Structural Requirements for Reverse Transcription by a Diversity-generating Retroelement

Sumit Handa[1,3,a], Tapan Biswas[1,a], Jeet Chakraborty[1], Gourisankar Ghosh[1], Blair G. Paul[2], and Partho Ghosh[1,b]

[1]Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, CA, USA

[2]Marine Biological Laboratory, Josephine Bay Paul Center, Woods Hole, MA, USA.

[3]Current Address: Singular Genomics, 3010 Science Park Rd, San Diego, CA, USA

[a]Contributed equally to the work.

[b]To whom correspondence should be addressed: pghosh@ucsd.edu

Abstract

Diversity-generating retroelements (DGRs) create massive protein sequence variation in ecologically diverse microbes. Variation occurs during reverse transcription of a protein-encoding RNA template coupled to misincorporation at adenosines. In the prototypical *Bordetella* bacteriophage DGR, the template must be surrounded by upstream and downstream RNA segments for cDNA synthesis by the reverse transcriptase bRT and associated protein Avd. The function of the surrounding RNA was unknown. Cryo-EM revealed that this RNA enveloped bRT and lay over barrel-shaped Avd, forming an intimate ribonucleoprotein (RNP). An abundance of essential interactions between RNA structural elements and bRT-Avd precisely positioned an RNA homoduplex for initiation of cDNA synthesis by *cis*-priming. Our results explain how the surrounding RNA primes cDNA synthesis, promotes processivity, terminates polymerization, and strictly limits mutagenesis to select proteins through mechanisms that are likely conserved in DGRs from distant taxa.

Introduction

Diversity-generating retroelements (DGRs) create massive protein sequence variation (up to 10^{30}) in ecologically diverse bacteria and archaea, and their viruses. A recent survey identified ~31,000 DGRs from > 1,500 bacterial and archaeal genera, constituting > 90 environment types (*1*). DGRs are especially enriched in the human gut microbiome (*1, 2*) and nano-sized microbes, which appear to comprise a major fraction of microbial life. The latter maintain DGRs despite having reduced genomes (*3, 4*). DGRs are also implicated in the emergence of multicellularity (*5, 6*). The prototypical DGR of *Bordetella* bacteriophage (BΦ) can produce ~10^{13} variants of its receptor-binding protein Mtd (*7, 8*), which enables the BΦ to use alternative receptors when its usual receptor is missing due to gene expression changes in *Bordetella.* This rapid adaptation to a dynamic environment is likely the general function of DGRs.

The BΦ DGR consists of three genes and two repeats (Fig. 1a). The genes encode the variable protein Mtd, a reverse transcriptase (bRT), and an accessory variability determinant (Avd) protein (*7*). The repeats are the variable region (*VR*), which encodes the variable amino acids of Mtd, and the template region (*TR*), which is nearly identical to *VR* except mainly at adenosines. Sequence variation arises during reverse transcription of an RNA copy of *TR* through frequent misincorporation at *TR* adenosines. *TR*-cDNA, containing errors at positions corresponding to adenosines, then homes to and replaces the extant copy of *VR* to give rise to an Mtd variant. The mechanism of homing is poorly understood. *TR* is sequence invariant and thus there are no limits on the number of rounds of adenosine (A)-mutagenesis that can occur.

Reverse transcription is carried out by a complex of bRT and Avd. Unlike other RTs, bRT lacks reverse transcription activity on its own and requires association with Avd (*9*). bRT belongs to a monophyletic clade of DGR reverse transcriptases (RTs) most closely related to group II intron RTs (*1, 10, 11*), and Avd forms a pentameric, positively-charged barrel-like structure similar to S23 ribosomal proteins (*12*). Avd-like proteins occur in the great majority of DGRs (*13*). The bRT-Avd complex has intrinsic A-mutagenesis activity on RNA or DNA templates (*9, 14*).

Significantly, *TR* RNA is insufficient for reverse transcription by bRT-Avd and must instead be surrounded by *cis*-acting elements (*9*). Most prominent is a 140-nucleotide (nt) noncoding RNA segment called Spacer (Sp) located immediately downstream of *TR* (Fig. 1a). Sp RNA supplies the priming site (^{Sp}A56) for cDNA synthesis (*9, 15*). The 2'-OH of ^{Sp}A56 is

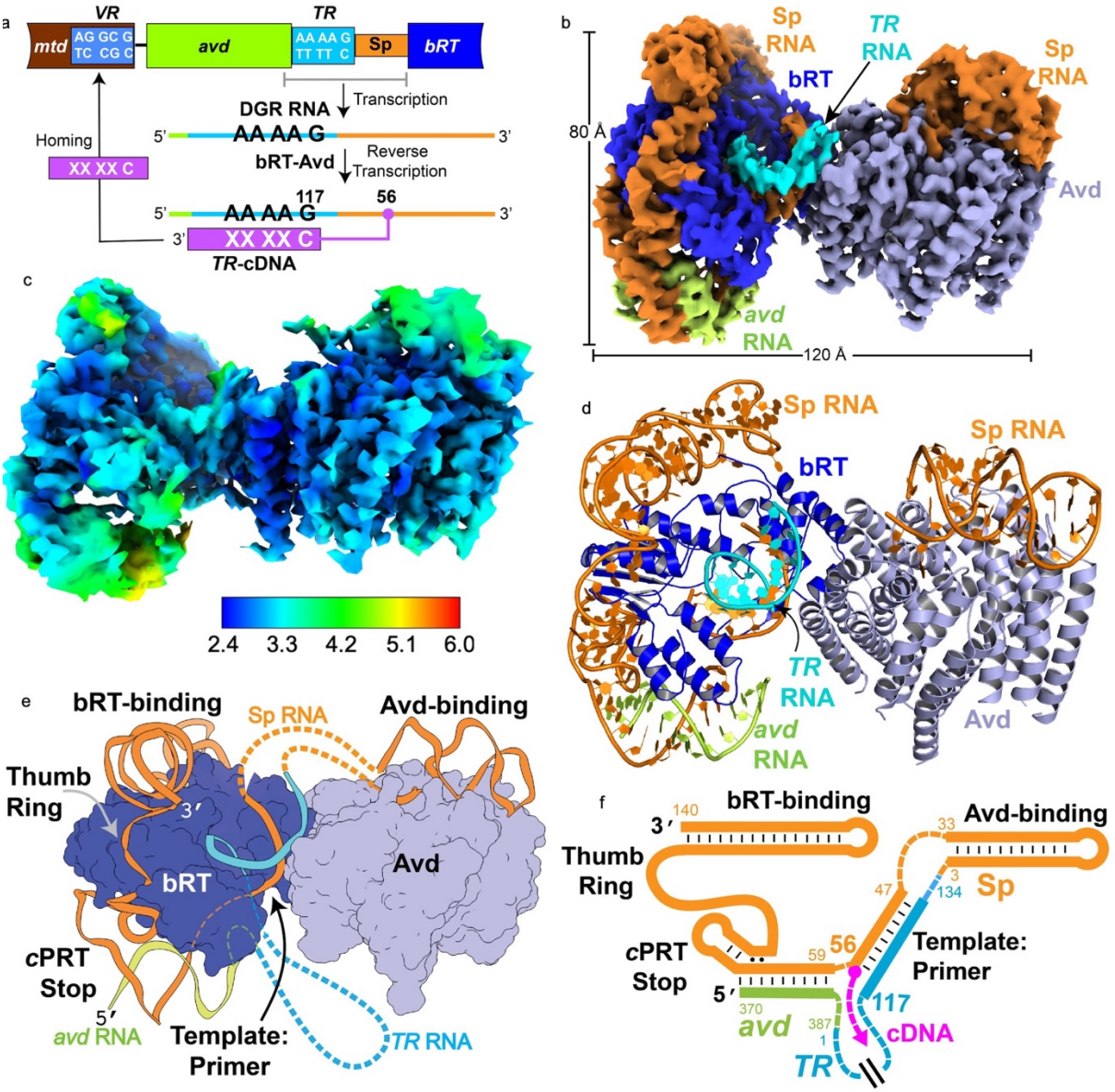

Fig. 1. DGR and bRT-Avd:RNA$^{\Delta 98}$ complex. a. Schematic of BΦ DGR. *VR* and *TR* (both strands depicted) differ mainly at adenosines in *TR* (top strand). An example sequence is shown for *VR*, in which three of the four adenosines in *TR* were mutated in a previous round of A-mutagenesis. 'X' denotes misincorporation errors at adenosines. Thin lines correspond to RNA, and boxes to DNA. **b-d** Cryo-EM density (b), local resolution map with resolution scale in Å below (c), and cartoon (d) of bRT-Avd:RNA$^{\Delta 98}$ (bRT, blue; Avd, slate; RNA$^{\Delta 98}$ light green for *avd*, cyan for *TR*, orange for Sp). **e and f.** Schematic of RNA elements in three-dimensional (e) and linear representation (f). Dotted lines indicate sections that were either too flexible to model (parts of *avd, TR*, or Sp) or were absent in the complex (^{TR}10-107 and cDNA).

necessary for cDNA synthesis (*9*), suggesting that it acts as the priming nucleophile, as occurs in retrons (*16*). Thus, reverse transcription is *cis*-primed and a covalently linked RNA-cDNA molecule is produced. The first nucleotide reverse transcribed is ^{TR}G117 (*9, 15*). A 20 nt region upstream of *TR*, consisting of the 3'-terminal codons of *avd*, is also required for *cis*-primed reverse transcription (*c*PRT). We use the term 'DGR RNA' to refer to the longer RNA required

