Research Article

Biochemical properties of bacterial reverse transcriptase-related (*rvt*) gene products: multimerization, protein priming, and nucleotide preference

Irina A. Yushenova and Irina R. Arkhipova

Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, MA, 02543, USA

Running title: *Properties of a bacterial reverse transcriptase*

Correspondence to: Dr. Irina Arkhipova, Marine Biological Laboratory, 7 MBL St., Woods Hole, MA 02543, USA. Tel. +1 508 289 7120. Email: iarkhipova@mbl.edu. ORCID: 0000-0002-4805-1339

Dr. Irina Yushenova, Marine Biological Laboratory, 7 MBL St., Woods Hole, MA 02543, USA. Email: iyushenova@mbl.edu. ORCID: 0000-0001-6291-6215

Abstract

Cellular reverse transcriptase-related (*rvt*) genes represent a novel class of reverse transcriptases (RTs), which are only distantly related to RTs of retrotransposons and retroviruses, but, similarly to telomerase RTs, are immobilized in the genome as single-copy genes. They have been preserved by natural selection throughout the evolutionary history of large taxonomic groups, including most fungi, a few plants and invertebrates, and even certain bacteria, being the only RTs present across different domains of life. Bacterial *rvt* genes are exceptionally rare but phylogenetically related, consistent with common origin of bacterial *rvt* genes rather than eukaryote-to-bacteria transfer. To investigate biochemical properties of bacterial RVTs, we conducted *in vitro* studies of recombinant HaRVT protein from the filamentous gliding bacterium *Herpetosiphon aurantiacus* (Chloroflexi). Although HaRVT does not utilize externally added standard primer-template combinations, in the presence of divalent manganese it can polymerize very short products, using dNTPs rather than NTPs, with a strong preference for dCTP incorporation. Further, we investigated the highly conserved N- and C-terminal domains, which distinguish RVT proteins from other RTs. We show that the N-terminal coiled-coil motif, which is present in nearly all RVTs, is responsible for the ability of HaRVT to multimerize in solution, forming up to octamers. The C-terminal domain may be capable of protein priming, which is abolished by site-directed mutagenesis of the catalytic aspartate and greatly reduced in the absence of the conserved tyrosine residues near the C-terminus. The unusual biochemical properties displayed by RVT *in vitro* will provide the basis for understanding its biological function *in vivo*.

Keywords: oligomer; reverse transcription; RNA-dependent DNA polymerase; coiled-coil; multimerization

Abbreviations: RT, reverse transcriptase; RVT, reverse transcriptase related; aa, amino acid; TERT, telomerase reverse transcriptase; RdRP, RNA-dependent RNA polymerase; TP, terminal protein; HBV, hepatitis B virus; IPTG, Isopropyl β-D-1-thiogalactopyranoside; IMAC, immobilized metal affinity chromatography; 6xHis, hexahistidine; MW, molecular weight; PCR, polymerase chain reaction; pI, isoelectric point; SDS-PAGE, sodium dodecyl sulfate polyacrylamide gel electrophoresis.

Introduction

Reverse transcription, i.e. polymerization of DNA on RNA templates, is a process characteristic of retrotransposons and reverse-transcribing viruses, which is performed by a group of enzymes called reverse transcriptases (RTs) (Eickbush and Jamburuthugoda 2008; Garfinkel, et al. 2016; Menendez-Arias, et al. 2017). It is generally believed that RTs have no specific function in host cells, since all of them originate from "selfish" mobile genetic elements, with the notable exception of telomerase reverse transcriptases (TERTs). TERTs can add short G-rich repetitive sequences to telomeres, thereby protecting the ends of linear chromosomes in eukaryotic cells (see (Lue and Autexier 2006; Arkhipova 2012) for review). They are usually encoded by single-copy genes, and do not increase in copy numbers, as do mobile elements. In 2011, we described another group of non-mobile cellular RTs (Gladyshev and Arkhipova 2011), named reverse transcriptase related (*rvt*) genes. While *rvt* genes were initially found in bdelloid rotifers *Adineta vaga* and *Philodina roseola* (phylum Rotifera, class Bdelloidea), a search of genome databases revealed that *rvt* genes are widespread and can be found in all kingdoms of life, albeit with patchy distribution. Finding the same type of RT in both prokaryotes and eukaryotes is so far unprecedented, and implies a function that should be applicable to both. In each host species, *rvt* is present either as a single-copy gene or as a gene family consisting of two or three members (Gladyshev and Arkhipova 2011). The overall structure of *rvt*-encoded proteins is highly conserved and significantly deviates from other RT types, especially in the N- and C-terminal regions, with the coiled-coil at the N-termini being the only recognizable motif other than RT (Gladyshev and Arkhipova 2011).

The *rvt* genes evolve under purifying selection, and were shown to contain spliceosomal introns in many eukaryotic lineages. Most are transcriptionally active; however, they may also undergo decay, pseudogenization, and loss. The purified NcRVT protein from the model filamentous fungus *Neurospora crassa* (phylum Ascomycota) is enzymatically active, and in the presence of Mn^{2+} can polymerize both dNTPs and NTPs with strong preference for the latter, exhibiting terminal transferase activity (Gladyshev and Arkhipova 2011). However, NcRVT was unable to perform canonical RNA-dependent DNA polymerization with standard exogenous template-primer combinations, as might be

expected from its sequence similarity to other RTs. Therefore, it was of interest to examine a representative from another kingdom, so that their properties could be compared directly.

Here, we investigate biochemical properties of the recombinant RVT protein of bacterial origin and compare it with its fungal counterpart, with special emphasis on the N- and C-terminal domains that distinguish RVTs from all other RTs. We find that it forms multimers in solution, and demonstrate that the N-terminal domain is directly responsible for its multimerization. We further show that RVT may employ protein priming to initiate polymerization, adding to the list of unusual properties of these enigmatic enzymes. Finally, we determine the nucleotide specificity of the bacterial enzyme, which displays a strong preference for dCTP incorporation, while showing little or no affinity for NTPs as opposed to dNTPs.

Materials and methods

Bacterial strains and growth conditions

Herpetosiphon aurantiacus DSM 785 (strain: ATCC 23779) was obtained from ATCC. *H. aurantiacus* was grown on CY-agar medium at 25ºC for 3-5 days. *Escherichia coli* Rosetta 2 (DE3) strain (Novagen) was used for expression of recombinant proteins. In all other cases, *E. coli* NEB5α (NEB) and Zymo5α (Zymo Research) chemically competent cells were used for transformation and plasmid propagation.

DNA manipulations

Genomic DNA from *H. aurantiacus* was extracted using UltraClean® Microbial DNA Isolation Kit (MoBio Labs). The *rvt* gene amplified by PCR was cloned into pET45b vector (Novagen) using BstBI and XhoI sites. The obtained construct carrying the N-terminal 6xHis tag (pEAG87, encoding HaRVT-N) was used to create a construct without the tag, followed by PCR-based ligation to introduce a C-terminal His-tag, yielding pEAG93 (HaRVT) (Table 1). All subsequent HaRVT mutants were obtained using GenEdit™ site-directed DNA mutagenesis kit (First Biotech). The resulting plasmids were verified by Sanger sequencing on the ABI3730XL at the W. M. Keck Ecological and Evolutionary Genetics Facility at the Marine Biological Laboratory. Secondary

structure predictions were done by Phyre2 (Kelley and Sternberg 2009) in the intensive modelling mode. All recombinant HaRVT protein variants obtained in this study are listed in Table 1.

Protein expression and visualization

NcRVT protein was obtained as described in (Gladyshev and Arkhipova 2011). Recombinant HaRVT versions were expressed in *E. coli* Rosetta 2 (DE3) in LB medium, Miller formulation (Amresco) supplied with 100 µg/ml ampicillin (Sigma), 34 µg/ml chloramphenicol (Acros Organic) and 660 mM D-sorbitol (Amresco). First, cells were grown at 37ºC, 200 rpm until OD=0.6. After that, expression of recombinant proteins was induced by supplying the growth medium with IPTG to 500 µM (Gold Bio), and the culture was grown for additional 4 h at 31ºC, 250 rpm. Bacterial cells were pelleted by centrifugation at 4ºC, 4000 g for 30 min and stored at -80ºC. Induction of recombinant proteins was confirmed by SDS-PAGE followed by Western blot hybridization.

