Structural Requirements for Reverse Transcription by a Diversity-generating Retroelement

Sumit Handa[1,3,a], Tapan Biswas[1,a], Jeet Chakraborty[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), which are pervasive among microbes, create massive protein sequence variation through reverse transcription of a protein-coding RNA template coupled to frequent misincorporation at template adenines. For cDNA synthesis, the template must be surrounded by up- and downstream sequences. Cryo-EM revealed that this longer RNA formed an integral ribonucleoprotein (RNP) with the DGR reverse transcriptase bRT and associated protein Avd. The downstream, noncoding (nc) RNA formed stem-loops enveloping bRT and laying over barrel-shaped Avd, and duplexes with the upstream and template RNA. These RNA structural elements were required for reverse transcription, and several were conserved in DGRs from distant taxa. Multiple RNP conformations were visualized, and no large structural rearrangements occurred when adenine replaced guanine as the template base, suggesting energetics govern misincorporation at adenines. Our results explain how the downstream ncRNA primes cDNA synthesis, promotes processivity, terminates polymerization, and stringently limits mutagenesis to DGR variable proteins.

Introduction

Diversity-generating retroelements (DGRs) produce the greatest level of protein sequence variation in the natural world (*2*), up to 10^{30} potential sequences in one case (*1*), surpassing variation in the vertebrate adaptive immune system. They occur in ecologically diverse bacteria, archaea, and their viruses, and are especially enriched in the human gut microbiome (*3, 4*) and nano-sized microbes, which appear to comprise a major fraction of microbial life and maintain DGRs despite reduced genomes (*5, 6*). A recent survey identified ~31,000 DGRs from >1,500 bacterial and archaeal genera, constituting >90 environment types (*3*). Half to 75% of these DGRs showed signs of recent activity, and in some organisms, were responsible for >10% of all amino acid changes. The prototypical DGR of *Bordetella* bacteriophage (BΦ) can produce ~10^{13} variants of its receptor-binding protein Mtd (*7, 8*), which enables 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, as it is with the adaptive immune response.

Variation by DGRs occurs through mutagenic retrohoming, in which protein-coding information in an RNA template region (*TR*) is reverse transcribed to cDNA, with errors made specifically at template adenines (Fig. 1a). The resulting adenine (A)-mutagenized *TR*-cDNA then replaces a homologous variable region (*VR*) in a DGR variable protein locus (e.g., *mtd*) (*7*), giving rise to a protein variant. *TR* is nearly identical to *VR*, differing almost exclusively at adenines, and serves as an invariant source of genetic information. Reverse transcription by the BΦ DGR is carried out by a complex of bRT and Avd proteins (*9*). bRT belongs to a monophyletic clade of DGR reverse transcriptases (RTs) most closely related to group II intron RTs (*3, 10, 11*). Unlike other RTs, bRT lacks reverse transcription activity on its own and requires association with Avd (*9*). Avd forms a pentameric, positively-charged barrel-like structure that is similar to S23 ribosomal proteins (*12*), and Avd-like proteins occur in the great majority of DGRs (*1*). The bRT-Avd complex A-mutagenizes both *TR* and non-cognate RNA templates, indicating it has intrinsic A-mutagenesis activity (*9*). The *in vitro* A-misincorporation frequency is remarkably high at ~50% as measured *in vitro* (*13*), similar to observations made *in vivo* (*7, 14*). Even in the case of correct incorporation, bRT-Avd has a ~30-fold lower catalytic efficiency at a template adenine as compared to guanine, cytosine, and uracil (*13*), further emphasizing its adenine-specificity.

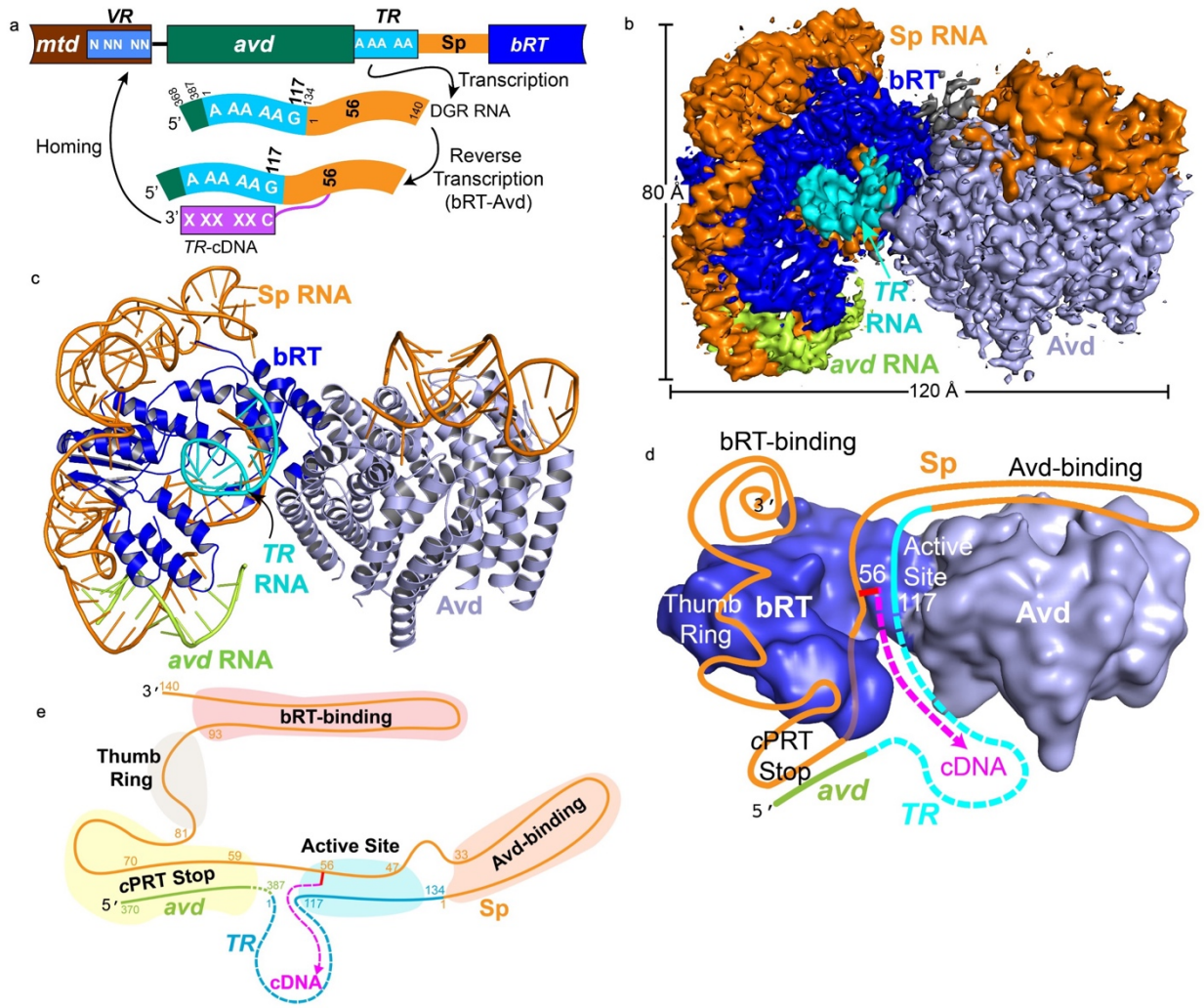

Fig. 1. DGR and bRT-Avd/RNA$^{\Delta 98}$ complex. a. Schematic of BΦ DGR. Transcription yields an RNA containing *TR* (134 nt) surrounded by up- and downstream sequences composed of at least 20 nt of *avd* and 140 of Sp, respectively. *TR* is reverse-transcribed by bRT-Avd, which introduces errors at template adenines. Reverse transcription initiates at ^{TR}G117 and ^{Sp}56A acts as the priming site. A-mutagenized *TR*-cDNA homes to the DGR variable protein locus (e.g., *mtd*) and replaces *VR*. **b and c.** Cryo-EM density (b) and cartoon (c) of bRT-Avd/RNA$^{\Delta 98}$ complex (bRT, blue; Avd, slate; RNA$^{\Delta 98}$ light green for *avd*, cyan for *TR*, orange for Sp). **d and e.** Schematic of RNA elements overlaid on a surface rendering of bRT-Avd (d) and a linear representation of RNA elements (e). Positions of the priming site (56 in Sp) and initiation site (117 in *TR*) are indicated. Dotted lines indicate sections that were either too flexible to model or were absent in the complex.