63 for *c*PRT, containing 20 *avd* and 140 Sp nts surrounding *TR*. This longer DGR RNA confers not
64 only *cis*-priming but also processive polymerization, resulting in ~120 nt cDNAs that span *TR*
65 (*9*), in contrast to the 5-30 nts cDNAs synthesized from non-cognate RNA templates using
66 exogenous primers (*9*).
67
68 **Structure Determination**
69 How the DGR RNA combines with bRT-Avd for cDNA synthesis was unknown and
70 pursued through cryo-EM. Initial maps of the bRT-Avd:RNA complex were limited to ~7 Å
71 resolution due to RNA flexibility and preferred particle orientation. To limit RNA flexibility, the
72 central 98 nts of *TR* (Δ^{TR}10-107) were deleted from DGR RNA (Fig. S1a). *TR* has an
73 informational rather than functional role (*17*), and the truncated RNA (called $RNA^{\Delta 98}$) has all the
74 essential *cis*-acting elements. Preferred orientation was unexpectedly alleviated by fusing
75 pentamer-forming *Methylobacterium extorquens* formaldehyde activating enzyme (Fae) to the
76 Avd C-terminus, in an attempt to increase the mass of the ~176 kDa bRT-Avd:$RNA^{\Delta 98}$ complex.
77 The bRT-(Avd-Fae):$RNA^{\Delta 98}$ (~258 kDa) complex yielded maps that extended to as high as ~3.1
78 Å resolution (Tables S1 and S2, Figs. S2-S5). Fusions of other pentamer-forming proteins did
79 not yield a similar improvement. bRT-(Avd-Fae) reverse transcribed $RNA^{\Delta 98}$ and
80 misincorporated at adenosine at the same frequency as bRT-Avd (Figs. S1b and c) (*14*). Fae was
81 positionally variable and masked out, and thus we refer to Avd rather than Avd-Fae hereafter.
82 The most interpretable map came from a complex that contained $RNA^{\Delta 98}$ with a
83 ^{TR}U114A substitution and had been incubated with dCTP, dATP, and ddGTP (Figs. 1b-d, Table
84 S1, Complex 1). The intent was to visualize a short cDNA strand and ^{TR}U114A at the template
85 site, but the population of such complexes was too sparse for class-averaging, consistent with <
86 2% of templates yielding cDNA (*14*). Instead, this sample yielded a complex that had not yet
87 commenced synthesis. Most of bRT and Avd were in density with a local resolution better than
88 the global 3.25 Å resolution (Figs. 1c and S6a, b), and the entirety of bRT and Avd, except for a
89 few N- and C-terminal amino acids in the latter, were modeled with an excellent fit (Table S2)
90 (*18*). Segments of RNA that contacted bRT or Avd were generally in < 3.25 Å resolution density
91 (Figs. 1c and 6b), while segments not in contact were often in ~4-5 Å resolution density. In total,
92 the RNA also had an excellent fit to the density (Table S2), and while there were a few breaks in
93 the chain, ~75% could be modeled.

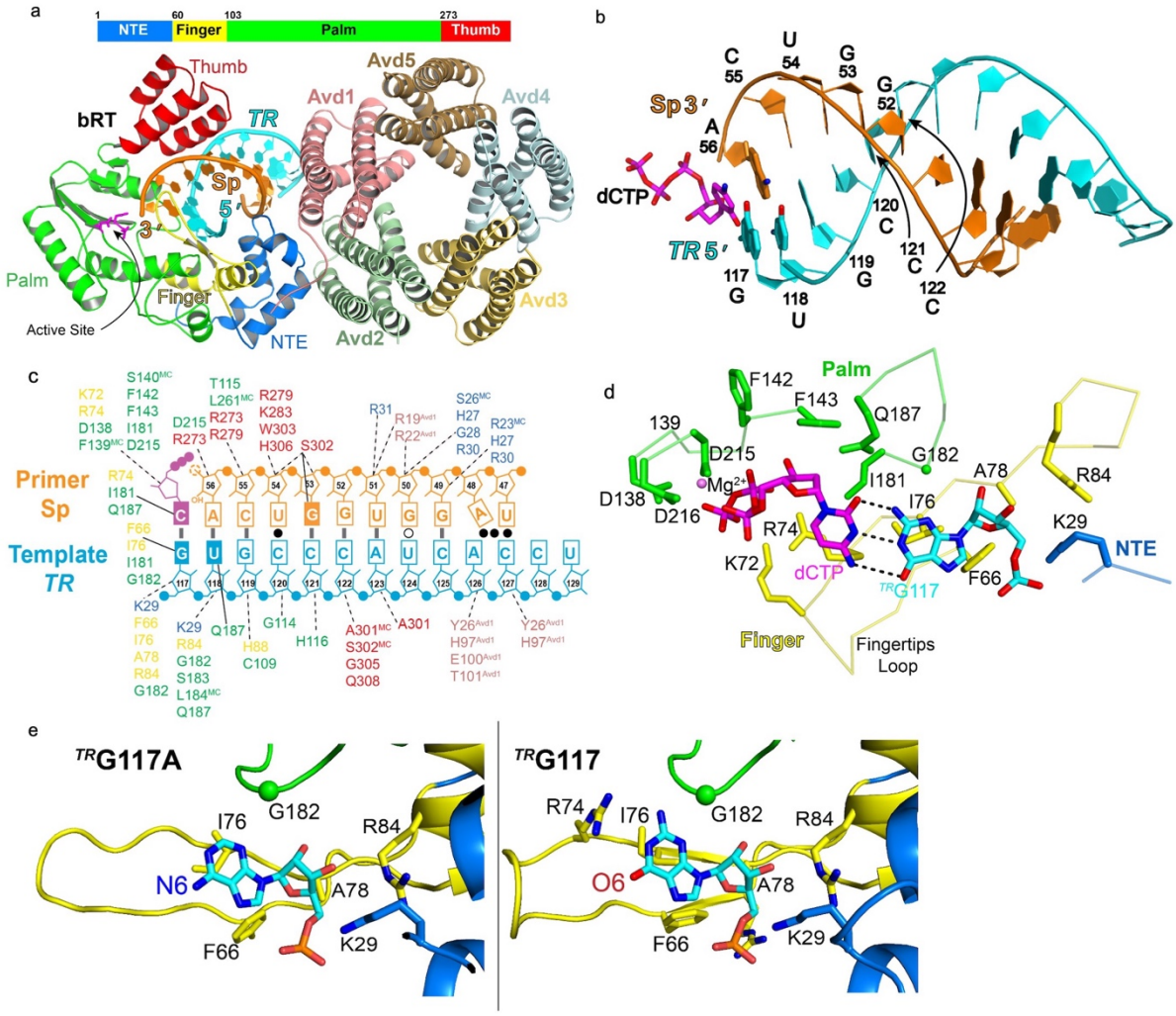

Fig. 2. bRT-Avd and Template:Primer Duplex. a. bRT (schematic and structure: NTE, blue; finger, yellow; palm, green; thumb, red), Avd (colored individually), and *TR*:Sp template:primer duplex (cyan and orange, respectively). bRT active site amino acids D138, D215, and D216 in magenta. **b.** *TR*:Sp template:primer duplex with dCTP in magenta. **c.** Schematic of *TR*:Sp template:primer duplex with base pairs shown (solid bar, Watson-Crick; open circle, wobble; and closed circle, non-standard). Base-specific (sold lines) and ribose or phosphate (dotted lines) contacts from bRT or Avd shown. Colors of labels correspond to coloring in panel a. The side chains of bRT Q187 and S301 are 3.9 Å from the O2 of ^{TR}U118 and 3.4 Å from the N3 of ^{Sp}53G, respectively. **d.** Interactions between ^{TR}G117 (carbon and phosphorus, cyan) at the T-site and dCTP (carbon and phosphorus, magenta) at the N-site (for both nucleotides, nitrogen blue, and oxygen red), and with bRT amino acids (colored as in panel a). Hydrogen bonds shown as black dotted lines. The putative Mg^{2+} was 2.2 and 2.3 Å from D138 and D215, respectively, and 3.3 Å from one of the oxygens of the α-phosphate. **e.** Interactions between (left) ^{TR}G117A and (right) ^{TR}G117 (carbon and phosphorus, cyan) at the T-site with bRT (carbon colored by subdomain, as in panel a; and oxygen red, and nitrogen blue for nucleotides and protein).

