Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Phage-encoded lineages of retron-related reverse transcriptases

Nicolas Toro

nicolas.toro@eez.csic.es

Department of Soil Microbiology and Symbiotic Systems, Estación Experimental del Zaidín, Consejo Superior de Investigaciones Científicas https://orcid.org/0000-0002-3614-807X

Research Article

Keywords: reverse transcriptase, retron, bacteriophage, jumbo phage, ubiquitin-like system, phage defense

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11790028/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Phage-encoded lineages of retron-related reverse transcriptases

Nicolás Toro
Department of Soil and Plant Microbiology, Estación Experimental del Zaidín, Consejo Superior de Investigaciones Científicas, Microbiome Engineering, Reverse-Transcriptases, and regulatory RNA (MIRTA), C/ Profesor Albareda 1, 18008 Granada, Spain.

nicolas.toro@eez.csic.es

Abstract
Retrons are bacterial anti-phage systems in which a reverse transcriptase (RT) copies part of a structured non-coding RNA, the msr-msd, into multicopy single-stranded DNA to control an effector. How the RTs carried by phages relate to retrons has not been examined systematically. Using profiles built from the RTs of nine phages with array-associated RTs (ARTs), we searched the VIRE virome database and found three groups related to retrons. PRL1, the largest (2,336 proteins), occurs in temperate and virulent phages and in proviruses. In a tree of 985 RTs it is the sister group of the retrons, and topology tests reject 139 of 141 placements inside them. Its RTs keep part of retron region X but have their own region Y. In 90% of species-level clusters an E1-like enzyme precedes the RT, usually with a JAMM/MPN protease, and more than half also carry a ubiquitin-fold protein, as in bacterial ubiquitin-like defense systems. In one clade of virulent phages a structured RNA with msr-msd-like topology follows the RT, its long hairpin supported by covariation. The second group, PRL2, contains the ARTs and occurs in large phages. Its RTs keep both retron motifs, but its loci have no counterpart among retrons: in its main lineages an array of short stem-loop repeats precedes the RT and a large orphan protein follows it. The third consists of known retron types carried by phages. Phages therefore carry complete retrons and two further lineages of retron-related RTs, one built around a ubiquitin-like module and the other around a repeat array.
Keywords reverse transcriptase, retron, bacteriophage, jumbo phage, ubiquitin-like system, phage defense

Introduction

Retrons are bacterial defense systems built around a reverse transcriptase (RT). The RT reverse-transcribes part of a structured non-coding RNA, the msr-msd, into a multicopy single-stranded DNA (msDNA), and the RT-msDNA complex holds in check one or more effector proteins. When a phage interferes with the complex, the effectors are activated and the infected cell dies or stops growing before the phage can spread [Millman et al., 2020a; Bobonis et al., 2022]. Thirteen retron types and several further clades have been defined from the phylogeny of the RT and the identity of the effector [Mestre et al., 2020; Toro, 2026], and almost all of them come from bacterial genomes.

Phages, however, are not only the targets of defense systems; they are also major carriers of them, and they encode RTs of their own, notably those of diversity-generating retroelements, which diversify the receptor-binding protein at the tips of the tail fibers [Doulatov et al., 2004], and those of group II introns, which were long thought to be rare in phages but have now been found across Caudoviricetes [Merk et al., 2025]. This raises a question that has not been examined systematically: do phages carry retrons, or retron-like systems of their own, and if so, how are they related to the bacterial retron types? Metagenome-derived virome resources now make it possible to answer it at scale.

We came to this from our ongoing survey of retron diversity, which began with a phylogenetic analysis of bacterial RTs [Toro and Nisa-Martínez, 2014; Toro et al., 2019], continued with the classification of retrons by their associated genes [Mestre et al., 2020] and of the related UG/Abi RTs [Mestre et al., 2022], and most recently mapped retrons across the SPIRE metagenomic resource [Schmidt et al., 2024], revealing candidate type XI-like lineages [Toro, 2026]. Our entry point into phage genomes was a recent preprint (hereafter, the ART preprint) in which autonomous AI agents, surveying RT loci in 1.9 billion metagenomic protein clusters, found a family of jumbo-phage RTs preceded by an array of repeated units of ~200 nt and followed by a dedicated partner gene, which the authors called array-associated RTs (ARTs) [Yoon et al., 2026]. They showed that the array is highly expressed and processed into discrete short RNAs during infection by phage SA1 and when the locus is expressed in *Escherichia coli*, and they proposed, as a hypothesis, that ART is a retron-like system that works with a bank of distinct RNAs instead of a single one. A later, independent computational preprint by a different author [Ye, 2026] went further and proposed that each repeat unit of an ART array is itself an msr-msd, so that ART loci would be retrons with multiple non-coding RNA genes.

These two studies focus on the ART loci themselves. Our question is broader: which retron-related RTs do phages carry, ARTs included, where do they branch relative to every described retron type, and what loci accompany them? We took as queries the RTs of eight Staphylococcus phages and of the Listeria jumbo phage LPJP1, which carries two distantly related RTs of the same group, all of them among the ART loci. The staphylococcal RTs and one of the two RTs of LPJP1 are preceded by 1.1-1.3 kb without annotated genes, where their arrays lie. We searched VIRE, which holds 1.8 million viral genomes from more than 80,000 environmental and host-associated samples [Nishijima et al., 2026], with a profile of the staphylococcal RTs and with the RT profiles of our retron survey, placed the RTs in a phylogeny that includes representatives of

all retron types, and tested alternative positions statistically. We then examined their gene neighborhoods, conserved RNA structures, and, for SA1, the published RNA-seq data. Because metagenomic collections contain many near-identical phage genomes, we counted phages by species- and genus-level clusters rather than by genome.
This approach revealed three groups of phage RTs related to retrons, each related to them in a different way. The largest, which we call PRL1, is not part of the ART group: it lies outside all retron types, as their sister group, and its loci carry a conserved ubiquitin-like module and, in one clade of virulent phages, a structured RNA downstream of the RT. A second group, PRL2, contains the ARTs together with other lineages and branches next to PRL1 and the retrons; its RTs have retron-like motifs, but its loci share an organization that has no counterpart among retrons. The third, VOTH, falls within the known retron types and keeps their effector where the type has one, showing that phages also carry retrons of known types. Finally, we test whether the ART arrays are tandem msr-msd units, as proposed in the preprint of Ye [Ye, 2026], and find that the evidence is better explained by the repeat structure of the arrays and by how they are transcribed. We therefore treat PRL1 and PRL2 as retron-related lineages rather than as new retron types.

Results

Three groups of phage RTs related to the retrons

We searched the VIRE phage database with a profile built from the eight staphylococcal RTs and with the RT profiles of the retron survey [Mestre et al., 2020; Toro, 2026], and grouped the hits by phylogeny and profile scores (**Methods**). The members of PRL2 were then collected with a second profile that also includes the two RTs of LPJP1. Three groups of phage RTs fall into the part of the RT tree that contains the retrons (**Table 1**, **Fig. 1**). We call two of them PRL1 and PRL2, for "phage retron-related lineage".

PRL1 comprises 2,336 RT proteins, which we reduced to 179 non-redundant representatives (clusters at 60% identity). A profile of the lineage separates its members from all other RT groups (below) and retrieves 143 further non-redundant sequences from bacterial and viral contigs of SPIRE and of NCBI ClusteredNR (hereafter NCBI); in the trees, the PRL1 clade therefore has 322 tips (179 + 143).

PRL2 is a smaller and more divergent group that includes the eight staphylococcal RTs and the two RTs of LPJP1. In VIRE, it comprises 247 proteins in 21 groups at 60% identity, and it is represented by 39 tips in the trees. In one of its lineages (RTr, named for the repeat array that precedes the RT), the RT is followed in the same locus by a second, distantly related RT (RTc, for companion); both come from a single genus-level cluster of phages and are represented by two tips each in the trees (below). The two RTc tips branch outside the clade formed by the other 37, which we take as the core of PRL2.

VOTH ("VIRE other") is the residual group: 151 proteins (51 representatives) retrieved by the retron profiles that belong neither to PRL1 nor to PRL2. All of them lie among the bacterial retrons in the trees, and most can be assigned to a described retron type or clade (below).

The three groups score very differently against the general retron RT profile of the retron survey (**Table 1**): almost all VOTH representatives pass it (49 of 51), only 40 of the 179 PRL1 representatives do (22%), and none of the 21 VIRE representatives of PRL2. Seven of the eight public staphylococcal RTs of PRL2 do pass it, so its VIRE members are more divergent from retrons than the publicly available ARTs.

Counts made per protein or per genome are dominated by a few clusters of near-identical phages. One 60%-identity cluster of PRL1 holds 1,503 of its 2,336 proteins, and the ten largest hold 78%; in PRL2, all the RTr and RTc proteins fall in a single genus-level cluster. We therefore present the results per representative, per tree tip, or per species- or genus-level cluster of the VIRE database, and state the unit each time.

Table 1. The groups of RTs in this study

Group	Proteins (non-redundant units)	Genomes	Pass the general retron RT profile*	Clade in the tree of 985 RTs (UFBoot: Q.pfam / LG+C60)
PRL1	2,336 (179 representatives; 322 tips with 143 SPIRE and NCBI hits)	VIRE phages of 25-158 kb (median 63 kb); proviruses and viral contigs in SPIRE and NCBI	40 of 179 (22%); median 149 bits	96 / 100
PRL2	247 in VIRE (21 groups at 60% identity) plus 8 *Staphylococcus* phages and LPJP1 (2 RTs); 39 tips	phages of median 232 kb (15 of 21 VIRE representatives of 200 kb or more)	0 of 21; 7 of 8 public RTs; median 115 bits	core of 37 tips: 98 / 100; with PRL1, the retrons and VOTH: 87 / 99; the RTc pair branches outside that clade
VOTH ("VIRE other")	151 (51 representatives)	VIRE phages	49 of 51 (96%); median 218 bits	all 51 inside the retron clade
Bacterial retrons	467 tips, covering all described retron types and clades (Methods)	bacterial genomes	49 of 59 references (83%); median 204 bits	with the 51 VOTH: 85 / 97
Non-retron RTs (UG/Abi, DGR, group II intron, G2L, RT-CRISPR)	106	various	1 of 106	root (DGR)

*General retron RT profile of the retron survey ($\geq$ 160 bits, E $\leq$ 1e-10).

PRL1 is a lineage separate from the retron types, and PRL2 branches outside both

To place the three groups among known RTs, we built maximum-likelihood trees with IQ-TREE [Minh et al., 2020] from an alignment of the RT domain (365 positions after trimming;

Q.pfam+F+R5), with branch support from 1,000 ultrafast bootstrap replicates (UFBoot). The main tree has 985 sequences and covers all described retron types and clades (**Fig. 1**; **Methods**); three smaller trees, with fewer retrons and different sets of PRL2 sequences, test whether the result depends on how densely the retrons are sampled or on which PRL2 sequences are included (**Supplementary Fig. S1**). The trees are rooted on the DGR RTs for display only; the relationships described below hold for any root placed among the non-retron RTs.

PRL1 forms a well-supported clade (UFBoot 96) that falls outside every retron type. In the main tree, its closest relative (sister group) is the clade that contains all 467 retrons and the 51 VOTH (UFBoot 88 for the node that joins them), and two of the smaller trees give the same arrangement (UFBoot 74 and 91). In the third (577 sequences), PRL1 instead joins a small group formed by retrons of types I-A, I-B1, I-B2, and II-A1 and 15 VOTH (UFBoot 72). Even in this tree, PRL1 is the sister of a group of types rather than part of any of them; in none of the trees does PRL1 fall inside a retron type.
Topology tests support this placement. We moved the whole PRL1 clade to each of the 179 branches of the main tree that leave at least eight tips on each side, and used the approximately unbiased (AU) test [Shimodaira, 2002] to ask whether each resulting tree fits the data significantly worse than the best one (**Methods**). The test rejects all but three alternatives (a placement is rejected when $p < 0.05$), among them 139 of the 141 positions inside the clade of retrons and VOTH; the three fit the data less well than the best placement, PRL1 as the sister of all retrons and VOTH ($p = 0.77$): as the sister of the clade of types II-A1 and I-A ($p = 0.10$), of type I-B2 ($p = 0.054$), or of the core of PRL2 (its 37 tips other than the RTc pair; $p = 0.17$). The smaller trees agree (**Table S1**), except that in the tree of 577 sequences a placement next to types I-A, I-B, and II-A1 comes first ($p = 0.62$) and the sister position second ($p = 0.37$). A model that better accounts for differences in amino-acid composition between sites (LG+C60+F+G, hereafter LG+C60; **Methods**) gives the same arrangement with stronger support: PRL1 is monophyletic (UFBoot 100) and the sister of the clade of retrons and VOTH (UFBoot 99 for the node that joins them).

PRL2 branches one step further out. Its core (UFBoot 98) is the sister of the clade formed by PRL1, the retrons, and VOTH (UFBoot 88), and together they form a clade (UFBoot 87) outside which the RTc pair branches (below). The smaller trees and the LG+C60 model give the same order (UFBoot 74-99). The AU test rejects every alternative placement of the core except as the sister of PRL1 alone ($p = 0.17$, against 0.83 for the best tree); only in the tree of 577 sequences are three further placements not rejected, all marginally ($p \approx 0.06$; **Table S1**).

Together, PRL1, PRL2, VOTH, and the retrons form a single clade of retron-related RTs that excludes the other RT groups, apart from one RT-CRISPR tip (UFBoot 100 in the main tree, 91-97 in the smaller trees, and 100 with LG+C60). Within this clade, the short alignment resolves the order of the branches poorly, and several retron types are not monophyletic: type II-A1, for example, lies in a clade of 83 tips together with type I-A and 15 VOTH. The trees therefore cannot say exactly where PRL1 branches relative to individual retron types, but they consistently show that it is a lineage separate from them, either as the sister of all retrons or of the group of types I-A, I-B, and II-A1, and that the core of PRL2 branches outside both.