Significantly, *TR* is insufficient for cDNA synthesis by bRT-Avd and must instead be surrounded by *cis*-acting elements (*9*). Most prominent is a 140-nucleotide (nt) noncoding (nc) RNA segment called Spacer (Sp) immediately downstream of *TR* (Fig. 1a). The first ~30 Sp nts bind Avd (*9*), and ^{Sp}A56 primes cDNA synthesis from ^{TR}G117 (*9, 14*). The 2'-OH of ^{Sp}A56 is necessary for cDNA synthesis (*9*), suggesting that it acts as the priming nucleophile, while other evidence argues for the 3'-OH in cleaved RNA (*14*). In either case, reverse transcription is *cis*-

64 primed and produces a covalently linked RNA-cDNA molecule. The role of the remaining
65 portions of Sp is unknown. Likewise, just upstream of *TR*, a 20 nt segment from *avd* is required
66 for *cis*-primed reverse transcription (*c*PRT), but its role is unknown. We use the term 'DGR
67 RNA' to refer to the minimal RNA required for *c*PRT, containing 20 *avd* and 140 Sp nts
68 surrounding *TR*. Complementary DNAs produced from the DGR RNA are up to ~120 nt in
69 length, spanning *TR* and terminating just within *avd* (*9*). In contrast, cDNAs synthesized from
70 non-cognate RNA templates require an exogenous primer and yield only short cDNAs of 5-30
71 nts (*9*). Complementary DNAs primed exogenously from positions other than ^{TR}G117 in the
72 DGR RNA or from the DGR RNA truncated at ^{Sp}A56 are similarly short. These results indicate
73 that the DGR RNA uniquely confers *cis*-priming and processive polymerization.
74
75 **Structure Determination**
76 It is not known how the DGR RNA combines with bRT and Avd to produce *TR*-cDNA.
77 To address this question, we turned to cryo-EM, but initial efforts were hampered by RNA
78 flexibility, preferred particle orientation, and particle denaturation. Nevertheless, initial low-
79 resolution maps provided clues for overcoming these problems. A portion of the RNA appeared
80 to be highly flexible, which we surmised corresponded to *TR* because of its informational rather
81 than functional role (*15*). Accordingly, the central 98 *TR* nts (^{TR}10-107) of the DGR RNA were
82 deleted (yielding RNA$^{\Delta 98}$). Results from 2D and 3D classification suggested increasing the mass
83 of the ~176 kDa complex as a means for achieving higher resolution, and thus fusions were
84 constructed to the Avd C-terminus, which had been visualized to be unhindered. These entailed
85 several pentameric barrel-forming proteins, with the best results coming from *Methylobacterium*
86 *extorquens* formaldehyde activating enzyme (Fae). The complex formed by bRT, Avd-Fae, and
87 RNA$^{\Delta 98}$ (~258 kDa) yielded maps that extended to as high as ~3.1 Å resolution limit (Table S1
88 and Figs. S1, S2). bRT-(Avd-Fae) was active in reverse transcribing RNA$^{\Delta 98}$, but interestingly,
89 Fae interfered with reverse transcription of the longer DGR RNA (Fig. S3a), suggesting that
90 intact *TR* clashed with Fae (Figs. S3b, c). Adenine-mutagenesis by bRT-(Avd-Fae) was
91 evaluated with RNA$^{\Delta 98}$ containing ^{TR}G117A, as RNA$^{\Delta 98}$ lacked *TR* adenines that could be
92 reverse transcribed. Complexes containing Avd or the Avd-Fae fusion had the same adenine
93 misincorporation frequency (Fig. S2d) (*13*). Fae, despite relieving the preferred particle

94 orientation problem, varied in position with respect to the rest of the complex and was masked
95 out (Fig. S3b). For this reason, we refer hereafter to Avd rather than Avd-Fae.
96
97 **Active Conformation**
98 The bRT-Avd/RNA$^{\Delta 98}$ complex was visualized in several conformational states. We start
99 with a state that was fortuitously visualized in a complex containing RNA$^{\Delta 98}$ with a ^{TR}U114A
100 substitution that had been incubated with dCTP, dATP, and ddGTP (Table S1, Complex 1). The
101 intent was to visualize a short cDNA strand and ^{TR}U114A at the template site, but the population
102 of such complexes was too sparse for class-averaging, consistent with <2% of templates yielding
103 cDNA (*13*). However, in this same sample, a complex in which cDNA synthesis had not yet
104 commenced and had dCTP bound was well populated, amenable to class-averaging, and
105 produced the best maps.
106 Most strikingly, in this and all complexes, bRT, Avd, and RNA$^{\Delta 98}$ formed an integral
107 ribonucleoprotein (RNP) with the Sp ncRNA segment playing a predominant role (Fig. 1). Sp
108 RNA formed stem-loops that enveloped bRT and lay over the narrow end of the Avd barrel,
109 interacting with these proteins specifically through bases flipped-out from the stem-loops. Sp
110 RNA also formed duplexes with *TR* and the upstream *avd* segment. As predicted (*11*), bRT most
111 closely resembled group II intron RTs (*16-18*) (Fig. S4a), and shared with these proteins and
112 non-LTR RTs (*19*) an N-terminal extension subdomain (NTE, amino acids 1-59). Following the
113 NTE subdomain were the canonical finger (60-102), palm (103-272), and thumb (273-328)
114 subdomains of nucleotide polymerases (Fig. S4b). The bRT NTE subdomain was responsible for
115 contacting two protomers of pentameric Avd through an extensive (~2300 Å^2 total buried surface
116 area) and mainly electrostatic interface. One Avd protomer, called Avd1, nestled its N-terminal
117 10 amino acids into a crook within the four helix-bundle bRT NTE (Fig. S4c). This N-terminal
118 region was disordered in other Avd protomers and in free Avd. Bound and free Avd were
119 structurally similar (1.4 Å rmsd, 456 Cα) (*12*). Avd R79 and R83 are required for bRT
120 interaction (*12*), and consistent with these results, these amino acids in the second protomer,
121 Avd2, formed salt bridges with bRT E40 (Fig. S4d). While Avd is required for reverse
122 transcription (*9*), it was positioned ~40 Å from the bRT catalytic site, suggesting that Avd has no
123 direct role in chemistry.

124 The bRT active site was occupied by an RNA duplex with one strand (^{TR}117-129)

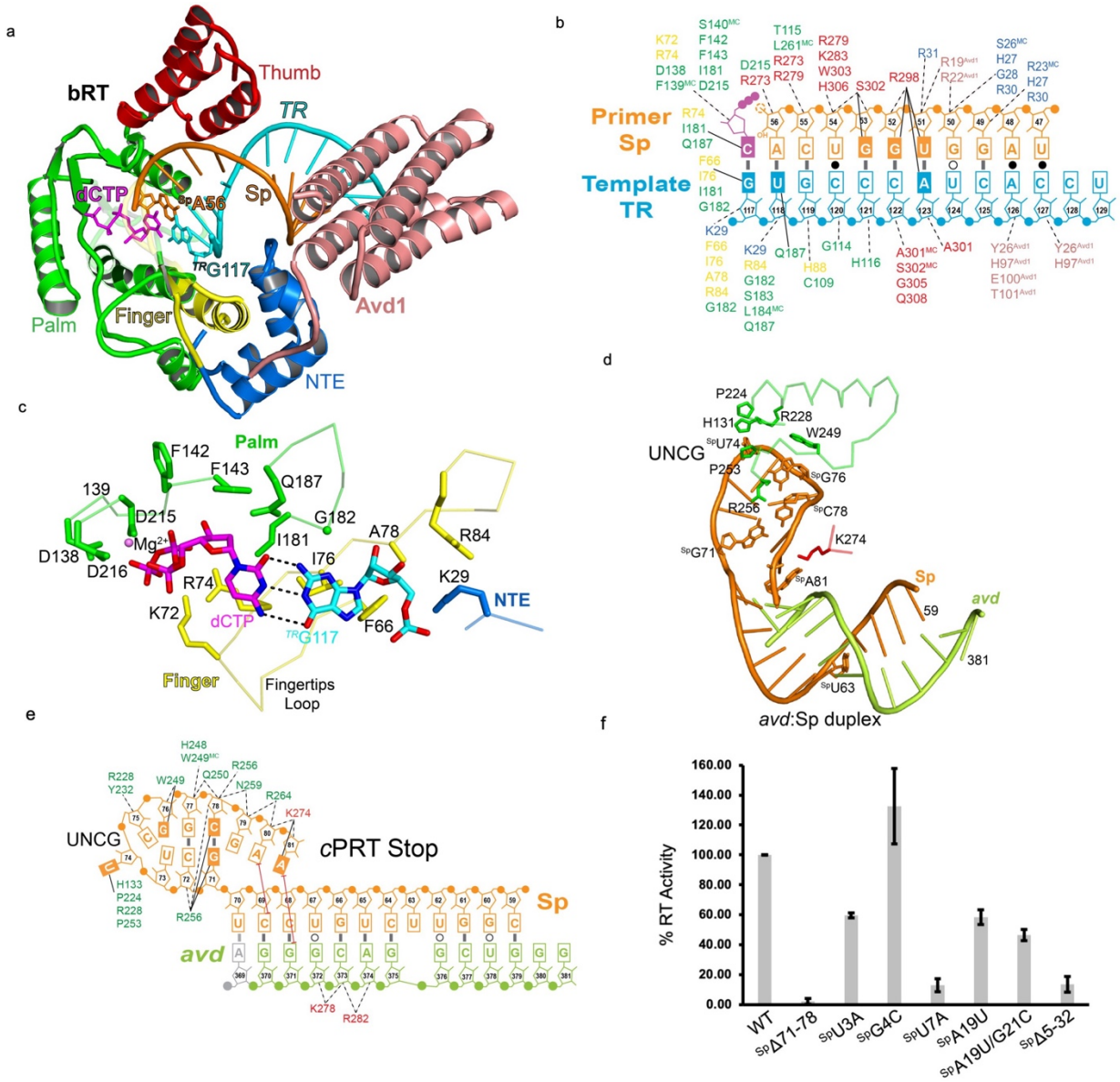

Fig. 2. Active Site Duplex and *c*PRT Stop. a. *TR*:Sp duplex (cyan and orange, respectively) and dCTP (magenta) in the active site of bRT (NTE, light blue; finger, yellow; palm, green; thumb, red) and contacted by Avd1 (salmon), shown in cartoon representation. Portions of the finger and palm subdomain were made partially transparent for clarity. **b.** Schematic of *TR*:Sp duplex with base pairs (solid bar, Watson-Crick; open circle, wobble; and closed circle, non-standard) shown. Base-specific (sold lines) and non-specific (dotted lines) contacts from bRT or Avd are shown. Color of labels correspond to coloring in panel a. **c.** Interactions between ^{TR}G117 (carbons and phosphorus colored cyan) at the T site and dCTP (carbons and phosphorus colored magenta) at the N site (for both nucleotides, nitrogens colored blue, and oxygens red), and with bRT amino acids (colored as in panel a), shown in bonds representation. Hydrogen bonds are shown as black dotted lines. **d.** Interactions between the *c*PRT stop (*avd*:Sp duplex and stem-tetraloop) and bRT shown in cartoon representation. ^{Sp}U63 and bases that are contacted by bRT are shown in bonds representation (2'- and 3'-OH of riboses omitted here and in other figures for clarity), as are the bRT amino acids responsible for these

