-control systems.

EX02. In vitro reaction of the purified RT with element RNA and analysis of the cDNA

Bears on: FP26. Status: under way; no result reported.

- **Design.** The purified RT, wild-type and YAAD variant, is incubated with no RNA, with the element RNA, with two shuffled element RNAs and with the reverse complement, with and without a primer, in two metal conditions and over a time course.
- **Readout.** Products are analysed on denaturing gels and by cDNA sequencing, and the repeat unit of the products is determined from the sequence reads.
- **What it cannot show.** The reaction shows what the RT can do with an RNA and not what it does in the cell. A negative result of the reaction would withdraw the in vitro prediction only and would not refute FP17, because partner proteins or the element may be required in the cell.
- **Needs.** The experiment needs a purified RT and element RNA, reaction conditions from the literature that have not been confirmed, and a positive-control RT.

EX03. 5' RACE, read-through RT-PCR and RNA size fractionation of the RT transcript

Bears on: FP25. Status: under way; no result reported.

- **Design.** Gene-specific primers inside the RT gene and the element are used for 5' RACE with and without treatments that distinguish primary 5' ends, and read-through RT-PCR across the junctions between adjacent genes is run with reactions without reverse transcriptase as controls. RNA is separated by size and element-derived RNA is looked for in the fractions, with uninduced or promoterless controls or a Bacteroidota host, because the vector promoter is read first in *Escherichia coli*.
- **Readout.** The position of the transcription start relative to the element, the continuity of the transcript across gene junctions, and the presence of an element-length RNA with its own ends are read.
- **What it cannot show.** The experiment does not show the role of the element, or the start sites in the natural host.
- **Needs.** The experiment needs constructs with the element, RNA from an expression host, and a host in which the natural start sites can be read.

EX04. Electrophoretic mobility shift assay of the +3 protein with the palindromic motif

Bears on: FP24. Status: under way; no result reported.

- **Design.** The purified +3 protein is incubated with labelled DNA that carries the palindromic motif, with DNA in which the core is mutated and with a shuffled control, with competition by unlabelled DNA and with a protein-DNA pair that binds as a positive control.
- **Readout.** The shift of the labelled DNA by the +3 protein and its dependence on the core sequence are read.
- **What it cannot show.** The assay does not show a function in the cell, or that the +3 protein is the protein that reads the site in the cell.
- **Needs.** The experiment needs a purified +3 protein, oligonucleotides and a positive-control protein-DNA pair.

EX05. Activity assay of the purified +3 N-terminal domain

Bears on: FP22. Status: under way; no result reported.

- **Design.** The purified N-terminal domain and an active control of the same fold are incubated with generic protease substrates, with and without EDTA and o-phenanthroline.
- **Readout.** Cleavage of the substrates and its inhibition by metal chelators are read.
- **What it cannot show.** The assay does not show what the domain does, or whether it has another activity.
- **Needs.** The experiment needs a purified domain, an active control of the same fold and substrates.

EX06. Junction PCR and amplicon sequencing on DNA of the source habitats

Bears on: FP27. Status: under way; no result reported.

- **Design.** Primers span the module junctions and the junctions between the cassette and its flanking host genes, designed from the assemblies of two source genomes, with a synthetic target as positive control. Amplicons are read by long-read sequencing, and occupied and empty sites are compared relatively by quantitative PCR.
- **Readout.** The amplicon size is compared with the size predicted in silico, and the sequence of the amplicon is compared with the assembly at the junction.
- **What it cannot show.** The assay does not give an absolute occupancy, because quantitative PCR favours short products and only relative comparisons are made, and a failure to amplify does not show absence of the sequence when the flanks diverge.
- **Needs.** The experiment needs DNA from the habitats of the source samples, primers and a synthetic target.

EX07. Bacterial two-hybrid and pull-down assays of free -1/+3 pairs with a MIDAS variant

Bears on: FP23. Status: under way; no result reported.

- **Design.** Wild-type and MIDAS-null -1 proteins are tested against the +3 protein of the same pair in a bacterial two-hybrid assay and in a pull-down with mass spectrometry, with cognate and non-cognate combinations, with empty-vector and positive-control pairs, and with fusions at both termini.
- **Readout.** The signal of each combination is read relative to an empty-vector control and to a positive-control pair.
- **What it cannot show.** The assays do not show binding in the natural host, or a function of the complex.
- **Needs.** The experiment needs -1 and +3 genes of free pairs from several orders, a MIDAS variant, two-hybrid and pull-down systems, and positive-control pairs.

What this note does not show

This note does not show that any prediction is true, that any experiment has been done, or anything about the function or expression of the RT or the cassette. A prediction that fails would lower the statement that it extends, and where predictions share a premise or a test, the failure or the null result of one can bear on the others, as the entries on linked predictions say for the main cases.

References

- 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, Helmecci 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.
- 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).

Table SI-24_COLUMN_DICTIONARY.tsv (28 rows x 6 columns)