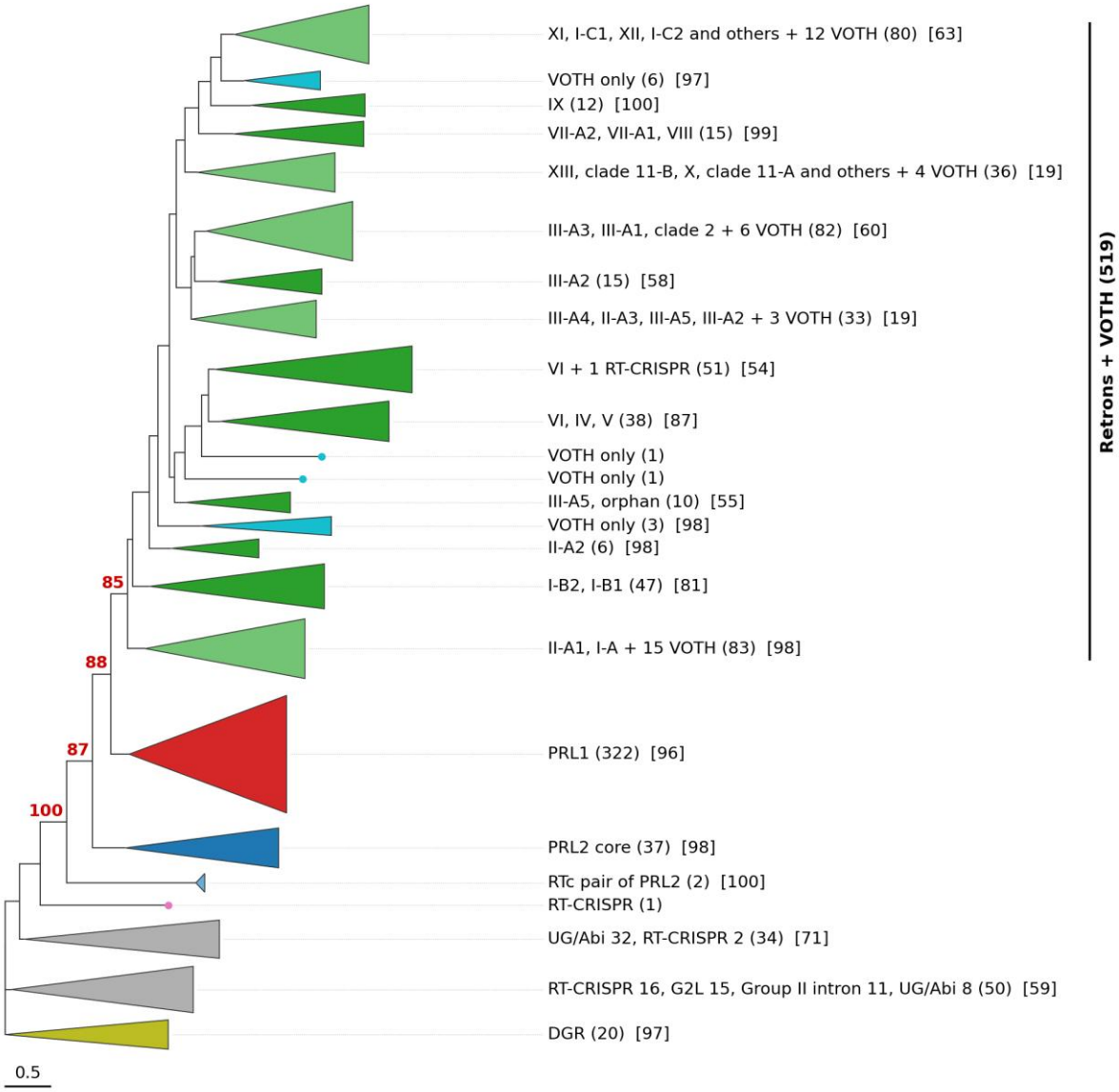

Figure 1. Phylogeny of retron-related phage RTs. Maximum-likelihood tree of 985 RT sequences inferred from an alignment of the RT domain trimmed to 365 positions (Q.pfam+F+R5; 1,000 UFBoot), shown rooted on the DGR RTs for display only. The tree contains PRL1 (322 tips: 179 VIRE representatives and 143 non-redundant hits from SPIRE and NCBI), PRL2 (39 tips, including the eight public staphylococcal RTs, the two RTs of LPJP1, and RTs from the Unified Human Gut Virome (UHGV) [Camargo et al., 2025] and SPIRE), 51 VOTH representatives, 467 bacterial retron RTs covering all described retron types and clades (**Methods**), and 106 non-retron RTs (UG/Abi, DGR, group II intron, G2L, and RT-CRISPR). Clades are collapsed into wedges whose length is the median depth of their tips and whose height grows with the number of tips (in parentheses); numbers in square brackets are the UFBoot values of the collapsed clades. The clade of retrons and VOTH (bar on the right) is shown as its subclades of at most 85 tips, labeled with the retron types they contain, most numerous first, and the number of VOTH (light green, subclades with VOTH; cyan, VOTH only). The non-retron RTs outside the clade of retron-related RTs are collapsed by branch and labeled with their

composition. Red numbers are UFBoot values: 85, the clade of retrons and VOTH; 88, the node that joins PRL1 to this clade; 87, the node that joins the core of PRL2 to PRL1, the retrons, and VOTH; and 100, the node that joins the RTc pair to all of these. Scale bar, substitutions per site.

PRL1 RTs keep only part of retron region X and have their own region Y

Retron RTs carry two short motifs that distinguish them from other RTs: region X (N-A-x-x-H) and region Y (V/I-T-G) [Mestre et al., 2020]. To see whether the phage RTs have them, we aligned each group on its own, so that divergent sequences are not forced into the pattern of the majority, and marked the columns of the two regions with eight retron RTs of known types added to each alignment (**Fig. 2**; **Methods**).

VOTH RTs have both retron motifs, as expected for retrons, and so do PRL2 RTs: in both, region X reads N-A-x-x-H as in retrons (closing histidine in 94-100% of sequences) and region Y keeps the conserved glycine (97-100%), with an invariant Arg four residues before the glycine in PRL2.

PRL1 RTs keep only part of them. In region X, the closing histidine is nearly invariant (99%), but the first positions differ: PRL1 RTs read A-[V/L/I]-K-R-H-Q, with Ala in place of the retron Asn (74%). Region Y differs altogether: PRL1 RTs read N-W-N-L-G-L-M-L, where the glycine is invariant but Leu replaces the retron Thr (90%) and an Asn is conserved four residues upstream. This region Y resembles the F-L-G of group II intron RTs more than the V-T-G of retrons.

The YADD catalytic motif, shared by most RTs, is present in 312 of the 322 PRL1 RTs (97%) and in 18 of the 32 PRL2 sequences, with YSDD, YCDD, or YLDD in the others.

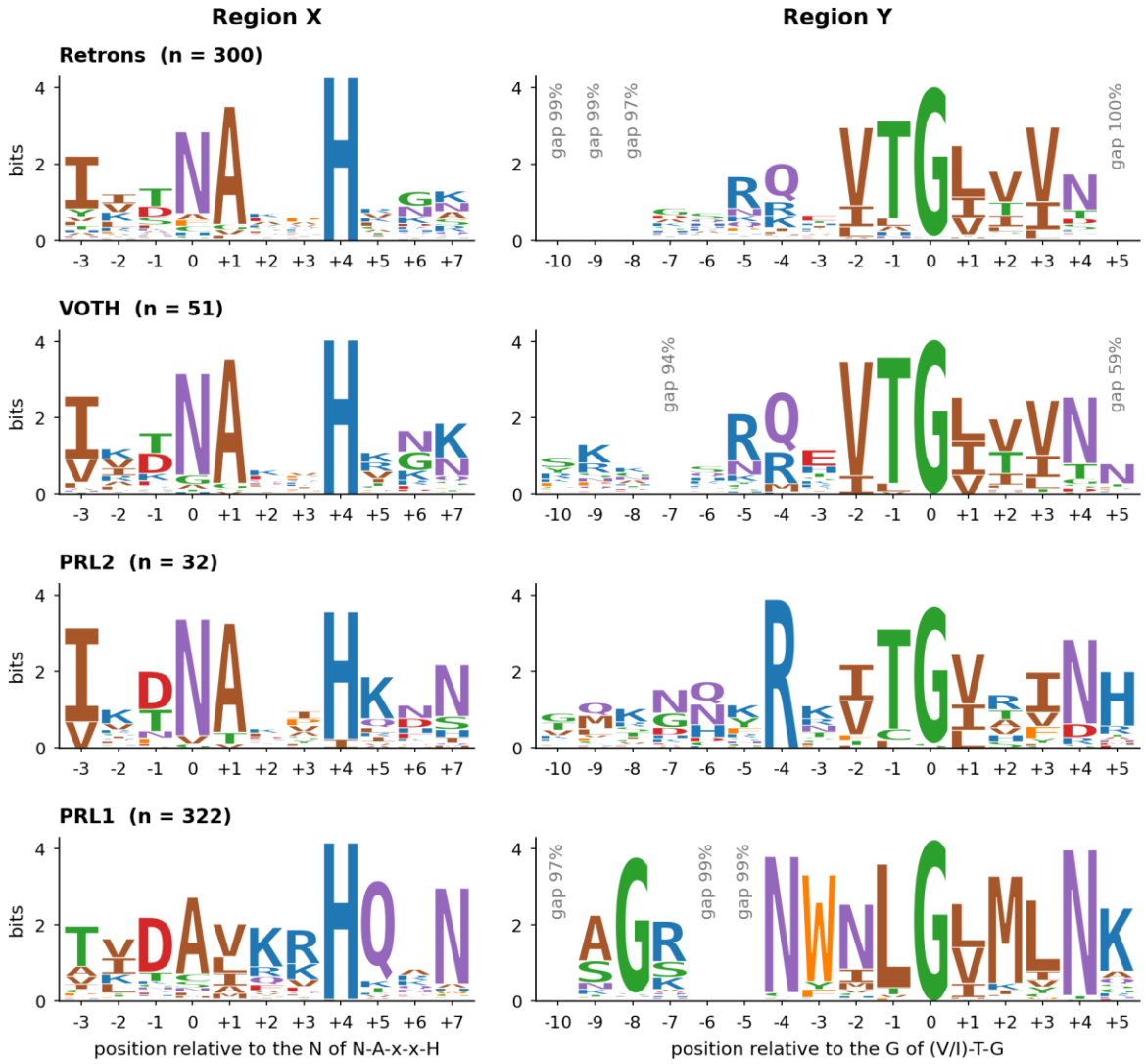

Figure 2. Regions X and Y in retrons and in the three groups of phage RTs. Sequence logos of regions X (left) and Y (right) from independent alignments of 300 retrons chosen at random from the reference set, 51 VOTH representatives, 32 PRL2 RTs (non-redundant at 90% identity), and 322 PRL1 RTs. VOTH and PRL2 RTs share both retron motifs, whereas PRL1 RTs keep only the histidine that closes region X and have a distinct region Y. Eight retron RTs of known types were added to each alignment to locate the columns and were not counted. Position 0 is the N of N-A-x-x-H (region X) or the conserved G of (V/I)-T-G (region Y). Letter height shows information content (bits, with small-sample correction) multiplied by column occupancy; grey labels give the percentage of gaps in columns that are less than 60% occupied. Colors: A/V/L/I/M brown, F/W/P orange, G/S/T/Y/C green, N/Q purple, K/R/H blue, D/E red.

The RT profiles of the retron survey show that PRL1 is related to the retrons but matches no retron type. Only 22% of its representatives pass the general retron RT profile (**Table 1**), and none of the 29 type-specific profiles assigns a type convincingly: the best type is II-A1 for 123

representatives, but with a median margin over the second-best type of 5 bits, against 61 bits for reference retrons.

The retron RNA models give the same answer. We searched 700-nt windows on each side of the PRL1 RTs with the 21 covariance models of retron msr-msd RNAs of the retron survey, the two models of the type XI-like lineages [Toro, 2026], and the Rfam model of the retron msr RNA (RF02462) [Ontiveros-Palacios et al., 2025] (**Methods**). None finds a convincing match: the best hit anywhere is a weak one of the type I-B1 model, in the downstream window of one genome (E = 9.3e-4), and RF02462 has no hit at all. A structured RNA does lie downstream of the RT in some PRL1 phages, but it is detected only by a model built from PRL1 genomes themselves (below) and resembles none of the published retron models.

PRL1 is carried by two kinds of uncharacterized phages and by proviruses

The PRL1 RTs (median 381 aa) are encoded by phage genomes of 25-158 kb (median 63 kb; none reaches 200 kb). A host is predicted for only 24 of the 179 representatives, mostly gut Bacillota_A (Faecalibacterium, Blautia, and Anaerostipes) and Bacteroidota. We annotated the genomes with Pharokka [Bouras et al., 2023] and compared their proteins with those of 5,375 RefSeq Caudoviricetes genomes (**Methods**). They fall into two kinds that coincide with the two major clades of the RT tree (**Table 2**): eight subclades of that tree (sub01 to sub08), also used for the RNA search below, hold 96 temperate and 54 virulent representatives, and the remaining 29 fall outside them. The BACPHLIP lifestyle score [Hockenberry and Wilke, 2021], computed for each genome on its own, agrees with this assignment for 143 of the 150 genomes and classifies 28 of the other 29 as virulent.

The temperate phages (n = 96) have genomes of ~100 kb with an integrase (96%), long non-contractile tails of the siphovirus type (tape-measure protein in 85%), and portal and DNA polymerase genes with RefSeq homologues (96% and 100%; median identity 35-37%); the RT lies a median of 29 genes from the integrase.

The virulent phages (n = 54) have genomes of ~50 kb with no integrase and a low temperate lifestyle score. Their portal and major head proteins have no homologues among RefSeq phage proteins, and their tail genes are rarely recognized; the terminase large subunit and the DNA polymerase have distant RefSeq homologues (median identity 24%), the DNA polymerase being most similar to that of Herelleviridae.

The PRL1 phages are all different species, and all are far from cultured phages. Their genomes are compared by Mash distance [Ondov et al., 2016], a genome-wide measure that ranges from 0 for identical genomes to 1 for genomes without shared sequence, values below 0.05 indicating the same species: no pair of PRL1 genomes falls below 0.05 (median 1.00), and the closest genome of INPHARED, a collection of cultured phages [Cook et al., 2021], lies at a median distance of 0.30. In a network that links genomes by the proteins they share, built with the 179 PRL1 phages and 5,375 RefSeq phage genomes, the PRL1 phages fall into four clusters, two of which contain only PRL1 phages, and no PRL1 genome shares more than 7% of its proteins with any RefSeq genome. PRL1 is therefore carried by diverse phages that are almost absent from the reference databases.

Table 2. The two kinds of phages that carry PRL1

Feature	Temperate (n = 96)	Virulent (n = 54)
Genome size (median)	~100 kb	~50 kb
Integrase	96%	0%
Tail genes	siphovirus-like (tape measure 85%)	rarely annotated
Portal / major head protein with a RefSeq homologue	96% / 75%	0% / 0%
Median fraction of proteins with a RefSeq hit	32%	20%
Nearest INPHARED genome (Mash distance, median)	0.30	0.30

Outside VIRE, PRL1 is found mostly in proviruses. A profile of PRL1 separates its members, all of which score 272 bits or more, from the retrons, PRL2, and the other RT groups (217 bits at most; **Methods**). It finds 59 loci in SPIRE bacterial genomes and 108 proteins in NCBI ClusteredNR, mostly from Clostridia. All 143 non-redundant hits fall inside the PRL1 clade when the trees are inferred with them included; when they are instead placed on the reference tree with EPA-ng [Barbera et al., 2019], 141 fall inside PRL1 and two next to a PRL2 tip.

Most of these loci are viral. geNomad [Camargo et al., 2024] either calls the whole contig viral or places the RT inside a predicted provirus in 72 of the 91 NCBI loci from non-viral records and in 40 of the 59 SPIRE loci, and most of the loci without a viral call are on short contigs. The exception is a clade of six RTs on long contigs (UFBoot 100, inside the PRL1 clade): four in SPIRE metagenome-assembled genomes classified as Clostridia and two in NCBI contigs without an assigned taxonomy. These six lack the conserved genes that precede the RT in the phages (below) and are surrounded by bacterial genes; the two hits placed next to PRL2 by EPA-ng are among them. We regard them as possible degraded prophages and did not examine them further.

An E1-like enzyme and a JAMM protease precede the PRL1 RT

The gene immediately upstream of the PRL1 RT encodes an E1-like protein of ~223 aa of the ThiF/MoeB family in 87% of the representatives and in 90% of the 548 species-level clusters of VIRE; in form A of the locus (below), it is the second gene upstream. Outside VIRE, it precedes the RT at the same distance in about half of the SPIRE and NCBI loci. This gene is absent from PRL2 (0 of the 157 loci examined) and rare next to the VOTH RTs (3 of 151 loci), and only 16 of 5,375 RefSeq phage genomes carry a ThiF homologue, none of them with a PRL1 RT. The genes next to it define two main forms of the locus (**Fig. 3A**, **Table 3**). In the commonest, form B (64% of species-level clusters), a JAMM/MPN gene lies before the E1-like gene, either immediately or after one to three small unannotated genes; in form A (17%; temperate phages

only), a protein of ~267 aa with a DUF6273 domain lies between the E1-like gene and the RT. Counted by genomes instead, form A would be the most common (53%), only because the largest cluster, of 1,503 temperate phages, has it; we therefore count by clusters.