SDS-PAGE was carried out in Mini-PROTEAN Tetra cell system (Bio-Rad) with Tris-Glycine buffer. Polyacrylamide 8-10% gels for protein separation were cast manually from custom-made solutions. After electrophoresis, proteins were visualized using Coomassie-based InstantBlue Protein Stain (Expedeon) or GelCode™ Blue Stain Reagent (Thermo Scientific), and gels were scanned either on Epson Perfection V500 scanner or using Amersham Imager 600 system (GE Healthcare). Western blotting was performed to PVDF membranes Amersham Hybond P (GE Healthcare) using wet transfer method. Membranes were incubated with the primary His-tag specific antibody (Aviva Systems Biology, OAEA00010, RRID:AB_10874637) at 1:5000 dilution and the secondary goat anti-mouse IgG-HRP (Santa Cruz Biotechnology, sc-2005, RRID:AB_631736) antibody at 1:10,000 dilution. His-tagged proteins were detected with SuperSignal West Dura Extended Duration Substrate (Thermo Scientific) using either blue sensitive X-ray film Amersham Hyperfilm ECL (GE Healthcare), or the Amersham Imager 600 chemiluminescence imager (GE Healthcare).

Protein purification

Lysates were prepared in lysis buffer (50 mM phosphate buffer, 300 mM NaCl, pH 7.0) for metal-affinity chromatography, or in buffer A (50 mM Tris-HCl, 100 mM NaCl, 1:1000 2-Mercaptoethanol, pH 7.5) for sucrose gradient separation; both were supplemented with Roche

cOmplete™ EDTA-free Protease Inhibitor Cocktail (Sigma) according to manufacturer's instructions. Frozen cell pellets containing induced protein were thawed in the respective buffer and sonicated on ice (10 s followed by 30-s pause for 3 times). Soluble proteins were separated from insoluble debris by centrifugation at 4ºC, 4000 g for 30 min.

Recombinant proteins were purified using Co^{2+}-based immobilized metal affinity chromatography (IMAC). We used HisTALON™ Gravity Column Purification Kit (Clontech) for HaRVT and TALON® Single Step Columns (Clontech) for other HaRVT mutant versions, following the manufacturers' protocols. Roche cOmplete™ EDTA-free Protease Inhibitor Cocktail (Sigma) was added to all buffers. Purified proteins were stored at 4ºC.

Analysis of oligomeric states

The bacterial lysate (2 ml) containing soluble proteins was loaded on top of a sucrose gradient made from 17 ml of 40% sucrose and 20 ml of 15% sucrose, with both stock solutions prepared in buffer A (50 mM Tris-HCl, 100 mM NaCl, pH 7.5) supplied with 1:1000 2-Mercaptoethanol. Gradients were centrifuged in a SW 32 Ti swinging bucket rotor (Beckman) for 24 h at 4ºC, 30,000 rpm. 1-ml fractions were collected with a peristaltic pump (Bio-Rad), starting from the bottom of each tube. Fractions were analyzed by SDS-PAGE followed by staining with GelCode™ Blue Stain Reagent (Thermo Scientific), and/or Western blotting with the above His-tag specific antibody. Thyroglobulin from bovine thyroid (Sigma), alcohol dehydrogenase from *Saccharomyces cerevisiae* (Sigma), catalase from bovine liver (Sigma), and bovine serum albumin (NEB) dissolved or diluted in buffer A were used as references to create a standard curve of protein size vs. the number of protein-containing fractions. We used R linear model function to estimate the size of HaRVT protein complexes. Coiled coil formation was predicted by PCOILS in the MPI toolkit (Alva, et al. 2016) using MTIDK matrix with weighting.

In vitro activity assays

For activity assays, proteins, either as purified HaRVT (0.5-3 µg) or as sucrose gradient fractions containing RVT variants (50-100 µg total protein), were mixed in a 10-µl reaction with 1 mCi/ml [α-^{32}P]dCTP (PerkinElmer) in buffer A (50 mM Tris-HCl, 100 mM NaCl, pH 7.5) supplied with 3 mM

$MnCl_2$, and incubated as a master-mix for 1 h at 25ºC. Aliquots were then placed in tubes with 1 µl of respective chases, i.e. NTPs/dNTPs 10 mM each nucleotide, or nuclease-free water, and incubated for 1 h at 25ºC. Nucleic acids were selectively removed, as needed, by treatment with 2 U of DNase I (NEB) or 10 U of RNase ONE™ Ribonuclease (Promega) at 37ºC for 30 min, or with 0.3 M NaOH at 65ºC for 20 min. After incubation, all reactions were supplied with 0.6 units of Proteinase K (Thermo Scientific) and incubated at 53ºC for 2 h or overnight. Loading buffer (95% formamide, 5 mM EDTA, 0.025% Bromophenol blue, 0.025% Xylene cyanol FF) was added to each reaction, followed by denaturation at 95ºC for 3 min. Reactions were separated on 12% or 20% denaturing PAGE in 1x TBE buffer (89 mM Tris-borate, 2 mM EDTA, pH 8.3). The gel percentage in each case was chosen based on the expected product size. Radioactive signals were detected using a storage phosphor imaging system Storm 860 (Amersham) or Typhoon FLA 7000 (GE Healthcare). A linear relationship between signal and image intensity was used in all cases.

To assay HaRVT nucleotide preference, a master mix was prepared as follows: purified recombinant HaRVT (1 µg) was mixed in a 10-µl reaction with 1 mCi/ml [α-^{32}P]dCTP or [α-^{32}P]UTP in 1x priming buffer (50 mM Tris-HCl, 50 mM KCl, 1 mM DTT, pH 8.0) supplied with 3 mM $MnCl_2$. After incubation at 25ºC for 30 min, an aliquot from each reaction was transferred to a tube containing 1 µl of the respective 10 mM nucleotide, or control nuclease-free water, and incubated at 25ºC for 1 h. Reactions were treated with 0.6 U Proteinase K, denatured in formamide buffer, and separated on 20% denaturing PAGE. Typhoon FLA 7000 (GE Healthcare) was used to detect ^{32}P-labeled products.

In the *trans*-complementation assay, amounts of HaRVT proteins were equalized according to Western hybridization data. In the reaction containing both Ha-D484A and Ha-YF, the amount of each protein mutant was the same as in reaction with respective mutant alone. Proteins were incubated with [α-^{32}P]dCTP in 1x priming buffer (50 mM Tris-HCl, 50 mM KCl, 1 mM DTT, pH 8.0) supplied with 3 mM $MgCl_2$ as described above. After incubation, reactions were mixed with 5x SDS-PAGE sample buffer (250 mM Tris-HCl, 5% SDS, 50% glycerol, 250 mM DTT, 0.01% bromophenol blue, pH 6.8, 1 mM PMSF), boiled for 3 min, separated on 10% SDS-PAGE, stained with InstantBlue, scanned, dried and phosphor imaged.

Assays of protein-nucleic acid linkage

RVT was incubated with [α-^{32}P]dCTP or [α-^{32}P]UTP in 1x priming buffer (50 mM Tris-HCl, 50 mM KCl, 1 mM DTT, pH 8.0) supplied with 3 mM $MgCl_2$ or 3 mM $MnCl_2$ as described above. After incubation, reactions were mixed with 5x SDS-PAGE sample buffer (250 mM Tris-HCl, 5% SDS, 50% glycerol, 250 mM DTT, 0.01% bromophenol blue, pH 6.8, 1 mM PMSF), boiled for 3 min, separated on 10% SDS-PAGE, stained with InstantBlue, scanned, dried and phosphor imaged. For in-gel analysis of nucleotide-amino acid linkage, reactions were separated on 10% SDS-PAGE, and the gels were treated with 1 M KOH at 65°C, as described in (Galligan, et al. 2011).