Intimate Ribonucleoprotein

Most strikingly, the structure revealed an intimate ribonucleoprotein (RNP) in which bRT was almost entirely enveloped by Sp RNA, which also lay over the narrow end of the Avd barrel (Fig. 1). In between, Sp RNA formed two adjacent duplexes, one with *TR* and the other with the

99 upstream *avd* segment. bRT most closely resembled group II intron RTs as predicted (*11, 19-21*)
100 (Figs. 2a and S6c), and shared with these proteins and non-LTR RTs (*22*) an N-terminal
101 extension subdomain (NTE). Following the NTE were the canonical finger, palm, and thumb
102 subdomains of nucleotide polymerases. bRT lacked additional subdomains found in related RTs,
103 such as DNA-binding, RNA-binding, or endonuclease. The bRT NTE contacted Avd (two of five
104 protomers) through an extensive and mainly electrostatic interface. Bound and free Avd were
105 structurally similar (1.4 Å rmsd, 456 Cα), and the bRT-Avd interface was consistent with prior
106 mutagenesis results (*12*) (Fig. S6d). While Avd is required for catalysis (*9*), its closest approach
107 to the bRT active site was ~40 Å, suggesting that it is not directly involved in chemistry.
108
109 **Template:Primer Duplex in Active Site**
110 The bRT active site was occupied by an RNA homoduplex in which one strand (^{TR}117-
111 129) contained the initiating nucleotide ^{TR}G117, and the other (^{Sp}47-56) the priming nucleotide
112 ^{Sp}A56 (Figs. 2 and S7). The bRT finger, palm, and thumb contacted the end of the
113 template:primer duplex proximal to the active site, while the NTE and an Avd protomer
114 contacted the distal end. This suggested that the role of Avd in catalysis is to enforce proper
115 positioning of the template:primer duplex either through direct contact or indirectly through the
116 NTE subdomain. The *TR*:Sp duplex formed seven canonical base pairs (bps), as predicted (*9,*
117 *15*), along with three non-standard pairs (Figs. 2b, c) (*9, 15*). The four base pairs closest to the
118 active site, spanning ^{Sp}53-56 and ^{TR}118-121, are functionally required whereas the others are not
119 (*9, 15*). Bases of the *TR*:Sp RNA homoduplex were not contacted closely by bRT or Avd, the
120 exception being the template base ^{TR}G117 (Figs. 2c and d), which was contacted by conserved
121 fingertips loop (F66 and I76) and palm (I181 and G182) amino acids. The NTE subdomain
122 contacted the 5' phosphate of ^{TR}G117 through conserved Lys 29, which was on a loop braced by
123 Avd (Fig. S7e). dCTP was at the incoming dNTP (N)-site, with cytosine forming canonical
124 hydrogen bonds with the guanine at the template (T)-site. The cytosine was contacted by
125 conserved fingertips loop (R74) and palm (I181 and Q187) amino acids, and stacked coaxially
126 against the base of the priming nucleotide ^{Sp}A56 (Fig. 2b). The 2'-OH of ^{Sp}A56 could not be
127 placed with chemical precision due to the resolution of the map, compounded by the lack of
128 density for the following nucleotide, which would otherwise constrain its 3'-OH. The catalytic
129 bRT aspartates 138, 215, and 216, and a cation, presumably Mg^{2+}, were positioned adjacent to

130 the dCTP phosphates (Fig. 2d). Thus, this complex appeared poised for catalysis and we termed
131 it the 'Active G:dCTP' conformation (Table S1, Complex 1).
132 As bRT and other DGR RTs misincorporate specifically at template adenosines, we
133 determined the 3.23 Å global resolution structure of an RNP that had a ^{TR}G117A substitution in
134 RNA$^{\Delta 98}$. In this RNP, the adenine at ^{TR}117 occupied the T-site, and while the nonhydrolyzable
135 dCTP analog dCpCpp had been incubated with it, only broken density was observed at the N-site
136 (Figs. 2e and S8, and Table S1, Complex 2). This RNP closely resembled one determined at 3.07
137 Å global resolution in which the T-site was occupied by guanine (^{TR}G117 in wild-type RNA$^{\Delta 98}$)
138 and, despite incubation with dCpCpp, the N-site was similarly empty (Figs 2e and S8, Table S1,
139 Complex 3). These two RNPs had an Active conformation (termed Active A:Empty and Active
140 G:Empty, respectively), but differed from the Active G:dCTP conformation in having greater
141 conformational variability. This was evident in the bRT finger and thumb subdomains and in the
142 RNA. The *TR*:Sp template:primer duplex in these Active N-empty conformations, either with
143 adenine or guanine at the T-site, occupied density of moderate local resolution (~4 Å), in
144 contrast to the higher local resolution in the Active G:dCTP conformation (Fig. S7). These
145 observations suggest that the Active conformation does not absolutely depend on N-site
146 occupancy but does seem to be stabilized by the correct incoming dNTP. Comparison of the N-
147 empty complexes showed similar contacts by bRT to adenine and guanine at the T-site (Fig. 2e).
148 Notably, neither bRT nor Avd contacted the substituent at the C6 position of these purines,
149 which is a major determinant of A-mutagenesis (*14*).
150
151 ***c*PRT Stop**
152 The template:primer duplex was immediately adjacent to a second duplex (Figs. 1e and f)
153 in which one strand consisted of ^{Sp}59-70 and the other the essential upstream *avd* segment
154 (^{avd}369-379) (Figs. 3 and S9). These strands formed 11 canonical bps, with ^{Sp}U63 unpaired but
155 accommodated within the coaxial stack. The *avd*:Sp duplex appeared to be part of a larger
156 structural assembly, with the Sp strand continuing into a short stem-loop (^{Sp}71-78) followed by
157 three unpaired nucleotides (^{Sp}79-81). bRT formed sparse contacts with the *avd*:Sp duplex,
158 reflected in the moderate local resolution (~4 Å) of its density, but numerous contacts to the
159 short stem-loop and unpaired bases, which were in density of < 3 Å local resolution (Figs. 3b and
160 c, S9). Two of the unpaired nucleotides (^{Sp}A80 and A81) were inserted into the minor groove of

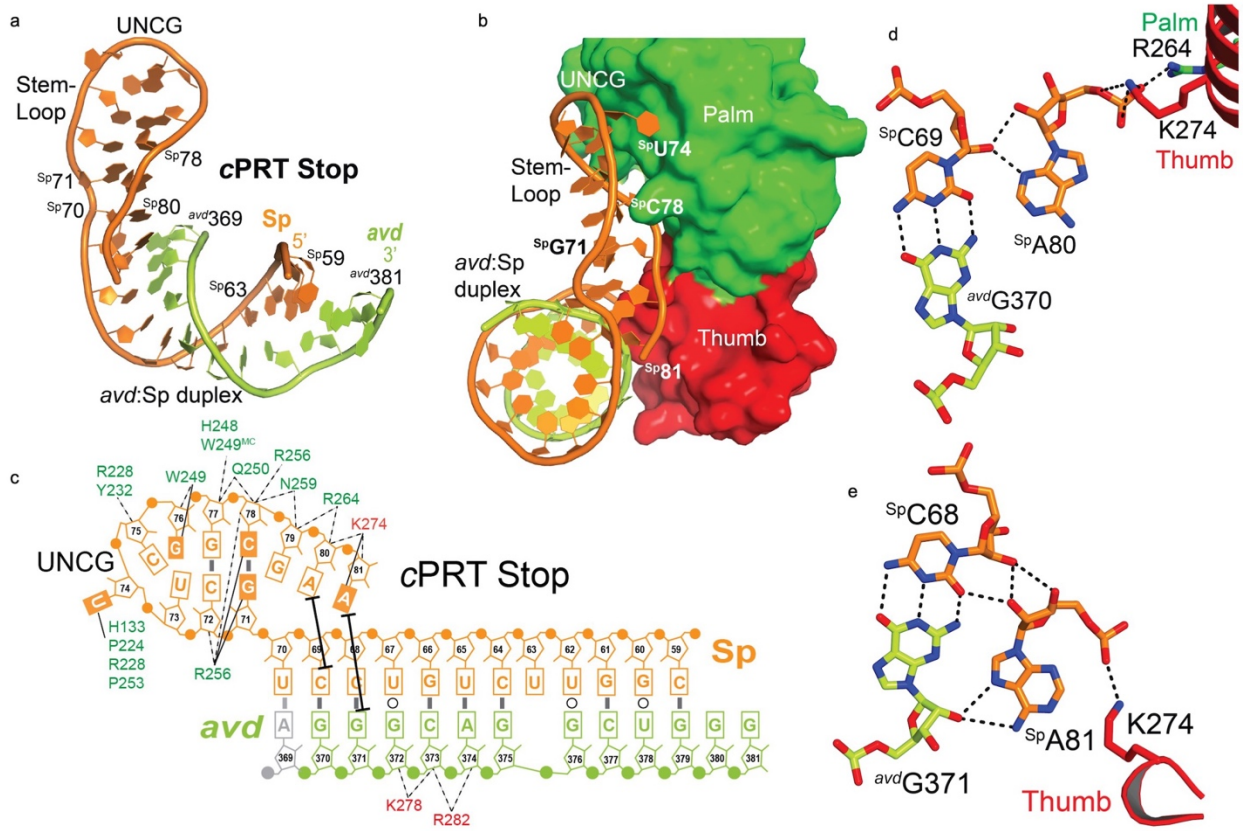

Fig. 3. *c*PRT Stop. a. *c*PRT Stop (*avd*:Sp duplex, stem-loop, and unpaired bases). The stem had 2 bps and an UNCG tetraloop. **b.** Interaction between *c*PRT Stop (in cartoon) and bRT (surface with palm in green and thumb in red). **c.** Schematic of *c*PRT Stop duplex, in same format as Fig. 2c. Solid thick black lines indicate tertiary A-minor motif contacts. ^{avd}A369 is a guanosine in RNA$^{\Delta 98}$, introduced by *in vitro* transcription, and forms a wobble base pair with ^{Sp}A70. The figure depicts the wild-type adenosine in gray. **d and e.** Tertiary A-minor motif interactions of ^{Sp}A80 with ^{Sp}C69:^{avd}G70 (d) and ^{Sp}A81 with ^{Sp}C68:^{avd}G371 (e). ^{Sp}A80 and ^{Sp}A81 are also contacted by bRT. Black dotted lines are hydrogen bonds.

G:C base pairs of the *avd*:Sp duplex, forming A-minor motifs and engaging in one of the few tertiary contacts in the RNA (Figs. 3d and e). The A-minor motif is one of the most prevalent tertiary interactions in RNA (*23, 24*). As observed here, it preferentially involves G:C base pairs, and often occurs as pairs or triples of adenosines. The unpaired adenosines were also contacted by bRT.