The forms follow the RT tree and the lifestyle of the phage, although not strictly. On the tree of the 179 representatives, the three states (A, B, and other) require 29 changes, against a mean of 37 when the forms are assigned at random (Fitch parsimony, 1,000 permutations; $p = 0.001$), and they differ between the temperate and the virulent subclades (Fisher's exact test, $p = 1.3 \times 10^{-4}$; **Supplementary Fig. S2**). The difference lies mainly in the other arrangements, which are almost restricted to temperate phages: 32 of the 96 temperate representatives have one, against 4 of the 54 virulent ones.

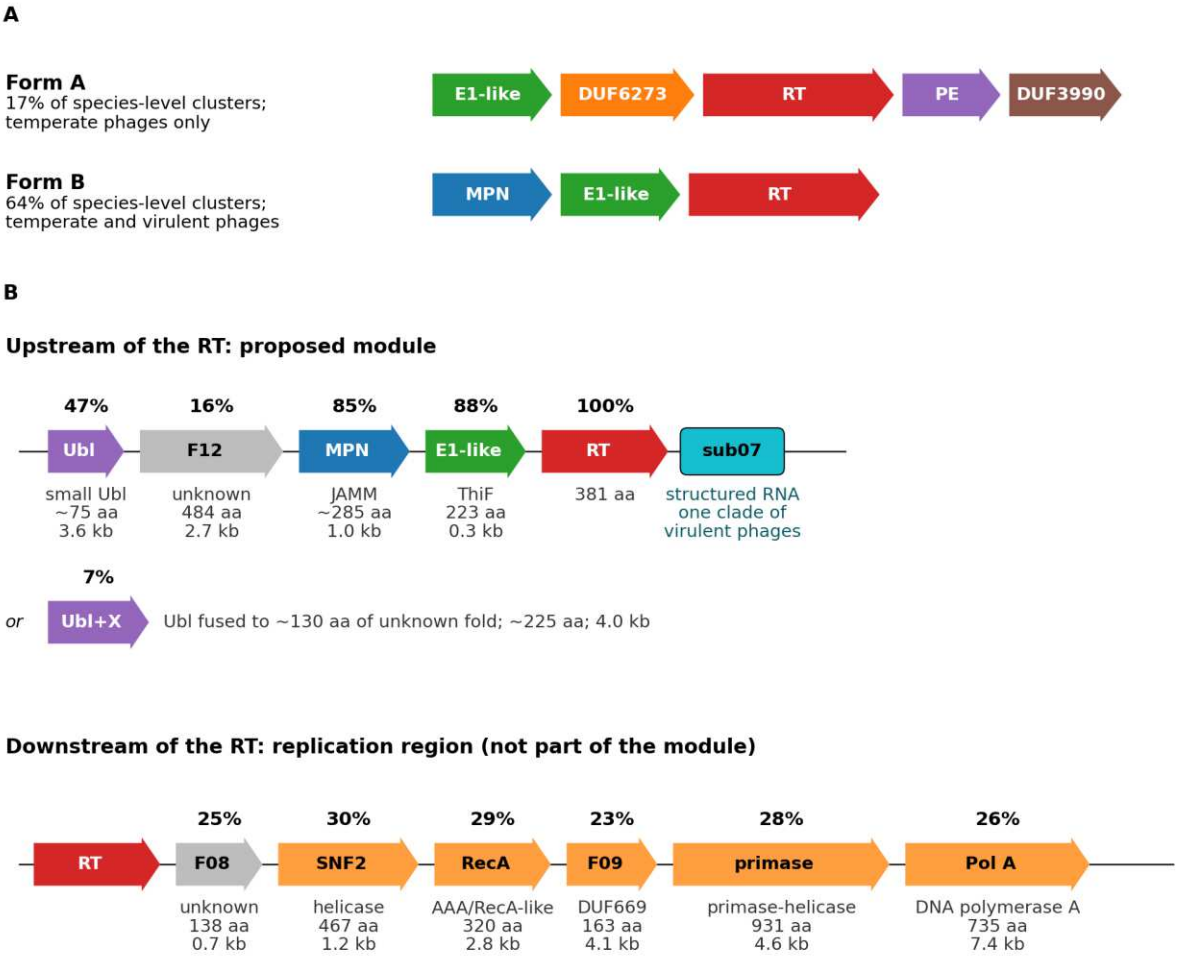

Figure 3. The PRL1 locus and the proposed module around the RT. A, the two main forms of the locus, with the share of species-level clusters in which each occurs; the E1-like (ThiF) gene precedes the RT in 90% of species-level clusters. MPN, JAMM/MPN protease (in form B, one to three small genes may lie between the MPN and E1-like genes); PE, phosphoesterase-like protein, at position +1 in 9% of species-level clusters; DUF6273 and DUF3990, proteins with

domains of unknown function (DUF3990 at +2 in 5%). Other arrangements (20% of species-level clusters) are not drawn. Genes are not to scale. B, genes found within 10 kb of the RT in the 307 tips of the PRL1 clade with an annotated window, all on the strand of the RT, in the order of their median position; top, genes upstream of the RT (the proposed module); bottom, genes downstream of it. Under each gene: annotation, median length of the family, and median distance from the RT; above, the percentage of tips that carry it. Gene widths grow with their median length; spacing is not to scale. The E1-like (ThiF) gene is at position -1 in 88% of tips, and a JAMM/MPN gene (four families of 27-134 tips) is at -2, or at -3 to -5 after small genes, in 85%; both are present in 85%. A protein with a ubiquitin-like domain lies 3-4 kb upstream in 170 of the 307 tips (55%): a small ubiquitin-like protein (Ubl, ~75 aa) in 143, a ~225-aa protein in which the ubiquitin-like domain is fused to a ~130-aa domain of unknown fold (Ubl+X) in 23 (one tip has both), and a longer protein with a ubiquitin-like domain in the other five. No E2 or E3 domain was detected. The sub07 RNA (~200 nt, starting a median of 21 nt before the end of the RT gene) is found in 28 of 87 virulent representatives (17 used to build the model, 3 confident and 8 weak independent detections) and in 2 of 92 temperate ones (weak). The downstream genes are enriched near the RT but form the usual replication region of these phages and are not proposed as part of the module. F12 and F08 are unannotated. Form A of the locus (E1-like gene followed by a DUF6273 gene; 9% of tips) is not drawn in B. Colors: violet, ubiquitin-like proteins; green, E1-like; blue, JAMM/MPN; gray, unannotated; orange, replication and recombination genes; cyan, sub07 RNA.

Table 3. Forms of the PRL1 locus by counting unit

Unit	N	Form A (E1-like, DUF6273, RT)	Form B (MPN gene before the E1-like gene, RT)*	E1-like, other arrangement	No E1-like at -2/-1
Genomes	2,336	53%	37%	5%	6%
Species clusters	548	17%	64%	10%	10%
Genus clusters	464	16%	63%	11%	10%
Representatives	179	2%	77%	other arrangements: 21%	

*Includes loci in which one to three small genes lie between the MPN and E1-like genes (18-25% of the units).

Structural predictions identify the two proteins that precede the RT (**Table 4**). HHsearch [Steinegger et al., 2019] relates the E1-like proteins to the small E1-like enzymes HcgE and TcdA and to ubiquitin E1 enzymes, and AlphaFold2 models of three representatives [Jumper et al., 2021; Mirdita et al., 2022] superpose on the small E1 and E1-like enzymes of the ThiF/MoeB family with TM-scores of 0.64-0.78, the closest being the SUMO E1 subunit SAE1, MoeB, ThiF, and Cap2, the E1-like enzyme of CBASS anti-phage systems; the match to the adenylation domain of the ubiquitin E1 Uba1 is weaker (0.56-0.57). Of the 158 representatives with an E1-like gene in the window, 153 retain the ATP-binding motif G-x-G-x-x-G, but only 2 have the

zinc-binding CxxC motif of the crossover loop of canonical E1 enzymes. In an alignment of the 270 E1-like proteins of the windows, one cysteine, in a conserved P-V-S-A-C motif near the C-terminus, is invariant; whether it is the catalytic cysteine of these compact enzymes remains to be determined. The E1-like proteins are therefore small adenylyltransferases of the HcgE type, whose substrate is unknown.

The gene before the E1-like gene in form B, which HHsearch labels "ubiquitin-like protein" because its best template is a complex of ubiquitin with the protease Rpn11, in fact encodes a JAMM/MPN metalloprotease, the family of deubiquitinases to which Rpn11 belongs [Ambroggio et al., 2004]. The model of a 286-aa representative superposes on the MPN domains of Rpn8/Rpn11, of archaeal JAMM deubiquitinases, and of the CBASS isopeptidase (TM-score 0.60-0.75; **Table 4**); 457 of the 459 MPN proteins of the loci carry the complete JAMM motif (E...H-x-H-x7-S-x2-D), and in the model the two histidines and the aspartate that bind zinc lie 3.2-4.3 Å apart, with the serine and glutamate in place, as in an active enzyme. An E1-like enzyme next to a JAMM protease is the organization of the Cap2-Cap3 pair of type II CBASS systems and of related bacterial ubiquitin-like antiviral systems [Millman et al., 2020b; Ledvina et al., 2023; Chambers et al., 2024]. The DUF6273 protein of form A is modeled with confidence but matches no known structure. None of these proteins is a known retron effector: the 35 effector profiles of the retron survey find a single weak hit among the 9,253 genes within two positions of the PRL1 RTs.

Table 4. Structural homologues of the accessory proteins of the PRL1 locus

Protein (models)	**pLDDT / pTM**	**Best structural homologues**	**TM-score (normalized by the length of the model; for MPN, by the length of the target MPN domain) or Foldseek probability**
E1-like (ThiF), 3 representatives	94.8-95.6 / 0.91-0.92	SAE1 (SUMO E1); MoeB; ThiF; Cap2 (CBASS); MccB; TcdA; Uba1 (ubiquitin E1)	0.73-0.78; 0.76-0.78; 0.72-0.75; 0.71-0.74; 0.68-0.74; 0.64-0.68; 0.56-0.57
MPN (JAMM), 1 representative	82.6 / 0.72	Rpn8/Rpn11; archaeal JAMM deubiquitinases (Pyrococcus, Caldiarchaeum); CBASS isopeptidase	0.60-0.75; 0.68-0.73; 0.61
DUF6273, 1 representative	88.6 / 0.86	none (E ≤ 0.001)	-
F10, small ubiquitin-like protein (tandem representative of 156 aa)	78 / 0.51	ISG15; FAT10; linear di-ubiquitin; each half: SUMO-like and ubiquitin-like proteins	probability 1.0; E 1e-6 (whole), 0.003-0.01 (halves)
F03, ubiquitin-like domain plus domain of unknown fold (225 aa)	89 / 0.60	N-terminal 80 residues: ubiquilin 3, SUMO2, Hub1, Smt3; rest: none	0.54-0.68

AlphaFold2 models compared with Foldseek against PDB100 and AlphaFold/SwissProt.

Ubiquitin-like proteins complete the module, but no E2 or E3 is detected

An E1-like enzyme and a deubiquitinase suggest a third component: a small protein with a ubiquitin fold that is activated and conjugated. As none was evident next to the RT, we examined all genes within 10 kb of the RT in the 322 tips of the PRL1 clade (7,341 proteins), grouped them into families, and compared the frequency of each family with that in 1,000 random windows of the same genomes (**Methods**). Of the 46 families present in at least 15 tips, 44 are enriched near the RT (empirical P = 0.001, the minimum possible). Besides the E1-like protein and four MPN families, the enriched families are the usual genes of the phage replication and packaging region (SNF2-type helicase, RecA-like ATPase, primase-helicase, DNA polymerase A, resolvase, integrase, terminase, and others) and three enzymes of NAD synthesis (phosphoribosyl pyrophosphokinase, nicotinamide phosphoribosyltransferase, and NAD synthetase).

Two of the unannotated families have a ubiquitin fold (**Table 4**). F10 (57 tips) is a protein of 76 aa, the length of ubiquitin, encoded a median of 3.6 kb upstream of the RT and always in the same orientation; one member made of two such units matches the double ubiquitin-like proteins ISG15 and FAT10, and each unit alone matches SUMO-like and ubiquitin-like structures. F03 (130 tips; about 4 kb upstream) combines an N-terminal ubiquitin-like domain of ~80 residues with a domain of ~130 residues of unknown fold, either fused in one protein of ~225 aa or encoded by two genes. With four further families of the same fold at the same position, 170 of the 307 tips (55%) carry a protein with a ubiquitin-like domain. A model of one 76-aa member matches SUMO bound to its E1 enzyme (TM-score 0.73) and has a disordered C-terminus, as in conjugatable modifiers, but none of these proteins ends in the Gly-Gly motif that ubiquitin uses for conjugation. No E2 (UBC) or E3 (RING, HECT, U-box) domain was found among the 3,495 protein families of the windows (Pfam [Mistry et al., 2021] and HHsearch; **Methods**), apart from a Prok-E2_D domain in one protein of one tip.

We therefore propose a module of three conserved components around the PRL1 RT (**Fig. 3B**): the E1-like enzyme at -1 (88% of tips), the JAMM protease at -2, or at -3 to -5 after small inserted genes (85%), and a ubiquitin-like protein 3-4 kb upstream (55%). In one clade of virulent phages, the structured RNA described next lies downstream of the RT. An E1-like enzyme, a JAMM protease, and a ubiquitin-like protein resemble the core of bacterial ubiquitin-like antiviral systems [Chambers et al., 2024], which also carry an E2 that we did not detect; whether this module conjugates anything, and to what, is unknown. Two cautions apply. First, the E1-like gene follows the RT phylogeny (Mantel r = 0.65), but no better than other conserved genes of the same phages (portal, DNA polymerase, terminase; r = 0.70-0.77), so the association can be explained by common descent and is not by itself evidence of a functional link. Second, the replication genes near the RT are also enriched, but such genes are expected in this region of any phage genome, so we do not include them in the module.

A structured RNA lies at the 3′ end of the RT gene in one clade of virulent phages

Retron msr-msd RNAs are usually located upstream of the RT. We searched both sides of the PRL1 RT for conserved structured RNAs with CMfinder [Yao et al., 2006], separately in the eight subclades defined above, and checked the candidates for base-pair covariation, the signature of a conserved RNA structure (R-scape [Rivas et al., 2017]). The 700 nt upstream of the RT give no conserved structure in any subclade, and the ~310-nt spacer between the E1-like gene and the RT is no more conserved than a shuffled sequence and matches neither Rfam nor the retron models. Downstream of the RT, in contrast, the search finds a robust element in the virulent subclade sub07; a second, different motif in the virulent subclade sub06 was not studied further.

We describe the structure of the sub07 element (~200 nt; range 123-250) with an iterative covariance model (r0) built from the sub07 genomes and with the alignment of its 32 confident detections (**Fig. 4**). The element has three parts: a closing stem of 11-12 bp that joins its two ends, a left domain with two hairpins, and a long hairpin of 29-34 bp. Just inside the closing stem, two guanosines face each other, one after its 5′ arm and one before its 3′ arm; they are present in 97% and 83% of the 30 copies in which the stem is aligned, and both together in 83%, whereas the neighboring positions are variable (G in 0-14%).

In this alignment, 32 of the 62 pairs of the structure covary significantly (R-scape, $E < 0.05$), against a single pair outside it, and neither random sequences forced onto the model nor the best hits in shuffled windows give more than one covarying pair. Because the model itself chose the sequences, however, part of this signal comes from the selection: the best hits in the upstream control windows still give 7 covarying pairs. We therefore rebuilt the alignment leaving out one part of the structure at a time and measured covariation in the part left out. Only the long hairpin keeps more covarying pairs than expected (12 of 33, against 8.7 expected), whereas the left domain (3 of 18; 3.1 expected) and the closing stem (1 of 11; 4.5 expected) do not. The pairing of the left domain and of the closing stem therefore remains hypothetical.