Results

rvt genes are exceedingly rare in bacteria but form a monophyletic clade

The vast majority of *rvt* genes were found in fungal genomes, with filamentous ascomycetes displaying the highest degree of lineage diversification, and the only sequenced *rvt*-containing prokaryote being *Herpetosiphon aurantiacus*, a filamentous gliding bacterium (Gladyshev and Arkhipova 2011). Their highly sporadic occurrence in bacteria led to a suggestion that it may represent an exceptionally rare case of eukaryote-to-bacteria transfer. Following a huge increase in sequenced bacterial and eukaryotic genomes since 2011, we sought to re-examine their distribution in different domains of life. In the fungal kingdom, their presence can be traced to the earliest-branching fungi, such as blastocladiomycetes and chytrids (Fig. 1; expanded version, Supplementary Fig. S1). None were found in archaea, and still surprisingly few can be detected in bacteria. The diversity of representation across bacterial phyla has nevertheless increased: in addition to the phylum Chloroflexi (represented by *Herpetosiphon aurantiacus*, a closely related *H. geysericola*, and an environmental sequence from activated sludge), we found an intact *rvt* in the phylum Bacteroidetes (marine bacterium AO1-C), fragmented copies in the phyla Cyanobacteria (*Scytonema tolypothrichoides*, a terrestrial hydrophobic filamentous cyanobacterium), Planctomyces (*Singulisphaera sp. GP187* from forest soil) and candidate phylum TM6 (bacterium UASB293 from upflow anaerobic sludge blanket reactor), and three highly similar transcripts of uncertain (possibly cyanobacterial) origin in a marine metatranscriptome sample. Interestingly, *rvt* sequences of bacterial origin are on average ~100 aa

shorter than their eukaryotic counterparts, and the conserved C-terminal domain lacks the 50-60-aa central part, which is present in all eukaryotic *rvt* genes. Phylogenetically, bacterial *rvt* genes, while highly diverse, form a clade (shown in Fig. 1 in cyan) that is separate from all eukaryotes, including fungi, rotifers, insects, and oomycetes, arguing against eukaryote-to-prokaryote transfer, and suggesting that their origin, which resulted from a successful combination of three domains (N-terminal, core RT, and C-terminal), is very ancient and may in fact predate the prokaryotic-eukaryotic divide. If so, the degree of their loss from bacterial genomes must be quite unprecedented.

Choice of expression system

A conservation of biochemical properties, or lack thereof, could potentially illuminate the biological functions of RVT proteins, if a bacterial representative is compared with a known eukaryotic counterpart. In this study, we chose to focus on HaRVT, the enzyme from *Herpetosiphon aurantiacus* (phylum Chloroflexi) (Holt and Lewin 1968). *H. aurantiacus* is the only *rvt*-carrying bacterium with a sequenced genome that is available as a cultured strain in the ATCC collection. The sequenced genome of *H. aurantiacus* (Kiss, et al. 2011) has a single *rvt* copy which is located on its circular chromosome (GenBank Acc. No. CP000875.1: 3377924-3380284), and a close homolog is present in a syntenic environment in a congeneric *H. geysericola* (Ward, et al. 2015); however, genetic tools for either species are lacking. Previously, we found that expression of NcRVT, an enzyme from the filamentous fungus *Neurospora crassa* (Ascomycota; Sordariomycetes), can be induced by exposing the growing culture to low concentrations of antibiotics which block protein synthesis (blasticidin S or cycloheximide) (Gladyshev and Arkhipova 2011), thereby permitting purification of the native enzyme upon induction. However, our preliminary experiments did not detect any significant induction of native HaRVT expression by RT-PCR after blasticidin S or cycloheximide treatment (data not shown), indicating that such inducibility is not a general feature of bacterial RVT proteins. To overcome the lack of inducibility in the natural host, we chose to express recombinant HaRVT in a heterologous host *E. coli*, with a well-developed expression system which provides an added advantage of rapid and efficient site-directed mutagenesis. Additionally, expression of a recombinant protein in a heterologous system may be expected to yield a product devoid of host-

specific templates, potentially permitting supply of external templates. To this end, we cloned a full-length recombinant HaRVT with a C-terminal 6xHis-tag, and induced its expression in *E. coli* with IPTG. An attempt to employ the N-terminal 6xHis-tag was unsuccessful, as the resulting proteins were less stable and showed signs of degradation (data not shown). In addition to the full-length HaRVT, we also created a series of mutants either with deletion of a specific domain of interest, or with site-specific mutations expected to interfere with domain functions. Table 1 lists these mutants, which will be described in more detail in the context of domain structure presented below.

HaRVT domain structure

HaRVT (GenBank Acc. No. ABX05271.1) is an 89.3-kDa protein 786 aa in length, with an isoelectric point 5.77. As described in (Gladyshev and Arkhipova 2011), all RVTs contain the conserved RT1-RT7 core motifs plus the thumb subdomain. A distinctive feature that sets them apart from all other RTs is a large insertion loop 2a between core RT motifs 2 and 3 (Gladyshev and Arkhipova 2011), which in HaRVT has a cysteine close to the N-boundary and an asparagine near the C-boundary, and might be a remnant of an intein, i.e. protein intron (Novikova, et al. 2014). In addition to the core RT with intact catalytic residues, HaRVT possesses several other discernible domains (Fig. 2). First, it harbors an N-terminal coiled-coil spanning 29 amino acids (I12-D40), which is highly conserved in virtually all RVT proteins. Additionally, the region between 178-229 aa has four alpha helices showing weak homology to an HTH DNA-binding domain. Finally, the C-terminal domain with no recognizable motifs comprises ~130 aa, beginning at the thumb domain boundary with the GGLG motif, which is shared between RVT and non-LTR retrotransposons (Gladyshev and Arkhipova 2011) and may act as a hinge between the core RT and the C-terminal domain.

HaRVT multimerizes via the N-terminal coiled-coil domain

We previously showed that NcRVT multimerizes in solution, forming ~1-MDa decameric complexes (Gladyshev and Arkhipova 2011). To find out whether HaRVT displays the same ability, we analysed the recombinant full-length HaRVT tagged with a C-terminal 6xHis affinity tag in *E. coli*

Rosetta 2 (DE3) strain. After cell lysis, soluble proteins were fractionated by sedimentation in a sucrose gradient. Position of the His-tagged protein in the gradient was identified by SDS-PAGE of 1-ml fractions, followed by Western blotting with a His-tag-specific antibody. Extrapolation of the calibration curve shows that HaRVT, similarly to NcRVT, tends to multimerize, and can form up to octamers (Fig. 3). In contrast to NcRVT, the number of subunits in oligomers is not well defined, but instead varies from 1 to 8, averaging at 4-5.

To identify domains involved in multimerization, we created an N-terminally truncated HaRVT mutant without the coiled-coil domain (Ha-ΔCC; Table 1). It is well known that coiled-coil domains can mediate oligomerization in a broad range of oligomeric states, the most common being parallel and antiparallel dimers, trimers, and tetramers, although higher-order oligomers are not uncommon (Lupas and Bassler 2016). In agreement with our expectation, the coiled-coil-deficient version lost the ability to multimerize, and was primarily recovered in the monomeric state (Fig. 3D). The HaRVT catalytic mutant (Ha-D484A), which carries a D-to-A mutation in the FVDD box of the core motif RT5, preserves the ability to multimerize in the same way as the full-length HaRVT (Fig. 3C). Unexpectedly, another HaRVT version with a mutation designed to disrupt the putative HTH-like DNA-binding domain (Ha-DBD) peaks in the dimer rather than the monomer range (Fig. 3E). Collectively, these results indicate that HaRVT oligomerizes primarily through its N-terminal coiled-coil, but multimerization can also be modulated by additional sites within the N-terminal domain.