In the Sp strand, only two disordered nucleotides (^{Sp}57 and 58) separated the *avd*:Sp duplex from the *TR*:Sp template:primer duplex. In the other strand, ~30 disordered nts separated the ends visible in RNA$^{\Delta 98}$ (between *avd*^{381} and *TR*^{117}); in intact *TR*, this span would be ~130 nts. These ~30 nts were likely looped out between the *avd*:Sp and template:primer duplexes. In this architecture, the bulk of *TR* would extend towards the wide end of the Avd barrel (Fig. S9d), to which Fae was fused, explaining why the RNP containing Fae reverse transcribed RNA$^{\Delta 98}$ with its truncated *TR* but not DGR RNA with its intact *TR* (Fig. S1b).

173 Significantly, *c*PRT of DGR RNA terminates within ^{avd}383-387 (*9*), just before the
174 *avd*:Sp duplex, suggesting that this duplex along with its associated elements (i.e., stem-loop and
175 unpaired nucleotides) create a physical barrier to polymerization. We therefore termed these
176 elements the '*c*PRT Stop'. Notably, deletion of the *avd* strand (Δ^{avd}368-377) results not in longer
177 cDNAs but instead no reverse transcription at all (*9*). This is not due to loss of RNA binding, as a
178 DGR RNA lacking the *avd* segment still bound bRT and Avd (Fig. S10). This result suggested
179 that the *c*PRT Stop not only terminates polymerization but is also required for *c*PRT. We asked
180 whether base pairing in the *avd*:Sp duplex was essential for <u>*c*PRT</u>. Substitutions that disrupted 12
181 bps between *avd* and Sp resulted in loss of *c*PRT, but surprisingly reversion through
182 complementary substitutions did not restore *c*PRT (Fig. S11). Disruption of just the six Watson-
183 Crick base pairs had a similar result (Fig. S11). These findings suggested that the function of the
184 *avd*:Sp duplex in *c*PRT is conformationally sensitive, possibly due to A-minor motif interactions,
185 which were affected by both sets of mutations. In support of this conclusion, removal of the short
186 stem-loop (Δ^{Sp}71-78) preceding the A-minor motif interactions also led to loss of *c*PRT (Figs. 3a
187 and S12). Together, these results suggested a structurally dynamic link between the *c*PRT Stop
188 element and template:primer duplex, with proper alignment between the two required for *c*PRT.
189

190 **Thumb Ring, bRT-binding, and Avd-binding RNA**

191 The *c*PRT Stop element continued in the Sp strand into two further RNA elements, both
192 of which contacted bRT and were in density of excellent local resolution (Figs. 4 and S13). The
193 first element (^{Sp}82-94) wrapped around the base of the bRT thumb and was therefore called the
194 'Thumb Ring', and was notable in being the longest stretch of single-stranded RNA in the RNP.
195 Eight of 12 unpaired bases in the Thumb Ring were contacted by bRT (Fig. 4b). Of these, ^{Sp}C92
196 and U94 were the most significant for *c*PRT (Figs. 4c and S13). Substitutions at other bases in
197 the Thumb Ring likewise reduced *c*PRT.

198 Sp continued from the Thumb Ring into a bRT-binding element composed of a long stem
199 (^{Sp}95-113 and 121-139) interrupted by two interior loops (^{Sp}100-103:130-131 and ^{Sp}108-
200 112:122-125), and capped by a heptaloop (^{Sp}114-120) (Figs. 4 and S12), the most common type
201 of loop after tetraloops (*25*). The interior loops mostly had non-standard base-base interactions.
202 Bases flipped-out from the stem (^{Sp}U110 and G131) and heptaloop (^{Sp}U120) were contacted by
203 bRT palm and NTE. Substitutions at these bases decreased *c*PRT, with the greatest effect at

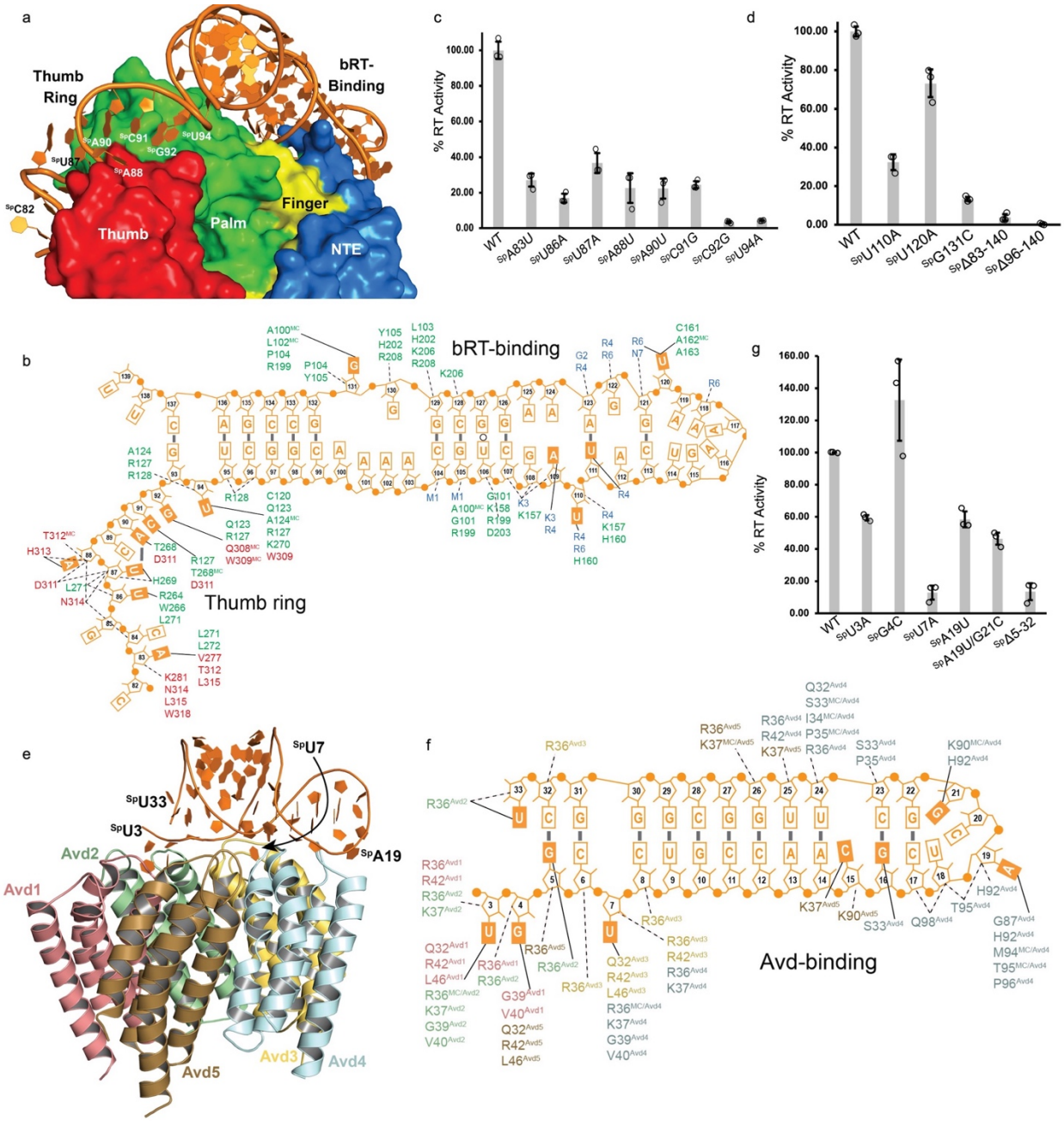

Fig. 4. Thumb Ring, bRT-binding, and Avd-binding RNA. a. Interaction of Thumb Ring and bRT-binding RNA with the surface of bRT. **b.** Schematic of Thumb Ring and bRT-binding RNA, in same format as Fig. 2c. **c and d.** *c*PRT activity of DGR RNA mutated in the Thumb Ring (c) or bRT-binding RNA element (d) relative to wild-type DGR RNA. Data here and below are from experiments that were repeated three independent times, and means and standard deviations shown. **e.** Interaction between Avd-binding RNA and pentameric Avd. **f.** Schematic of Avd-binding RNA, in same format as Fig. 2c. **g.** *c*PRT activity of DGR RNA mutated in the Avd-binding RNA element relative to wild-type DGR RNA.