For detection, we used a shorter covariance model (comb287, with a 7-bp closing stem), which finds the element in 17 of the 18 training genomes and in none of 1,780 shuffled windows (**Table S2**); 17 of those genomes form a clade of virulent phages in the final RT tree (UFBoot 88). Outside the training set, it finds the element with confidence ($E \leq 10^{-5}$) in three independent genomes, also of virulent phages; weaker hits are as frequent as in the control window upstream of the RT and cannot be distinguished from background. The hits begin a median of 21 nt before the end of the RT gene, on the same strand. Among the 179 representatives, the element is found with confidence in 20 of the 87 virulent phages and only weakly in 2 of the 92 temperate ones; because the model was trained on virulent phages, this difference partly reflects its sensitivity, and a different form of the element may exist in the temperate phages. The element lies downstream rather than upstream of the RT and is not detected by any of the retron msr-msd models, although its overall topology recalls them: a closing stem enclosing internal hairpins, with two facing guanosines just inside that stem, at the positions of the branching guanosine and of the opposing guanosine in a retron msr-msd.

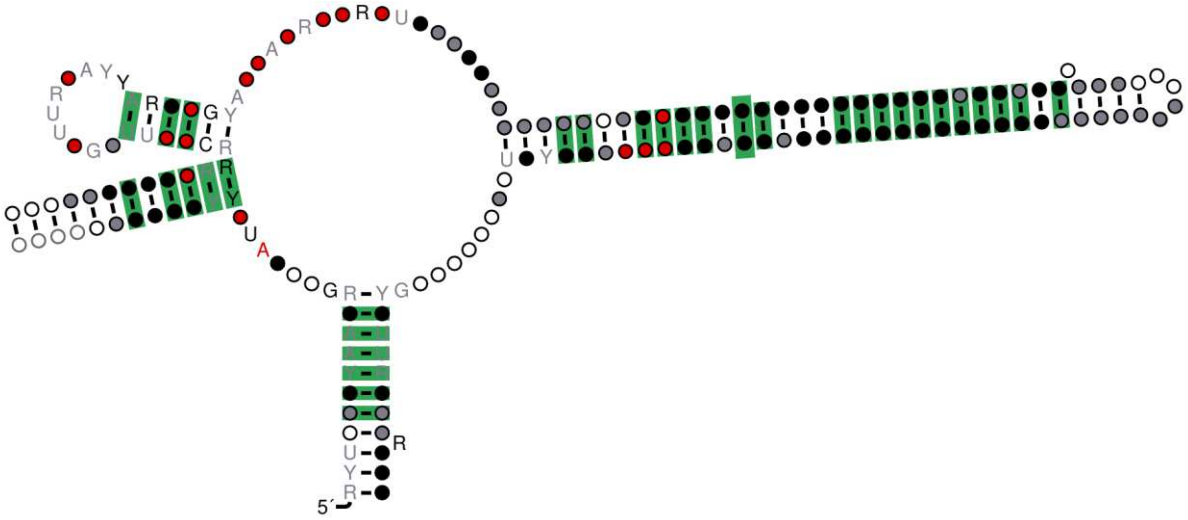

Figure 4. Consensus secondary structure of the sub07 element. R2R drawing of the alignment of its 32 confident detections by the iterative covariance model r0: 17 sub07 genomes, 8 genomes of other VIRE clades, and 7 regions of SPIRE and NCBI genomes. Green boxes, base pairs with significant covariation (R-scape, E < 0.05; 32 of the 62 pairs); red nodes, highly conserved nucleotides. Because the model chose the sequences, part of this covariation comes from the selection; when each part of the structure is left out of the alignment used to select the sequences, only the long hairpin keeps more covarying pairs than expected (12 of 33; 8.7 expected), whereas the left domain (3 of 18; 3.1 expected) and the closing stem (1 of 11; 4.5 expected) do not. The two guanosines that face each other inside the closing stem, at the base of the structure, are present in 29 and 25 of the 30 copies in which the closing stem is aligned (97% and 83%).

The VOTH RTs are bacterial retron types carried by phages

Unlike the PRL1 RTs, the VOTH RTs score like retrons with the retron profiles (**Table 1**), and all 51 lie among the bacterial retrons in the tree. We assigned each representative to the type that dominates the smallest clade of the tree of 985 sequences that contains it and at least three retrons (**Table 5**); they are spread over 11 types and clades and three clades of mixed types. Six belong to the type XI-like lineage TXI_C2like described in our retron survey [Toro, 2026]: they score higher with the profile of that lineage than with that of type XI (278-341 against 224-247 bits) and, placed on the reference tree of that work, fall inside the TXI_C2like clade with short pendant branches.

Most VOTH are single tips among bacterial retrons. Twenty of the 51 form six phage-only clades of two to six tips (UFBoot 85 or more), five of them subclades of bacterial types or of TXI_C2like (**Table 5**). The sixth is formed by three VOTH (UFBoot 98) that branch in a large clade of mixed types and are not close to any type; these three resemble neither the retrons nor PRL2 clearly (123-132 bits with either profile), and we mention them only as an example of the diversity that a screen of phage genomes can retrieve.

Retron types are defined by the effector that accompanies the RT. We scanned the genes within two positions of the RT in the 151 VOTH loci with the 35 effector profiles of the retron survey: 71 genes in 54 loci (36%) match one. Among the 27 VOTH representatives in types whose bacterial members usually carry an effector of the panel, 17 have it (**Table 5**): all 15 in the clades of types II-A1 and I-A and two of three in type III-A1, whereas it is absent in types II-A3, VI, and X. The 21 VOTH in types and lineages without an effector in the panel and the three in the large clade of mixed types have none, but three of the six TXI_C2like loci have the WYL regulator characteristic of that lineage, and one VOTH locus carries a CD-NTase-like protein of the type that makes cyclic oligonucleotides in CBASS systems. The VOTH are therefore bacterial retron systems carried by phages, which keep the effector in the types where it is the rule in bacteria. The scan covers only two genes on each side and the effectors of the panel, so effectors farther away or of other families would be missed, and all VOTH come from VIRE: we did not search SPIRE or NCBI for retrons on viral contigs.

Table 5. Retron type of each VOTH representative and effectors next to the RT

Type or lineage of the clade	VOTH	Phage-only subclade (UFBoot)	Effector expected (share of bacterial references with it)	With it	With another effector or regulator	Without
II-A1	8	none (single tips)	NDT (81%)	8	0	0
Mixture of I-A and II-A1	2	none (single tips)	NDT or ATPase + HNH	2 (NDT)	0	0
I-A	5	one of 3 (85)	ATPase + HNH (89%)	5	0	0
III-A1	3	none (single tips)	PRTase / WH (67%)	2	0	1
II-A3	3	one pair (55)	NDT (86%)	0	2 (PRTase, WH)	1
VI	2	none	HTH / sp (63%)	0	0	2
X	4	one of 3 (99)	cluster 2273 (92%)	0	0	4
Lineage TXI_C2like (clade dominated by type XI)	6	one of 6 (97)	none identified (WYL regulator)	-	3 (WYL)	3

Type or lineage of the clade	VOTH	Phage-only subclade (UFBoot)	Effector expected (share of bacterial references with it)	With it	With another effector or regulator	Without
I-C1 (4), I-C2 (6), I-C3 (1), XI (1), clade 2 (3)	15	I-C2: one of 3 (97); I-C1: one pair (91)	none in the panel	-	1 (HTH)	14
Large clade of mixed types (type VI, 21%)	3	one of 3 (98)	-	-	0	3
Total	51			17	6	28

The type is that of the clade of the tree of 985 sequences that contains the representative; effectors were searched within two genes of the RT.

PRL2: phage RTs of large genomes in several lineages

PRL2 contains the RTs reported as ARTs [Yoon et al., 2026] and those called NART in the same work (non-array-associated RTs, the sister clade of the ARTs in their tree), which correspond to one of our lineages, RTp (defined below); all eight staphylococcal RTs of the ART preprint fall in it. The initial search of VIRE recovered 207 PRL2 proteins, whose 21 representatives and the eight staphylococcal RTs form the 29 PRL2 tips of the tree of 567 sequences. Because the profile of the staphylococcal RTs scored the VIRE members poorly (median 111 bits), we built a new profile, FAMILY31, from these 29 sequences and the two RTs of LPJP1, and searched VIRE and four other collections again (**Methods**, **Table S3**). FAMILY31 separates PRL2 from all other RT groups (median 408 bits for the VIRE members and 269-276 for the public RTs, against 199 bits at most for any other sequence of the tree), also when each of the 29 sequences is left out of the profile, except for the RTc (below).

The new profile added 39 RTc proteins in VIRE that the initial search had missed, bringing PRL2 to 247 proteins, and a few more in the other collections: two further *Staphylococcus* phages and LPJP1 in INPHARED, three RTs on bacterial contigs of SPIRE, seven RTs of Bacillaceae genomes in NCBI, and one RT of a 294-kb jumbo phage in the Unified Human Gut Virome (UHGV) (**Table S3**). In the final trees, PRL2 has 39 tips. Many of these proteins come from closely related phages: the RTr and the RTc each fall in a single genus-level cluster, so that their 179 proteins form only 11 clusters at 95% identity. We therefore count lineages and clusters rather than proteins, and say so where a count per protein is given.

In contrast to PRL1, PRL2 is carried by large phages: the median genome size of the VIRE representatives is 232 kb (15 of 21 are 200 kb or more), the staphylococcal phages have 258-274 kb, and LPJP1 has 224 kb.

Three features organize PRL2: the clade of the RT, the N-terminal region that precedes the RT domain (below), and the protein encoded immediately after the RT, which we call its partner (the "partner gene" of the ART preprint), without implying a function. Together they define the lineages of **Table 6** (**Figs. 5** and **6**), three of which have working names. The RTr, introduced above, is preceded by a repeat array in 81 of its 90 proteins; the RTc occurs only in RTr genomes, about 4.4 kb downstream of the RTr. The RTp (nine tips; UFBoot 100) is named for the polymerase: in seven of them, the RT is followed by a homologue of the β-like subunit of the jumbo-phage non-virion RNA polymerase (nvRNAP; gp123 of phiKZ [de Martín Garrido et al., 2021]), the gene next to which these RTs were first noticed in the ART preprint [Yoon et al., 2026]; the RTs of the UHGV phage and of two SPIRE contigs also fall in this lineage.

The region before the RT falls into three classes: A, an intergenic gap of at least 1 kb with a repeat array; A2, the same gap without repeats; and B, genes or a short gap between genes. We refer to an RT by the class of this region (for example, the class A and class B RTs of LPJP1). The classes do not follow the phylogeny: class B occurs in most lineages, including the RTc, the RTp lineage, and three RTr proteins, and all six A2 proteins are RTr.

The other lineages are the short class B lineage (five tips; RTs of 283-322 aa without an N-terminal region, next to which the bacterial RTs of ClusteredNR and one SPIRE RT are placed), the eight staphylococcal public RTs, and small lineages of one or two tips, named by the VIRE identifier of their RT (genome and gene number; for example, 1629013_162 is gene 162 of VIRE genome 1629013): the 538531_117 group, 1629013_162 and 1250369_48, 604347_41, and 1461883_135. The names are labels, not definitions: nine RTr proteins have no repeats, arrays also occur in other lineages, and two RTp tips lack the nvRNAP partner. The two RTs of LPJP1, which lie in one locus, are not assigned to any lineage (below).

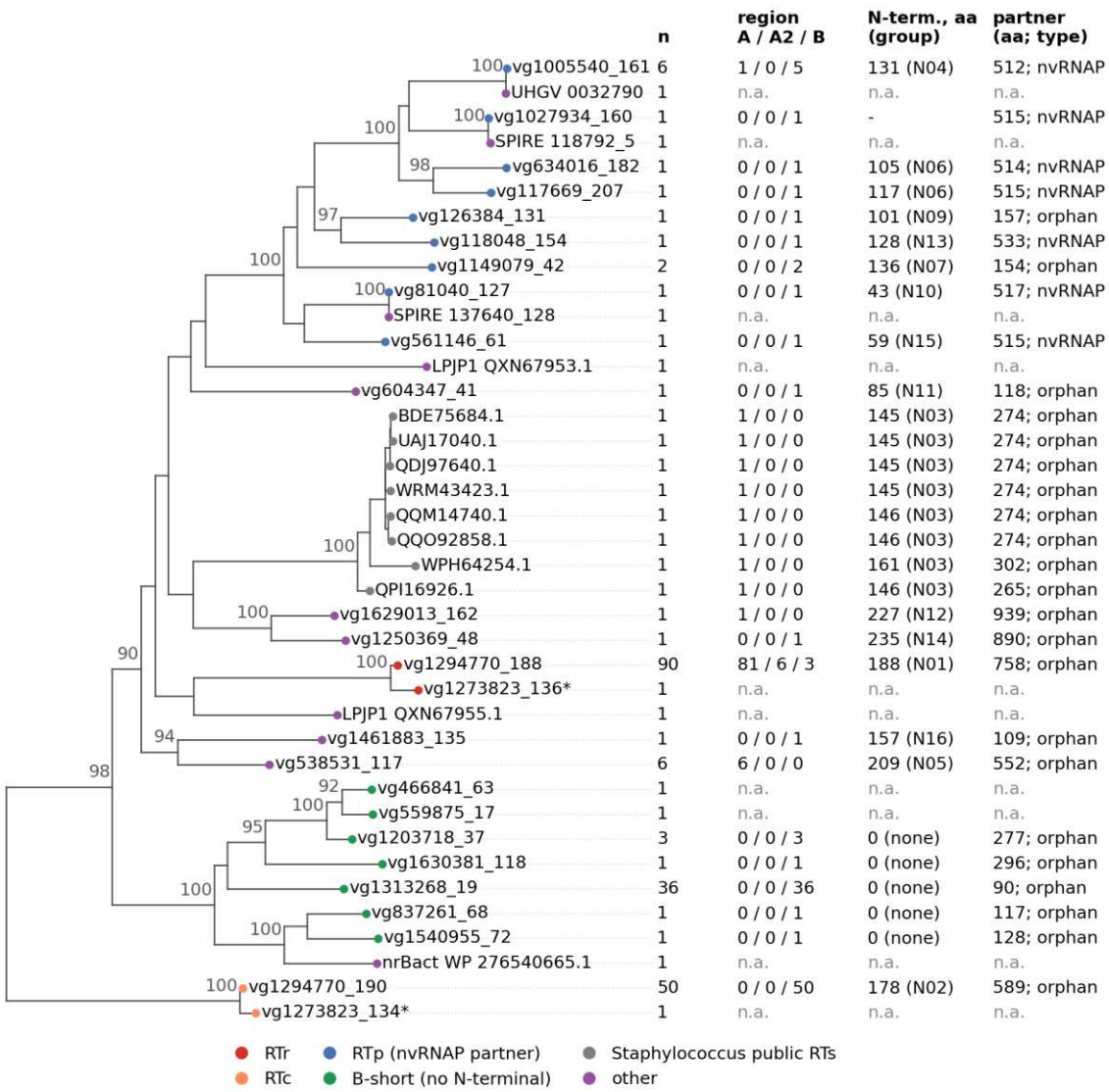

Figure 5. Lineages of PRL2. Subtree of the 39 PRL2 tips of the maximum-likelihood tree of 985 RTs (Q.pfam+F+R5; shown rooted on the DGR RTs; all other tips pruned; numbers are UFBoot values of 90 or more). The two RTc tips are not part of the clade of the other 37 tips (UFBoot 98): in the complete tree, they branch immediately outside the clade that those 37 tips form with PRL1, the retrons, and VOTH (see text). Tip labels: vg, VIRE RTs, named by genome and gene number as in the text; SPIRE and UHGV, RTs from these collections; nrBact, the bacterial RT from ClusteredNR; other labels are GenBank protein accessions of the public RTs. For each tip: number of proteins it represents (n), class of the region before the RT (A / A2 / B), N-terminal region (length in aa and group at 30% identity), and length and type of the partner. Colors: RTr (red), RTc (orange), RTp (blue), short class B (green), staphylococcal public RTs (grey), and other small lineages (purple). Tips added to the 29 original ones (UHGV, SPIRE, LPJP1, the bacterial RT, two short class B RTs, and the RTr and RTc of genome 1273823) have no values in these columns (n.a.); asterisks mark the RTr and RTc of genome 1273823, from the second 90%-identity cluster of the RTc.