contacts. **e.** Schematic of *c*PRT stop, consisting of an *avd*:Sp duplex and stem-UNCG tetraloop. Depiction is as in panel b. The base preceding ^{avd}G370 is ^{avd}A369, which can base pair with ^{Sp}U70, but was a guanine in RNA$^{\Delta 98}$ for technical reasons; for this reason, ^{avd}A369 is colored gray. The red lines indicate tertiary contacts between the stem-loop and duplex. **f.** *c*PRT activity of mutant DGR RNA relative to wild-type DGR RNA. Data are from experiments that were repeated three independent times, and means and standard deviations shown.

125 containing the initiating nucleotide ^{TR}G117 and the other (^{Sp}47-56) the priming nucleotide ^{Sp}A56
126 (Figs. 2a and 2b). The *TR*:Sp duplex was grasped by NTE, finger, and thumb subdomains, and
127 rested on the palm subdomain. Avd1 also contacted the active site duplex, and may thus function
128 in reverse transcription by enforcing proper positioning of the duplex or the NTE subdomain, or
129 both. Connections between the active site duplex and the rest of the RNA lacked cryo-EM
130 density, likely due to flexibility. The *TR*:Sp duplex formed seven canonical base pairs (bps), as
131 predicted (*9, 14*), along with three non-standard pairs (Fig. 2b). Only three of these base pairs (at
132 ^{Sp}56, 55, and 52) are essential for *c*PRT (*9, 14*). In addition, a guanine at ^{Sp}53 is required for
133 *c*PRT, and the structure revealed a specific contact to this base by bRT S302, which is highly
134 conserved in DGR RTs (*1*).

135 ^{TR}G117 occupied the template (T) site in bRT and formed hydrogen bonds to dCTP
136 which occupied the incoming dNTP (N) site (Figs. 2c and S5a). dCTP stacked coaxially against
137 the priming nucleotide ^{Sp}A56, adjacent to catalytic bRT aspartates 138, 215, and 216. Several
138 base-specific contacts were made to ^{TR}G117 and incoming dCTP by highly conserved fingertips
139 loop (F66, R74, and I76) and palm (I181, G182, and Q187) amino acids. In addition, the NTE
140 subdomain contacted the phosphate of ^{TR}G117 through the conserved amino acid K29. A cation,
141 presumably Mg^{2+}, was positioned between the α- and β-phosphates of the incoming dCTP and
142 coordinated by bRT D138 and D215 (Fig. 2c). Thus, this complex appeared poised for catalysis
143 and we termed it the 'Active' conformation (Table S1, Complex 1). Ala-substitution of R74 and
144 Asn-substitution of I181 decrease A-mutagenesis, while Ile-substitution of A78 increases A-
145 mutagenesis (*13*). These effects are consistent with non-specific stabilization of nucleotides,
146 involving either electrostatic interactions by R74 with dNTP phosphates, or hydrophobic
147 contacts by I181 and A78 to N and T site riboses, respectively.

148 The Active conformation also occurred when the N site was empty or poorly occupied
149 (Fig. S5b). This particular conformation was seen in RNPs containing wild-type RNA$^{\Delta 98}$ and
150 incubated with the nonhydrolyzable dCTP analog dCpCpp (Table S1, Complex 3). The Active,
151 N-empty conformation differed from the Active, N-occupied conformation in having an even

152 greater mobility in the two most mobile bRT subdomains, the fingertips loop and thumb, as
153 assessed by Q-scores (*20*) (Fig. S5c and Table S2). ^{TR}G117 was also more mobile with the N site
154 empty (Fig. S5c). These observations suggest that the Active conformation does not absolutely
155 depend on N site occupancy but does seem to be stabilized by the correct incoming dNTP.
156
157 **Noncoding Sp RNA**
158 The nc Sp RNA formed functionally significant structures along its entire length (Figs. 1d
159 and e). The location and flexibility of some these structures depended on the conformational
160 state of the RNP. We describe the path of the RNA in the Active conformation first and
161 differences in other conformations afterwards.
162
163 ***cPRT Stop***. The 5' end of the DGR RNA composed of the upstream sequence from *avd* (^{avd}370-
164 379) formed an 11-bp duplex with ^{Sp}59-70 (Figs. 2d and e), just beyond the priming site. ^{Sp}U63
165 was unpaired but accommodated within the coaxial stack. Immediately following in Sp was a
166 short stem-tetraloop (^{Sp}71-78) and unpaired bases (^{Sp}80-81) that contacted the *avd*:Sp duplex.
167 The tetraloop belonged to the highly abundant UNCG class (*21*). This duplex and stem-tetraloop
168 structure, together termed '*c*PRT Stop', appears to create a physical block to polymerization
169 beyond *TR*, readily explaining why *c*PRT terminates within ^{avd}383-387 (*9*).
170 Deletion of the *avd* strand (^{avd}368-379) eliminates *c*PRT (*9*). We asked whether this was
171 due to loss of DGR RNA binding. We found that wild-type and Δ^{avd}368-379 DGR RNA were
172 gel-shifted by bRT-Avd, and that the gel-shifted bands contained both bRT and Avd (Fig. S6),
173 indicating no significant alteration in protein binding strength. This result tallies with the
174 observation that the *avd*:Sp duplex is hardly contacted by bRT (Fig. 2d). We next asked whether
175 base pairing between *avd* and Sp was functionally sufficient. Substitutions that disrupted 12
176 potential base pairs between *avd* and Sp resulted in loss of *c*PRT, but reversion of the base pairs
177 through complementary substitutions failed to restore *c*PRT (Fig. S7). A similar result was seen
178 for disruption of just six Watson-Crick base pairs (Fig. S7), suggesting that the structure of the
179 *c*PRT Stop is more functionally crucial than base pairing *per se*.
180 Unlike the *avd*:Sp duplex, the adjacent stem-tetraloop was contacted by bRT through
181 numerous non-specific and base-specific interactions. Deletion of the stem-tetraloop (Δ^{Sp}71-78)
182 in DGR RNA reduced *c*PRT to 2% of wild-type (Figs. 2f and S8). Together these results