HaRVT polymerizes short products in vitro in the presence of Mn^{2+} ions

As a reverse transcriptase, HaRVT is expected to polymerize DNA on RNA templates. However, we previously showed that NcRVT displays a terminal transferase activity *in vitro* and can polymerize both NTPs and dNTPs, with a strong preference for NTPs (Gladyshev and Arkhipova 2011). This activity is supported by Mn^{2+} ions acting as a co-factor. We sought to find out whether HaRVT is also an active polymerase and, if so, to verify the nature of the resulting product, i.e. DNA or RNA. Purified recombinant HaRVT and induced natural NcRVT proteins were incubated with [α-^{32}P]dCTP in the presence of 3 mM Mn^{2+}, followed by chase with an excess of "cold" NTPs or dNTPs. We observed the same results with NcRVT as described in (Gladyshev and Arkhipova 2011), serving as

an internal control (Fig. 4). However, HaRVT displayed only the ability to polymerize dNTPs at a very low level, yielding short extension products (Fig. 4). This difference may be an intrinsic property of HaRVT, although it may also result from the presence of a C-terminal His-tag. Sucrose gradient fractions containing the His-tagged HaRVT protein exhibited the same activity as the IMAC purified protein (Supplementary Figs. S2 and S3). HaRVT mutants lacking the coiled-coil domain (Ha-ΔCC) or DNA-binding domain (Ha-DBD) display the same dNTP polymerization pattern as the full-length protein (Supplementary Figs. S3 and S4), yielding very short products after the dNTP but not NTP chase. Importantly, replacement of one of the catalytic aspartates in the FVDD box with alanine (Ha-D484A) leads to complete elimination of polymerase activity of HaRVT (Supplementary Fig. S5), ruling out the possibility that the ^{32}P-labeled nucleotide is simply being captured in the binding pocket for reasons unrelated to catalysis.

Both HaRVT and NcRVT bind dCTP

For all recombinant HaRVT versions, with notable exception of the D484A catalytic mutant, broad bands with apparent size of *ca*. 30 nt and 40 nt, regardless of NTP/dNTP chase, could be clearly detected on denaturing PAGE (Fig. 4 and Fig. 9, bands marked with asterisks). Treatment of reaction mixes with DNase, RNase, or 0.3 M NaOH, as well as pre-treatment of the protein with RNase, did not eliminate those bands and did not affect their size (Supplementary Fig. S6). However, ethanol precipitation of the reaction mix removes those bands completely, which is indicative of their proteinaceous nature (Supplementary Figs. S2 and S6). To test this hypothesis, we separated the same reaction mixes, but without Proteinase K treatment, by SDS-PAGE. Indeed, bands corresponding to full-length HaRVT in size, and showing the presence of His-tagged protein after Western blot hybridization, were clearly detected (Figs. 5 and 6). Chasing the reaction with dNTPs slightly increases the size of the protein band, indicating that more than one deoxynucleotide becomes bound to the protein, i.e. that HaRVT-dCMP represents an initiation complex that can be elongated (Fig. 6, left panels). Pre-treatment of HaRVT with RNase shows no effect on the priming reaction, and neither it is affected by post-treatment with DNase or RNase (Fig. 6).

When we performed the priming assay with sucrose gradient fractions containing full-length HaRVT, rather than IMAC-purified protein, the signal is visible only for the protein band corresponding to HaRVT, but not for any other protein in the mix (Fig. 6). Thus, for further assays we used sucrose gradient fractions of the corresponding mutants, in order to confirm that no protein other than HaRVT derivatives can form nucleotide-protein linkages. Both mutant HaRVT versions lacking the multimerization pattern (Ha-ΔCC and Ha-DBD) display binding of labelled dCTP, indicating that neither the coiled-coil nor the putative DNA-binding domain are involved in priming (Supplementary Fig. S7). The reaction was performed in the presence of 3 mM Mn^{2+}, and the lack of a DNA-binding domain or the absence of any natural template attached to HaRVT is not expected to play a role in priming, since the initiation stage of polymerization may be template-independent in the presence of Mn^{2+} ions, as was shown for HBV RT (Urban, et al. 1998). Importantly, the catalytic mutant Ha-D484A, although clearly visible on a Western blot, shows no trace of protein-bound nucleotides, indicating that the addition of the first nucleotide is performed by the core catalytic RT domain of HaRVT (Fig. 6, A and D). As with the sucrose gradient fractions of Fig. 6, the IMAC-purified Ha-D484A mutant shows no signs of a dCTP-HaRVT bond formation (data not shown), reiterating the catalytic nature of dCTP-protein linkage.

Interestingly, the native NcRVT from *N. crassa*, which was induced by blasticidin and purified as in (Gladyshev and Arkhipova 2011), also shows the in-gel labeling pattern consistent with formation of a covalent protein-nucleotide linkage (Supplementary Fig. S8), indicating that the ability for protein priming may be a general RVT feature. Nevertheless, this activity in NcRVT may be outcompeted by formation of long non-templated chains, with only trace amounts of amino acid-dCMP products visible in Fig. 4 (right panel), as well as in Fig. 3D from (Gladyshev and Arkhipova 2011).

Effect of different cations on HaRVT priming activity

We checked the effect of Mg^{2+} and Mn^{2+} on protein-nucleic acid linkage. Attachment of labeled dCTP to HaRVT can be detected at low levels in the absence of divalent cations (Fig. 7), probably because of residual ions remaining after protein purification. However, addition of 3 mM Mg^{2+} or 3

mM Mn^{2+} dramatically increases the amounts of dCMP linked to HaRVT (Fig. 7). Substitution of 100 mM NaCl with 75 mM or 50 mM KCl does not significantly affect the priming ability (compare Fig. 4 using Buffer A with 100 mM NaCl to Fig. 9 using priming buffer with 50 mM KCl). The effect of monovalent cations on HaRVT remains a subject for further investigation, since it could affect conformation of putative natural templates.

HaRVT and NcRVT may be linked to dCTP through a tyrosine residue in the C-terminal domain

Since protein priming could be initiated by a hydroxyl group-containing serine, threonine, or tyrosine residue of a protein (Salas, et al. 1996), it was of interest to find out which domain carries the priming residue, and which amino acid is used by RVT for priming. It is known that the phosphotyrosine linkage is resistant to 1 M hot KOH treatment, but the phosphoserine and phosphothreonine linkages are not (Cooper and Hunter 1981; Galligan, et al. 2011). After KOH treatment at 65°C, both HaRVT and NcRVT still show the labeled nucleotide linked to the protein, indicating that both may use a Tyr residue to prime polymerization (Fig. 6C and Supplementary Fig. S8), and narrowing the search to a conserved Tyr. Comparison of HaRVT and NcRVT protein sequences revealed four conserved Tyr residues: Y335, located in the loop region between the RT2 and RT3 motifs of the RT domain; Y626, located in the αJ helix of the RT thumb; and Y741 and Y749, both located in the C-terminal domain of HaRVT (Fig. 2; numbering refers to HaRVT protein). We created the corresponding HaRVT mutants, substituting each Tyr with the structurally similar Phe lacking the hydroxyl group (Table 1; Ha-Y335F, Ha-Y626F, and Ha-YF). Since Y741 and Y749 are located very close to each other and could potentially substitute each other for protein priming purposes, we have mutated both of them simultaneously to avoid possible compensatory effects (Fig. 2C). As shown in Fig. 8A, very little labeling was observed in the double-Tyr mutant Ha-YF, suggesting that the C-terminal domain may serve the function of the so-called terminal protein (TP) (Salas 1991) (see Discussion). Nevertheless, it is also possible that even the conservative Tyr-to-Phe replacement in the C-terminus negatively affects catalysis. We cannot rule out the involvement of a loop tyrosine either, since there are three other, less conserved Y296, Y308 and Y361 residues in loop 2a which may potentially substitute for Y335.