Table 6. Lineages of PRL2

Lineage	Tips (proteins)	N-terminal region, aa (group)	Partner	Region before the RT	Genomes
RTr	1 (90)	188 (N01)	758 aa, orphan; 8-nt overlap	A 81, A2 6, B 3	VIRE; 89 of 90 also carry an RTc
RTc	1 (89)	178 (N02)	589 aa, orphan; 8-nt overlap	B	the RTr genomes
RTp	9 (15), plus the UHGV RT and two SPIRE RTs	43-136	nvRNAP β-like subunit, 512-533 aa, in 7 tips; 154-157 aa in 2	B (one A)	128-340 kb, median 287 kb
Short class B	5 (43)	none	90-296 aa, orphan	B	VIRE; bacterial relatives in ClusteredNR and SPIRE
Staphylococcal public RTs	8 (8)	146 (N03)	265-302 aa, orphan	A (2-4 copies)	*Staphylococcus* phages, 258-274 kb
1629013_162 and 1250369_48	2 (2)	227 and 235	939 and 890 aa, orphan	A (10 copies) and B	VIRE
538531_117 group, 604347_41, 1461883_135	3 (8)	209, 85, 157	552, 118, 109 aa	A (11 copies), B, B	VIRE
LPJP1 (not assigned)	2 RTs	108 and 149	481 and 661 aa, orphan	A (gap of 1,208 nt, four copies) and B	*Listeria* phage, 223.6 kb

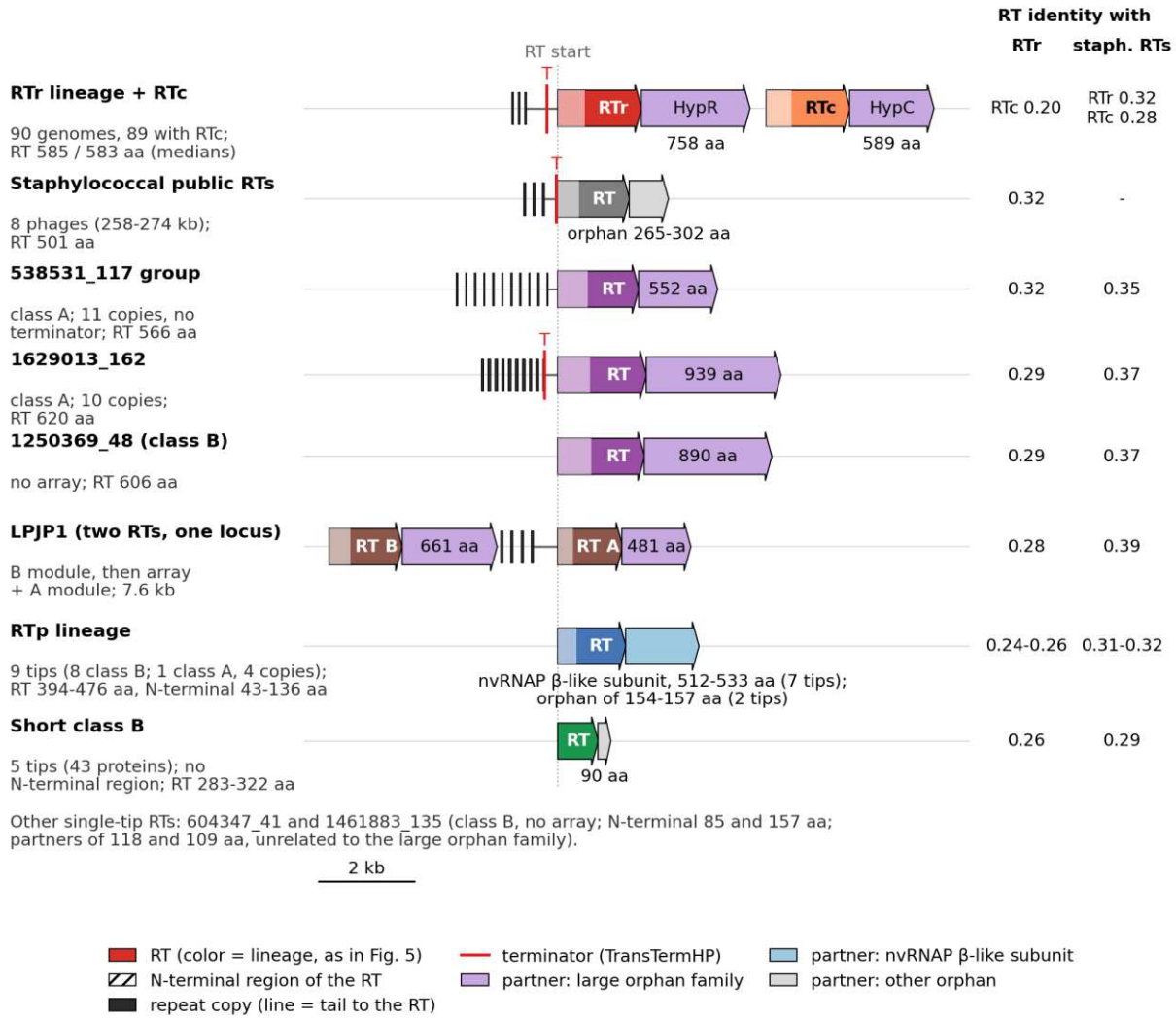

Figure 6. Architectures of the lineages of PRL2. Loci drawn to scale from the RT start codon (3 nt per residue; typical copy number; medians). RT colors follow **Fig. 5**; the lighter part of each RT is its N-terminal region; black boxes are repeat copies, and the line is the tail; red marks are terminators. Partners: lilac, large orphan family that shares a ~150-aa domain; light blue, nvRNAP β-like subunit; grey, other orphans. HypR and HypC are the partners of the RTr and RTc, which overlap them by 8 nt. LPJP1 is shown as one locus of 7.6 kb: the class B RT and its partner are followed, 11 nt after the partner, by the array and the class A RT. The last two columns give the median identity of the RT (365 aligned positions) with the RTr and with the staphylococcal public RTs. Scale bar, 2 kb.

The RTr-RTc locus, the partners, and the N-terminal region

In the 89 genomes that carry both the RTr and the RTc, the two genes are on the same strand and, in 88, in the same order: array, RTr, partner of 758 aa, RTc, partner of 589 aa (**Supplementary Fig. S3**). In 86 genomes, the partner of the RTr is the only gene between the two RTs, and the RTc starts a median of 4.4 kb after the RTr. Each RT overlaps the start of its partner by exactly 8 nt in all 90 RTr and all 89 RTc loci, which suggests that the two genes of

each pair are translated together. The 340 nt between the end of the RTr partner and the start of the RTc are nearly identical in all genomes (mean pairwise identity 98.3%) and fold into a series of long helices with a hairpin close to the start codon (z = -7.0 against shuffled sequence), but they vary too little to test the structure by covariation. These genomes are closely related phages of one genus-level cluster; the pairing is therefore one observation seen in many related genomes, not 89 independent ones.

The partners of the RTr and RTc, which we call HypR and HypC, respectively, are conserved within their lineage and have no Pfam domain, no significant match in PDB70, and no homologue in UniRef30 (**Table S4**). They are distantly related to each other and share a region of about 150 residues (33% identity, E = 1e-21). AlphaFold2 models of this region, built with alignments of the 265 partners of PRL2, are confident and superpose on each other (TM-score 0.81-0.84) but match no structure in PDB100; with alignments against UniRef30, which finds no homologue, the same regions are predicted with low confidence, so the confidence comes from the covariation in the PRL2 alignment, which contains few independent sequences, and the agreement of the two models is the stronger result. AlphaFold3 models of the RTr, RTc, and their partners in one genome agree (**Table S5**): the shared region is a compact domain in both partners and superposes on the AlphaFold2 models (TM-score 0.88-0.92). The C-terminal region of HypR resembles a GCN5-related N-acetyltransferase (GNAT) fold (Foldseek [van Kempen et al., 2024]) but lacks the acetyl-CoA binding motif, so we treat this as a possible fold rather than a function. The rest of both partners is predicted with low confidence.

The family of large orphan proteins to which HypR and HypC belong (group I in **Table 7**) also follows the RTs of the 538531_117 group, 1629013_162 and 1250369_48, and both RTs of LPJP1 (481-939 aa; for example, 31% identity over 486 residues between the RTr partner and that of the 538531_117 group). The partners of the staphylococcal RTs, RTp lineage, and short class B lineage are unrelated to it (below).

Most PRL2 RTs have, before the RT domain, an N-terminal region of up to ~300 residues, whose limit we defined from profile-profile alignments with group II intron RTs, the closest RTs of known structure (**Methods**). Its length is characteristic of each lineage, from 43 residues in the RTp lineage to 188 in the RTr, and it is absent in the short class B lineage (**Table 6**). At 30% identity, the regions form 17 groups that coincide with the lineages, and none has a Pfam domain or a convincing match in PDB70. In AlphaFold2 models of nine full-length RTs from five lineages and LPJP1 (**Table S6**), the RT domain is predicted with confidence whereas the N-terminal regions are not, and they match no known structure; in AlphaFold3 models of the RTr and RTc, the position of the N-terminal region relative to the RT domain is not defined (**Table S5**). Only the N-terminal regions of SA1 and of one RTp protein have a sizeable confident part, without a match in PDB100. The N-terminal region is therefore a marker of each lineage, whose structure and function are unknown.

How the lineages are related, and the place of the RTc

In the tree of 985 sequences, the 37 core tips of PRL2 form a well-supported clade (UFBoot 98) with two branches (**Fig. 5**; **Supplementary Fig. S4**): one holds the short class B lineage and a bacterial RT from ClusteredNR (UFBoot 100), and the other all the remaining lineages (UFBoot

90). In the second branch, the 538531_117 group and 1461883_135 split off first (UFBoot 94); the order of the other lineages is poorly resolved, although the RTp lineage, which includes the UHGV and SPIRE RTs, is a robust clade (UFBoot 100) nested among lineages with repeat arrays. RT identity agrees: the RTr is most similar to the staphylococcal RTs and the 538531_117 group (median identity 0.32) and least similar to the RTp lineage (0.24-0.26), and the lineages closest to it tend to have repeat arrays and partners of the large orphan family (group I).

The partners, in contrast, do not follow the RT tree. A profile-profile comparison of 26 representatives gives probabilities of homology that are either 97-100% or below 50%, and sorts them into four homology groups and seven unrelated proteins (**Table 7**, **Supplementary Fig. S4**). Group I follows the RT in six lineages that are scattered in the RT tree, whereas the staphylococcal RTs and the RTp lineage, which lie among them, carry proteins of groups III and II. The gene that follows the RT was therefore either replaced as these lineages diversified or diverged beyond recognition. The staphylococcal RTs are not RTr that lost their companion: their genomes contain no remnant of the RTc, HypR, or HypC (tblastn [Camacho et al., 2009], E ≤ 1e-3), and their RTs are as close to the RTr as to the 538531_117 group and LPJP1 (27-29% identity over the whole protein).

Table 7. Homology groups of the partners (profile-profile alignments of 26 representatives)

Group	Partners	Lineages	Relationship between members (probability of homology)
I, large orphan family	758 (RTr), 589 (RTc), 552 (538531_117 group), 939 (1629013_162), 890 (1250369_48), 661 and 481 aa (LPJP1)	RTr, RTc, 538531_117 group, 1629013_162, 1250369_48, LPJP1	97-100% for every pair; all align with the ~150-aa region shared by HypR and HypC (137-141 of its 141 residues), four of them over 311-658 residues of HypR
II, nvRNAP β-like subunit	512-533 aa (7 representatives)	RTp (7 tips)	100% among them; none to group I
III, staphylococcal	265, 274 and 302 aa	staphylococcal public RTs (8)	100% among them; none to the others
IV, small orphan	128 and 117 aa	short class B (1540955_72, 837261_68)	99% between them

Group	Partners	Lineages	Relationship between members (probability of homology)
Unrelated to any other	90 (37 proteins), 277 (3 proteins), 296 aa; 154 and 157 aa; 118 and 109 aa	short class B (1313268_19, 1203718_37, 1630381_118), RTp with orphan partner (1149079_42, 126384_131), 604347_41, 1461883_135	below 50% with every other partner

LPJP1 carries two PRL2 modules in one locus of 7.6 kb (**Fig. 6**): a class B RT (509 aa) with its partner, followed 11 nt later by an array of four copies, and then a class A RT (454 aa) with its partner. The two RTs are 35% identical to each other and 25-33% to the RTr and RTc, and both partners belong to group I. In the tree, the class B RT groups with the RTr and the class A RT with the RTp lineage, both weakly (UFBoot 59 and 65), so neither can be assigned to a lineage. Two RTs in one locus therefore occur twice in PRL2: as the RTr-RTc pair, in one genus-level cluster, and in LPJP1. The four group II introns of LPJP1 [Merk et al., 2025] are unrelated to these modules: the intron-encoded RT (QXN67909) matches neither the PRL1 nor the PRL2 profile, and the nearest intron lies more than 15 kb from either PRL2 RT.

The RTc is a separate, deep lineage. It is as distant from the RTr (identity 0.19-0.22) as from PRL1, the retrons, or non-retron RTs (0.19-0.22), whereas the RTr is closer to the other lineages of PRL2 (0.24-0.33). Its partner is equally distinct: in a tree of the 190 partners that carry the shared region, the HypC proteins form a separate branch (UFBoot 100), whereas the HypR proteins group with the partners of other lineages. The position of the RTc is the least stable feature of the trees (**Table S1**). In all five trees it branches just outside the clade formed by the core of PRL2, PRL1, the retrons, and VOTH (UFBoot 87-99), but in a tree of the RT domains of the 179 RTr and RTc proteins alone it is the outermost branch of PRL2 (UFBoot 90). The AU test does not settle the question: 14 of the 267 positions tested are not rejected, including placements next to or inside the core of PRL2, next to non-retron RTs, and inside the retrons.

The RTc is nevertheless linked to PRL2 by its locus, where it accompanies the RTr with a group I partner, and by the PRL2 profile, which scores it at 297 bits when it is part of the model but at only 128 bits when it is left out (best non-PRL2 sequence, 185 bits). We therefore treat the RTc as a separate lineage that accompanies the RTr and leave its relationship to the PRL2 clade open: it may have diverged before the rest of PRL2 or evolve faster, and a long branch attracted toward the outgroup could explain its position.

A repeat array of a short stem-loop precedes the RT

The region before the RT that first drew our attention to the public RTs is, in several lineages of PRL2, an array of short repeated copies followed by a tail of several hundred nucleotides that ends at an intrinsic terminator (**Supplementary Fig. S5**, **Table 8**).