^{Sp}G131 (Fig. 4d and S13). Deletion of the Thumb Ring along with the bRT-binding stem-loop (Δ^{Sp}83-140) or the bRT-binding stem-loop alone (Δ^{Sp}96-140) eliminated *c*PRT almost entirely. The latter also resulted in loss of RNA binding to bRT while binding to Avd was maintained

207 (Fig. S10). bRT associates with Avd in the absence of RNA (*9, 12*), but this last result shows that
208 this interaction is strengthened by DGR RNA.
209 The final RNA element consisted of the 5' end of Sp (^{Sp}3-33), which formed a 11-bp
210 stem capped by a UNCG tetraloop (Figs. 4e, f and S14). This stem-loop was contacted by all five
211 Avd protomers, consistent with RNase protection experiments (*9*). Bases flipped-out from the
212 stem-loop were inserted into crevices between protomers (Fig. S14). The same arsenal of Avd
213 amino acids but from different protomers contacted the flipped-out bases ^{Sp}U3, G4, and U7. ^{Sp}U7
214 made the most crucial contacts as its substitution reduced *c*PRT to the same level as deletion of
215 the entire Avd-binding segment (Δ^{Sp}5-32) (Figs. 4g and S11). Substitution of other flipped-out
216 bases decreased *c*PRT less dramatically.
217
218 **RNA Conformational Variability**
219 Several RNA elements — the template:primer duplex, *c*PRT Stop, and Thumb Ring —
220 were conformationally variable, whereas others — bRT- and Avd-binding stem-loops — were
221 stable across several alternative RNP conformations. These alternative conformations were
222 evident in multiple samples and independent of incubation with dNTPs or analogs.
223 An RNP conformation in which the thumb was collapsed into the active site, called
224 'Resting', was visualized in three independent samples (Figs. 5a and S15, Table S1, Complexes
225 4-6), including one in which no dNTP or analog was added. The thumb was highly mobile and
226 occupied both N- and T-sites, having displaced the disordered fingertips loop. The
227 template:primer duplex skirted the surface of bRT rather than entering the active site. The bRT-
228 and Avd-binding stem-loops were in the same positions as in the Active conformation, but
229 density was lacking for the *c*PRT Stop and Thumb Ring in two of the samples, although in the
230 third (Complex 5) these could be modeled into density of low local resolution (4-5 Å), which
231 revealed that these RNA structures were disengaged from bRT.
232 A second alternative RNP conformation, in which the *TR*:Sp duplex was in the active site
233 but ^{Sp}G57 occupied the N-site and ^{Sp}C58 joined the *c*PRT Stop duplex, was visualized in three
234 samples (Figs. 5b and S16, Table S1, Complexes 7-9), again including one in which no dNTP or
235 analog was added. ^{Sp}G57 and ^{Sp}C58 were disordered in the Active conformation but here were
236 seen to directly connect the *c*PRT stop and template:primer duplexes. ^{Sp}G57, which is required
237 for *c*PRT (*9*), formed a guanine-guanine pair with ^{TR}G117 through their Watson-Crick faces.

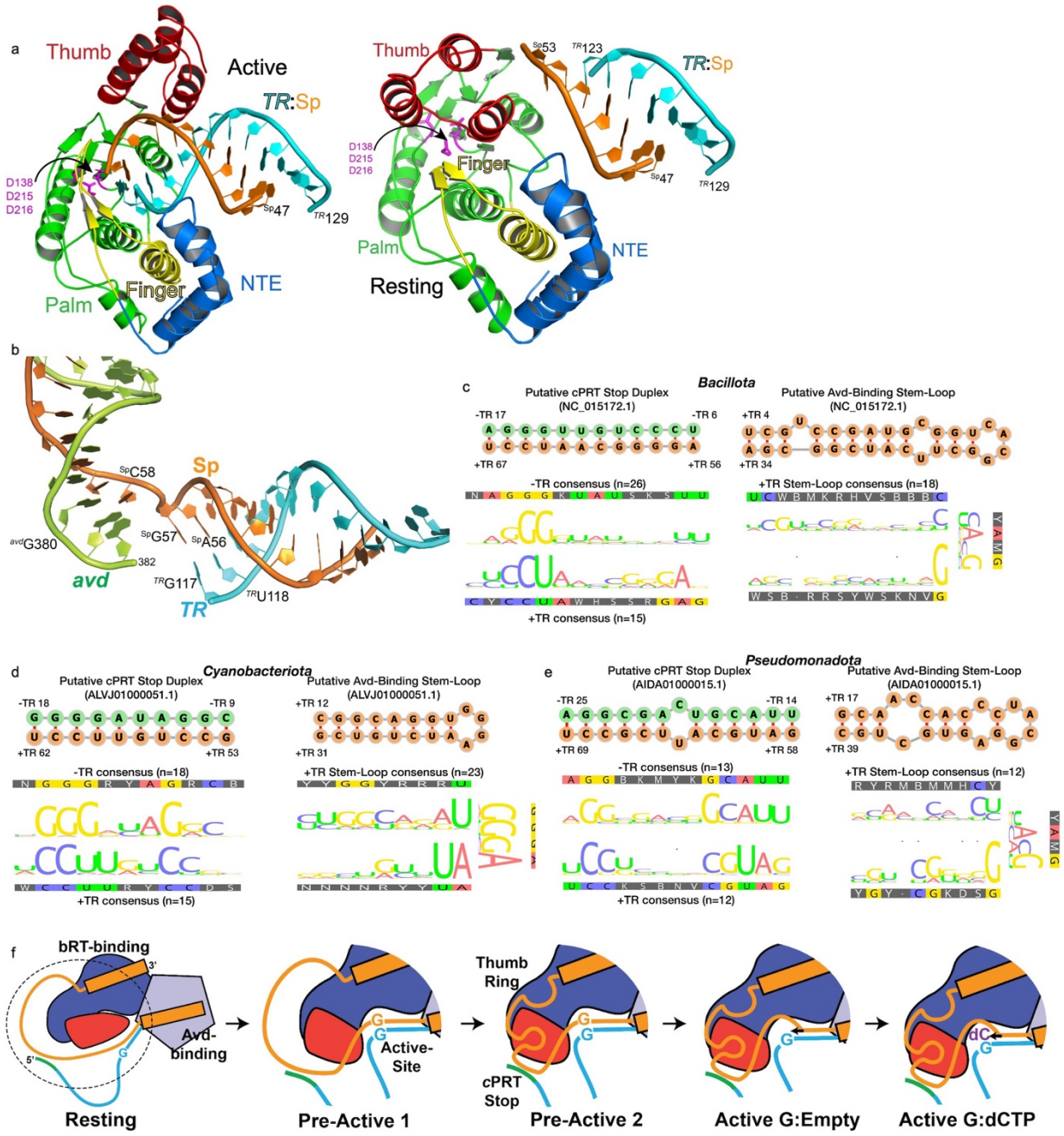

Fig. 5. RNA Conformational Variability. a. (Left) Active and (right) Resting conformations, with bRT and *TR*:Sp template:primer duplex shown (coloring as in Fig. 2a). The locations of catalytic D138, D215, and D216 identify the active site. The thumb subdomain was translated ~19 Å and rotated ~117° in the Resting conformation relative to the Active conformation. **b.** *TR* and Sp RNA in the Pre-Active 1 conformation. **c-e.** Representative RNA structure predictions from *Bacillota* (c), *Cyanobacteriota* (d), and *Pseudomonadota* (e) DGRs. Putative *c*PRT Stop duplexes and Avd-binding stem-loops are represented by example sequences (sequences upstream and downstream of *TR* are green and orange, respectively), with corresponding sequence alignment consensus profiles. The number of genomes in each alignment is indicated. **f.** Schematic of bRT-Avd:RNA conformational states, shown as a pathway from Resting to Active G:dCTP. The circled area in the Resting conformation is magnified in the remaining panels.

While such interactions are rare, a similar pairing is observed in *E. coli* 16S rRNA (*26*). The Avd- and bRT-binding elements were in the same positions as in the Active conformation, but

240 the stem-loop of the *c*PRT Stop was disengaged from bRT and the Thumb Ring was too flexible
241 to be modeled. This conformation was called 'Pre-Active 1' to distinguish it from a variant in
242 which ^{Sp}G57 still occupied the N-site but the *c*PRT Stop stem-loop and Thumb Ring were
243 positioned as in the Active conformation and engaged with bRT; this latter variant was called
244 'Pre-Active 2' (Fig. S17, Table S1, Complex 9). The Pre-Active 2 conformation appeared to be
245 transitional between the Pre-Active 1 and Active conformations, as the density for ^{Sp}G57 in the
246 T-site was weaker in Pre-Active 2 than in Pre-Active 1 and ^{Sp}C58 was too flexible to be
247 modeled, consistent with ^{Sp}G57 exiting the N-site and making it available for an incoming dNTP.
248

249 **Conservation of RNA elements**

250 DGRs from microbial genomes spanning *Pseudomonadota*, *Bacillota*, *Cyanobacteriota*,
251 *Nanoarchaeota*, and members of the Candidate Phylum Radiation (CPR) among others had RNA
252 structural elements similar to ones observed in the BΦ DGR. Of 278 DGRs that have Avd
253 homologs (*13*), 207 had putative Avd-binding stem-loops located in sequences just downstream
254 of *TR* (Figs. 5c-e and Fig. S18, and Table S3). The stems most commonly had 9 bps, contained
255 potentially flipped-out bases, and were capped by tetraloops. GNRA tetraloops were prevalent
256 among *Cyanobacteriota* DGRs and UNCG tetraloops occurred frequently in *Pseudomonadota*
257 DGRs. The same 207 DGRs had putative *c*PRT Stop duplexes, involving pairing of distal
258 sequences (upstream and downstream of *TR*). The duplexes had an average of 13 bps, with
259 sequences especially conserved in *Cyanobacteriota* DGRs. The *c*PRT Stop element is complex,
260 involving not only a duplex but also a stem-loop followed by adenosines in the sequence
261 downstream of *TR*. Remarkably, 47 DGRs had predicted stem-loops adjacent to the duplex, with
262 all but two having multiple adenosines following, suggesting the presence of A-minor motif
263 interactions. These results strongly suggested that the mechanism of DGR reverse transcription
264 was conserved across distant taxa.
265

266 **Discussion**

267 We sought to understand how the DGR RNA with its essential RNA segments
268 surrounding *TR* combines with bRT and Avd to initiate reverse transcription. Most RTs require a
269 primer for cDNA synthesis, and the source of that primer varies widely. For example, group II
270 intron RTs and non-LTR retrotransposons use nicked target DNA as a primer (*21, 27-30*), HIV-1

271 RT uses a host tRNA (*31, 32*), and telomerase uses an associated RNA (*33*). These primers along
272 with the corresponding templates are recognized directly through sequence, structure, or both.
273 Our work revealed an unprecedented mechanism for the BΦ DGR, in which the primer and
274 template were not recognized directly, but rather placed precisely in the bRT active site through
275 extensive, intricate, and intimate interaction of RNA structural elements surrounding the
276 template.