Finally, we sought to find out whether the reduced activity of the Ha-YF mutant could be rescued in *trans* by addition of the catalytically dead enzyme with an intact C-terminus. However, combining Ha-D484A with Ha-YF did not improve labeling, indicating that trans-complementation does not occur (Fig. 8B). Overall, our results are consistent with the view that the catalytic center is in the closed conformation provided by the greatly extended loop 2a (Zhao, et al. 2017) and is largely inaccessible to external substrates, but is capable of adding dNTP onto suitable hydroxyl groups in the vicinity of the catalytic center (either short oligos captured during enzyme preparation, or accessible hydroxyls in the protein itself).

HaRVT displays a strong preference for dCTP

Usually, the choice of nucleotide for protein priming is directed by the 3' end of a specific template (Salas, et al. 1996). Even when a recombinant protein expressed in a heterologous system is expected to be free of a host-specific template, it may still utilize a template bearing some resemblance to the natural template. To find out whether recombinant HaRVT displays any nucleotide preferences, we incubated the HaRVT protein with [α-^{32}P]dCTP, and then chased with dNTP, NTP, or each of the four nucleotides (Fig. 9). It is clearly seen that addition of dCTP yields a prominent 13-nt band, which is also visible at a much lower intensity with dNTP chase. This result indicates that HaRVT strongly prefers to incorporate dCTP over all other dNTPs, which may in turn hint at the presence of a G-rich template in the natural host. Nevertheless, in the presence of 3 mM Mg^{2+}, [α-^{32}P]UTP could also attach to HaRVT protein to a certain extent (Supplementary Fig. S9). However, in contrast with [α-^{32}P]dCTP, the [α-^{32}P]UTP-amino acid adducts are barely visible (Supplementary Fig. S10), even in the presence of 3 mM Mn^{2+}, consistent with the preference for deoxy- over ribonucleotides. In this experiment, when dCTP was used for chase, a diffuse area corresponding to putative extension products appears on denaturing PAGE, again indicating preference for deoxycytidine incorporation (Supplementary Fig. S10).

Overall, our findings contrast sharply with *in vitro* properties of NcRVT, which does not strongly discriminate between individual bases and sugars, in fact showing a slight preference for NTP incorporation over dNTP, acting in a non-templated fashion. While the underlying reasons for base

discriminations may vary, we favor the hypothesis that the preference for dCTP displayed by HaRVT may be indicative of its ability to perform RNA-dependent DNA polymerization, as expected of an RT, and that it may be explained by affinity for a G-rich RNA template in its natural host.

Discussion

RVT: the only RT with cross-domain presence

RVTs are encoded by chromosomal genes and can be found in all kingdoms of life, implying a biological function applicable to both prokaryotes and eukaryotes. Their prevalence in the fungal kingdom (as per our most recent inventory, only 30 out of more than 250 RVT sequences did not originate from fungi), and their conspicuous absence from most metazoans, except for collembolan insects and bdelloid rotifers, seems to favour the possibility that RVTs act autonomously and can occasionally undergo horizontal transfers without undermining their biological function.

The presence of *rvt* genes across the major domains of life, Eukarya and Bacteria, may be regarded as one of their most unusual properties. Every other known RT type can be assigned either to Bacteria/Archaea, or to Eukarya. In our initial study (Gladyshev and Arkhipova 2011), only a single sequenced bacterium (and two environmental samples) harboured an *rvt* gene, which could be most parsimoniously explained by a single eukaryote-to-prokaryote horizontal transfer event. In this work, we conducted an extensive search of prokaryotic genomes and metagenomes, and identified six sequences of assigned bacterial origin and five environmental metagenome/metatranscriptome sequences of presumably bacterial origin, which displayed clear homology to *rvt*. The prokaryotic sequences, which come from diverse phyla (Chloroflexi, Cyanobacteria, Bacteroidetes, Planctomycetes), exhibit more similarity to each other than to their eukaryotic counterparts. Thus, the monophyly of bacterial *rvt* genes may be best interpreted in terms of their shared ancestry, with intradomain rather than interdomain horizontal exchange, combined with a strong trend for elimination from bacterial genomes. Indeed, three bacterial contigs harbored only partial *rvt* fragments.

The unusual properties of RVTs may bear relevance to the broader question of biological significance of reverse transcription in living cells. Reverse transcription is usually utilized by viruses,

retrotransposons, or retroplasmids, i.e. serves the proliferative needs of autonomous genetic elements which are not involved in major cellular processes. TERT is an important exception which performs chromosome end maintenance in most eukaryotes by reverse-transcribing a highly-specialized unlinked RNA template. RVTs represent the only other exception, however their cellular role(s) remain unknown. It is worth noting that in ascomycete genomes, two out of five RVT lineages (L and M, Fig. 1) underwent further specialization associated with apparent loss of reverse transcription, which is manifested in the loss of catalytic residues by all members of the lineage (data not shown). The *N. crassa* genome harbors a single lineage (N) coding for a non-essential ORF, which retains all of the catalytic motifs required for activity, and can be purified as a decamer (Gladyshev and Arkhipova 2011).

RVT as a multimer

An interesting feature of RVT proteins, conserved across different kingdoms, is its ability to form multimers. Many polymerases, including RTs, are known to act as dimers, although in some cases they can also act as monomers, e.g. different retroviral RTs, or TERTs from certain species (Nowak, et al. 2013; Sandin and Rhodes 2014). In principle, multimerization above the dimeric state could be a mechanism preventing uncontrolled RVT action in the host. On the other hand, many viral RNA-dependent RNA polymerases (RdRPs) can form oligomers *via* an N-terminal domain and even form lattices without loss of functionality (O'Reilly and Kao 1998; te Velthuis 2014; Ferrer-Orta, et al. 2015), although they lack any coiled-coil motifs. Neither are such motifs typically found in RTs. Computational predictions of the oligomeric state for RVT proteins by Logicoil (Vincent, et al. 2013) can vary, but the most probable state is usually a tetramer or a trimer (data not shown). Our results identify the N-terminal coiled-coil motif of HaRVT as the primary determinant for multimerization, since its removal shifts the protein from the multimeric to the monomeric state. In addition, the oligomerization process can also be affected by a mutation in the presumptive DNA-binding domain, which may provide an extra contact surface. Priming activity, however, appears to be unaffected in the Ha-ΔCC mutant. While the functional significance of multimerization *in vivo* remains to be determined, its conservation in all RVTs, except for the M lineage lacking the catalytic aspartate in

motif A, suggests a connection to RVT enzymatic properties. We hypothesize that multimerization may serve to prevent RVT from acting on host templates in an uncontrollable fashion: since the putative DBD is adjacent to a region also prone to coiled-coil formation (Fig. 2B), it may become exposed upon unzipping of the coiled-coils and made available for target binding.

Protein priming by reverse transcriptases

Our data indicate that RVT may be the first case of a chromosomal RT with protein priming ability. It is still a mystery why the product of a domesticated RT gene would utilize protein priming, which has so far been reserved for virus-borne, plasmid-borne, or mobile element-associated polymerases, the sole exception being the maintenance of linear chromosomes in the bacterial genus Streptomyces (Yang, et al. 2002). Use of a protein to prime polymerization, especially in eukaryotic species with enough options for solving the problems with replication of linear DNA ends, would be very unusual and may imply a non-trivial function.

Most polymerases, including RTs, require a primer, i.e. a free 3'-OH group, with the exception of *de novo* initiation by a mitochondrial plasmid-encoded RT in Neurospora (Chen and Lambowitz 1997) and the non-canonical RdRP activity of TERT (Maida, et al. 2016), although *de novo* initiation is often utilized by viral and cellular RdRPs (van Dijk, et al. 2004). Primers are classified into four groups, viz. 1) 3' hydroxyl termini of DNA complementary to the RNA template; 2) nascent RNA chains; 3) tRNA molecules that anneal to specific sequences in the RNA genomes, as in retroviruses; and 4) deoxyribonucleoside monophosphate which is covalently attached to a specific serine, threonine, or tyrosine residue of a protein through its hydroxyl group (Salas, et al. 1996). The latter mode, called protein priming, is utilized by bacteriophages (e.g. φ29, PRD1, Cp-1); plasmids (e.g. pGKLl, pGKL2, pFOXC); certain animal viruses with linear DNA genomes (e.g. adenovirus); reverse-transcribing viruses such as hepadnaviruses (e.g. hepatitis B virus, or HBV); and bacteria from the genus *Streptomyces* with linear chromosomes (Salas, et al. 1996). Thus, protein priming can allow full autonomous replication of linear DNAs that cannot form circular structures through cohesive ends, hairpin structures at their termini, *etc*.