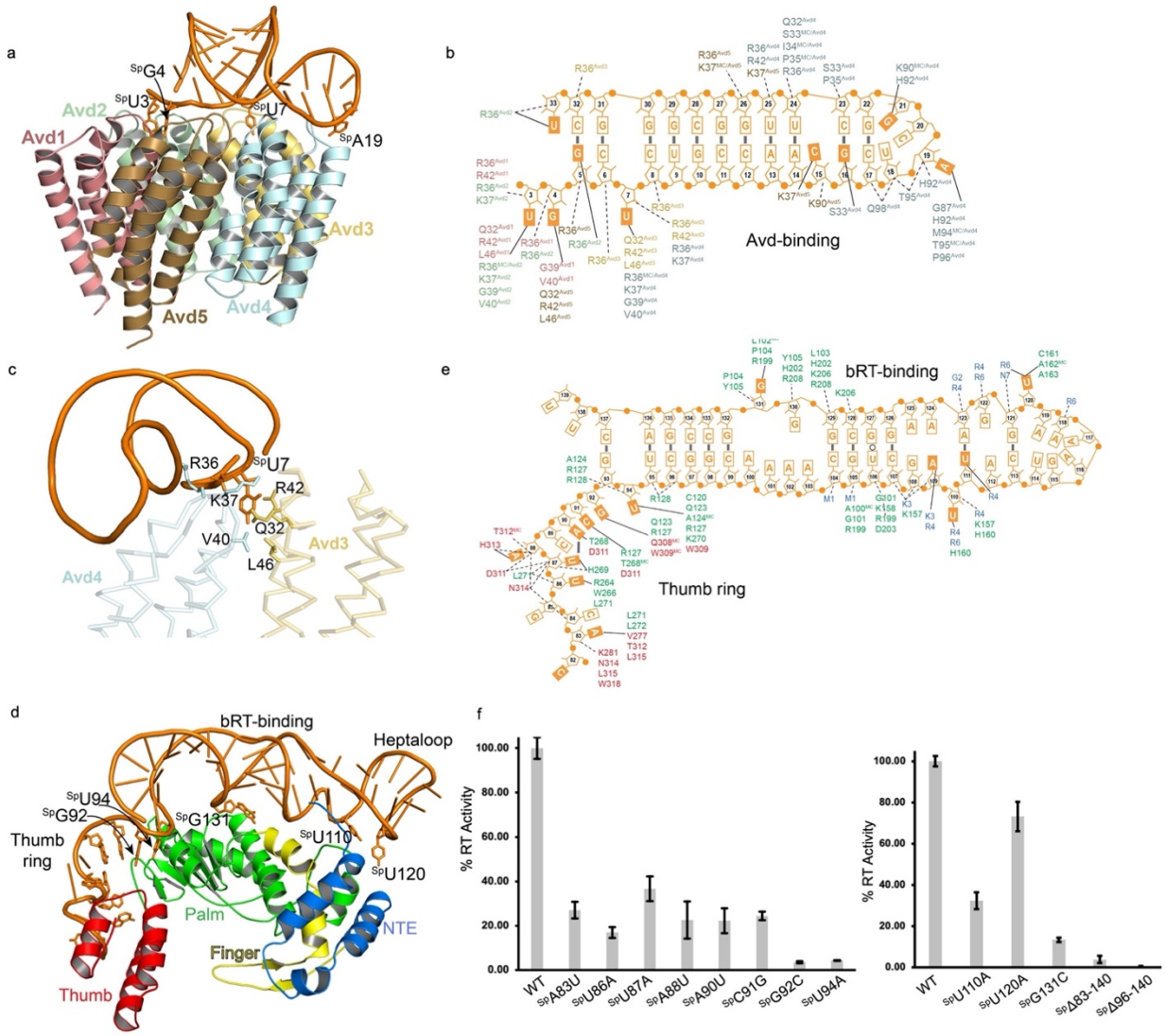

Fig. 3. Avd-binding, Thumb Ring, and bRT-binding RNA. a. Avd-binding RNA (orange) in contact with Avd protomers 1-5, in cartoon representation. Flipped-out bases shown in bonds representation. **b.** Schematic of Avd-binding RNA, consisting of Sp stem-UNCG tetraloop. Depiction is as in Fig. 2b. **c.** Interaction between flipped-out base ^{Sp}U7 (bonds representation) in Avd-binding RNA (orange, cartoon representation) and Avd protomers 3 (light yellow) and 4 (light blue), both in ribbons representation. Side chains of Avd that contact ^{Sp}U7 are in bonds representation. **d.** Interaction of Thumb ring and bRT-binding RNA elements (orange, cartoon representation) with bRT (cartoon representation coloring as in Fig. 2a). Bases in the Thumb ring that contact bRT and flipped-out bases in the bRT-binding element are shown in bonds representation. **e.** Schematic of Thumb ring and bRT-binding RNA elements. Depiction is as in Fig. 2b. **f.** *c*PRT activity of mutant DGR RNAs relative to wild-type DGR RNA. Data are from experiments that were repeated three independent times, and means and standard deviations shown.

suggested that the *avd*-Sp duplex and stem-tetraloop elements not only functioned as a *c*PRT Stop but were important for proper positioning of the DGR RNA.

186 ***Avd-binding.*** The first 30 nts of Sp (^{Sp}3-33) formed an 11-bp stem capped by a UNCG tetraloop,
187 with the stem-loop extensively engaged by all five Avd protomers (Fig. 3). Several of these
188 protomers supplied a common set of amino acids that interacted with bases that were flipped out
189 from the stem-loop (Fig. 3c). For example, a crevice between Avd1 and Avd2 used Q32, R36, and
190 K37 among others to bind the flipped-out base ^{Sp}U3 (Fig. 3b), while a crevice between Avd3 and
191 Avd4 provided the same amino acids to bind the flipped-out base ^{Sp}U7. This last base made the
192 most crucial contacts, as evidenced by ^{Sp}U7A decreasing *c*PRT in the DGR RNA to 13% of
193 wild-type (Figs. 2e and S8). This reduction was similar to one resulting from a deletion of the
194 entire Avd-binding segment (Δ^{Sp}5-32) (Figs. 2e and S8), indicating that the Avd-binding duplex
195 enhances *c*PRT efficiency but is not absolutely required. This deletion also showed that the
196 RNA-binding and catalytic functions of Avd are dissociable. Substitution of other flipped-out
197 bases decreased *c*PRT less dramatically (i.e., ^{Sp}U3A, ^{Sp}A19U, and ^{Sp}A19U/G21C), and one
198 increased *c*PRT (^{Sp}G4C) (Figs. 2e and S8).
199
200 ***Thumb Ring***. Following the *c*PRT Stop, a single-stranded portion of Sp (^{Sp}82-94) wrapped
201 around the base of the bRT thumb and was called the 'Thumb Ring' (Fig. 3d). Nine of 13
202 consecutive bases in the Thumb Ring were contacted by bRT, with ^{Sp}C92 and U94 being the
203 most significant. Substitutions at these bases, which were contacted by the conserved bRT amino
204 acid R127, reduced *c*PRT in the DGR RNA to 4% of wild-type (Figs. 3f and S8). Other bases in
205 the Thumb Ring were also contacted by well conserved bRT amino acids, and substitutions at
206 these likewise reduced *c*PRT (average of 25% of wild-type, Figs. 3f and S9).
207
208 ***bRT-binding.*** Sp continued from the Thumb Ring into a bRT-binding element composed of a
209 long stem (^{Sp}95-113 and 121-139) interrupted by two interior loops (^{Sp}100-103:130-131 and
210 ^{Sp}108-112:122-125), and capped by a heptaloop (^{Sp}114-120) (Figs. 3d and e), the most common
211 type of loop after tetraloops (*21*). The interior loops mostly had non-canonical base-base
212 interactions. Bases flipped-out from the stem and heptaloop were contacted by bRT palm and
213 NTE subdomains. Substitutions at these bases (^{Sp}U110 and G131 in stem, ^{Sp}U120 in heptaloop)
214 decreased *c*PRT, with the greatest effect at ^{Sp}G131 (Fig. 3f, 13%). Deletion of the Thumb Ring
215 along with the bRT-binding stem-loop (Δ^{Sp}83-140) or the bRT-binding stem-loop alone (Δ^{Sp}96-

216 140) eliminated *c*PRT almost completely (Fig. 3f). The latter also resulted in loss of bRT-binding
217 while Avd-binding was maintained (Fig. S6).
218
219 ***RNA Dynamics.*** To determine whether the DGR RNA was inherently ordered, selective 2'-
220 hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) (*22*)
221 was carried out in the absence and presence of bRT-Avd. In free RNA, about half of the
222 nucleotides had low reactivity (<0.3), indicative of inherent structuring (*23*) (Fig. S9a). bRT-Avd
223 had little effect on the overall RNA reactivity (0.90 correlation coefficient ± bRT-Avd). In
224 general, nucleotides with low flexibility were interspersed with those with high flexibility, with
225 some exceptions. The Avd-binding region in the free RNA was much less flexible than average,
226 and even less so when bound to bRT-Avd, especially ^{Sp}U7. In addition, the connection (^{Sp}39-46)
227 between the Avd-binding region and the active site duplex, which was not clearly visible in the
228 structure, as well as the Thumb Ring were much more flexible than average in both free and
229 bound RNA. We looked for correlation between reactivity and misincorporation frequency at *TR*
230 adenines, as the latter varies between 12-75% depending on the position (*13*), but found no clear
231 relationship (Fig. S9b), suggesting that factors other than RNA flexibility are at play.
232
233 ***Conservation of RNA elements.*** We asked whether RNA structures observed for the BΦ DGR
234 existed in other DGRs. We examined 274 DGRs from complete genomes that had predicted Avd
235 homologs (*1*), and identified 32 with equivalents of the Avd-binding stem-loop in the sequence
236 immediately downstream of *TR* (Fig. S10 and Table S3). The 32 DGRs spanned *Proteobacteria*,
237 *Firmicutes*, *Cyanobacteria*, *Nanoarchaeota*, and members of the Candidate Phylum Radiation
238 (CPR). The stems had at least one bulged nucleotide, as in the BΦ DGR, and most loops had the
239 sequence (C/U)ACG (UNCG class). We examined 29 of these DGRs, excluding three redundant
240 CPR ones, and found in all, structures similar to the *c*PRT Stop duplex formed by sequences
241 upstream and downstream of *TR* (*avd* and Sp equivalents, respectively) (Fig. S10 and Table S4).
242 These duplexes had an average of 10 bps. Additionally, equivalents of the *TR*:Sp active site
243 duplex, which sets the cDNA start site and contains mostly paired but also some mismatched
244 bases, were found in all but two of the 29 DGRs (Table S4).
245

Resting and Pre-Active Conformations

Two RNP conformations in addition to the Active conformation were visualized. In one, the bRT thumb subdomain was collapsed into the active site and occupied both N and T sites,