277 Crucial structures and functions were distributed across the entire length of Sp RNA. At
278 its 5' and 3' ends, Sp formed stem-loops that acted as grapples for Avd and bRT, respectively,
279 binding to these proteins through flipped-out bases. The grapples were positioned similarly
280 across multiple RNP conformations. These stable grapples provide a ready explanation for the
281 processivity endowed by the DGR RNA, enabling *TR* to maintain association with bRT-Avd
282 through multiple polymerization cycles. Notably, bRT and other DGR RTs lack the processivity-
283 promoting α-loop found in group II intron RTs (Fig. S18) (*34*). In between the ends, the DGR
284 RNA formed the template:primer duplex, *c*PRT Stop, and Thumb Ring. The latter two elements
285 were especially mobile across multiple RNA conformations, and in the Active conformation
286 interacted extensively with bRT, appearing to provide a conformational guide for the placement
287 of the template:primer duplex in the bRT active site.

288 The multiple RNP conformations observed lay along a plausible pathway from initial
289 association of the DGR RNA with bRT-Avd to initiation of reverse transcription. We surmise
290 that the Resting conformation corresponds to initial association, with the template:primer duplex
291 on the surface of bRT, blocked from entering the active site by the collapsed thumb subdomain,
292 and the *c*PRT Stop and Thumb Ring disorganized and disengaged from bRT. A similar
293 conformation with the thumb collapsed into the active site in the absence of a template:primer
294 duplex occurs in HIV-1 RT (*35*). The Resting conformation likely proceeds to the Pre-Active 1
295 conformation, with the thumb receding and the template:primer duplex entering the active site,
296 but with ^{Sp}57G in the N-site. We had previously found that ^{Sp}G57 was required for *c*PRT (*9*), but
297 the basis for this was unexplained. Our current results suggest that guanosine at ^{Sp}57 is necessary
298 because it stabilizes the N-site prior to complete organization of Sp RNA into the Active
299 conformation. As noted above, the Pre-Active 2 conformation likely represents a transition to the
300 Active conformation, with the RNA completely organized and ^{Sp}57G at the N-site but with
301 greater apparent mobility. Displacement of ^{Sp}57G from the N-site would result in the Active N-

302 empty conformation, and finally, binding of an incoming dNTP would stabilize the Active
303 conformation, enabling chemistry to occur. An N-empty conformation with adenosine at the T-
304 site was nearly indistinguishable with one with guanosine. No contact by bRT or Avd was made
305 to the substituent at the C6 position of A or G, which is a crucial determinant of A-mutagenesis
306 (*14*). These observations suggest that A-mutagenesis is unlikely a result of specific protein
307 contacts or major structural rearrangements, and more likely attributable to dynamics.
308 In summary, we discovered that reverse transcription in the BΦ DGR depends on an
309 abundance of essential interactions between RNA structural elements and bRT-Avd, resulting in
310 the precise positioning of a template:primer duplex for *cis*-priming and strictly restricting A-
311 mutagenesis to DGR variable protein. Strikingly, one of the more complex RNA structural
312 elements, the *c*PRT Stop, appeared to be present in DGRs from distant taxa, suggesting an
313 extraordinary conservation of the mechanism of reverse transcription.
314
315 **Materials and Methods**
316 **Protein expression and purification, RNA synthesis, and reverse transcription**
317 The coding sequence for the Avd fusion partner *Methylobacterium extorquens*
318 formaldehyde activating enzyme (Fae, PDB 1Y60, 170 amino acids) was synthesized with
319 codons optimized for *E. coli* expression (Genewiz). The coding sequence, starting from the
320 codon for Fae amino acid 4, was inserted after the Avd codon for amino acid 124. bRT and Avd
321 or Avd-fusion protein were co-expressed in *E. coli* BL21 Gold (DE3) and purified as described
322 (*9*). RNA was produced through *in vitro* transcription using T7 RNA polymerase and gel-
323 purified, and reverse transcription reactions with bRT and Avd or Avd-fusion protein were
324 carried out, as described (*9*). RNA and cDNA were examined by denaturing gel electrophoresis,
325 as described (*9*), and in some cases, cDNA was subjected to short-read next-generation
326 sequencing (Amplicon EZ, Genewiz), as described (*14*).
327
328 **Cryo-EM sample preparation and data collection**
329 For assembly of complexes, 150 µL of $RNA^{\Delta 98}$ (14 µM) was incubated in 1.2 mL of 75
330 mM KCl, 3 mM $MgCl_2$, 10 mM DTT, 50 mM HEPES, pH 7.5, and 10 % glycerol at 37 °C for
331 15 min. bRT-Avd-Fae (150 µL of 14 µM) was added and the mixture was further incubated at
332 37 °C for 2 h. The concentration of the complex was determined using A_{280} measured with a

333 NanoDrop (Thermo Scientific) and a calculated molar extinction coefficient of 206,960
334 $M^{-1}cm^{-1}$. The sample was passed through a 0.22 µm filter and concentrated to 100 µL using a 30-
335 kDa molecular weight cutoff membrane. The complex was resolved by gel filtration at 4 °C with
336 a Superdex 200 10/300 column (GE Healthcare) equilibrated with 75 mM KCl, 30 mM
337 $(NH_4)_2SO_4$, 50 mM Tris, pH 7.5, 3 mM $MgCl_2$, and 2 mM DTT. Eluted fractions were analyzed
338 by SDS-PAGE to visualize proteins, and denaturing gel electrophoresis (8 M urea containing 8
339 % polyacrylamide gels in 1× TBE buffer) to visualize RNA. Fractions containing bRT-Avd-
340 Fae/$RNA^{\Delta 98}$ complexes were concentrated to ~2 µM using a 30-kDa molecular weight cutoff
341 membrane. The concentration of bRT-Avd-Fae/$RNA^{\Delta 98}$ was determined by measuring
342 absorbance at A_{260} with a NanoDrop (Thermo Scientific), assuming a 1:5:1 bRT:Avd-
343 Fae:$RNA^{\Delta 98}$ stoichiometry. The concentration of the complex was adjusted to 1 µM and left
344 either unincubated or incubated with the following dNTP analogs or dNTPs at 37 °C for 2 h. For
345 complexes containing wild-type $RNA^{\Delta 98}$, 100 µM of non-hydrolyzable dCTP (2'-deoxycytidine-
346 5'-[(α,β)-methyleno]triphosphate, Jena Bioscience) was used. For complexes containing *TR*
347 G117A $RNA^{\Delta 98}$, 2 mM non-hydrolyzable dCTP (as above) or 2 mM non-hydrolyzable dTTP (2'-
348 deoxythymidine-5'-[(α,β)-methyleno]triphosphate, Jena Bioscience) was used. For complexes
349 containing *TR* U114A $RNA^{\Delta 98}$, 2 mM each of dCTP, dATP, and ddGTP was used. Complexes
350 (4 µL) were then applied manually 2-3x in quick succession to Quantifoil™ R 1.2/1.3 Au grids,
351 which had been freshly plasma cleaned in Gatan Solarus II (under Ar/O_2 mixture) for 10 sec.
352 After application, grids were blotted with varying force and times using a Vitrobot Mark IV
353 (Thermo Fisher) at 4 °C and 100% humidity, and then immediately plunge-frozen in liquid
354 ethane.
355 Cryo-EM data were collected on either a 300 keV Titan Krios cryo-transmission electron
356 microscope at the Pacific Northwest Center for Cryo-EM (PNCC) equipped with GIF quantum
357 energy filters attached to either a K3 electron counting direct detector (Gatan) using Serial EM
358 software, or a 300 keV Titan Krios cryo-transmission electron microscopes (FEI Company) at
359 the University of California San Diego Cryo-EM facility equipped with GIF quantum energy
360 filters attached to a K2 electron counting direct detector (Gatan) using EPU (Thermo Fisher).
361 Images were acquired at a magnification of ~130,000, corresponding to a calibrated raw pixel
362 size of ~0.5 Å, in EFTEM mode with an energy filter slit width of 10-20 eV. Movie image stacks
363 with a total electron dose of ~55 $e^-/Å^2$ were saved in non-super resolution counting mode.