The mechanism of protein priming in DNA replication was reviewed in (Salas 1991). In general, specific initiation proteins interact with the origin of replication, initiating the unwinding of the double helix and exposing single-stranded DNA. The molecule of the terminal protein (TP) forms a complex with a specific DNA polymerase, and interacts with the origin of replication via recognition between TP and specific sequences at the DNA end. In the presence of the dNTP corresponding to the 5'-terminus, DNA polymerase catalyzes formation of a covalent bond between the dNMP and the OH group of a specific serine, threonine or tyrosine in the TP. After that, chain elongation occurs by a strand-displacement mechanism, with the concomitant removal of the initiation proteins and binding of the single-strand binding protein to the parental single-strand that is being displaced. The whole process also needs a topoisomerase or helicase activity. In most cases (e.g. phi29, adenovirus), TP is represented by a separate protein, but TP can also be located on the same polypeptide chain with the polymerase, as in HBV (Salas 1991).

So far, protein priming is not known to be exploited by eukaryotic host polymerases, and it is still unclear what cellular function could possibly utilize this essentially viral mechanism of initiation of polymerization. The only protein-primed RT-related protein with a biological function is the plasmid-borne AbiK from the bacterium *Lactococcus lactis*, which also exhibits terminal transferase but not RT activity (Wang, et al. 2011). Unlike RVT, AbiK is encoded by a native plasmid, and implies a non-replicative role for a polymerase in abortive phage infection, although the exact mechanism of its action remains unknown. A linear mitochondrial pFOXC plasmid in the fungus *Fusarium oxysporum* encodes an RT which is thought to employ protein priming for self-replication, although it can also initiate DNA synthesis by using snapped-back RNA templates or loosely matching DNA primers (Galligan, et al. 2011). The final and the best-studied example of RT utilizing protein priming is provided by hepadnaviruses, such as HBV. The HBV viral polymerase contains four domains, including TP (~175 aa), spacer, RT and RNase H. In HaRVT, the presumptive TP domain is also located on the same polypeptide chain with RT, while φ29 or other DNA viruses use a TP that is separate from polymerase. Importantly, we show that substitution of the catalytic aspartate to alanine in HaRVT abolishes formation of the protein-nucleotide bond, demonstrating that the RT

catalytic domain of RVT is involved in initiation of polymerization. However, additional modes of initiation cannot be ruled out, since the linkage is detectable only in radioactive assays and may be present in minor quantities, and the bulk of the enzyme molecules did not reveal the expected product in a mass-spec analysis of a tryptic digest (data not shown), as was also the case for pFOXC (Galligan, et al. 2011) (for AbiK, mass spec analysis was not performed).

The presumptive TP domain in HaRVT is relatively small (<130 aa), although in eukaryotic RVTs the respective domain can reach up to 200 aa in length. The TPs of Streptomyces are about 185 aa long and are believed to be the smallest TPs (Yang, et al. 2002), while TPs of the φ29-like phages are 230–270 aa in length, and the adenoviral TPs are the largest (600–660 aa, derived from even larger precursors). Our finding that tyrosine replacement with a chemically similar phenylalanine in the C-terminal domain of HaRVT results in drastic reduction of the radioactive signal suggests the involvement of a phosphotyrosine linkage, as initially inferred from KOH treatment, although it is still possible that even a chemically similar replacement in the C-terminus leads to a conformational change affecting functionality in an unpredictable way. While Y741 is shared between HaRVT and NcRVT, the Y749 residue exhibits better overall conservation (Fig. 2C). Nevertheless, both residues are conserved in most RVTs, with very few exceptions. In the future, high-resolution structural studies of the purified HaRVT should demonstrate which residues are better positioned for priming.

Modes of polymerization

In contrast to the fungal NcRVT, which prefers to polymerize NTPs but can also polymerize dNTPs, we find that the bacterial HaRVT polymerizes mostly dNTPs, as may be expected for an RT-related enzyme. Further experiments should clarify whether the observed difference in biochemical properties is related to functional differences between bacterial and fungal proteins, which may have undergone lineage-specific changes. A notable difference is also the extremely short length of polymerization products produced by HaRVT, while NcRVT synthesizes up to 80-nt products (Gladyshev and Arkhipova 2011).

For HBV, the initiation step of polymerization, which also yields very short products, occurs in the presence of either Mg^{2+} or Mn^{2+} divalent cations (Lin, et al. 2008). However, its terminal

transferase activity, i.e. template-independent polymerization, requires Mn^{2+}, which promotes conformational changes in the protein, while Mg^{2+} does not provide a sufficient degree of freedom (Jones and Hu 2013). Both NcRVT and HaRVT exhibit terminal transferase activity in the presence of Mn^{2+}. In the presence of Mg^{2+}, NcRVT shows a significant decrease in length and amounts of polymerization products when compared to Mn^{2+} (Gladyshev and Arkhipova 2011). The same apparently applies to HaRVT, with the difference that it yields extremely short products, which are difficult to visualize even in high-percentage polyacrylamide gels. In the absence of a natural template, HaRVT may be unable to complete the transition from the priming mode to the elongation mode *in vitro*, or perhaps to perform template jumps, giving rise to products of very limited length, which are protected from the action of DNase, as is also seen in HBV. The HBV RT synthesizes very short (up to 10 nt) DNA oligomers, either in a templated or a non-templated Mn^{2+}-dependent fashion, and then undergoes a conformational change to enter the "generic" elongation mode, using the viral RNA template to yield longer DNase-sensitive products (Jones and Hu 2013). NcRVT, in contrast, can apparently enter the elongation mode and proceed as a terminal transferase to yield much longer nucleotide chains, in a manner reminiscent of AbiK (Wang, et al. 2011). It is worth mentioning that the terminal transferase activity in the presence of Mn^{2+} can also be exhibited by TERTs (Lue, et al. 2005) and by a few other RTs, and may therefore represent one of the ancestral RT activities.

The most logical next step towards understanding RVT biological function would be identification of a cellular template for the enzyme in its native host. While overexpression in a heterologous host is not expected to yield an enzyme associated with a specialized host template, our data indicate that HaRVT favors attachment of deoxycytidine, consistent with the preferred template RNA starting with guanosine, and its strong preference for dCTP incorporation implies that such template may be G-rich. Interestingly, NcRVT showed preference for CTP and TTP, but was unable to incorporate GTP (Gladyshev and Arkhipova 2011), which may fit the concept of a G-rich template. It is of note that neither HaRVT nor NcRVT were capable of *in vitro* extension of several exogenously supplied primer-template combinations, which were readily extended by commercially available RTs (data not shown). We hypothesize that the extended loop region, which upon modelling

blocks the entrance to the catalytic center, may serve as the accessibility gate for exogenous templates and may undergo conformational changes *in vivo* to regulate the catalytic properties.

If we were to make a statement that *rvt* is the rarest bacterial gene, it would not be far from truth. The fact that, of all sequenced bacterial genomes, now in the hundreds of thousands, only a few representatives harbor *rvt* genes indicates that these RTs in bacteria are highly specialized and can be lost at astonishingly high rates. In addition to template-dependent polymerization, non-canonical *rvt* functions, such as modulation of protein activity by deoxynucleotydilation, could be entertained. It is also possible that the activities observed *in vitro* represent only a subset of all possible RVT activities, of which protein priming may be just an evolutionary remnant. We hope that further studies in natural hosts will uncover more canonical RNA-dependent polymerization properties and, eventually, a biological function of the enigmatic RVT proteins.