table	column	definition	values_or_format	unit	blank_rule
SI-24_predictions.tsv	prediction_id	Identifier of the prediction, FP01 to FP27.	FP followed by two digits	none (identifier)	never blank
SI-24_predictions.tsv	group	A = tested once under a rule fixed before the result; B = needs further analysis or new input; C = wet-laboratory test (under way).	A, B or C	none (code letter)	never blank
SI-24_predictions.tsv	family	Topic of the prediction.	classification, co-inheritance, culture, lesions, reads, persistence, divergence, arrangement, selection or wet-laboratory	none (label)	never blank
SI-24_predictions.tsv	title	Short title of the prediction.	text	none (text)	never blank
SI-24_predictions.tsv	basis	The observation that the prediction extends, in short form, with the level of support given in the text of the note.	text	none (text; a count or a bound carries its own unit in the text)	never blank; text that starts with See points to the card or note that gives the observation, and text that starts with No premise observation says that none exists
SI-24_predictions.tsv	prediction	The prediction, written as a statement that data or an experiment can contradict.	text	none (text)	never blank
SI-24_predictions.tsv	refuted_if	The observation that refutes the prediction (or, for one prediction, shows that it is not supported).	text	none (text)	never blank
SI-24_predictions.tsv	returns_no_information_if	The conditions under which the test cannot decide, which is neither support nor refutation.	text	none (text)	never blank
SI-24_predictions.tsv	test_status	Status of the test, one of the controlled phrases listed in the values column. Outcomes of tests already run are not given here.	Tested once under a rule fixed before the result Not testable at present Planned; not run Planned; not completed Tested once with low power; repeat planned Under way; no result reported	none (controlled phrase)	never blank
SI-24_predictions.tsv	status_note	What is open and where the outcome and its limits are reported.	text	none (text)	never blank
SI-24_predictions.tsv	needs	Inputs, data, materials or methods that the test needs and that are not yet in hand.	text	none (text)	never blank
SI-24_predictions.tsv	related	Cards, notes, figures and experiments that belong to the prediction.	text	none (text)	never blank
SI-24_predictions.tsv	linked_predictions	The main other predictions that share a premise or a test with this one, so that a refutation or a null result of one can bear on the other; the list is not exhaustive.	text	none (text)	never blank; the word None is written where no such link has been identified
SI-24_predictions.tsv	limits_carried	Limits that travel with the prediction.	text	none (text)	never blank
SI-24_experiments.tsv	experiment_id	Identifier of the experiment under way, EX01 to EX07.	EX followed by two digits	none (identifier)	never blank
SI-24_experiments.tsv	name	Name of the experiment.	text	none (text)	never blank
SI-24_experiments.tsv	tests_predictions	Predictions of this note that the experiment can decide.	FP identifiers	none (identifier list)	never blank
SI-24_experiments.tsv	design	Material and conditions of the design.	text	none (text; a number carries its own unit in the text)	never blank
SI-24_experiments.tsv	readout	What is measured.	text	none (text)	never blank
SI-24_experiments.tsv	cannot_show	What the experiment cannot show even if it succeeds.	text	none (text)	never blank
SI-24_experiments.tsv	needs	Materials and access on which the experiment depends.	text	none (text)	never blank
SI-24_experiments.tsv	status	Under way; no result reported.	fixed phrase	none (fixed phrase)	never blank
SI-24_references.tsv	citation_in_text	The citation as it stands in the statements (author and year).	text	none (text)	never blank
SI-24_references.tsv	authors	Authors of the cited work.	text	none (text)	never blank

table	column	definition	values_or_format	unit	blank_rule
SI-24_references.tsv	year	Year of the cited work.	four digits	calendar year	never blank
SI-24_references.tsv	title	Title of the cited work, quoted from it.	text	none (text)	never blank
SI-24_references.tsv	venue	Journal or preprint server, with volume and pages where they exist.	text	none (text)	never blank

Table SI-24_experiments.tsv (8 rows x 8 columns)

experiment_id	name	tests_predictions	design	readout	cannot_show	needs	status
EX01	Expression of the cassette in Escherichia coli deletion strains with a catalytic-box variant of the RT	FP17, FP18, FP19, FP20, FP21	The cassette core, from the -1 gene to the +3 gene and including the element, is expressed from an arabinose-inducible plasmid in the wild-type strain BW25113 and in strains deleted for recB, recC, recD or sbcB, with and without Gam or gp5.9. Each condition is run with the wild-type RT and with the YAAD variant and against an empty vector. The design uses two independently assembled plasmid clones, several independent transformants or cultures, and two alleles from different source genomes of the genus. In-frame deletions of the -1, +1, +2 or +3 gene are tested one at a time in the conditions that give a phenotype.	Growth is followed as the area under the growth curve over a fixed time window and as colony-forming units, against the empty vector under the same induction.	The experiment does not show whether the cassette has a function in its natural host, or the identity of an effector.	The experiment needs synthesis of the cassette core and its variants from sequence files, strains with the deletions, vectors and positive-control systems.	Under way; no result reported.
EX02	In vitro reaction of the purified RT with element RNA and analysis of the cDNA	FP26	The purified RT, wild-type and YAAD variant, is incubated with no RNA, with the element RNA, with two shuffled element RNAs and with the reverse complement, with and without a primer, in two metal conditions and over a time course.	Products are analysed on denaturing gels and by cDNA sequencing, and the repeat unit of the products is determined from the sequence reads.	The reaction shows what the RT can do with an RNA and not what it does in the cell. A negative result of the reaction would withdraw the in vitro prediction only and would not refute FP17, because partner proteins or the element may be required in the cell.	The experiment needs a purified RT and element RNA, reaction conditions from the literature that have not been confirmed, and a positive-control RT.	Under way; no result reported.
EX03	5' RACE, read-through RT-PCR and RNA size fractionation of the RT transcript	FP25	Gene-specific primers inside the RT gene and the element are used for 5' RACE with and without treatments that distinguish primary 5' ends, and read-through RT-PCR across the junctions between adjacent genes is run with reactions without reverse transcriptase as controls. RNA is separated by size and element-derived RNA is looked for in the fractions, with uninduced or promoterless controls or a Bacteroidota host, because the vector promoter is read first in Escherichia coli.	The position of the transcription start relative to the element, the continuity of the transcript across gene junctions, and the presence of an element-length RNA with its own ends are read.	The experiment does not show the role of the element, or the start sites in the natural host.	The experiment needs constructs with the element, RNA from an expression host, and a host in which the natural start sites can be read.	Under way; no result reported.
EX04	Electrophoretic mobility shift assay of the +3 protein with the palindromic motif	FP24	The purified +3 protein is incubated with labelled DNA that carries the palindromic motif, with DNA in which the core is mutated and with a shuffled control, with competition by unlabelled DNA and with a protein-DNA pair that binds as a positive control.	The shift of the labelled DNA by the +3 protein and its dependence on the core sequence are read.	The assay does not show a function in the cell, or that the +3 protein is the protein that reads the site in the cell.	The experiment needs a purified +3 protein, oligonucleotides and a positive-control protein-DNA pair.	Under way; no result reported.
EX05	Activity assay of the purified +3 N-terminal domain	FP22	The purified N-terminal domain and an active control of the same fold are incubated with generic protease substrates, with and without EDTA and o-phenanthroline.	Cleavage of the substrates and its inhibition by metal chelators are read.	The assay does not show what the domain does, or whether it has another activity.	The experiment needs a purified domain, an active control of the same fold and substrates.	Under way; no result reported.
EX06	Junction PCR and amplicon sequencing on DNA of the source habitats	FP27	Primers span the module junctions and the junctions between the cassette and its flanking host genes, designed from the assemblies of two source genomes, with a synthetic target as positive control. Amplicons are read by long-read sequencing, and occupied and empty sites are compared relatively by quantitative PCR.	The amplicon size is compared with the size predicted in silico, and the sequence of the amplicon is compared with the assembly at the junction.	The assay does not give an absolute occupancy, because quantitative PCR favours short products and only relative comparisons are made, and a failure to amplify does not show absence of the sequence when the flanks diverge.	The experiment needs DNA from the habitats of the source samples, primers and a synthetic target.	Under way; no result reported.
EX07	Bacterial two-hybrid and pull-down assays of free -1/+3 pairs with a MIDAS variant	FP23	Wild-type and MIDAS-null -1 proteins are tested against the +3 protein of the same pair in a bacterial two-hybrid assay and in a pull-down with mass spectrometry, with cognate and non-cognate combinations, with empty-vector and positive-control pairs, and with fusions at both termini.	The signal of each combination is read relative to an empty-vector control and to a positive-control pair.	The assays do not show binding in the natural host, or a function of the complex.	The experiment needs -1 and +3 genes of free pairs from several orders, a MIDAS variant, two-hybrid and pull-down systems, and positive-control pairs.	Under way; no result reported.