364

365 **Cryo-EM data processing and analysis**

366 The collection and processing of raw data is outlined in detail in Fig. S2 and Table S1.
367 Briefly, raw movies of each independent dataset were imported in cryoSPARC 3.1.1 (*36*),
368 RELION 3.0, or both (*37*); motion-corrected using patch motion correction (multi) of
369 cryoSPARC or MotionCor2; and CTF estimated using patch CTF estimation (multi) in
370 cryoSPARC, or GCTF or CTFFIND4 in RELION. A list of micrographs for each dataset
371 (removing poor quality micrographs using various criteria e.g., CTF fit of < 8 Å, astigmatism <
372 1200, motion distribution < 60, and motion curvature < 40) was prepared. Data were collected at
373 a relative ice thickness of ~1-1.2 μm, as thin ice led to particle denaturation whereas thick ice led
374 to high background noise. Particles were picked from curated micrographs using various
375 methods and iterative strategies outlined in Fig. S2a. Duplicate particles were removed in
376 cryoSPARC, and binned or un-binned particles were extracted from micrographs depending on
377 the purpose and stage of analysis. Multiple rounds of 2D classification were performed to
378 eliminate poor quality particles that could not be classified (Fig. S3). Selected particles from 2D
379 classes were used for initial 3D classification (with different number of classes) using *ab initio*
380 reconstruction (Fig. S4). Particles belonging to poorly resolved *ab initio* 3D class averages were
381 discarded. Particles from 3D classes which indicated the presence of all domains were used for
382 heterogeneous classification with various plausible models. Good classes from heterogeneous
383 refinements were fed back into refined template picking using programs outlined in Fig. S2.
384 Iteration of the above processes provided a refined set of particles, which were classified into
385 probable 3D classes (judged from shape profiles of 3D models of heterogenous classification),
386 and eventually particles from final 3D class sets were extracted without binning with a 320-pixel
387 box size for consecutive refinement using homogeneous, non-uniform, and local refinement
388 algorithms in cryoSPARC with appropriate masks. Per-particle defocus refinement was also
389 performed, although resulting in little to no discernible improvement in map quality. Trials of
390 particle subtraction did not indicate any improvement of maps and thus this step was omitted
391 from refinement. The total number of movies, number of particles used for final refinement of
392 each class, orientation distribution, half-map-based correlation, global resolution estimate, and
393 several other parameters are provided in Table S1. Density sharpened versions of the maps were
394 used for model building.

395

396 **Model building and refinement**

397 The model building pipeline is depicted in detail in Fig. S2b. In brief, an initial model of
398 bRT was built manually using PDB templates of poliovirus RNA-dependent RNA polymerase
399 (3OL9) and group II intron RT GsI-IIC (6AR1), and the initial model of Avd was based on the
400 structure of the free form of the pentamer (PDB 4DWL). For modeling RNA, the secondary
401 structure for DGR RNA was first predicted through iterative rationalization using RNAstructure
402 (*38*), incorporating refined SHAPE-MaP data and biochemical results, and finalized using initial
403 map features (e.g., base pairing in the *TR*:Sp template:primer duplex) (*9, 15*). The dot bracket
404 files defining plausible RNA stem regions were used with simRNA (*39*) and RNAComposer (*40*)
405 to obtain initial 3D templates of RNA stems, which were docked in the sharpened maps. Rigid
406 body fitted models of bRT subdomains, Avd protomers, RNA stems and the connecting single
407 stranded RNA regions were iteratively (re)built in real space in Coot (*41*). In addition, a
408 pentameric model of Fae (PDB 1Y60) was placed into the diffuse density contiguous to the
409 density for the Avd pentamer. The models of complexes composed of bRT, Avd, and $RNA^{\Delta 98}$
410 were finally refined in CCPEM (*42*). Model geometry was validated using MolProbity package
411 within the Phenix suite (*43*). This final model displays good stereochemistry (Table S1), except
412 for regions of high flexibility. RNA structure was analyzed using RNAPdbee (*44*). Q-scores (*45*)
413 were computed using Chimera (*46*). Derived B-factors were calculated as 150 • (1 – Q) (*18*),
414 with the scale factor of 150 empirically identified to provide the highest map correlation
415 coefficient, as calculated with Phenix (*43*) (Table S2). Figures of molecular models were made
416 with PyMOL (The PyMOL Molecular Graphics System, Schrödinger, LLC) and ChimeraX (*47*).

417

418 **SHAPE-MaP**

419 SHAPE-MaP experiments were carried out using 10 mM N-methylisatoic anhydride
420 (NMIA, Millipore-Sigma). Thirty-five µL of DGR RNA (500 ng/µL) in water was denatured at
421 95 °C for 3 min and then immediately placed on ice for 5 min. The RNA was refolded by
422 incubation at 37 °C for 20 min in 110 mM KCl, 70 mM HEPES, pH 7.5, 4 mM $MgCl_2$ in a total
423 volume of 270 µL. Five µL of 300 mM $(NH_4)_2SO_4$, 20 mM HEPES, pH 7.5, 5 mM $MgCl_2$, 2
424 mM DTT, and 10% glycerol containing bRT-Avd (2 mg/mL) was added to 40 µL of refolded
425 RNA, whereupon the sample was incubated at 37 °C for 1 h. Five µL NMIA (100 mM in

426 DMSO) was then added to the samples, which were further incubated at 37 °C for 1 h. After this
427 time, NMIA was quenched by the addition of DTT to a final concentration of 50 mM. A negative
428 control for the reaction was carried out similarly, but with 5 µL of DMSO instead of NMIA.
429 Reactions were then incubated with 2 µL of proteinase K (Ambion, 0.4 mg/ml) at 50 °C for 30
430 min, and proteinase K was then quenched with a final concentration of 1 mM PMSF. The
431 resulting samples were purified with a desalting column (G-25).
432 A denaturing control was carried out by incubating 2 µL of RNA (500 ng/µL), 10 µL
433 formamide, 2 µL of 10x denaturing buffer (500 mM HEPES, pH 8.0, and 40 mM EDTA) in a
434 total volume of 18 µL. The sample was incubated at 95 °C for 1 min, and 2 µL of NMIA (100
435 mM) was added and further incubated at 95 °C for 1 min. The sample was placed on ice for 5
436 min, and then purified with a desalting column (G-25) after adjusting the volume to 50 µL.
437 Reverse transcription reactions were carried out by first annealing 8 µL of purified RNA
438 from the previous step with 1 µL of reverse transcription primer P1 (2 µM, GACTGGAGTTC
439 AGACGTGTGCTCTTCCGATCTGAAGTCGGCCCCGCCTTT) and 3 µL water at 65 °C for 5
440 min and cooling on ice for 5 min. To these annealed reactions, 8 µL of 2.5x MaP buffer (50 mM
441 Tris, pH 8.0, 75 mM KCl, 6 mM MnCl2, 10 mM DTT, and 0.5 mM dNTPs) and 1 µL of
442 ProtoScript® II Reverse Transcriptase (NEB) was added. These reactions were incubated at 42
443 °C for 3 h and then at 70 °C for 15 min. These samples were then purified with a desalting
444 column (G-25) after adjusting the volume to 50 µL.
445 The cDNAs resulting from the reverse transcription reactions were amplified in a 100 µL
446 PCR reaction using 10 µL of the purified sample from the previous step, 50 µL of Q5® High-
447 Fidelity 2X Master Mix (NEB), 5 µL each of 10 µM primer P1 and P2
448 (ACACTCTTTCCCTACACGACGCTCTTCCGATCTAAGGGCA GGCTGGGAAAT). The
449 PCR conditions used for amplification were: 98 °C for 30 s; 30 cycles of 98 °C for 10 s, 60 °C
450 for 30 s, and 72 °C for 20s; followed by 72 °C for 2 min. The amplified DNA was gel purified
451 and the samples were deep-sequenced in multiplexes using the Illumina MiSeq platform (2 × 250
452 bp, paired-end, Genewiz). Demultiplexed fastq files were processed using ShapeMapper 2.0
453 pipeline (*48*).
454
455 **EMSA and Mass Spectrometry**

456 Reverse transcription was carried out at 37 °C for 20 min in 20 µL volume containing 1.4
457 µM RNA, varying concentrations of bRT-Avd (0-2.0 µM), and 20 units RNase inhibitor (NEB)
458 in 75 mM KCl, 3 mM $MgCl_2$, 10 mM DTT, 50 mM HEPES, pH 7.5, 10% glycerol. The reaction
459 contents were centrifuged at 12,000 x *g* for 10 min, and 15 µL was loaded on a native 6%
460 polyacrylamide gel and resolved for 90 min at 100 V. The gel was stained with GreenGlo™ Safe
461 DNA Dye (Thomas Scientific) and imaged on a Bio-Rad Chemidoc™ system. Gel-shifted bands
462 (Fig. S13, black arrow) were extracted from the gel, and subjected to mass spectrometry on an
463 Orbitrap Fusion Lumos MS at the UCSD Biomolecular and Proteomics Mass Spectrometry
464 facility. Mass spectrometry data were analyzed using Peaks Studio 10.6 software.
465
466 **Analysis of RNA elements**
467 The subset of DGR sequences from Wu et al. (*13*) that contain an Avd-like gene along
468 with five archaeal genomes were analyzed. For each sequence, 40-nt upstream of *TR* was
469 extracted and concatenated to 100-nt downstream of *TR* (i.e. spanning most of the putative Sp).
470 RNA fold predictions were carried out with MxFold2 tool v. 2.4.18 (*49*). The resulting structural
471 predictions were inspected for putative *cPRT* Stop duplexes (complementarity between
472 sequences up- and downstream of *TR*) and Avd-binding stem-loops (in sequences downstream of
473 *TR*).
474
475 **Acknowledgments**
476 Use of Pacific Northwest Center for Cryo-EM (PNCC) supported by NIH grant U24GM129547.
477 We thank R. Baker and A. Leschziner for initial cryo-EM characterization of complexes, T.
478 Baker for computational resources, and J. Meyers for data collection at PNCC.
479
480 **Funding**
481 This work was supported by the National Institutes of Health R01 GM132720 (PG), AI163327
482 (GG), and R01 GM033050-35 (T. Baker). BGP was supported by the Gordon and Betty Moore
483 Foundation, the G. Unger Vetlesen Foundation, and the Owens Family Foundation.
484
485 **Conflict of Interest.**
486 No conflicts to declare.