Acknowledgements: We thank Eugene Gladyshev for constructing pEAG87 and pEAG93 plasmids at the early stages of this work and for valuable comments on the manuscript, and Tatsiana Mello for technical assistance with pHa-Y336F and pHa-YF plasmid construction.

Funding: This work was supported by the U.S. National Science Foundation grant MCB-1121334 to I.A.

Conflict of interest: The authors declare that they have no conflict of interest.

Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors.

Author contributions: IY conducted the experiments; IA and IY designed the experiments, analysed the data, reviewed the results, wrote the manuscript, and approved the final version.

References

Alva V, Nam S-Z, Söding J, Lupas AN. 2016. The MPI bioinformatics Toolkit as an integrative platform for advanced protein sequence and structure analysis. Nucleic Acids Res 44:W410-W415.

Arkhipova IR. 2012. Telomerase, retrotransposons, and evolution. In: Lue NF, Autexier C, editors. Telomerases: Chemistry, Biology, and Clinical Applications. Hoboken, NJ: John Wiley & Sons, Inc. p. 265-299.
Chen B, Lambowitz AM. 1997. De novo and DNA primer-mediated initiation of cDNA synthesis by the mauriceville retroplasmid reverse transcriptase involve recognition of a 3′ CCA sequence. J Mol Biol 271:311-332.
Cooper JA, Hunter T. 1981. Changes in protein phosphorylation in Rous sarcoma virus-transformed chicken embryo cells. Mol Cell Biol 1:165-178.
Eickbush TH, Jamburuthugoda VK. 2008. The diversity of retrotransposons and the properties of their reverse transcriptases. Virus Res 134:221-234.
Ferrer-Orta C, Ferrero D, Verdaguer N. 2015. RNA-dependent RNA polymerases of picornaviruses: From the structure to regulatory mechanisms. Viruses 7:4438-4460.
Galligan JT, Marchetti SE, Kennell JC. 2011. Reverse transcription of the pFoxC mitochondrial retroplasmids of Fusarium oxysporum is protein primed. Mob DNA 2:1.
Garfinkel DJ, Tucker JM, Saha A, Nishida Y, Pachulska-Wieczorek K, Blaszczyk L, Purzycka KJ. 2016. A self-encoded capsid derivative restricts Ty1 retrotransposition in Saccharomyces. Curr Genet 62:321-329.
Gladyshev EA, Arkhipova IR. 2011. A widespread class of reverse transcriptase-related cellular genes. Proc Natl Acad Sci U S A 108:20311-20316.
Holt JG, Lewin RA. 1968. Herpetosiphon aurantiacus gen. et sp. n., a new filamentous gliding organism. J Bacteriol 95:2407-2408.
Jones SA, Hu J. 2013. Protein-primed terminal transferase activity of hepatitis B virus polymerase. J Virol 87:2563-2576.
Kelley LA, Sternberg MJ. 2009. Protein structure prediction on the Web: a case study using the Phyre server. Nat Protoc 4:363-371.
Kiss H, Nett M, Domin N, Martin K, Maresca JA, Copeland A, Lapidus A, Lucas S, Berry KW, Glavina Del Rio T, et al. 2011. Complete genome sequence of the filamentous gliding predatory bacterium Herpetosiphon aurantiacus type strain (114-95T). Stand Genomic Sci 5:356-370.
Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Mol Biol Evol 33:1870-1874.
Lin L, Wan F, Hu J. 2008. Functional and structural dynamics of hepadnavirus reverse transcriptase during protein-primed initiation of reverse transcription: effects of metal ions. J Virol 82:5703-5714.
Lue NF, Autexier C. 2006. The structure and function of telomerase reverse transcriptase. Annu Rev Biochem 75:493-517.
Lue NF, Bosoy D, Moriarty TJ, Autexier C, Altman B, Leng S. 2005. Telomerase can act as a template- and RNA-independent terminal transferase. Proc Natl Acad Sci U S A 102:9778-9783.
Lupas AN, Bassler J. 2016. Coiled coils - a model system for the 21st century. Trends Biochem Sci 42:130-140.
Maida Y, Yasukawa M, Masutomi K. 2016. De novo RNA synthesis by RNA-dependent RNA polymerase activity of telomerase reverse transcriptase. Mol Cell Biol 36:1248-1259.
Menendez-Arias L, Sebastian-Martin A, Alvarez M. 2017. Viral reverse transcriptases. Virus Res 234:153-176.
Novikova O, Topilina N, Belfort M. 2014. Enigmatic distribution, evolution, and function of inteins. J Biol Chem 289:14490-14497.
Nowak E, Potrzebowski W, Konarev PV, Rausch JW, Bona MK, Svergun DI, Bujnicki JM, Le Grice SFJ, Nowotny M. 2013. Structural analysis of monomeric retroviral reverse transcriptase in complex with an RNA/DNA hybrid. Nucleic Acids Res 41:3874-3887.
O'Reilly EK, Kao CC. 1998. Analysis of RNA-dependent RNA polymerase structure and function as guided by known polymerase structures and computer predictions of secondary structure. Virology 252:287-303.

Salas M. 1991. Protein-priming of DNA replication. Annu Rev Biochem 60:39-71.
Salas M, Miller JT, Leis J, DePamphilis ML. 1996. Mechanisms for priming DNA synthesis. In: DePamphilis ML, editor. DNA Replication in Eukaryotic Cells. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press. p. 131-176.
Sandin S, Rhodes D. 2014. Telomerase structure. Curr Opin Struct Biol 25:104-110.
te Velthuis AJW. 2014. Common and unique features of viral RNA-dependent polymerases. Cell Mol Life Sci 71:4403-4420.
Urban M, McMillan DJ, Canning G, Newell A, Brown E, Mills JS, Jupp R. 1998. In vitro activity of hepatitis B virus polymerase: requirement for distinct metal ions and the viral epsilon stem-loop. J Gen Virol 79:1121-1131.
van Dijk AA, Makeyev EV, Bamford DH. 2004. Initiation of viral RNA-dependent RNA polymerization. J Gen Virol 85:1077-1093.
Vincent TL, Green PJ, Woolfson DN. 2013. LOGICOIL--multi-state prediction of coiled-coil oligomeric state. Bioinformatics 29:69-76.
Wang C, Villion M, Semper C, Coros C, Moineau S, Zimmerly S. 2011. A reverse transcriptase-related protein mediates phage resistance and polymerizes untemplated DNA in vitro. Nucleic Acids Res 39:7620-7629.
Ward LM, Hemp J, Pace LA, Fischer WW. 2015. Draft genome sequence of Herpetosiphon geysericola GC-42, a nonphototrophic member of the Chloroflexi Class Chloroflexia. Genome Announc 3:pii: e01352-01315.
Yang C-C, Huang C-H, Li C-Y, Tsay Y-G, Lee S-C, Chen CW. 2002. The terminal proteins of linear Streptomyces chromosomes and plasmids: a novel class of replication priming proteins. Mol Microbiol 43:297-305.
Zhao C, Liu F, Pyle AM. 2017. An ultra-processive, accurate reverse transcriptase encoded by a metazoan group II intron. RNA 24:183-195.

Figure Legends

Fig. 1. A neighbour-joining phylogram of 200 representative amino acid RVT sequences (listed by clade in Supplementary Dataset 1). Clade support values exceeding 50% are shown; clade topology is unchanged in a maximum-likelihood phylogram (not shown). Catalytically inactive clades are marked by asterisks; representative genera are given in parentheses. Clades were compressed in MEGA7.0.18 (Kumar, et al. 2016) and color-coded by taxonomy: gray, ascomycetes; purple, basidiomycetes; brown, basal fungi; olive, oomycetes; green, mosses; magenta, insects; red, rotifers; cyan, bacteria. Scale bar, amino acid substitutions per site.