Table SI-24_predictions.tsv (28 rows x 14 columns)

prediction_id	group	family	title	basis	prediction	refuted_if	returns_no_information_if	test_status	status_note	needs	related	linked_predictions	limits_carried
FP01	A	classification	The missing family assignment of the RT reflects its lineage and not the reach of the classifiers	The classifier does not call the cassette RTs, its models are biased toward Proteobacteria, and the RT position is unresolved.	Known UG17 and UG3 members that are held out of the profile-defining sets and matched to the cassette RT in distance from their nearest reference are called by DefenseFinder and PADLOC clearly more often than the cassette RTs are, so that the missing family assignment reflects the lineage of the RT and not the reach of the classifiers.	the matched known members are not called significantly more often than the cassette RTs by the tools as shipped (Fisher exact P at or above 0.05); profiles rebuilt with the members held out do not call the matched known members significantly more often than the cassette RTs (P at or above 0.05)	the tools cannot be run, fewer than a pre-set minimum number of matched units exist, or only one of the two comparisons, with the cassette RTs and with random RTs, is significant	Tested once under a rule fixed before the result	This prediction was tested once under a rule fixed before the result, and the outcome and its limits are given in Supplementary Note S2, card P01. Matching used identity alone. A version matched also on alignment score and coverage, a version for the Bacteroidota-associated subtype of UG17 and a run of the shipped tools on held-out members remain open.	Reference members at a greater distance than the present pool offers, the marker list of the UG17 subtype, and the training-set member lists of the tools.	Supplementary Note S2, card P01 (H1, H2); Extended Data Figure 9, panel A	None	All cassette RTs come from one genus; the comparison is an emulation inside that genus and not a run of the shipped tools on held-out members; P values computed over units assume independent units.
FP02	A	co-inheritance	The RT and its neighbouring families share one history within the genus	The cassette RT clade is confined to one genus, and the loci come from closely related genomes that are not independent samples.	At the same loci, the gene trees of the RT and of each neighbouring family are more congruent with one another than with a null in which locus labels are permuted, and no supported topological conflict separates the RT tree from a neighbour tree.	a supported branch of the RT tree conflicts with a supported branch of a neighbour tree, support meaning that the branch meets a pre-set branch-support rule, and the refutation then holds for that neighbouring family	fewer than a pre-set minimum number of units are shared between two trees, the trees are too poorly resolved to beat the null of label permutations at P at or below 0.05, or the positive control of the rule fails at the threshold used	Tested once under a rule fixed before the result	This prediction was tested once under a rule fixed before the result, and the outcome and its limits are given in the main text and in Extended Data Figure 13, panel C. A repeat needs a rule whose positive control passes at its threshold, a nucleotide-level screen for recombination breakpoints and additional close relatives of the loci.	A rule that passes its own positive control, a nucleotide-level breakpoint screen, and more close relatives of the loci.	Extended Data Figure 13, panel C; main text (Discussion)	None	The five trees re-use the same loci and are not independent samples; the effective independent unit is one genus.
FP03	A	culture	The cassette of the Beijing culture series lies on a host genome contig	Cassette-matching reads occur in the culture series of one Beijing study, a single study and a single culture lineage.	Reads of the DNA of the culture series assemble into a contig that holds the cassette modules together with single-copy host genes of a Pollutibacter-like bacterium.	the cassette contig carries no host single-copy gene and carries only plasmid-type or phage-type genes, while the host marker genes assemble on other contigs	no cassette-positive contig reaches the pre-set contig-length floor, a cassette contig carries too few host genes, or it carries neither host genes nor plasmid-type or phage-type genes	Tested once under a rule fixed before the result	This prediction was tested once under a rule fixed before the result, and the outcome belongs to the results on the culture series and is not repeated in this note. A decisive version needs cassette contigs that reach the length floor of the rule, for which an isolate of the culture or a longer-read data set of the same series would serve.	Cassette contigs of sufficient length from the culture series, which would need an isolate or longer reads of the same culture.	Culture-series analyses of this work	None	The prediction concerns one culture series of one study, and the independent unit is one genus.