487

488 **Author Contributions.**

489 Conceptualization: SH and PG

490 Investigation: SH, TB, JC, BGP, PG

491 Visualization: SH, TB, BGP, PG

492 Funding acquisition: PG, GG

493 Project Administration: PG

494 Supervision: PG

495 Writing – original draft: SH, TB, BGP, PG

496 Writing – review & editing: SH, TB, GG, BGP, PG
497

498 **Supplementary Information**

499 Supplementary Figures S1-S19.

500 Supplementary Tables S1-S3.

501

502 **References**

503 1. S. Roux *et al.*, Ecology and molecular targets of hypermutation in the global microbiome.
504 *Nat Commun* **12**, 3076 (2021).
505 2. S. Nayfach *et al.*, Metagenomic compendium of 189,680 DNA viruses from the human
506 gut microbiome. *Nat Microbiol* **6**, 960-970 (2021).
507 3. B. G. Paul *et al.*, Retroelement-guided protein diversification abounds in vast lineages of
508 Bacteria and Archaea. *Nat Microbiol* **2**, 17045 (2017).
509 4. C. T. Brown *et al.*, Unusual biology across a group comprising more than 15% of domain
510 Bacteria. *Nature* **523**, 208-211 (2015).
511 5. H. Dore *et al.*, Targeted hypermutation of putative antigen sensors in multicellular
512 bacteria. *Proc Natl Acad Sci U S A* **121**, e2316469121 (2024).
513 6. G. Kaur, A. M. Burroughs, L. M. Iyer, L. Aravind, Highly regulated, diversifying NTP-
514 dependent biological conflict systems with implications for the emergence of
515 multicellularity. *Elife* **9**, (2020).
516 7. M. Liu *et al.*, Reverse transcriptase-mediated tropism switching in Bordetella
517 bacteriophage. *Science* **295**, 2091-2094 (2002).
518 8. S. A. McMahon *et al.*, The C-type lectin fold as an evolutionary solution for massive
519 sequence variation. *Nat Struct Mol Biol* **12**, 886-892 (2005).
520 9. S. Handa *et al.*, Template-assisted synthesis of adenine-mutagenized cDNA by a
521 retroelement protein complex. *Nucleic Acids Res* **46**, 9711-9725 (2018).
522 10. S. Doulatov *et al.*, Tropism switching in Bordetella bacteriophage defines a family of
523 diversity-generating retroelements. *Nature* **431**, 476-481 (2004).
524 11. B. G. Paul *et al.*, Targeted diversity generation by intraterrestrial archaea and archaeal
525 viruses. *Nat Commun* **6**, 6585 (2015).

526 12. M. Alayyoubi *et al.*, Structure of the essential diversity-generating retroelement protein
527 bAvd and its functionally important interaction with reverse transcriptase. *Structure* **21**,
528 266-276 (2013).
529 13. L. Wu *et al.*, Diversity-generating retroelements: natural variation, classification and
530 evolution inferred from a large-scale genomic survey. *Nucleic Acids Res* **46**, 11-24
531 (2018).
532 14. S. Handa, A. Reyna, T. Wiryaman, P. Ghosh, Determinants of adenine-mutagenesis in
533 diversity-generating retroelements. *Nucleic Acids Res* **49**, 1033-1045 (2021).
534 15. S. S. Naorem *et al.*, DGR mutagenic transposition occurs via hypermutagenic reverse
535 transcription primed by nicked template RNA. *Proc Natl Acad Sci U S A* **114**, E10187-
536 E10195 (2017).
537 16. S. Inouye, M. Y. Hsu, A. Xu, M. Inouye, Highly specific recognition of primer RNA
538 structures for 2'-OH priming reaction by bacterial reverse transcriptases. *J Biol Chem*
539 **274**, 31236-31244 (1999).
540 17. H. Guo *et al.*, Diversity-generating retroelement homing regenerates target sequences for
541 repeated rounds of codon rewriting and protein diversification. *Mol Cell* **31**, 813-823
542 (2008).
543 18. G. Pintilie, W. Chiu, Validation, analysis and annotation of cryo-EM structures. *Acta*
544 *Crystallogr D Struct Biol* **77**, 1142-1152 (2021).
545 19. C. Zhao, A. M. Pyle, Crystal structures of a group II intron maturase reveal a missing link
546 in spliceosome evolution. *Nat Struct Mol Biol* **23**, 558-565 (2016).
547 20. J. L. Stamos, A. M. Lentzsch, A. M. Lambowitz, Structure of a Thermostable Group II
548 Intron Reverse Transcriptase with Template-Primer and Its Functional and Evolutionary
549 Implications. *Mol Cell* **68**, 926-939 e924 (2017).
550 21. D. B. Haack *et al.*, Cryo-EM Structures of a Group II Intron Reverse Splicing into DNA.
551 *Cell* **178**, 612-623 e612 (2019).
552 22. F. J. Blocker *et al.*, Domain structure and three-dimensional model of a group II intron-
553 encoded reverse transcriptase. *RNA* **11**, 14-28 (2005).
554 23. P. Nissen, J. A. Ippolito, N. Ban, P. B. Moore, T. A. Steitz, RNA tertiary interactions in
555 the large ribosomal subunit: the A-minor motif. *Proc Natl Acad Sci U S A* **98**, 4899-4903
556 (2001).
557 24. S. F. Torabi *et al.*, RNA stabilization by a poly(A) tail 3'-end binding pocket and other
558 modes of poly(A)-RNA interaction. *Science* **371**, (2021).
559 25. P. Danaee *et al.*, bpRNA: large-scale automated annotation and analysis of RNA
560 secondary structure. *Nucleic Acids Res* **46**, 5381-5394 (2018).
561 26. J. A. Dunkle *et al.*, Structures of the bacterial ribosome in classical and hybrid states of
562 tRNA binding. *Science* **332**, 981-984 (2011).
563 27. K. Chung *et al.*, Structures of a mobile intron retroelement poised to attack its structured
564 DNA target. *Science* **378**, 627-634 (2022).
565 28. M. E. Wilkinson, C. J. Frangieh, R. K. Macrae, F. Zhang, Structure of the R2 non-LTR
566 retrotransposon initiating target-primed reverse transcription. *Science* **380**, 301-308
567 (2023).
568 29. P. Deng *et al.*, Structural RNA components supervise the sequential DNA cleavage in R2
569 retrotransposon. *Cell* **186**, 2865-2879 e2820 (2023).
570 30. A. Thawani, A. J. F. Ariza, E. Nogales, K. Collins, Template and target-site recognition
571 by human LINE-1 in retrotransposition. *Nature* **626**, 186-193 (2024).

31. K. P. Larsen *et al.*, Architecture of an HIV-1 reverse transcriptase initiation complex. *Nature* **557**, 118-122 (2018).
32. K. Das, S. E. Martinez, J. J. DeStefano, E. Arnold, Structure of HIV-1 RT/dsRNA initiation complex prior to nucleotide incorporation. *Proc Natl Acad Sci U S A* **116**, 7308-7313 (2019).
33. B. Liu *et al.*, Structure of active human telomerase with telomere shelterin protein TPP1. *Nature* **604**, 578-583 (2022).
34. C. Zhao, F. Liu, A. M. Pyle, An ultraprocessive, accurate reverse transcriptase encoded by a metazoan group II intron. *RNA* **24**, 183-195 (2018).
35. Y. Hsiou *et al.*, Structure of unliganded HIV-1 reverse transcriptase at 2.7 A resolution: implications of conformational changes for polymerization and inhibition mechanisms. *Structure* **4**, 853-860 (1996).
36. A. Punjani, J. L. Rubinstein, D. J. Fleet, M. A. Brubaker, cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat Methods* **14**, 290-296 (2017).
37. S. H. Scheres, RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J Struct Biol* **180**, 519-530 (2012).
38. J. S. Reuter, D. H. Mathews, RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* **11**, 129 (2010).
39. M. Magnus, M. J. Boniecki, W. Dawson, J. M. Bujnicki, SimRNAweb: a web server for RNA 3D structure modeling with optional restraints. *Nucleic Acids Res* **44**, W315-319 (2016).
40. M. Popenda *et al.*, Automated 3D structure composition for large RNAs. *Nucleic Acids Res* **40**, e112 (2012).
41. P. Emsley, B. Lohkamp, W. G. Scott, K. Cowtan, Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* **66**, 486-501 (2010).
42. K. Yamashita, C. M. Palmer, T. Burnley, G. N. Murshudov, Cryo-EM single-particle structure refinement and map calculation using Servalcat. *Acta Crystallogr D Struct Biol* **77**, 1282-1291 (2021).
43. D. Liebschner *et al.*, Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr D Struct Biol* **75**, 861-877 (2019).
44. T. Zok *et al.*, RNApdbee 2.0: multifunctional tool for RNA structure annotation. *Nucleic Acids Res* **46**, W30-W35 (2018).
45. G. Pintilie *et al.*, Measurement of atom resolvability in cryo-EM maps with Q-scores. *Nat Methods* **17**, 328-334 (2020).
46. E. F. Pettersen *et al.*, UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605-1612 (2004).
47. E. C. Meng *et al.*, UCSF ChimeraX: Tools for structure building and analysis. *Protein Sci* **32**, e4792 (2023).
48. S. Busan, K. M. Weeks, Accurate detection of chemical modifications in RNA by mutational profiling (MaP) with ShapeMapper 2. *RNA* **24**, 143-148 (2018).
49. K. Sato, M. Akiyama, Y. Sakakibara, RNA secondary structure prediction using deep learning with thermodynamic integration. *Nat Commun* **12**, 941 (2021).