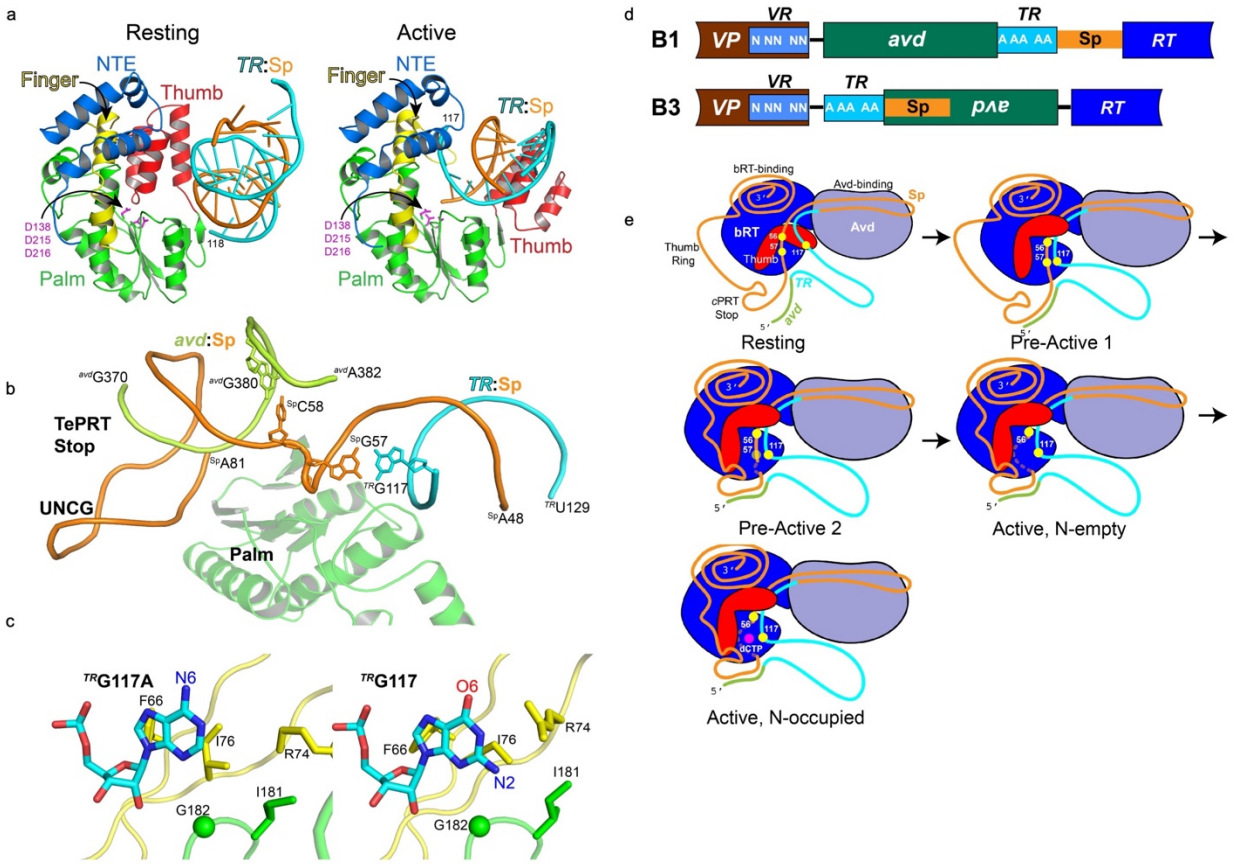

Fig. 4. Resting and Pre-active Conformations, and template adenine. a. (Left) Resting and (right) Active conformation, with bRT and active site *TR*:Sp duplex shown in cartoon representation (coloring as in Fig. 2a). The locations of catalytic D138, D215, and D216 identify the active site. The thumb subdomain was translated ~19 Å and rotated ~117° relative to the active conformation. **b.** Pre-Active 1 conformation, with bRT palm subdomain and active site *TR*:Sp duplex and *c*PRT stop elements shown in cartoon representation, and interacting bases in bonds representation. **c.** Interactions between (left) ^{TR}G117A and (right) ^{TR}G117, in bonds representation (carbons and phosphates cyan, oxygens red, and nitrogens blue) at the T site in the Active conformation. bRT main and side chains shown in cartoons and bonds representation, respectively. **d.** Schematic of DGR architectures, as in Wu et al. (*1*). *VP*, variable protein; *RT*, reverse transcriptase. In the B3 architecture, *avd* is encoded on the opposite strand. **e.** Schematic of bRT-Avd/RNA conformational states, shown as a pathway from Resting to Active. In the Resting conformation, the thumb subdomain (red) blocks the active site and the *TR*:Sp duplex is localized to the bRT surface. Other RNA elements, besides the Avd- and bRT-binding elements, are disorganized. In Pre-Active 1, the *TR*:Sp duplex is in the active site but ^{Sp}57 occupies the N site while ^{TR}G117 occupies the T site. In Pre-Active 2, the *c*PRT stop and Thumb ring are properly positioned, and ^{Sp}57 is more mobile. In the Active, N-empty conformation ^{Sp}57 is dissociated from the N site and flexible, and in the Active, N-occupied conformation dCTP occupies the N site.

249 having displaced the fingertips loop, which was mostly disordered (Figs. 4a and S11). The *TR*:Sp
250 duplex, instead of entering the active site, skirted the surface of bRT. This conformation, called
251 'Resting', was present in the population of RNPs incubated with no dNTPs (Table S1, Complex
252 8). Density for ^{TR}G117 was poor but the rest of *TR* through its junction with Sp could be
253 modeled, as could the entirety of Sp. The bRT thumb and palm were highly flexible compared to
254 the Active conformation, as was the bRT-binding RNA (Fig. S11c). The *c*PRT Stop and Thumb
255 Ring RNA were also in different locations than in the Active conformation (Fig. S11d).
256 In the second additional conformation, the *TR*:Sp duplex was within the bRT active site
257 but had ^{Sp}G57 occupying the N site and forming a base-base interaction with ^{TR}G117 (Figs. 4b
258 and S12, Table S1, Complexes 2 and 4). In the Active conformation ^{Sp}G57 was disordered as was
259 ^{Sp}C58, which instead here formed a base pair with ^{avd}G380, slightly extending the *avd*:Sp duplex
260 in the *c*PRT Stop structure (Fig. 4b). Additionally, the *c*PRT stop was rotated 19° and translated
261 30 Å away from bRT relative to the Active conformation (Figs. S12d, e), such that the UNCG
262 tetraloop no longer interacted with bRT, and the Thumb Ring was too flexible to be modeled.
263 This conformation was called 'Pre-Active 1' to distinguish it from a variant of this conformation
264 in which ^{Sp}G57 occupied the N site but the *c*PRT Stop and Thumb Ring were positioned as in the
265 Active conformation, interacting with bRT; this latter variant was called 'Pre-Active 2' (Table
266 S1, Complex 7). The Pre-Active 2 conformation appeared to be transitional to the Active
267 Conformation, as the density for ^{Sp}G57 was weaker than in Pre-Active 1 (Fig. S12f) and ^{Sp}C58
268 was too flexible to be modeled, consistent with the N site becoming free for occupancy by a
269 dNTP. Interestingly, cytosine-substitution of ^{Sp}G57 eliminates *c*PRT (*8*). This loss occurs with
270 either guanine or adenine at ^{TR}117, indicating that base pairing or lack thereof between ^{Sp}57 and
271 ^{TR}117 is not important (*8*). The occurrence of the Pre-Active conformation suggests that ^{Sp}G57 is
272 functionally required because it stabilizes the N site prior to complete organization of Sp into the
273 Active conformation.
274
275 **Template Adenine**
276 To visualize the effect of an adenine at the T site, we examined a complex containing
277 RNA$^{\Delta 98}$ in which ^{TR}G117 was substituted with adenine. An Active conformation was seen, with
278 ^{TR}G117A at the T site (Figs. 4c and S13; Table S1, Complex 5). The RNP had been incubated
279 with dCpCpp, and broken density was observed at the N site but could not be confidently

280 modeled as dCpCpp or ^{Sp}G57 (Fig. S13a). Adenine at the T site formed contacts with bRT that
281 were very similar to those formed by guanine, however adenine appeared to be more disordered
282 than guanine (Fig. S13d). A Resting conformation was also visualized for ^{TR}G117A (Table S1,
283 Complexes 6 and 9), which was not substantially different from the one described above for
284 wild-type RNA (Table S1, Complex 8), except for higher flexibility in certain regions of the
285 RNA which could not be modeled (*c*PRT stop, Thumb ring, and parts of *TR*) (Fig. S13e).
286
287 **Discussion**
288 We sought to understand how the essential DGR components bRT, Avd, and DGR RNA
289 combine for reverse transcription of *TR*, and especially why *cis*-acting sequences surrounding *TR*
290 are required for cDNA synthesis. Our work revealed a paramount role for the nc Sp RNA, which
291 had essential functions distributed across its entire length. At its 5' and 3' ends, Sp formed stem-
292 loops that acted as protein grapples for Avd and bRT, respectively, binding specifically to these
293 proteins through flipped-out bases. In between, Sp formed duplexes with *TR* to prime synthesis
294 and with *avd* to terminate synthesis (i.e., *c*PRT Stop). Structures resembling the Avd-binding
295 stem-loop and *c*PRT Stop and active site duplexes were identified in DGRs from distant taxa and
296 with varying genetic architectures (Fig. 4d) (*1*). In the B1 DGR architecture of BΦ DGR and
297 some *Proteobacteria* DGRs (Fig. 4d), a segment of *avd* is both information-carrying (encoding
298 Avd) and structure-forming (*c*PRT Stop duplex). More remarkably, in the B3 architecture of
299 CPR bacteria, the predicted Avd-binding stem-loop is encoded by the reverse complement of
300 *avd*, suggesting strict constraints on this sequence due to its dual information-carrying and
301 structure-forming roles (Fig. 4d).
302 The Avd- and bRT-binding grapples of Sp were structurally invariant across all
303 conformations visualized and provide a plausible means for initial association of the DGR RNA
304 with bRT-Avd, as observed in the Resting conformation (Fig. 4e). In this conformation, the Avd-
305 binding grapple is in place as is the bRT-binding grapple but more mobile, while other RNA
306 elements, such as the conformationally variable *c*PRT Stop and Thumb Ring, are disorganized
307 and the *TR*:Sp duplex blocked from entering the bRT active site by the thumb subdomain. The
308 grapples provide a ready explanation for the processivity endowed by the DGR RNA, enabling
309 the RNA to maintain association with bRT-Avd through multiple polymerization cycles.
310 Notably, bRT and other DGR RTs lack the processivity-promoting α-loop found in group II