Supplementary Note S1: Predictions and wet-laboratory tests under way (plans, not results)

prediction_id	group	family	title	basis	prediction	refuted_if	returns_no_information_if	test_status	status_note	needs	related	linked_predictions	limits_carried
FP04	B	classification	The RT lies on a branch beside UG3 and UG17	The RT position among UG3 and UG17 is unresolved: distances are saturated and junction support is weak.	The cassette RT is not nested within UG17, including its Bacteroidota-associated subtype, or within UG3, but lies on a branch beside the pair, so that approximately unbiased (AU) tests on a saturation-reduced, structure-guided alignment with UG17 and UG3 references reject nesting within each group.	the AU tests do not reject nesting within UG17 or within UG3 at P below 0.05, or maximum likelihood places the cassette RT inside the Bacteroidota-associated UG17 subtype with ultrafast bootstrap support at or above a pre-set level	the alignment remains saturated after reduction, or the labelled references of the UG17 subtype are not available	Not testable at present	Not testable at present, because the labelled references of the UG17 subtype are not among the inputs and the alignment is saturated. The related comparison of distance-matched held-out members was run once and is reported in Supplementary Note S2, card P01.	Labelled UG17 subtype and UG3 reference sequences, and a saturation-reduced, structure-guided alignment.	Supplementary Note S2, card P01 (H2)	None	The prediction concerns the RT of one genus, and the K family is a post hoc definition.
FP05	B	lesions	Disruption stays rare in new copies of the neighbouring families	None of the observable copies of the neighbouring families is disrupted in the assembled contigs of one genus; the upper limits follow from the counts.	Among new observable copies of the neighbouring families in cassette-bearing Pollutibacter genomes outside these loci, the fraction disrupted by a frameshift, premature stop or truncating deletion stays below 7.9 per cent.	enough new observable copies are disrupted for an exact one-sided test to reject the unrounded upper limit of 7.9 per cent at a pre-set level	the host baseline of the same detector is also zero (FP06), because the census then cannot tell the absence of lesions from a floor of the detector	Planned; not run	Planned and not run, because the detector has not yet been applied to cassette-bearing genomes outside the first census, which are the first source of new copies.	New observable copies of the neighbouring families from cassette-bearing genomes outside the first census.	Figure 6; Supplementary Data S1; Supplementary Table S6	FP06 tests the host baseline of the same detector, and a host baseline of zero is the condition under which this prediction returns no information.	The copies belong to one genus, divergence saturation inflates the neutral expectation, the host baseline is not yet available, and the exact rule has low power for modest true rates.
FP06	B	lesions	The detector of the census recovers lesions in host sequence	The detector recovered the disruptions that were planted as a positive control, and no host pseudogene baseline exists yet.	The same detector, applied to pseudogene-prone host sequence of the same assemblies, recovers at least one lesion, so that the lower bound of the 95% Clopper-Pearson interval of the lesion rate is above zero and a count of zero for the cassette is not a floor of the detector.	the host baseline recovers no lesion, so that the lower bound of the interval is not above zero	too few host gene families are evaluable in the assemblies for a rate to be estimated	Planned; not completed	Planned; the test has not been completed, and no output of it is used in this work.	A host baseline set of gene families with candidate lesions, analysed with the same detector and classification rules as the census.	Figure 6; Supplementary Data S1; Supplementary Table S6	FP05 returns no information if the host baseline of the same detector is also zero, which is the refutation condition of this prediction. FP08 also takes pseudogene-prone host sequence of the same assemblies as its baseline, so that a host baseline that cannot be interpreted here bears on it as well.	The baseline concerns assembled contigs of one genus and has no read-level support, and a single recovered lesion meets the rule without showing how large the host rate is.
FP07	B	lesions	The constraint on the neighbouring families was exerted inside the genus	The omega of the cassette modules is far from neutral and higher than that of single-copy host orthologs.	If the constraint on the neighbouring families was exerted inside the genus, the divergence of these families between genomes of one close group follows the divergence of single-copy host genes of the same genomes.	the divergence of the neighbouring families between genomes of one close group exceeds that of single-copy host genes of the same genomes, as expected if alleles that were already selected had been introduced from a donor population	divergence between the genomes is saturated, or the close groups hold too few cassette-bearing genomes for a comparison	Planned; not run	Planned and not run; it needs cassette-bearing genomes grouped by close relationship and a divergence measure that is not saturated.	Cassette-bearing genomes grouped by close relationship, and a divergence measure that is not saturated.	Figure 6; Supplementary Data S1; Supplementary Table S6	None	Cassette divergence and host divergence both come from genomes of one genus.
FP08	B	lesions	Minor alleles with lesions are rare in the reads of the source runs	No premise observation: the assembled-contig result does not address minor alleles.	In a read-level census of the 16 source runs of the cassette MAGs, minor alleles that carry a frameshift or a premature stop in the neighbouring families are rarer than in pseudogene-prone host sequence of the same assemblies.	the fraction of lesion-carrying minor alleles in the neighbouring families is equal to or higher than in the host baseline	the read depth at the module positions stays below a pre-set floor, or the numbers of lesion-carrying minor alleles are too small for a comparison	Planned; not completed	Planned; the test has not been completed, and the reads are taken from the public archive.	The source reads of the 16 MAGs from the public archive.	Figure 6; Supplementary Data S1; Supplementary Table S6	FP06 also takes pseudogene-prone host sequence of the same assemblies as its baseline, so that a host baseline that cannot be interpreted there bears on this prediction as well.	Each run is one sample of one genus, and the census depends on the read depth of the cassette contigs.