In the RTr lineage, each repeat copy is 32 nt long. The consensus of the 231 copies that could be aligned, out of the 235 detected in the 81 class A regions, is ACGUAAAGGUACAAUGAGAGGAUCUCUCAUCA, which folds into a hairpin with a stem of 7 bp and a loop of 3 nt, preceded by an unpaired segment of 13 nt. The copy is the conserved unit: the most frequent base accounts for 0.94 of its positions on average, against 0.65-0.67 in the 40 nt on either side, and conservation drops sharply at its borders (**Supplementary Fig. S6**). The motif GUAAAGG is almost invariant. One stem pair shows compensatory changes that keep the pairing, but the stem is not more conserved than the unpaired positions ($p = 0.56$), so the structure of a single copy is weakly supported.

The copies sit on a fixed grid: separated by spacers of 110 nt, they recur with a period of 142 nt at -966, -827, -689, and -545 nt from the RT start codon (**Supplementary Fig. S7**). Genomes with two, three, or five copies differ by whole units of this grid, and the tail between the last copy and the start codon shortens by one unit for each added copy. The spacers are specific to their position: the spacer at a given position is nearly identical across genomes (identity 0.92), whereas the spacers at different positions of the same array are unrelated (0.45, against 0.39 for random sequence). The array is therefore neither a CRISPR array, whose spacers are short, unique, and acquired, nor a diversity-generating retroelement, which has two long near-identical repeats.

Copy number changes repeatedly among nearly identical RTr sequences: at least 20 changes are needed on the tree of the 89 distinct RTr sequences, fewer than for shuffled states (28.1; $p = 0.0005$), although whether arrays grow or shrink cannot be told. The repeat is confined to the locus: in 86 complete RTr genomes, 203 of 214 matches to the copy lie in the array window; the other 11 lie in four genomes, where in three they extend the same grid upstream (about -2 kb) and in the fourth, the only genome without an RTc, lie elsewhere on the contig.

Other lineages have their own repeats with the same layout, but the copy and the spacing are specific to each (**Table 8**). Measured as the distance between the starts of consecutive copies, the period is 142 nt in the RTr lineage, 189 in the 538531_117 group, about 193 in the staphylococcal RTs, 144 in 1629013_162, and 57 in the class A protein of the RTp lineage (a mean over each array, except in the RTr lineage, because the other spacers are more irregular). The consensus copies of all lineages fold into one or two stem-loops (**Table S7**), but their sequences are unrelated, so the arrays share an organization rather than a sequence. The arrays are also nearly identical within lineages: the 95 class A and A2 proteins and the eight staphylococcal RTs form only ten clusters at 90% identity, so the number of independent arrays is approximately ten.

Table 8. Repeat arrays by lineage

Lineage or group	Proteins	Copies	Copy length	Spacer	First copy	End of last copy	Tail	Terminator	Tail z*

Lineage or group	Proteins	Copies	Copy length	Spacer	First copy	End of last copy	Tail	Terminator	Tail z*
Staphylococcal public RTs	8	2-4	51 (34-65)	140 (128-162)	-717	-281	280 (269-282)	-40	-7.1
RTr, class A	81	2-5	32 (28-49)	110 (75-116)	-964	-660	659 (513-803)	-232	-7.9
RTr, class A2	6	0	-	-	gap of 1,293	-	-	-232	-
538531_117 group	6	11	31	121	-2,124	-203	202	none found	-3.1
RTp, class A protein	1	4	31	26	-1,220	-1,018	1,017	-219	-8.2
1629013_162	1	10	57	105	-1,607	-259	258	-284	-3.6
LPJP1, RT of 454 aa	1	4	49 (41-60)	143 (108-234)	-1,197	-513	513	-50	-8.4**
Machias, AH12 (INPHARED; not in the VIRE set)	2	2	34	163	-509	-280	280	-40	not computed

Values are medians, with ranges in parentheses; positions are in nt from the RT start codon. Folding z-score of the tail against 60 dinucleotide-shuffled sequences (RNAfold). Of the 81 RTr class A tails, 18 were scored: ten longer than 700 nt and eight of the 71 of 500-700 nt. *Mononucleotide shuffling; not comparable with other values.

The tail that follows the last copy ends at an intrinsic terminator whose position is fixed within each lineage: 232 ± 2 nt before the start codon in all 87 class A and A2 RTr proteins, including the six without repeats, 40 nt in the staphylococcal RTs, and 50 and 130 nt before the class A and class B RTs of LPJP1 (**Table 8**). The tails fold better than shuffled sequences, but no conserved structure is supported by covariation (RTr tails: 2 covarying pairs where 12.4 ± 2.8 would be expected; staphylococcal tails: 0 of 38 predicted pairs, with little power), and none of the 21 msr-msd covariance models matches the region.

The only transcriptomic data for a PRL2 member are RNA-seq data of Staphylococcus lentus infected by the jumbo phage SA1 [Zhang et al., 2022] (**Fig. 7**). On the strand of the RT, the region before the RT is covered by two transcripts with sharp 3′ ends: the first ends 13 nt after the second repeat copy, and the second starts at the third copy and runs through the tail to 6 nt before the terminator. Both peak at 15 min after infection (42 and 59 times the genome mean) and fall by 55 min, whereas the RT itself is barely covered (0.2-1.4 times the mean), and only 0.2% of the fragments run from the tail into the RT. No σ70-like promoter lies upstream of the first or third copy, but the copies contain TATAAT-like elements. These discrete, abundant transcripts agree with the processed array RNAs described for the same data in the ART preprint [Yoon et al., 2026]. The tail is therefore a transcribed RNA of unknown function, not a recognizable msr-msd. Because SA1 has no RTc, these data cannot tell whether the RTr and RTc are transcribed together.

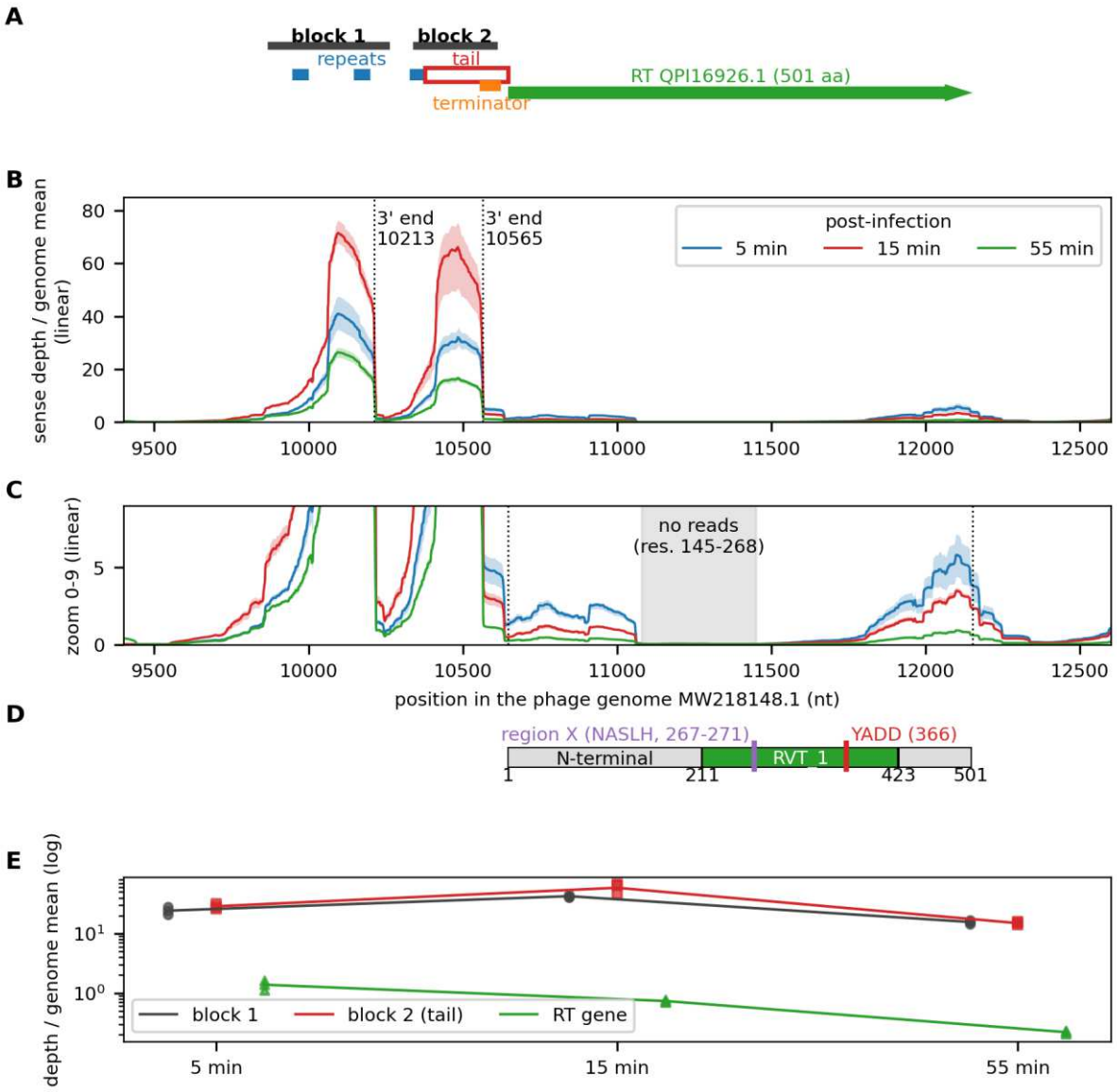

Figure 7. Transcription of the region before the RT in phage SA1. RNA-seq of PRJNA836150 (host Staphylococcus lentus JTB1-3), sense strand. A, the region before the RT: repeat copies (50-54 nt), tail (269 nt), terminator, and the two transcribed blocks. B, sense-strand depth relative to the genome mean at 5, 15, and 55 min after infection (mean of three replicates and range); dotted lines, 3′ ends of the two transcripts. C, the same data on a 0-9 scale; the region without reads corresponds to residues 145-268 of the RT. D, positions of the RT protein on the genome (N-terminal region, RVT_1 domain, region X, and YADD). E, time course of the two blocks and of the RT gene (each point is one replicate; lines join the means): the blocks peak at 15 min, whereas the RT stays at or below the genome mean and falls late. The samples without phage are not shown.

Discussion

Three ways of being related to the retrons

The three groups of phage RTs are related to the retrons in different ways. The VOTH are retrons: they fall within the bacterial retron types and usually carry the effector where the type has one (**Table 5**), so phages carry complete retron systems, as they carry many other defense systems. PRL1 and PRL2, in contrast, are not assigned to any retron type. PRL1 is the sister group of the retrons in three of the four trees and under both substitution models, and the AU tests reject almost every position inside them; the remaining alternatives place it next to a group of types I-A, I-B, and II-A1 or next to the core of PRL2. The core of PRL2 branches one step further out, and the RTc outside it. Their motifs also differ: PRL2 RTs keep both retron motifs, whereas PRL1 RTs keep only the closing histidine of region X and have their own region Y. The trees cannot tell in which direction these RTs moved between bacteria and phages: PRL1 and PRL2 may descend from ancestors shared with the retrons that persisted in phages, or may have left the retrons early and diverged. Because our collection is centered on phages and the alignment is short, we describe them as phage RT lineages related to the retrons, not as new retron types.

Is PRL1 a retron-like system?

Three observations suggest that PRL1 is a retron-like system. Its RT is the sister of the retron RTs and keeps the catalytic motif and the closing histidine of region X; in a clade of virulent phages, a structured RNA of ~200 nt, whose long hairpin is supported by covariation, overlaps the end of the RT gene; and the RT is accompanied, at a constant distance and in very different phages, by a module of the kind found in anti-phage systems: an E1-like enzyme with the fold of Cap2 and a JAMM protease with an intact active site, as in type II CBASS and related bacterial ubiquitin-like systems, and, a few kilobases upstream, a protein with the ubiquitin fold. However, four observations argue against it or remain unexplained: the RT has its own region Y, closer to that of group II intron RTs than to that of retrons; the RNA lies downstream of the RT, whereas the msr-msd of most characterized retrons lies upstream, it matches no known msr-msd model, and it has been found in one clade only; no known retron effector lies next to the RT, and the module lacks typical features of conjugation systems (the C-terminal Gly-Gly of the ubiquitin-like proteins, the CxxC loop of canonical E1 enzymes, and any detectable E2 or E3); and the E1-like gene follows the RT phylogeny no better than other conserved phage genes, so their association may reflect shared descent.

The RNA nevertheless has the topology of an msr-msd. A stem joins its two ends, as the a1/a2 repeats do in a retron; two guanosines face each other just inside that stem, where the branching guanosine and the opposing guanosine of a retron sit (present in 97% and 83% of the copies, and both in 83%); and the interior is a short stem-loop domain followed by a long hairpin, the arrangement of msr and msd. What is missing is not the shape but independent support: the closing stem covaries in the alignment drawn in Fig. 4 (6 of its 11 pairs), but not when it is left

out of the model used to select the sequences (1 of 11, where 4.5 would be expected), and the element lies after rather than before the RT.

PRL2 and the ART loci

Our observations on PRL2 agree with those of the ART preprint [Yoon et al., 2026], which describes the same RTs from a different starting point: an array of repeats with a palindromic core before the RT, with spacers kept in order between related phages; an N-terminal extension of unknown fold; an intact YxDD motif; and no *cas* genes. Its partner types I and II correspond to our group I, which includes HypR and HypC and the GNAT-like fold of HypR, and to our staphylococcal group III. Both analyses place the group next to the retrons and outside them. They differ on two points: in their tree, the NART RTs are the sister clade of the ARTs, whereas in ours the RTp lineage is nested, with weak support, among lineages with arrays; and the ARTs are described as largely confined to jumbo phages, whereas PRL2 also includes RTs of phages of 128-340 kb and of bacterial contigs.

PRL2 does not resemble a retron. Instead of an msr-msd, it has an array of copies of a short stem-loop on a fixed grid, followed by a tail that is transcribed in SA1 but has no structure supported by covariation; its RT carries an N-terminal region of unknown structure; and the partner differs between lineages. None of these features has a counterpart among retrons, and the array is neither a CRISPR array nor the template-target pair of a diversity-generating retroelement. The ART preprint proposes, as a hypothesis, that each RNA unit of the array forms its own complex with the RT and the partner, so that one enzyme would work with a bank of distinct RNAs; the abundance and processing of the array RNAs fit this model, but msDNA has not been examined. We therefore leave open whether PRL2 is retron-like; its RTs may also work by a different mechanism.

A stronger version of this idea proposes that each repeat unit of an ART array is itself an msr-msd [Ye, 2026], on the basis of a scan for complementary arms of at least 8 bp separated by 100-300 nt, with a guanosine at a fixed position. Two observations argue against it. First, this criterion does not distinguish msr-msd units from repeats. With a simplified version of it, we count 18-45 such inverted repeats in the 900 nt before the RT of the five public loci with arrays, two to three times more than in shuffled sequence (6.6-15.1); masking the two to four repeat copies of each array reduces the count to 4-26, below what masking the same length of random sequence leaves (11-36), and the class B RT of LPJP1, which has no array, has only 5. The excess thus comes from the stem arms of neighboring copies, which lie one repeat period (100-300 nt) apart. Second, transcription does not follow the proposed units: in SA1, the first half of a candidate element is covered at 31-55 times the genome mean, but its second half, which would contain the msd, only at 1.2-3.6 times, the transcript ending 13 nt after the repeat copy. Whether the repeat units are msr-msd genes will have to be resolved experimentally, by testing whether they are reverse-transcribed into msDNA.

What could these RTs do for a phage?