311 intron RTs (Fig. S14) (*24*). The Resting conformation likely proceeds to the Pre-Active 1
312 conformation, with the thumb receding and *TR*:Sp entering the active site, but with ^{Sp}57G
313 apposed to ^{TR}G117 in the N and T sites, respectively. The Pre-Active 2 conformation then results
314 from organization of the *c*PRT Stop and Thumb Ring to contact bRT. ^{Sp}57G remains at the N site
315 in this conformation but with greater apparent mobility. The transition to the Active
316 conformation requires full displacement of ^{Sp}57G from the N site (N-empty), freeing it for an
317 incoming dNTP (N-occupied) and for chemistry to occur. Thus, reverse transcription of *TR*
318 appears to depend on a multitude of precise interactions between bRT-Avd and DGR RNA
319 structural elements. The extreme specificity thereby conferred likely serves to stringently limit
320 A-mutagenesis to the DGR RNA and curtail potentially damaging mutagenesis of host RNAs.
321 A complex with adenine at the template was nearly indistinguishable from one with
322 guanine at the template site (i.e., Active, N-empty). This indicates that A-mutagenesis is unlikely
323 to be due to major structural rearrangements. What then leads to A-mutagenesis? Our previous
324 experiments using nucleobase analogs showed that the substituent at the purine C6 position is
325 crucial (*13*). An amine at C6 (as in adenine) is functionally equivalent to having no substituent at
326 this position (i.e., purine), resulting in significant misincorporation, while replacement of the C6
327 amine with a carbonyl (i.e., hypoxanthine) leads to largely correct incorporation. Notably, the C6
328 substituent is not contacted directly by bRT (Fig. 4c). In contrast, bRT does contact the C2 amine
329 in guanine (Fig. 4c), and installation of an amine at the C2 position of adenine (2-amino adenine)
330 leads to more homogenous incorporation, albeit with a preference for wobble over Watson-Crick
331 pairing.
332 How does bRT sense the substituent at the C6 position if no atoms are in direct contact?
333 While several possibilities exist, a strong one is the dipole moment. Infidelity-promoting adenine
334 and purine have low dipole moments (μ of 2.5 and 3.7, respectively), while fidelity-promoting
335 guanine and hypoxanthine have high dipole moments (6.8 and 5.6, respectively) (*25, 26*). A high
336 dipole moment is likely to present an energetic cost at the largely hydrophobic T site, but one
337 that can be paid by hydrogen bonding to the correct incoming dNTP, which would favor the
338 Active conformation and stabilization of the catalytic site to promote chemistry. A mispaired
339 dNTP would be unable to reduce the energetic cost, and therefore most likely dissociate before
340 chemistry occurred. In contrast, a low dipole moment would incur a lower energetic cost,
341 enabling a mispaired dNTP to remain bound sufficiently long enough for misincorporation to

342 occur. Supporting this hypothesis is our observation that 6-methyladenine, which increases the
343 hydrophobicity at the C6 substituent, increases misincorporation frequency (*13*). A low dipole
344 moment would also mean that adenine is mostly unoriented in the T site, consistent with the
345 greater flexibility observed for adenine compared to guanine and explaining why even correct
346 incorporation for adenine is more catalytically inefficient than for other bases. Like guanine,
347 cytosine and uracil have high dipole moments (7.0 and 4.7, respectively) (*25, 26*), and are thus
348 likely to be ordered in the active site as well. In light of this, it appears significant that adenine
349 has only one exocyclic group while the other RNA bases have two, enabling easier evolution of
350 base-specific misincorporation. With the structure of bRT in hand, it should be possible to
351 identify the set of amino acids that lead to A-mutagenesis and differentiate bRT and other DGR
352 RTs from high-fidelity RTs.
353

354 **Materials and Methods**

355 **Protein expression and purification, RNA synthesis, and reverse transcription**

356 Coding sequences for the Avd fusion partner *Methylobacterium extorquens* formaldehyde
357 activating enzyme (Fae) was synthesized with codons optimized for *E. coli* expression
358 (Genewiz). The coding sequence, starting from the codon for Fae amino acid 4, was inserted
359 after the Avd codon for amino acid 124. bRT and Avd or Avd-fusion protein were co-expressed
360 in *E. coli* BL21 Gold (DE3) and purified as described (*9*). RNA was produced through *in vitro*
361 transcription using T7 RNA polymerase and gel-purified, and reverse transcription reactions with
362 bRT and Avd or Avd-fusion protein were carried out, as described (*9*). RNA and cDNA were
363 examined by denaturing gel electrophoresis, as described (*9*), and in some cases, cDNA was
364 subjected to short-read next-generation sequencing (Amplicon EZ, Genewiz), as described (*13*).
365

366 **SHAPE-MaP**

367 SHAPE-MaP experiments were carried out using 10 mM N-methylisatoic anhydride
368 (NMIA, Millipore-Sigma). Thirty-five µL of core DGR RNA (500 ng/µL) in water was
369 denatured at 95 °C for 3 min and then immediately placed on ice for 5 min. The RNA was
370 refolded by incubation at 37 °C for 20 min in 110 mM KCl, 70 mM HEPES, pH 7.5, 4 mM
371 $MgCl_2$ in a total volume of 270 µL. Five µL of 300 mM $(NH_4)_2SO_4$, 20 mM HEPES, pH 7.5, 5
372 mM $MgCl_2$, 2 mM DTT, and 10% glycerol, either containing bRT-Avd (2 mg/mL) or not, was

373 added to 40 µL of refolded RNA, whereupon the sample was incubated at 37 °C for 1 h. Five µL
374 NMIA (100 mM in DMSO) was then added to the samples, which were further incubated at
375 37 °C for 1 h. After this time, NMIA was quenched by the addition of DTT to a final
376 concentration of 50 mM. A negative control for the reaction was carried out similarly, but with 5
377 µL of DMSO instead of NMIA. Reactions were then incubated with 2 µL of proteinase K
378 (Ambion, 0.4 mg/ml) at 50 °C for 30 min, and proteinase K was then quenched with a final
379 concentration of 1 mM PMSF. The resulting samples were purified with a desalting column (G-
380 25).
381 A denaturing control was carried out by incubating 2 µL of RNA (500 ng/µL), 10 µL
382 formamide, 2 µL of 10x denaturing buffer (500 mM HEPES, pH 8.0, and 40 mM EDTA) in a
383 total volume of 18 µL. The sample was incubated at 95 °C for 1 min, and 2 µL of NMIA (100
384 mM) was added and further incubated at 95 °C for 1 min. The sample was placed on ice for 5
385 min, and then purified with a desalting column (G-25) after adjusting the volume to 50 µL.
386 Reverse transcription reactions were carried out by first annealing 8 µL of purified RNA
387 from the previous step with 1 µL of reverse transcription primer P1 (2 µM, GACTGGAGTTC
388 AGACGTGTGCTCTTCCGATCTGAAGTCGGCCCCGCCTTT) and 3 µL water at 65 °C for 5
389 min and cooling on ice for 5 min. To these annealed reactions, 8 µL of 2.5x MaP buffer (50 mM
390 Tris, pH 8.0, 75 mM KCl, 6 mM MnCl2, 10 mM DTT, and 0.5 mM dNTPs) and 1 µL of
391 ProtoScript® II Reverse Transcriptase (NEB) was added. These reactions were incubated at 42
392 °C for 3 h and then at 70 °C for 15 min. These samples were then purified with a desalting
393 column (G-25) after adjusting the volume to 50 µL.
394 The cDNAs resulting from the reverse transcription reactions were amplified in a 100 µL
395 PCR reaction using 10 µL of the purified sample from the previous step, 50 µL of Q5® High-
396 Fidelity 2X Master Mix (NEB), 5 µL each of 10 µM primer P1 and P2
397 (ACACTCTTTCCCTACACGACGCTCTTCCGATCTAAGGGCA GGCTGGGAAAT). The
398 PCR conditions used for amplification were: 98 °C for 30 s; 30 cycles of 98 °C for 10 s, 60 °C
399 for 30 s, and 72 °C for 20s; followed by 72 °C for 2 min. The amplified DNA was gel purified
400 and the samples were deep-sequenced in multiplexes using the Illumina MiSeq platform (2 × 250
401 bp, paired-end, Genewiz). Demultiplexed fastq files were processed using ShapeMapper 2.0
402 pipeline (*27*).
403