Supplementary Note S1: Predictions and wet-laboratory tests under way (plans, not results)

prediction_id	group	family	title	basis	prediction	refuted_if	returns_no_information_if	test_status	status_note	needs	related	linked_predictions	limits_carried
FP09	B	reads	Cassette reads recur in host-rich long-read runs of further studies	7 of 44 host-rich long-read runs carry a strict cassette read, 4 of them from one study (runs treated as independent; without that study 3 of 35 and a bound of 0.018); the host-read threshold is post hoc.	In host-rich non-amplicon freshwater long-read runs of studies other than the Texas river study that have not yet been screened, the per-run probability of at least one strict cassette read is not below 0.066, the lower 95% bound of 7 positive runs among 44.	44 consecutive host-rich runs from several studies other than the Texas river study carry no strict cassette read, which would reject a per-run probability of 0.066 or more at the 5% level	the runs are amplicon-like, hold fewer than the minimum number of host reads, come from the Texas river study, or come from studies that were already screened	Planned; not run	Planned and not run; it needs new run files from the public archive.	New long-read run files from studies not yet screened.	Supplementary Data S2	None	The bound is a lower limit for a per-run probability among runs selected after the fact and is not a prevalence; it treats the runs as independent although 4 of the 7 positive runs come from one study, and the same bound without that study is 0.018, whose refutation at the same level would take 165 consecutive runs; reads do not show the host genus; PacBio and sediment runs are not informative.
FP10	B	persistence	The same cassette-bearing clone is recovered in more than one year	A pond time series gave no information on persistence.	In multi-year series from one water body, the same cassette-bearing ANI95 clone is recovered in more than one collection year.	cassette-bearing clones occur only at single time points although the clone itself is recovered at several time points	no series records the same clone in more than one year, or collection dates are missing or too coarse to separate years	Planned; not run	Planned and not run; it needs the collection date and the clone assignment of each cassette-bearing genome from multi-year series, and the dates of the source samples are compiled in Supplementary Table S9.	Collection dates and clone assignments of cassette-bearing genomes from multi-year series.	Supplementary Table S9	None	Collection dates are recorded to different precision, and cassette-bearing genomes are a sample selected for carrying the cassette.
FP11	B	divergence	The divergence of the cassette tracks the divergence of the host	Over 66 pairs the RT divergence is rank-correlated with host-marker divergence, and the focal sample pair is the one pair that stands out.	In further pairs of cassette-bearing genomes, the divergence of the cassette tracks the divergence of host markers irrespective of the continent of the samples, with the focal sample pair as the one known deviation.	a pre-set number of further pairs have an RT identity of at least 98 per cent while their host-marker identity lies below a threshold that is fixed before any new pair is scored	too few further pairs with host-marker orthologs exist, or continent and clade stay confounded in the new pairs	Planned; not run	Planned and not run on further pairs; the threshold for the host-marker identity is fixed before new pairs are scored, with the focal pair as the reference case.	Further cassette-bearing genomes with host-marker orthologs, and a host-marker identity threshold fixed before scoring.	Figure 8; Supplementary Table S9	None	The rank correlation and the pair divergence come from the same sequences, partial ORF pairs break the rule fixed for the comparison, and continent is confounded with clade, so the correlation is not independent evidence.
FP12	B	arrangement	The combination of cassette parts and the K-lineage RT recurs in a further genus	The combination is found in one genus within the range scanned, and the detector sensitivity for diverged loci is unproven.	A pre-specified, calibrated scan of public catalogues beyond the GTDB representatives finds in a genus other than <i>Pollutibacter</i> either a –1-type gene next to an RT of the K lineage or an RT-bearing segment between a –1-type and a +3-type gene, confirmed by <i>rpoB</i> and <i>gyrB</i> identity.	a scan that meets its pre-set genome number and sensitivity thresholds finds no such locus. The opposite reading, that the combination arose once and stays within the genus, is refuted if such a locus is confirmed in another genus.	the sensitivity of the scan at the identities of the target loci is not shown	Planned; not run	Planned and not run, because the simulated reach of the present detectors is low at the identities of the RTs outside the genus, so that a scan with them would be uninformative (Supplementary Note S2, card P03).	A detector with calibrated sensitivity at the identities of the target loci, and catalogues that have not yet been scanned.	Supplementary Note S2, card P03 (H1, H2); Extended Data Figure 9, panel B; Extended Data Figure 13, panel D	None	The detectors derive from the cassette, and genera that carry only one of the parts are not evidence about the combination.