Defense systems carried by prophages often protect the lysogen against other phages, and the temperate half of PRL1 fits this role. That the same RT and E1-like core occurs in virulent

phages of ~50 kb is more surprising: these phages could use it against competing phages or against host defenses. The E1-like and JAMM pair, together with a ubiquitin-like protein, suggests the conjugation and removal of a small protein; if the RT and its RNA acted as a sensor and the E1-like module as the effector arm, as in many tripartite defense systems, the arrangement would be a defense architecture, but this remains a hypothesis. PRL2 is carried by large phages, and its RTs are followed by large orphan proteins in several lineages and by a second RT in one; whether they act in defense or in regulating the phage's own genes cannot be decided from these data.

Limitations

The RT alignment is short (365 columns), so several deep nodes have moderate support (UFBoot 85-88 in the main tree), several retron types are not monophyletic, the position of PRL1 relative to individual retron types and that of the RTc are not fixed, and the largest tree samples 467 of the 1,928 reference retrons. The collection is centered on phages, so the direction of transfer between phages and bacteria cannot be inferred. Lifestyles and hosts are computational predictions, and hosts are known for few representatives. The ncRNA model was trained on one clade and gives few independent detections. Structural annotations rely on predicted models, some of low confidence, and a fold match does not establish a function. Counts by genome are dominated by a few large clusters, which is why we report counts by cluster; apart from the RTr and RTc, each lineage of PRL2 is represented by one to a few independent phages.

Conclusions

This study maps the retron-related RTs carried by phages and shows that they are related to the retrons in three different ways: phages carry complete retron systems of known types (VOTH); a sister lineage of the retrons, PRL1, combines its RT with a conserved ubiquitin-like module and, in one clade, a structured RNA with the topology of an msr-msd; and a more distant group, PRL2, which includes the ARTs, has a locus organization with no counterpart among retrons, with lineage-specific repeat arrays, partner proteins, and, in two independent cases, a second RT. The profiles, covariance models, and phylogenetic placements developed here allow these lineages to be recognized in new genomes, and the analysis of the arrays shows that their repeat structure and transcription do not support the proposal that they are tandem msr-msd units. The remaining questions are functional and require experiments: whether representative loci of PRL1 and PRL2 produce msDNA and protect against phages, and whether the E1-like, JAMM, and ubiquitin-like proteins of PRL1 form a conjugation system.

Methods

Sequence collections

The eight *Staphylococcus* phage RTs used as queries (QQM14740.1, QPI16926.1, QQO92858.1, WRM43423.1, BDE75684.1, QDJ97640.1, UAJ17040.1, and WPH64254.1) are those listed in the ART preprint [Yoon et al., 2026] and were retrieved from GenBank; the two RTs of LPJP1 (MZ422438), a phage that is also among the ART loci of that preprint, were found in INPHARED with the ART8 profile. VIRE phage genomes and proteins [Nishijima et al., 2026], with the CheckV quality, BACPHLIP lifestyle [Hockenberry and Wilke, 2021], and host

predictions distributed with the resource; the complete VIRE proteome was searched as 93,155 per-biosample protein files (the metadata table of the resource lists 1,784,509 viral genomes from 80,254 biosample names and 732 studies; the 93,155 protein files are the ones distributed with the resource). SPIRE predicted proteomes and MAG proteomes [Schmidt et al., 2024]. NCBI ClusteredNR. INPHARED [Cook et al., 2021], in two snapshots: that of April 14, 2025 (34,076 genomes; proteins predicted with pyrodigal in metagenome mode) for the searches for PRL2 members, and that of August 9, 2025 (37,563 genomes; the Mash sketch distributed with Pharokka) for the comparison of phage genomes. Representatives of the Unified Human Gut Virome (UHGV; Zenodo 17402089; 7,426,124 proteins) [Camargo et al., 2025]. RefSeq Caudoviricetes genomes of 20-400 kb (5,375). The retron reference set of 1,928 RTs with their type assignments [Mestre et al., 2020; Toro, 2026]. Non-retron RTs from published classifications: UG/Abi [Mestre et al., 2022], DGR [Doulatov et al., 2004], group II introns [Lambowitz and Zimmerly, 2011], G2L [Toro and Nisa-Martínez, 2014; Kojima and Kanehisa, 2008], and RT-CRISPR [Silas et al., 2016].

Searches and definition of the groups

VIRE proteins were searched with hmmsearch (HMMER [Eddy, 2011]) using a profile of the eight staphylococcal RTs (ART8; 3,734 hits at E ≤ 1e-3, 2,705 with at least 50 bits) and the 30 RT profiles of the retron survey. Hits were clustered with MMseqs2 [Steinegger and Söding, 2017] at 60% identity and placed in a tree with retrons and non-retron RTs; PRL1 and PRL2 were defined as clades of that tree, and VOTH as the remaining VIRE RTs retrieved by the profiles. The PRL1 profile (391 positions; hmmbuild from the 179 representatives) was used with a threshold of 250 bits to search SPIRE and ClusteredNR. The hits were made non-redundant with MMseqs2 (90% identity, 80% coverage), and the 143 representatives were placed with EPA-ng [Barbera et al., 2019] on a reference tree of the RT groups that did not include them, before being added to the trees. PRL2 profile FAMILY31 was built from a MAFFT L-INS-i alignment of the 21 VIRE representatives, the eight staphylococcal RTs, and the two RTs of LPJP1; it was evaluated by leaving out each of the original 29 sequences against the other sequences of the reference tree, and used with hmmsearch (E ≤ 1e-3 on VIRE and E ≤ 1e-5 on the other collections, then 200 bits) on VIRE, INPHARED, SPIRE (23 chunks of genome proteins) and UHGV, and as BLASTp queries (29 sequences, E ≤ 1e-10) against ClusteredNR [Camacho et al., 2009]. Candidates with 100-199 bits and no match to the PRL1 profile were placed on the reference tree with EPA-ng (LG+FC+G4m; sequences added with MAFFT --addfragments --keeplength and trimmed to the reference columns).

Alignments and trees

RT sequences were aligned with MAFFT L-INS-i [Katoh and Standley, 2013] and trimmed to 365 positions of the RT domain. Maximum-likelihood trees were inferred with IQ-TREE 2.3.2 [Minh et al., 2020] under Q.pfam+F+R5 [Minh et al., 2021], selected by ModelFinder [Kalyaanamoorthy et al., 2017] under BIC among LG, Q.pfam, WAG, and JTT with 4-category gamma or 5-category free-rate heterogeneity on the initial alignment and kept for all trees; branch support was estimated with 1,000 ultrafast bootstrap replicates with NNI refinement (UFBoot [Hoang et al., 2018]). The tree of 567 sequences contains the 179 PRL1 representatives, the 143 non-redundant hits from SPIRE and NCBI retrieved with the PRL1 profile (90% identity, 80% coverage), 29 PRL2 sequences (21 VIRE representatives and the eight staphylococcal RTs), 51 VOTH, 59 reference retrons (two RTs chosen at random from each

retron type, subtype, and clade of the reference set [Mestre et al., 2020], and one orphan retron RT), and 106 non-retron RTs (UG/Abi 40, DGR 20, RT-CRISPR 20, G2L 15, group II introns 11). The tree of 573 sequences adds six PRL2 sequences (the two RTs of LPJP1, the UHGV RT, two SPIRE RTs, and one representative of the ClusteredNR bacterial RTs), and the tree of 577 sequences adds four more (two short class B RTs and the RTr and RTc of genome 1273823, the second 90%-identity cluster of the RTc). The main tree of 985 sequences adds to these 577 one representative of each of the 408 clusters obtained by clustering, with MMseqs2 (50% identity, 70% coverage), the 1,869 retrons of the reference set of 1,928 [Mestre et al., 2020] that are not among the 59 reference retrons; its 467 retrons cover types I-A to XIII and clades 2 and 11. Sequences were added to the alignment with MAFFT --addfragments --keeplength, keeping the same 365 positions, and each tree was inferred from scratch. Because protein names are not unique across SPIRE assemblies, the SPIRE hits were verified by sequence with the PRL1 profile, and 13 proteins of other assemblies that carry the name of a hit but match neither the PRL1 profile nor Pfam RVT_1 were excluded. No outgroup is available for the RT family; the trees were rooted on the DGR RTs for display, and the relationships among PRL1, PRL2, VOTH, and retrons reported here require only that the root lies among the non-retron RTs. As a model control, the alignment of 985 sequences was analyzed with LG+C60+F+G using the posterior mean site frequency approximation (PMSF [Wang et al., 2018]; guide tree from the Q.pfam analysis; 1,000 ultrafast bootstrap replicates without NNI refinement). A tree of the RT domains of all RTr and RTc proteins (11 clusters at 95% identity and 80% coverage, added to the 39 PRL2 tips, the 179 PRL1 representatives, the VOTH, the 59 reference retrons, and the non-retron RTs; 445 sequences) was inferred with the same alignment procedure and model. The tree of the partners (190 proteins carrying the shared region; 168 columns) was inferred with LG+G (1,000 UFBoot).

Topology tests

For the approximately unbiased (AU) tests [Shimodaira, 2002], the clade of interest was pruned and regrafted onto every branch of the remaining tree that leaves at least eight tips on each side (one of these placements reproduces the maximum-likelihood tree); the branch lengths of each tree were optimized in IQ-TREE under the same model, and the trees were compared using 10,000 RELL replicates (-zb 10000 -au). Placements with $p < 0.05$ were considered rejected. In the tree of 985 sequences, we tested the PRL1 clade (322 tips; 179 placements), the core of PRL2 (PRL2 without the RTc proteins; 37 tips; 255 placements), and the RTc pair (267 placements); in the trees of 567, 573, and 577 sequences, the PRL1 clade (71, 74, and 78 placements) and the core of PRL2 (28, 34, and 37 tips; 152, 149, and 152 placements).

Motifs and retron types

Regions X and Y were compared in four independent alignments built with MAFFT L-INS-i from full-length proteins: the 322 PRL1 sequences, 32 PRL2 sequences (non-redundant at 90% identity), the 51 VOTH representatives, and 300 retrons chosen at random from the reference set of 1,928 [Mestre et al., 2020]. Eight retron RTs of known types (four of type XI and one each of types I-A, II-A1, III-A1, and VI) were added to each alignment only to locate the columns of regions X and Y and were not counted. Logos show information content (bits, with small-sample correction) multiplied by column occupancy. Relatedness to retrons was scored with the HMM profiles of the retron survey [Toro, 2026] using hmmsearch: the general retron RT profile (threshold 160 bits and $E \leq 1e\text{-}10$) and the 29 type-specific profiles, for which we recorded the

best-scoring type of each sequence and its margin over the second-best type. Retron msr-msd RNAs were searched with cmsearch (Infernal [Nawrocki and Eddy, 2013]; default settings, both strands, E ≤ 10 as the reporting threshold) using 24 covariance models: the 21 models of retron msr-msd RNAs of the retron survey [Toro, 2026], the two models of the type XI-like lineages TXI_C2like and TXI_noncan_h [Toro, 2026], and the Rfam model of the retron msr RNA (RF02462) [Ontiveros-Palacios et al., 2025]. The search was run in two 700-nt windows around the RT of each PRL1 representative: upstream, from 600 nt before the start codon to 100 nt into the gene, and downstream, from 100 nt before the end of the gene to 600 nt after it. One of the 179 representatives lies too close to a contig end and has no window, so 178 were searched.

Phages

Genomes were annotated with Pharokka [Bouras et al., 2023] (PHROGs [Terzian et al., 2021]); proteins were compared with RefSeq Caudoviricetes proteins using MMseqs2; distances were computed with Mash [Ondov et al., 2016] (k = 21) against INPHARED; the gene-sharing network of 5,554 genomes used hypergeometric edge weights and MCL clustering. The complete contigs of SPIRE and NCBI hits were classified using geNomad 1.12.0 [Camargo et al., 2024] (end-to-end, with score calibration, default database); the CheckV [Nayfach et al., 2021] and geNomad values of the VIRE genomes are those distributed with the resource.

Gene neighborhood of the PRL1 RT

The genes within two positions of the RT were annotated with Pfam (hmmscan, GA thresholds [Mistry et al., 2021]) and HHblits/HHsearch (UniRef30, PDB70 [Steinegger et al., 2019]) and grouped in forms by protein-cluster membership; counts were made per genome, per species-level, and genus-level cluster of VIRE, and per representative. The association of the forms with the RT tree was tested by Fitch parsimony against 1,000 permutations of the tip states, and with lifestyle (temperate and virulent subclades) by Fisher's exact test, using the Freeman-Halton extension for a 2 × 3 table; the co-phylogeny of the E1-like protein and the RT and of other conserved genes was tested by the rank correlation of pairwise alignment distances (Mantel statistic on p-distances; 158 representatives), with the 53 single-copy gene families present in at least 20 representatives as controls. For a wider context, all genes within 10 kb of the RT in the 322 tips of the PRL1 clade were clustered with MMseqs2 (30% identity) into families, and, for families present in at least five VIRE genomes, the number of VIRE genomes in which the family lies within 10 kb of the RT was compared with its number in 1,000 sets of random 20-kb windows, one per genome, placed at least 12 kb away from the RT in VIRE genomes of 50 kb or more (empirical P = (k + 1)/(N + 1), minimum 0.001). Families were annotated with Pfam (E < 0.01) and PHROGs, and for the 273 families of 60-300 aa present in 3-14 tips, HHblits and HHsearch were used.

Protein structure

Models were built with AlphaFold2 (ptm model type) through ColabFold [Jumper et al., 2021; Mirdita et al., 2022], run locally on the CPU with three recycles and with alignments supplied by us (the ColabFold server was not used). The proteins of the PRL1 locus (one representative of each family of the windows of +/-10 kb) were modeled with one model each, from HHblits alignments of that representative against UniRef30_2023_02 (two iterations, E ≤ 1e-3). PRL2 RTs (full length), their N-terminal regions (representative of each group at 30% identity), the 156-aa F10 protein, and the partner regions were modeled with two models each, from HHblits

alignments of the query against UniRef30_2023_02 (two or three iterations, E ≤ 1e-3). For the partner regions, we also built alignments with jackhmmer (four iterations, E ≤ 1e-3) against the 265 partner proteins of PRL2 and the public partners (132 and 137 sequences after removing duplicates and sequences covering less than half of the region); the confidences quoted in Results come from these alignments unless stated. The RTr, the RTc, and their partners in genome 1294770 were also modeled with the AlphaFold Server (AlphaFold3 [Abramson et al., 2024]; one seed, five models, templates allowed). Models were compared with PDB100 and AlphaFold/SwissProt with Foldseek [van Kempen et al., 2024] and with each other by TM-score. Foldseek TM-scores are normalized by the length of the query model, except for the MPN protein, whose model extends beyond the MPN domain and for which they are normalized by the length of the target domain; where no TM-score is given, the Foldseek probability and E-value are reported. The JAMM motif was located with a regular expression and its geometry measured on the model. The N-terminal region of PRL2 RTs was defined as the residues before the start of the Pfam RVT_1 envelope minus 64, the offset at which profile-profile homology with group II intron RTs begins in four proteins (uncertainty about ±10 residues).

Effectors and VOTH types in the study

Genes within two positions of each VOTH RT were scanned with the 35 effector and fusion profiles of the retron survey (hmmscan, E ≤ 1e-5); genes without annotation were searched using HHsearch against PDB70. Each VOTH representative was assigned a retron type that dominated the smallest clade of the tree of 985 sequences containing it and at least three retrons. For the type XI-like lineages, profiles of TXI_C2like, TXI_noncan_h, type XI, and type IX were built from the alignment of [Toro, 2026] (204 columns; 126, 24, 94, and 26 sequences), and the VOTH were placed on a tree of those 270 sequences with EPA-ng (LG+F+G4).