404 **Cryo-EM sample preparation and data collection**

405 For assembly of complexes, 150 µL of $RNA^{\Delta 98}$ (14 µM) was incubated in 1.2 mL of 75
406 mM KCl, 3 mM $MgCl_2$, 10 mM DTT, 50 mM HEPES, pH 7.5, and 10 % glycerol at 37 °C for
407 15 min. bRT-Avd-Fae (150 µL of 14 µM) was added and the mixture was further incubated at
408 37 °C for 2 h. The concentration of the complex was determined using A_{280} measured with a
409 NanoDrop (Thermo Scientific) and a calculated molar extinction coefficient of 206,960
410 $M^{-1}cm^{-1}$. The sample was passed through a 0.22 µm filter and concentrated to 100 µL using a 30-
411 kDa molecular weight cutoff membrane. The complex was resolved by gel filtration at 4 °C with
412 a Superdex 200 10/300 column (GE Healthcare) equilibrated with 75 mM KCl, 30 mM
413 $(NH_4)_2SO_4$, 50 mM Tris, pH 7.5, 3 mM $MgCl_2$, and 2 mM DTT. Eluted fractions were analyzed
414 by SDS-PAGE to visualize proteins, and denaturing gel electrophoresis (8 M urea containing 8
415 % polyacrylamide gels in 1× TBE buffer) to visualize RNA. Fractions containing bRT-Avd-
416 Fae/$RNA^{\Delta 98}$ complexes were concentrated to ~2 µM using a 30-kDa molecular weight cutoff
417 membrane. The concentration of bRT-Avd-Fae/$RNA^{\Delta 98}$ was determined by measuring
418 absorbance at A_{260} with a NanoDrop (Thermo Scientific), assuming a 1:5:1 bRT:Avd-
419 Fae:$RNA^{\Delta 98}$ stoichiometry. The concentration of the complex was adjusted to 1 µM and left
420 either unincubated or incubated with the following dNTP analogs or dNTPs at 37 °C for 2 h. For
421 complexes containing wild-type $RNA^{\Delta 98}$, 100 µM of non-hydrolyzable dCTP (2'-deoxycytidine-
422 5'-[(α,β)-methyleno]triphosphate, Jena Bioscience) was used. For complexes containing *TR*
423 G117A $RNA^{\Delta 98}$, 2 mM non-hydrolyzable dCTP (as above) or 2 mM non-hydrolyzable dTTP (2'-
424 deoxythymidine-5'-[(α,β)-methyleno]triphosphate, Jena Bioscience) was used. For complexes
425 containing *TR* U114A $RNA^{\Delta 98}$, 2 mM each of dCTP, dATP, and ddGTP was used. Complexes
426 (4 µL) were then applied manually 2-3x in quick succession to Quantifoil™ R 1.2/1.3 Au grids,
427 which had been freshly plasma cleaned in Gatan Solarus II (under Ar/O_2 mixture) for 10 sec.
428 After application, grids were blotted with varying force and times using a Vitrobot Mark IV
429 (Thermo Fisher) at 4 °C and 100% humidity, and then immediately plunge-frozen in liquid
430 ethane.

431 Cryo-EM data were collected on either a 300 keV Titan Krios cryo-transmission electron
432 microscope at the Pacific Northwest Center for Cryo-EM (PNCC) equipped with GIF quantum
433 energy filters attached to either a K3 electron counting direct detector (Gatan) using Serial EM
434 software, or a 300 keV Titan Krios cryo-transmission electron microscopes (FEI Company) at

435 the University of California San Diego Cryo-EM facility equipped with GIF quantum energy
436 filters attached to a K2 electron counting direct detector (Gatan) using EPU (Thermo Fisher).
437 Images were acquired at a magnification of ~130,000, corresponding to a calibrated raw pixel
438 size of ~0.5 Å, in EFTEM mode with an energy filter slit width of 10-20 eV. Movie image stacks
439 with a total electron dose of ~55 $e^-/Å^2$ were saved in non-super resolution counting mode.
440
441 **Cryo-EM data processing and analysis**
442 The collection and processing of raw data is outlined in detail in Fig. S1a and Table S1.
443 Briefly, raw movies of each independent dataset were imported in cryoSPARC 3.1.1 (*28*),
444 RELION 3.0, or both (*29*); motion-corrected using patch motion correction (multi) of
445 cryoSPARC or MotionCor2; and CTF estimated using patch CTF estimation (multi) in
446 cryoSPARC, or GCTF or CTFFIND4 in RELION. A list of micrographs for each dataset
447 (removing poor quality micrographs using various criteria e.g., CTF fit of <8 Å, astigmatism
448 <1200, motion distribution <60, and motion curvature <40) was prepared. Data were collected at
449 a relative ice thickness of ~1-1.2 μm, as thin ice led to particle denaturation whereas thick ice led
450 to high background noise. Particles were picked from curated micrographs using various
451 methods and iterative strategies outlined in Fig. S1a. Duplicate particles were removed in
452 cryoSPARC, and binned or un-binned particles were extracted from micrographs depending on
453 the purpose and stage of analysis. Multiple rounds of 2D classification were performed to
454 eliminate poor quality particles that could not be classified. Selected particles from 2D classes
455 were used for initial 3D classification (with different number of classes) using *ab initio*
456 reconstruction. Particles belonging to poorly resolved *ab initio* 3D class averages were discarded.
457 Particles from 3D classes which indicated the presence of all domains were used for
458 heterogeneous classification with various plausible models. Good classes from heterogeneous
459 refinements were fed back into refined template picking using programs outlined in Fig. S1.
460 Iteration of the above processes provided a refined set of particles, which were classified into
461 probable 3D classes (judged from shape profiles of 3D models of heterogenous classification),
462 and eventually particles from final 3D class sets were extracted without binning with a 320-pixel
463 box size for consecutive refinement using homogeneous, non-uniform, and local refinement
464 algorithms in cryoSPARC with appropriate masks. Per-particle defocus refinement was also
465 performed, although resulting in little to no discernible improvement in map quality. Trials of

466 particle subtraction did not indicate any improvement of maps and thus this step was omitted
467 from refinement. The total number of movies, number of particles used for final refinement of
468 each class, orientation distribution, half-map-based correlation, global resolution estimate, and
469 several other parameters are provided in Fig. S1 and Table S1. Density sharpened versions of the
470 maps were used for model building.
471
472 **Model building and refinement**
473 The model building pipeline is depicted in detail in Fig. S1b. In brief, an initial model of
474 bRT was built manually using PDB templates of poliovirus RNA-dependent RNA polymerase
475 (3OL9) and group II intron RT GsI-IIC (6AR1), and the initial model of Avd was based on the
476 structure of the free form of the pentamer (PDB 4DWL). For modeling RNA, the secondary
477 structure for DGR RNA was first predicted through iterative rationalization from initial map
478 features and finalized using RNAstructure (*30*) incorporating refined SHAPE-MaP data and
479 biochemical results (e.g., base pairing in the *TR*:Sp active site duplex) (*9, 14*). The dot bracket
480 files defining plausible RNA stem regions were used with simRNA (*31*) and RNAComposer (*32*)
481 to obtain initial 3D templates of RNA stems, which were docked in the sharpened maps. Rigid
482 body fitted models of bRT subdomains, Avd protomers, RNA stems and the connecting single
483 stranded RNA regions were iteratively (re)built in real space in Coot (*33*). In addition, a
484 pentameric model of Fae (PDB 1Y60) was placed into the diffuse density contiguous to the
485 density for the Avd pentamer. The models of complexes composed of bRT, Avd, and $RNA^{\Delta 98}$
486 were finally refined in CCPEM (*34*). Model geometry was validated using MolProbity package
487 within the Phenix suite (*35*). This final model displays good stereochemistry (Table S1), except
488 for regions of high flexibility. RNA structure was analyzed using RNAPdbee (*36*). Q-scores (*37*)
489 were computed using Chimera (*38*). Derived B-factors were calculated as 150 • (1 – Q), with the
490 scale factor of 150 empirically identified to provide the highest map correlation coefficient, as
491 calculated with Phenix (*35*). Figures of molecular models were made with PyMOL (The PyMOL
492 Molecular Graphics System, Schrödinger, LLC).
493
494 **EMSA and Mass Spectrometry**
495 Reverse transcription was carried out at 37 °C for 20 min in 20 µL volume containing 1.4
496 µM RNA, varying concentrations of bRT-Avd (0-2.0 µM), and 20 units RNase inhibitor (NEB)