Supplementary Note S1: Predictions and wet-laboratory tests under way (plans, not results)

prediction_id	group	family	title	basis	prediction	refuted_if	returns_no_information_if	test_status	status_note	needs	related	linked_predictions	limits_carried
FP13	B	arrangement	The -1 and +3 genes were an adjacent pair before the RT-bearing segment lay between them	The order in which the pair and the RT-bearing segment arose is unresolved.	If the -1- and +3-type genes were an adjacent pair before the RT-bearing segment lay between them, the cassette -1 and +3 genes fall within the diversity of the free genes in outgroup-rooted trees with support values. In that case the pair also lies at the cassette insertion site in cassette-free Pollutibacter genomes, and remnants of the segment are not enriched at the junction beyond a control.	the cassette -1 and +3 genes fall into different free lineages with support, or the pairs of cassette-free Pollutibacter genomes lie only elsewhere. The reading is weakened if remnants of the segment are enriched at the junction above the control (genus-level permutation P below a pre-set level).	outgroup-rooted trees stay saturated, so that the root cannot be placed	Not testable at present	Not testable at present for the tree part, because outgroup-rooted trees of the -1 and +3 families are saturated (Supplementary Note S2, card P03), while the junction parts are planned. The reading assumes that the cassette -1 and +3 genes belong to the free pair families, an assumption that is nearly circular.	Outgroup-rooted trees that are not saturated, and cassette-free Pollutibacter genomes that carry the pair and the insertion site.	Supplementary Note S2, card P03 (H3)	FP15 shares the assumption that the cassette -1 and +3 genes belong to the free pair families.	The cassette -1 and +3 genes are assumed to belong to the free pair families, an assumption that is nearly circular.
FP14	B	selection	The N-terminal region of the +3 protein is not under weaker purifying selection than the rest	The direction of selection on the +3 family is purifying, and its function is not determined.	In a repeat of the comparison on a larger alignment of free +3-type sequences, the N-terminal region does not carry a lower fraction of negatively selected codons than the C-terminal region. A decision rule written before the pre-specified re-analyses were run, with a margin set post hoc, decides whether the prediction is treated as not contradicted.	the fraction of negatively selected codons is lower in the N-terminal region than in the C-terminal region with a two-sided Fisher exact P at or below 0.05, as the same decision rule requires	the sample is too small to tell a weaker N-terminal region from none, or selection strength and rate heterogeneity cannot be separated	Tested once with low power; repeat planned	The comparison was run once on the alignment of the cassette +3 sequences, with low power for small differences, and it is described in Supplementary Note S2, card P05; a repeat needs a larger alignment of free +3-type sequences.	A larger alignment of free +3-type sequences with enough variable sites to give adequate power.	Supplementary Note S2, card P05 (H3); Supplementary Table S6	FP22 concerns the same N-terminal region, and a refutation of this prediction, which would show weaker purifying selection on that region, bears on the statement that the region is maintained.	The comparison is post hoc, the boundary between the regions is arbitrary, and a negatively selected codon may reflect a structural constraint; the first design had low power for small differences.
FP15	B	selection	Free -1/+3 pairs outside the genus are maintained as coding units	RT-free -1/+3 pairs recur in held-out genera, and the recurrence is circular because the pair models derive from the cassette.	In genera outside the cassette lineage that carry RT-free -1/+3 pairs, the pairs are maintained as coding units and not carried as decaying remnants. Accordingly the other member of a pair stays adjacent more often than a null of adjacency loss along the genus tree allows, and the disrupted fraction of the pair lies below that of insertion-sequence ORFs in the same genomes. Both genes of the pair have an omega below a pre-set limit in the codon alignment of each genus, as a necessary condition only.	in a genus the adjacency of the pair is not above the null of adjacency loss, or the disrupted fraction of the pair is at or above that of insertion-sequence ORFs, or omega is at or above a pre-set higher limit in a pre-set number of genera, and the unit reading is then dropped for that genus	the pair definition cannot be rebuilt from seeds independent of the cassette, or too few genera carry both an informative pair and a codon alignment with enough sequences	Planned; not run	Planned and not run; the thresholds are fixed before the result, and the same test is repeated after the pair definition is rebuilt from seeds independent of the cassette.	A null of adjacency loss along the genus tree, a baseline of insertion-sequence ORFs, codon alignments for each genus, and a pair definition rebuilt from seeds independent of the cassette.	Supplementary Table S10; Extended Data Figure 9, panel B	FP13 shares the assumption that the cassette -1 and +3 genes belong to the free pair families.	The pair models derive from the cassette, the genera are not phylogenetically independent, and an omega below the threshold is met by most host core genes, so that it is only a necessary condition.
FP16	B	selection	Constraint concentrates at the catalytic box and the motif columns of the RT	The RT keeps its catalytic box in every evaluable ORE, which does not separate a retained activity from a retained fold or binding.	In the codon alignment of the RT, the fraction of negatively selected codons at pre-defined motif columns, the YxDD box and the RVT_1 motifs, is higher than at the remaining columns of the same alignment, judged against random column sets of the same size (P below 0.05, with an odds ratio above one).	the fraction at the motif columns is equal to or lower than at the remaining columns, and the reading of constraint concentrated at the motifs is then dropped	too few variable motif columns exist for the permutation test to have power	Planned; not run	Planned and not run; it needs the set of motif columns to be defined before the test.	A pre-defined set of motif columns of the RT alignment, and the codon-model output of the alignment.	Supplementary Table S5; Supplementary Table S6	None	The test is a necessary condition for selection on catalysis and does not show catalytic activity, and the alignment comes from one genus.