Non-coding RNA of the PRL1 locus

Motifs were searched with CMfinder 0.4 [Yao et al., 2006] in the 700-nt windows upstream and downstream of the RT of eight subsets of the representatives defined as clades of the first version of the RT tree (sub01 to sub08; in the final tree of 567 sequences, sub06 and sub08 remain clades, 17 of the 20 members of sub07 form a clade, and sub02 is split), refined with LocARNA [Will et al., 2007] and RNAalifold, and tested for covariation with R-scape [Rivas et al., 2017], including the leave-one-module-out test. RNA folding used ViennaRNA [Lorenz et al., 2011]. Covariance models (the iterative model r0 for the structure and the covariation tests, and the combined model comb287 for the detection statistics) were built, calibrated, and searched with Infernal 1.1, with dinucleotide-shuffled windows and the upstream window as controls. The consensus structure was drawn with R2R [Weinberg and Breaker, 2011].

PRL2 loci

The partner of each RT was the next gene in the direction of the RT (VIRE gene coordinates; pyrodigal for public genomes). Partners and N-terminal regions were clustered with MMseqs2 (90% and 60% identity with 80% coverage; 30% identity with 70% coverage) and annotated with Pfam (GA), HHblits (UniRef30), and HHsearch (PDB70); relationships between partners were tested with blastp (E ≤ 1e-5) and with HHalign of HHblits profiles of 26 representatives. Repeat copies were detected in the region before the RT by self-comparison with blastn-short (word size 7, E ≤ 0.01, direct strand; alignments of at least 20 nt and 75% identity merged into intervals of 24-90 nt); copies of each lineage were aligned with MAFFT and summarized by a

50% consensus. Conservation of the copy and of its flanks was computed on the 183 copies of the 61 RTr genomes with three copies (most frequent base per column) and the per-column variability on 231 copies aligned over 32 columns. Consensus copies and tails were folded with RNAfold, with z-scores against 300 dinucleotide-shuffled sequences (esl-shuffle -d) unless stated; tails were scored against 60 dinucleotide shuffles generated by a random Eulerian path (Altschul-Erickson), all tails longer than 700 nt and eight of the 71 RTr class A tails of 500-700 nt being scored, and the SA1 transcript and LPJP1 against mononucleotide shuffles. Terminators were predicted with TransTermHP [Kingsford et al., 2007]. For the INPHARED genomes (Machias, AH12, and LPJP1), the region before the RT (2.8 kb) was extracted from the genome, copies were detected as above, and TransTermHP was run with the RT as the only annotated gene and a placeholder gene at the start of the region. Copy number was mapped on a tree of the 89 distinct RTr sequences (MAFFT L-INS-i; IQ-TREE LG+G4, rooted with RTc proteins) and compared with 2,000 permutations of the tip states (Fitch parsimony); the root state was estimated by Sankoff parsimony. The 32-nt consensus was searched in 86 complete RTr genomes and in 10 shuffled genomes of the same composition (blastn-short; hits of at least 24 nt and 75% identity).

RNA-seq of phage SA1

Twelve runs of PRJNA836150 [Zhang et al., 2022] were mapped to the SA1 genome using minimap2 (-ax sr). Per-strand depth was computed using samtools for a reverse-stranded library and normalized by the mean depth of the genome in each library. Fragment ends were obtained from read pairs. Promoters were searched as σ70-like -35/-10 elements.

Software

IQ-TREE 2.3.2, MAFFT 7.526, MMseqs2 18.8cc5c, HMMER 3.4, HH-suite 3.3.0, Foldseek 10.941cd33, EPA-ng 0.3.8, Infernal 1.1.5, CMfinder 0.4, LocARNA 1.9.2, R-scape 2.0.4.a, ViennaRNA 2.7.0, TransTermHP 2.09, BLAST+ 2.17.0, Mash 2.3, pyrodigal 3.7.1, geNomad 1.12.0, Pharokka 1.9.1, ColabFold 1.5.5 (alphafold-colabfold 2.3.6, JAX 0.4.26), samtools 0.1.19, minimap2 2.31; analyses in Python 3.12 with Biopython 1.87, DendroPy 5.0.8, NumPy, pandas, SciPy and matplotlib (versions and conda environments in the data package).

Use of AI tools

Claude Code (Anthropic) was used to assist with writing and running the analysis scripts (sequence searches, alignments, tree inference and topology tests, structural and RNA analyses, and figures). The study design, the choice of methods, criteria and thresholds, the evaluation of the results and their biological interpretation were performed by the author, who reviewed all code and outputs and takes responsibility for all content. Phylogenetic, profile, covariance-model and structural analyses were carried out with the established programs listed under Software.

Data availability

The sequences, alignments, trees, profiles, covariance and structural models, and the scripts used in this study are deposited at Zenodo (https://doi.org/10.5281/zenodo.24351680).

Acknowledgements

The author thanks the members of the Microbiome Engineering, Reverse-transcriptases, and regulatory RNA (MIRTA) laboratory for helpful discussions.

Funding

This work was supported by grant PID2023-147707NB-I00 to N.T. from the MCIN/AEI (10.13039/501100011033).

Competing interests

The author declares no competing financial interests.

References

Abramson J, Adler J, Dunger J, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 2024;630:493-500. https://doi.org/10.1038/s41586-024-07487-w

Ambroggio XI, Rees DC, Deshaies RJ. JAMM: a metalloprotease-like zinc site in the proteasome and signalosome. PLoS Biol 2004;2:e2. https://doi.org/10.1371/journal.pbio.0020002

Barbera P, Kozlov AM, Czech L, et al. EPA-ng: massively parallel evolutionary placement of genetic sequences. Syst Biol 2019;68:365-369. https://doi.org/10.1093/sysbio/syy054

Bobonis J, Mitosch K, Mateus A, et al. Bacterial retrons encode phage-defending tripartite toxin-antitoxin systems. Nature 2022;609:144-150. https://doi.org/10.1038/s41586-022-05091-4

Bouras G, Nepal R, Houtak G, et al. Pharokka: a fast scalable bacteriophage annotation tool. Bioinformatics 2023;39:btac776. https://doi.org/10.1093/bioinformatics/btac776

Camacho C, Coulouris G, Avagyan V, et al. BLAST+: architecture and applications. BMC Bioinformatics 2009;10:421. https://doi.org/10.1186/1471-2105-10-421

Camargo AP, Baltoumas FA, Ndela EO, et al. A genomic atlas of the human gut virome elucidates genetic factors shaping host interactions (Unified Human Gut Virome, UHGV). bioRxiv 2025, https://doi.org/10.1101/2025.11.01.686033.

Camargo AP, Roux S, Schulz F, et al. Identification of mobile genetic elements with geNomad. Nat Biotechnol 2024;42:1303-1312. https://doi.org/10.1038/s41587-023-01953-y

Chambers LR, Ye Q, Cai J, et al. A eukaryotic-like ubiquitination system in bacterial antiviral defence. Nature 2024;631:843-849. https://doi.org/10.1038/s41586-024-07730-4

Cook R, Brown N, Redgwell T, et al. INfrastructure for a PHAge REference Database: identification of large-scale biases in the current collection of cultured phage genomes (INPHARED). PHAGE 2021;2:214-223. https://doi.org/10.1089/phage.2021.0007

de Martín Garrido N, Orekhova M, Lai Wan Loong YTE, et al. Structure of the bacteriophage PhiKZ non-virion RNA polymerase. Nucleic Acids Res 2021;49:7732-7739. https://doi.org/10.1093/nar/gkab539

Doulatov S, Hodes A, Dai L, et al. Tropism switching in Bordetella bacteriophage defines a family of diversity-generating retroelements. Nature 2004;431:476-481. https://doi.org/10.1038/nature02833

Eddy SR. Accelerated profile HMM searches. PLoS Comput Biol 2011;7:e1002195. https://doi.org/10.1371/journal.pcbi.1002195

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: improving the ultrafast bootstrap approximation. Mol Biol Evol 2018;35:518-522. https://doi.org/10.1093/molbev/msx281

Hockenberry AJ, Wilke CO. BACPHLIP: predicting bacteriophage lifestyle from conserved protein domains. PeerJ 2021;9:e11396. https://doi.org/10.7717/peerj.11396

Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596:583-589. https://doi.org/10.1038/s41586-021-03819-2

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods 2017;14:587-589. https://doi.org/10.1038/nmeth.4285

Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 2013;30:772-780. https://doi.org/10.1093/molbev/mst010

Kingsford CL, Ayanbule K, Salzberg SL. Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake (TransTermHP). Genome Biol 2007;8:R22. https://doi.org/10.1186/gb-2007-8-2-r22

Kojima KK, Kanehisa M. Systematic survey for novel types of prokaryotic retroelements based on gene neighborhood and protein architecture. Mol Biol Evol 2008;25:1395-1404. https://doi.org/10.1093/molbev/msn081

Lambowitz AM, Zimmerly S. Group II introns: mobile ribozymes that invade DNA. Cold Spring Harb Perspect Biol 2011;3:a003616. https://doi.org/10.1101/cshperspect.a003616

Ledvina HE, Ye Q, Gu Y, et al. An E1-E2 fusion protein primes antiviral immune signalling in bacteria. Nature 2023;616:319-325. https://doi.org/10.1038/s41586-022-05647-4

Lorenz R, Bernhart SH, Höner zu Siederdissen C, et al. ViennaRNA Package 2.0. Algorithms Mol Biol 2011;6:26. https://doi.org/10.1186/1748-7188-6-26

Merk LN, Jones TA, Eddy SR. Presence of group II introns in phage genomes. Nucleic Acids Res 2025;53:gkaf761. https://doi.org/10.1093/nar/gkaa1149

Mestre MR, Gao LA, Shah SA, et al. UG/Abi: a highly diverse family of prokaryotic reverse transcriptases associated with defense functions. Nucleic Acids Res 2022;50:6084-6101. https://doi.org/10.1093/nar/gkac467

Mestre MR, González-Delgado A, Gutiérrez-Rus LI, Martínez-Abarca F, Toro N. Systematic prediction of genes functionally associated with bacterial retrons and classification of the encoded tripartite systems. Nucleic Acids Res 2020;48:12632-12647. https://doi.org/10.1093/nar/gkaa1149

Millman A, Bernheim A, Stokar-Avihail A, et al. Bacterial retrons function in anti-phage defense. Cell 2020;183:1551-1561. https://doi.org/10.1016/j.cell.2020.09.065

Millman A, Melamed S, Amitai G, Sorek R. Diversity and classification of cyclic-oligonucleotide-based anti-phage signalling systems. Nat Microbiol 2020;5:1608-1615. https://doi.org/10.1038/s41564-020-0777-y

Minh BQ, Dang CC, Vinh LS, Lanfear R. QMaker: fast and accurate method to estimate empirical models of protein evolution. Syst Biol 2021;70:1046-1060. https://doi.org/10.1093/sysbio/syab010

Minh BQ, Schmidt HA, Chernomor O, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol 2020;37:1530-1534. https://doi.org/10.1093/molbev/msaa015

Mirdita M, Schütze K, Moriwaki Y, et al. ColabFold: making protein folding accessible to all. Nat Methods 2022;19:679-682. https://doi.org/10.1038/s41592-022-01488-1

Mistry J, Chuguransky S, Williams L, et al. Pfam: the protein families database in 2021. Nucleic Acids Res 2021;49:D412-D419. https://doi.org/10.1093/nar/gkaa913

Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 2013;29:2933-2935. https://doi.org/10.1093/bioinformatics/btt509

Nayfach S, Camargo AP, Schulz F, et al. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nat Biotechnol 2021;39:578-585. https://doi.org/10.1038/s41587-020-00774-7

Nishijima S, Fullam A, Schmidt TSB, Kuhn M, Bork P. VIRE: a metagenome-derived, planetary-scale virome resource with environmental context. Nucleic Acids Res 2026;54:D902-D911. https://doi.org/10.1093/nar/gkaf1225

Ondov BD, Treangen TJ, Melsted P, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol 2016;17:132. https://doi.org/10.1186/s13059-016-0997-x

Ontiveros-Palacios N, Cooke E, Nawrocki EP, et al. Rfam 15: RNA families database in 2025. Nucleic Acids Res 2025;53:D258-D267. https://doi.org/10.1093/nar/gkae1023

Rivas E, Clements J, Eddy SR. A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs (R-scape). Nat Methods 2017;14:45-48. https://doi.org/10.1038/nmeth.4066

Schmidt TSB, Fullam A, Ferretti P, et al. SPIRE: a Searchable, Planetary-scale mIcrobiome REsource. Nucleic Acids Res 2024;52:D777-D783. https://doi.org/10.1093/nar/gkad943

Shimodaira H. An approximately unbiased test of phylogenetic tree selection. Syst Biol 2002;51:492-508. https://doi.org/10.1080/10635150290069913

Silas S, Mohr G, Sidote DJ, et al. Direct CRISPR spacer acquisition from RNA by a natural reverse transcriptase-Cas1 fusion protein. Science 2016;351:aad4234. https://doi.org/10.1126/science.aad4234

Steinegger M, Meier M, Mirdita M, et al. HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics 2019;20:473. https://doi.org/10.1186/s12859-019-3019-7

Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol 2017;35:1026-1028. https://doi.org/10.1038/nbt.3988

Terzian P, Olo Ndela E, Galiez C, et al. PHROG: families of prokaryotic virus proteins clustered using remote homology. NAR Genom Bioinform 2021;3:lqab067. https://doi.org/10.1093/nargab/lqab067

Toro N. Landscape of retron diversity across the SPIRE microbial metagenome resource reveals candidate novel type XI-like lineages. mSystems 2026;11:e00732-26. https://doi.org/10.1128/msystems.00732-26

Toro N, Martínez-Abarca F, Mestre MR, González-Delgado A. Multiple origins of reverse transcriptases linked to CRISPR-Cas systems. RNA Biol 2019;16:1486-1493. https://doi.org/10.1080/15476286.2019.1639310

Toro N, Nisa-Martínez R. Comprehensive phylogenetic analysis of bacterial reverse transcriptases. PLoS ONE 2014;9:e114083. https://doi.org/10.1371/journal.pone.0114083

van Kempen M, Kim SS, Tumescheit C, et al. Fast and accurate protein structure search with Foldseek. Nat Biotechnol 2024;42:243-246. https://doi.org/10.1038/s41587-023-01773-0

Wang HC, Minh BQ, Susko E, Roger AJ. Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic estimation. Syst Biol 2018;67:216-235. https://doi.org/10.1093/sysbio/syx068

Weinberg Z, Breaker RR. R2R: software to speed the depiction of aesthetic consensus RNA secondary structures. BMC Bioinformatics 2011;12:3. https://doi.org/10.1186/1471-2105-12-3

Will S, Reiche K, Hofacker IL, Stadler PF, Backofen R. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering (LocARNA). PLoS Comput Biol 2007;3:e65. https://doi.org/10.1371/journal.pcbi.0030065

Yao Z, Weinberg Z, Ruzzo WL. CMfinder: a covariance model based RNA motif finding algorithm. Bioinformatics 2006;22:445-452. https://doi.org/10.1093/bioinformatics/btk008

Ye Y. Computational evidence that array-associated reverse transcriptase loci are retron systems with multiple ncRNA genes. Research Square 2026, https://doi.org/10.21203/rs.3.rs-11367685/v1 (preprint, posted 15 October 2026).

Yoon PH, Athukoralage JS, Ameisen E, Kauderer-Abrams E, Perry NT, Durrant MG. Autonomous AI agents discover reverse transcriptases with tandem repeat arrays. bioRxiv 2026, https://doi.org/10.64898/2026.10.05.755809 (preprint, posted 11 October 2026).

Zhang B, Xu J, He X, Tong Y, Ren H. Interactions between jumbo phage SA1 and *Staphylococcus*: a global transcriptomic analysis. Microorganisms 2022;10:1590. https://doi.org/10.3390/microorganisms10081590

Supplementary Files

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

- ToroPRLSupplementary.pdf