497 in 75 mM KCl, 3 mM $MgCl_2$, 10 mM DTT, 50 mM HEPES, pH 7.5, 10% glycerol. The reaction
498 contents were centrifuged at 12,000 x *g* for 10 min, and 15 µL was loaded on a native 6%
499 polyacrylamide gel and resolved for 90 min at 100 V. The gel was stained with GreenGlo™ Safe
500 DNA Dye (Thomas Scientific) and imaged on a Bio-Rad Chemidoc™ system. Gel-shifted bands
501 (Fig. S5, black arrow) were extracted from the gel, and subjected to mass spectrometry on an
502 Orbitrap Fusion Lumos MS at the UCSD Biomolecular and Proteomics Mass Spectrometry
503 facility. Mass spectrometry data were analyzed using Peaks Studio 10.6 software.
504
505 **Conservation of RNA elements**
506 The subset of DGR sequences from Wu et al. (*1*) that contain an Avd-like gene and a
507 non-coding gap between *TR* and RT were examined for a structure similar to the BΦ DGR Avd-
508 binding stem-loop. RNA folds were predicted with Vienna RNAFold tool v. 2.4.18 (*39*) with
509 default parameters. Structural predictions were parsed and inspected using a custom python
510 script, resulting in 32 DGRs with putative Avd-binding stem-loops. Twenty-nine of these DGRs,
511 excluding three near-redundant CPR sequences from the 32, were examined for structures
512 similar to *avd*:Sp duplex in the BΦ DGR *c*PRT Stop. Potential base pairing was examined
513 between 20 nt upstream of *TR* (in some cases 40 nt, when only a few base pairs were found at the
514 5' end of the segment) and 100 nt downstream of *TR* (i.e., non-coding gap between *TR* and RT).
515 For this, the *TR* sequence was deleted and upstream and downstream sequences were
516 concatenated into 120 or 140 nt segments, and potential base pairing in these segments was
517 predicted using Mfold (*40*). To identify the putative active site duplex, a search for
518 complementary sites was conducted in the 29 DGRs that was predicted to have a *c*PRT stop
519 duplex. Two 30 bp windows were extracted from each DGR RNA sequence, consisting of the 3'-
520 end of *TR* and the start of the putative Sp (15 bp + 15 bp), and the reverse complement of a 30 bp
521 region immediately upstream of the putative *c*PRT stop duplex site in Sp. The two 30bp
522 windows were then locally aligned and assessed for potential duplex pairing, as above.
523

Acknowledgments

Use of Pacific Northwest Center for Cryo-EM (PNCC) supported by NIH grant U24GM129547. We thank R. Baker and A. Leschziner for initial cryo-EM characterization of complexes, T. Baker for computational resources, and J. Meyers for data collection at PNCC.

Funding

This work was supported by the National Institutes of Health R01 GM132720 (to P.G.). BGP was supported by the Gordon and Betty Moore Foundation, the G. Unger Vetlesen Foundation, and the Owens Family Foundation.

Conflict of Interest.

No conflicts to declare.

Author Contributions.

Conceptualization: SH and PG
Investigation: SH, TB, JC, BGP, PG
Visualization: SH, TB, BGP, PG
Funding acquisition: PG
Project Administration: PG
Supervision: PG
Writing – original draft: SH, TB, BGP, PG
Writing – review & editing: SH, TB, BGP, PG

Supplementary Information

Supplementary Figures S1-S14.
Supplementary Tables S1-S4.

References

1. L. Wu *et al.*, Diversity-generating retroelements: natural variation, classification and evolution inferred from a large-scale genomic survey. *Nucleic Acids Res* **46**, 11-24 (2018).
2. B. R. Macadangdang, S. K. Makanani, J. F. Miller, Accelerated Evolution by Diversity-Generating Retroelements. *Annu Rev Microbiol* **76**, 389-411 (2022).

557 3. S. Roux *et al.*, Ecology and molecular targets of hypermutation in the global microbiome.
558 *Nat Commun* **12**, 3076 (2021).
559 4. S. Nayfach *et al.*, Metagenomic compendium of 189,680 DNA viruses from the human
560 gut microbiome. *Nat Microbiol* **6**, 960-970 (2021).
561 5. B. G. Paul *et al.*, Retroelement-guided protein diversification abounds in vast lineages of
562 Bacteria and Archaea. *Nat Microbiol* **2**, 17045 (2017).
563 6. C. T. Brown *et al.*, Unusual biology across a group comprising more than 15% of domain
564 Bacteria. *Nature* **523**, 208-211 (2015).
565 7. M. Liu *et al.*, Reverse transcriptase-mediated tropism switching in Bordetella
566 bacteriophage. *Science* **295**, 2091-2094 (2002).
567 8. S. A. McMahon *et al.*, The C-type lectin fold as an evolutionary solution for massive
568 sequence variation. *Nat Struct Mol Biol* **12**, 886-892 (2005).
569 9. S. Handa *et al.*, Template-assisted synthesis of adenine-mutagenized cDNA by a
570 retroelement protein complex. *Nucleic Acids Res* **46**, 9711-9725 (2018).
571 10. S. Doulatov *et al.*, Tropism switching in Bordetella bacteriophage defines a family of
572 diversity-generating retroelements. *Nature* **431**, 476-481 (2004).
573 11. B. G. Paul *et al.*, Targeted diversity generation by intraterrestrial archaea and archaeal
574 viruses. *Nat Commun* **6**, 6585 (2015).
575 12. M. Alayyoubi *et al.*, Structure of the essential diversity-generating retroelement protein
576 bAvd and its functionally important interaction with reverse transcriptase. *Structure* **21**,
577 266-276 (2013).
578 13. S. Handa, A. Reyna, T. Wiryaman, P. Ghosh, Determinants of adenine-mutagenesis in
579 diversity-generating retroelements. *Nucleic Acids Res* **49**, 1033-1045 (2021).
580 14. S. S. Naorem *et al.*, DGR mutagenic transposition occurs via hypermutagenic reverse
581 transcription primed by nicked template RNA. *Proc Natl Acad Sci U S A* **114**, E10187-
582 E10195 (2017).
583 15. H. Guo *et al.*, Diversity-generating retroelement homing regenerates target sequences for
584 repeated rounds of codon rewriting and protein diversification. *Mol Cell* **31**, 813-823
585 (2008).
586 16. C. Zhao, A. M. Pyle, Crystal structures of a group II intron maturase reveal a missing link
587 in spliceosome evolution. *Nat Struct Mol Biol* **23**, 558-565 (2016).
588 17. J. L. Stamos, A. M. Lentzsch, A. M. Lambowitz, Structure of a Thermostable Group II
589 Intron Reverse Transcriptase with Template-Primer and Its Functional and Evolutionary
590 Implications. *Mol Cell* **68**, 926-939 e924 (2017).
591 18. D. B. Haack *et al.*, Cryo-EM Structures of a Group II Intron Reverse Splicing into DNA.
592 *Cell* **178**, 612-623 e612 (2019).
593 19. F. J. Blocker *et al.*, Domain structure and three-dimensional model of a group II intron-
594 encoded reverse transcriptase. *RNA* **11**, 14-28 (2005).
595 20. G. Pintilie, W. Chiu, Validation, analysis and annotation of cryo-EM structures. *Acta*
596 *Crystallogr D Struct Biol* **77**, 1142-1152 (2021).
597 21. P. Danaee *et al.*, bpRNA: large-scale automated annotation and analysis of RNA
598 secondary structure. *Nucleic Acids Res* **46**, 5381-5394 (2018).
599 22. M. J. Smola, G. M. Rice, S. Busan, N. A. Siegfried, K. M. Weeks, Selective 2'-hydroxyl
600 acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for
601 direct, versatile and accurate RNA structure analysis. *Nat Protoc* **10**, 1643-1669 (2015).

23. K. E. Deigan, T. W. Li, D. H. Mathews, K. M. Weeks, Accurate SHAPE-directed RNA structure determination. *Proc Natl Acad Sci U S A* **106**, 97-102 (2009).
24. C. Zhao, F. Liu, A. M. Pyle, An ultraprocessive, accurate reverse transcriptase encoded by a metazoan group II intron. *RNA* **24**, 183-195 (2018).
25. J. Sponer, J. Leszczynski, P. Hobza, Electronic properties, hydrogen bonding, stacking, and cation binding of DNA and RNA bases. *Biopolymers* **61**, 3-31 (2001).
26. P. Shih, L. G. Pedersen, P. R. Gibbs, R. Wolfenden, Hydrophobicities of the nucleic acid bases: distribution coefficients from water to cyclohexane. *J Mol Biol* **280**, 421-430 (1998).
27. S. Busan, K. M. Weeks, Accurate detection of chemical modifications in RNA by mutational profiling (MaP) with ShapeMapper 2. *RNA* **24**, 143-148 (2018).
28. 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).
29. S. H. Scheres, RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J Struct Biol* **180**, 519-530 (2012).
30. J. S. Reuter, D. H. Mathews, RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* **11**, 129 (2010).
31. 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).
32. M. Popenda *et al.*, Automated 3D structure composition for large RNAs. *Nucleic Acids Res* **40**, e112 (2012).
33. P. Emsley, B. Lohkamp, W. G. Scott, K. Cowtan, Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* **66**, 486-501 (2010).
34. 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).
35. 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).
36. T. Zok *et al.*, RNApdbee 2.0: multifunctional tool for RNA structure annotation. *Nucleic Acids Res* **46**, W30-W35 (2018).
37. G. Pintilie *et al.*, Measurement of atom resolvability in cryo-EM maps with Q-scores. *Nat Methods* **17**, 328-334 (2020).
38. E. F. Pettersen *et al.*, UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605-1612 (2004).
39. A. R. Gruber, R. Lorenz, S. H. Bernhart, R. Neubock, I. L. Hofacker, The Vienna RNA websuite. *Nucleic Acids Res* **36**, W70-74 (2008).
40. M. Zuker, Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31**, 3406-3415 (2003).