Supplementary Note S1: Predictions and wet-laboratory tests under way (plans, not results)

prediction_id	group	family	title	basis	prediction	refuted_if	returns_no_information_if	test_status	status_note	needs	related	linked_predictions	limits_carried
FP17	C	wet-laboratory	Any phenotype of the cassette depends on the catalytic box of the RT	The RT keeps its catalytic box in every evaluable ORF and its omega is far from neutral.	In a validated expression system, every phenotype that the wild-type cassette produces disappears when the catalytic box of the RT is altered from YADD to YAAD, so that the phenotype depends on RT catalytic activity.	the box-altered variant keeps a phenotype of the wild-type cassette in a validated system, and the refutation then holds for that phenotype	the wild-type cassette produces no phenotype in any tested condition, or the expression system is not validated	Under way; no result reported	Under way; no wet-laboratory result is reported in this paper.	Synthesis of the cassette and its catalytic-box variant, an expression host with the validity controls, and positive-control systems.	EX01; Supplementary Note S2, card P01	FP19, FP20, FP21 and FP26 assume that the phenotype or the product is lost with the YAAD variant, so that a refutation or a null result for this prediction bears on them. FP18, FP19, FP20 and FP21 use the same expression system as EX01. The reaction of the purified RT (EX02) is a separate test, and a negative result there is not a refutation condition of this prediction.	A null result in a heterologous host does not show that the cassette has no function in its natural host.
FP18	C	wet-laboratory	The five gene families act as one unit	The five families are present and intact where a contig spans the array, which does not show cooperation.	An in-frame deletion of any one of the -1, +1, +2 or +3 genes changes the phenotype of the cassette in a readout that is defined before the experiment.	all deletion strains equal the wild type in the readout, and the unit reading is then dropped for that readout	the wild-type cassette gives no phenotype in the readout, or the deletions are not confirmed to be in frame	Under way; no result reported	Under way; no wet-laboratory result is reported in this paper.	As for EX01, with the in-frame deletion constructs.	EX01	FP17, FP19, FP20 and FP21 use the same expression system as EX01, so that a failed validation returns no information for all of them.	A null result in a heterologous host does not show that the families do not cooperate in the natural host.
FP19	C	wet-laboratory	ExoI logic: a growth defect only when ExoI is absent	The DRT literature motivates an ExoI logic, and the general test of a RecBCD logic on our data did not meet its rule.	Conjecture, motivated by the DRT literature (Sánchez-Nieto et al. 2026) and not by our data: the cassette RT follows an ExoI logic, so that in Escherichia coli the cassette causes a growth defect only when ExoI (sbcB) is absent and not when recB, recC or recD is absent, and the defect disappears with the YAAD variant.	a defect appears in recB, recC or recD deletions, with Gam or gp5.9, or in the wild-type host in which ExoI is present, or validated expression gives no defect in any deletion strain	the expression system is not validated, or the deletion strains and the inhibitors are not confirmed to show the expected phenotypes	Under way; no result reported	Under way; no wet-laboratory result is reported in this paper.	As for EX01: synthesis of the cassette and its variants, strains with the deletions, vectors and positive-control systems.	EX01; Supplementary Note S2, card P02 (H1)	FP17 bears on this prediction, because the defect is expected to be lost with the YAAD variant. FP18, FP20 and FP21 use the same strains as EX01, and FP20 and FP21 are the other readings that the same deletion strains distinguish.	The test concerns Escherichia coli and not the natural host.
FP20	C	wet-laboratory	RecBCD logic: a growth defect when RecBCD is absent or inhibited	See FP19; the DRT3 literature motivates this logic.	Conjecture, motivated by the DRT literature (Wang et al. 2026) and not by our data: the cassette RT follows a RecBCD logic, with a growth defect when RecBCD is absent or inhibited (deletion of recB, recC or recD, or expression of Gam or gp5.9), not changed further by an sbcB deletion, and lost with the YAAD variant.	no defect appears in recB, recC or recD deletions, with Gam or with gp5.9, or a defect appears only in the sbcB deletion	the expression system is not validated, or Gam and gp5.9 are not confirmed to inhibit RecBCD in the strains used	Under way; no result reported	Under way; no wet-laboratory result is reported in this paper.	As for EX01.	EX01; Supplementary Note S2, cards P02 (H2) and P10 (H1)	FP17 bears on this prediction, because the defect is expected to be lost with the YAAD variant. FP18, FP19 and FP21 use the same strains as EX01, and FP19 and FP21 are the other readings that the same deletion strains distinguish.	The test concerns Escherichia coli and not the natural host, and the RecD-like gene beside the cassette cannot be told from the host RecD.
FP21	C	wet-laboratory	Mosaic reading: no deletion strain changes a validated cassette phenotype	See FP19; the mosaic reading is the alternative to both logics.	Conjecture: the cassette lineage follows neither logic, so that no deletion strain changes a validated cassette phenotype.	a defect that disappears with the YAAD variant appears in one deletion strain	the wild-type cassette gives no validated phenotype in any strain	Under way; no result reported	Under way; no wet-laboratory result is reported in this paper.	As for EX01.	EX01; Supplementary Note S2, card P02 (H3)	FP17 bears on this prediction, because a defect that is lost with the YAAD variant is part of its refutation condition. FP18, FP19 and FP20 use the same strains as EX01, and FP19 and FP20 are the two logics to which this reading is the alternative.	The mosaic reading is the alternative to both logics and is not exclusive with them.

Supplementary Note S1: Predictions and wet-laboratory tests under way (plans, not results)

prediction_id	group	family	title	basis	prediction	refuted_if	returns_no_information_if	test_status	status_note	needs	related	linked_predictions	limits_carried
FP22	C	wet-laboratory	The N-terminal domain of the +3 protein has no zinc-dependent protease activity	See Supplementary Note S2, card P05.	The purified N-terminal domain of the +3 protein shows no zinc-dependent proteolytic activity under conditions in which an active control of the same fold is active.	the purified domain shows proteolysis that EDTA or o-phenanthroline inhibits, under conditions in which the active control is active	the active control is inactive or the domain cannot be purified in a folded state	Under way; no result reported	Under way; no wet-laboratory result is reported in this paper. The observations behind this prediction are given in Supplementary Note S2, card P05.	A purified N-terminal domain, an active control of the same fold, and substrates.	EX05; Supplementary Note S2, card P05 (H1, H2)	FP14 tests whether the same N-terminal region is under weaker purifying selection, and a refutation there bears on the statement that the region is maintained.	A purified domain tested in vitro does not show what the domain does in the cell.
FP23	C	wet-laboratory	The MIDAS motif of the -1 protein is needed for binding the +3 protein in free pairs	Every evaluable -1 ORF carries a MIDAS motif by structure mapping, and no metal binding is shown.	In RT-free adjacent pairs of a -1-type and a +3-type gene from several orders, the wild-type pair binds in a bacterial two-hybrid assay and in a pull-down that both pass their positive controls, a MIDAS-null -1 protein binds the +3 protein less, and non-cognate combinations bind less than cognate pairs.	an expressed and folded MIDAS-null -1 protein binds the +3 protein as well as the wild-type -1 protein in both assays, the 90% interval of the signal ratio lying within a pre-set margin. The premise of binding is rejected in this system if a wild-type pair does not bind under positive controls that pass, which is not a conclusion about the natural host.	the proteins are not expressed or folded, or the positive controls of the assays fail	Under way; no result reported	Under way; no wet-laboratory result is reported in this paper.	-1 and +3 genes of free pairs from several orders, a MIDAS variant, two-hybrid and pull-down systems, and positive-control pairs.	EX07	None	The prediction concerns adjacent free pairs and does not claim that +3 proteins require -1 proteins at the family level, and carrying it over to the cassette genes, which are separated by the RT and two further genes, is a separate step.
FP24	C	wet-laboratory	The +3 protein binds the palindromic motif found upstream of free -1 genes	See Supplementary Note S2, card P06.	The +3 protein binds a DNA fragment that carries the core of the palindromic motif found upstream of free -1 genes, and the binding depends on the core sequence.	an electrophoretic mobility shift assay with a validated positive control shows no core-dependent binding of the +3 protein. The prediction is weakened if members of the same -1 lineage that lack the +3 gene keep the site.	the +3 protein cannot be purified in a folded state or the positive control does not bind	Under way; no result reported	Under way; no wet-laboratory result is reported in this paper. The observations behind this prediction are given in Supplementary Note S2, card P06.	A purified +3 protein, oligonucleotides, and a protein-DNA pair that binds as a positive control.	EX04; Supplementary Note S2, card P06 (H1)	None	The motif is a candidate recognition site that has not been tested, and binding in vitro would not show a function in the cell.
FP25	C	wet-laboratory	The element is the 5' region of the RT transcript	The element abuts the RT start codon in a minority of source windows, and its kind remains open.	The upstream element is the 5' region of the RT transcript: a transcription start lies at or within the 5' end of the element, and no stable element-length RNA with its own ends accumulates apart from the RT mRNA.	a stable element-length RNA with its own 5' and 3' ends accumulates apart from the RT mRNA, as shown by 5' RACE and RNA size fractionation	the RNA of the construct is not detected, or the promoter of the vector alone accounts for the start sites in the host used	Under way; no result reported	Under way; no wet-laboratory result is reported in this paper.	Constructs with the element, RNA from an expression host, and a host in which the natural start sites can be read.	EX03; Supplementary Note S2, card P09 (H1)	FP26 states the competing reading of the element, a separate RNA used as template or scaffold, so that a result for one of the two predictions bears on how the other is read.	A start site found in a heterologous host need not be the start site in the natural host.
FP26	C	wet-laboratory	The element serves as a template or scaffold for RT-made cDNA	See Supplementary Note S2, card P09.	The element is a separate RNA used as template or scaffold: the active RT, and not the YAAD variant, makes element-derived cDNA in a validated system.	the active RT does not make more element-derived cDNA than the YAAD variant in a validated system	the system is not validated or the positive-control RT is inactive under the conditions used	Under way; no result reported	Under way; no wet-laboratory result is reported in this paper. The observations behind this prediction are given in Supplementary Note S2, card P09.	A purified RT and element RNA, reaction conditions that have not been confirmed from the literature, and a positive-control RT.	EX02; Supplementary Note S2, card P09 (H2)	FP17 bears on this prediction, because both tests use the YAAD variant as the box-altered control. FP25 states the competing reading of the element, the 5' region of the RT transcript with no stable element-length RNA apart from the RT mRNA, so that a result for one of the two predictions bears on how the other is read.	In vitro synthesis shows what the RT can do with an RNA and not what it does in the cell.

prediction_id	group	family	title	basis	prediction	refuted_if	returns_no_information_if	test_status	status_note	needs	related	linked_predictions	limits_carried
FP27	C	wet-laboratory	The module junctions and the flanking junctions are present in the source DNA	See Supplementary Note S2, cards P07 and P08; read support and junction confirmation are needed before synthesis.	On DNA from the habitats of the source samples, PCR with primers that span the module junctions and the junction between the cassette and its flanking host genes gives amplicons of the sizes predicted from the assemblies, and their sequences match the assemblies.	a junction gives no product of the predicted size, or a product whose sequence differs from the assembly at the junction, with DNA of the source habitat in a reaction whose synthetic positive control amplifies	sample DNA is not available, the primers do not match the sequences of the population because the flanks diverge, or the positive control fails	Under way; no result reported	Under way; no wet-laboratory result is reported in this paper. The observations behind this prediction are given in Supplementary Note S2, cards P07 and P08.	DNA from the habitats of the source samples, primers, and a synthetic target.	EX06; Supplementary Note S2, cards P07 and P08	None	The assemblies come from metagenomes, and the sample DNA must come from the same habitat.

Table SI-24_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)