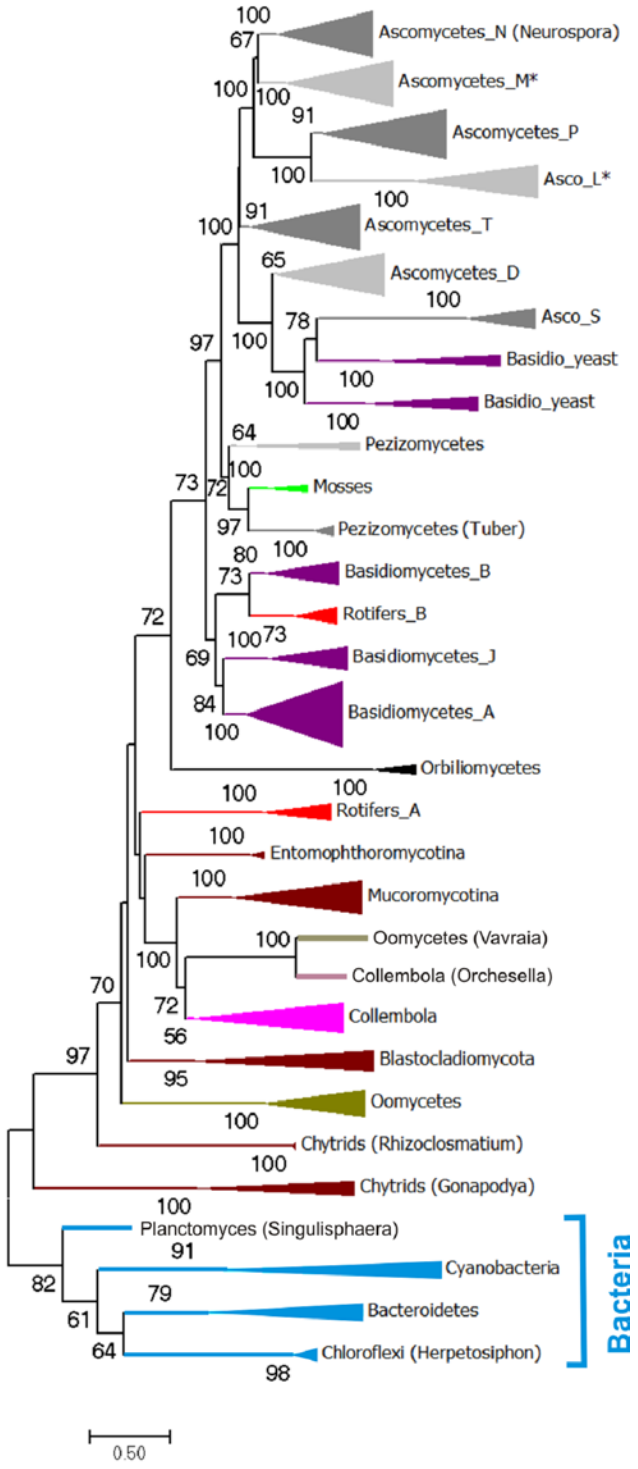

Fig. 2. Domain structure of HaRVT protein. (**A**) Diagram of the full-length HaRVT. Intertwined helices, coiled-coil motif (CC); yellow hexagon, putative HTH DNA-binding domain (DBD); gray rectangle, core RT with motifs 1-7 (indicated). Shown are the positions of the catalytic D,DD triad (with the mutagenized catalytic aspartate in the RFVDD motif underlined); the GGLG motif shared between RVT and non-LTR retrotransposons; the C and N residues at the loop boundaries; and the conserved Y residues mentioned in the text. Amino acid conservation at each position, calculated from an alignment of 200 RVT sequences, is plotted on the top, with predicted secondary structure elements such as alpha-helices (red) and beta-sheets (green) shown underneath. The yellow trapezoids are guiding to the respective regions in panels B and C. (**B**) Probability of coiled-coil formation in the alignment of 200 RVT N-termini, as predicted by COILS/ PCOILS (Alva, et al. 2016); according to predictions in (A), the N-terminal region harbors only α-helices. (**C**) Fragment of HaRVT and NcRVT secondary structure prediction by Phyre2 showing the C-terminal domain with two conserved Tyr residues (asterisks). The sequence logo generated from 200 RVT sequences was colored by hydrophobicity.

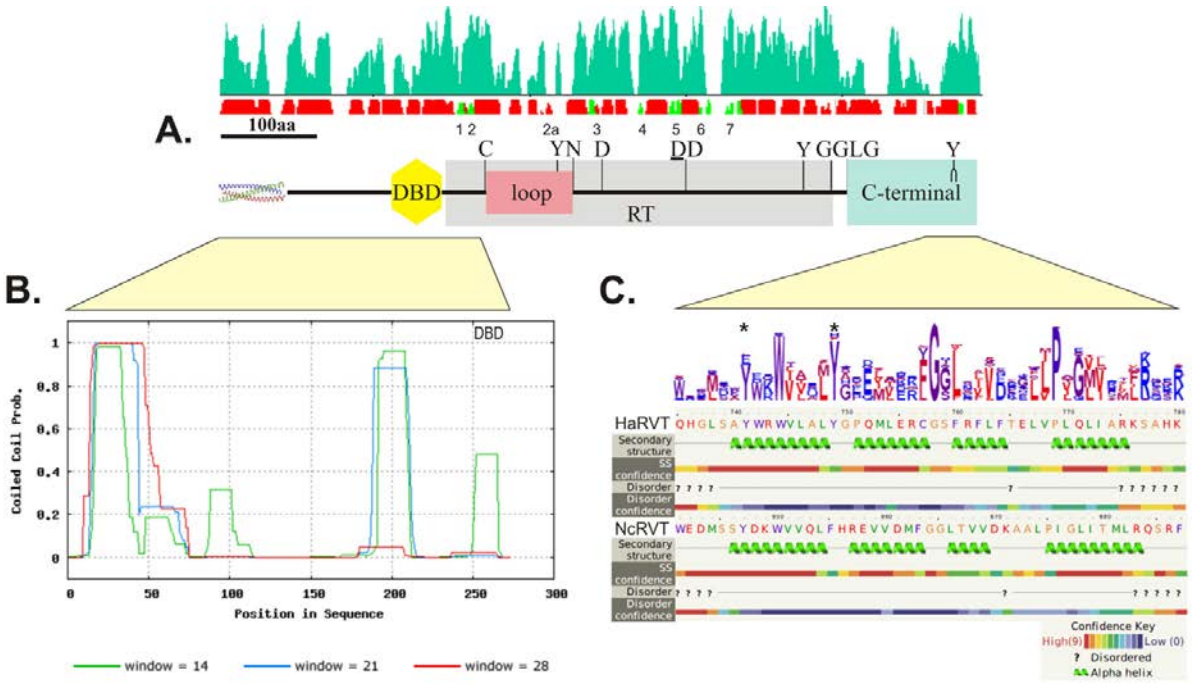

Fig. 3. HaRVT multimerization. (**A**) Relative positions of HaRVT complexes and molecular mass standards in sucrose gradient fractions. Protein standards are abbreviated as follows: Tg, thyroglobulin, 660 kDa; Cat, catalase, 250 kDa; ADH, alcohol dehydrogenase, 141 kDa; BSA - bovine serum albumin, 66.5 kDa. Positions of HaRVT monomer (90.1 kDa), dimer, tetramer, hexamer, octamer, and decamer in the gradient, as expected from the standard curve, are indicated schematically by the number of subunits on the graph. Distribution of fractions containing HaRVT, Ha-D484A, Ha-DBD, and truncated Ha-ΔCC (87.3 kDa as a monomer), as determined by Western blotting, is shown under the graph. (**B-E**) Distribution of HaRVT proteins in sucrose gradient fractions. Numbers on gel lanes correspond to fraction numbers. The diagram of each protein mutant (top) and western blot with anti-His-tag antibody (bottom) are shown for HaRVT (**B**), Ha-D484A (**C**), Ha-ΔCC (**D**) and Ha-DBD (**E**).

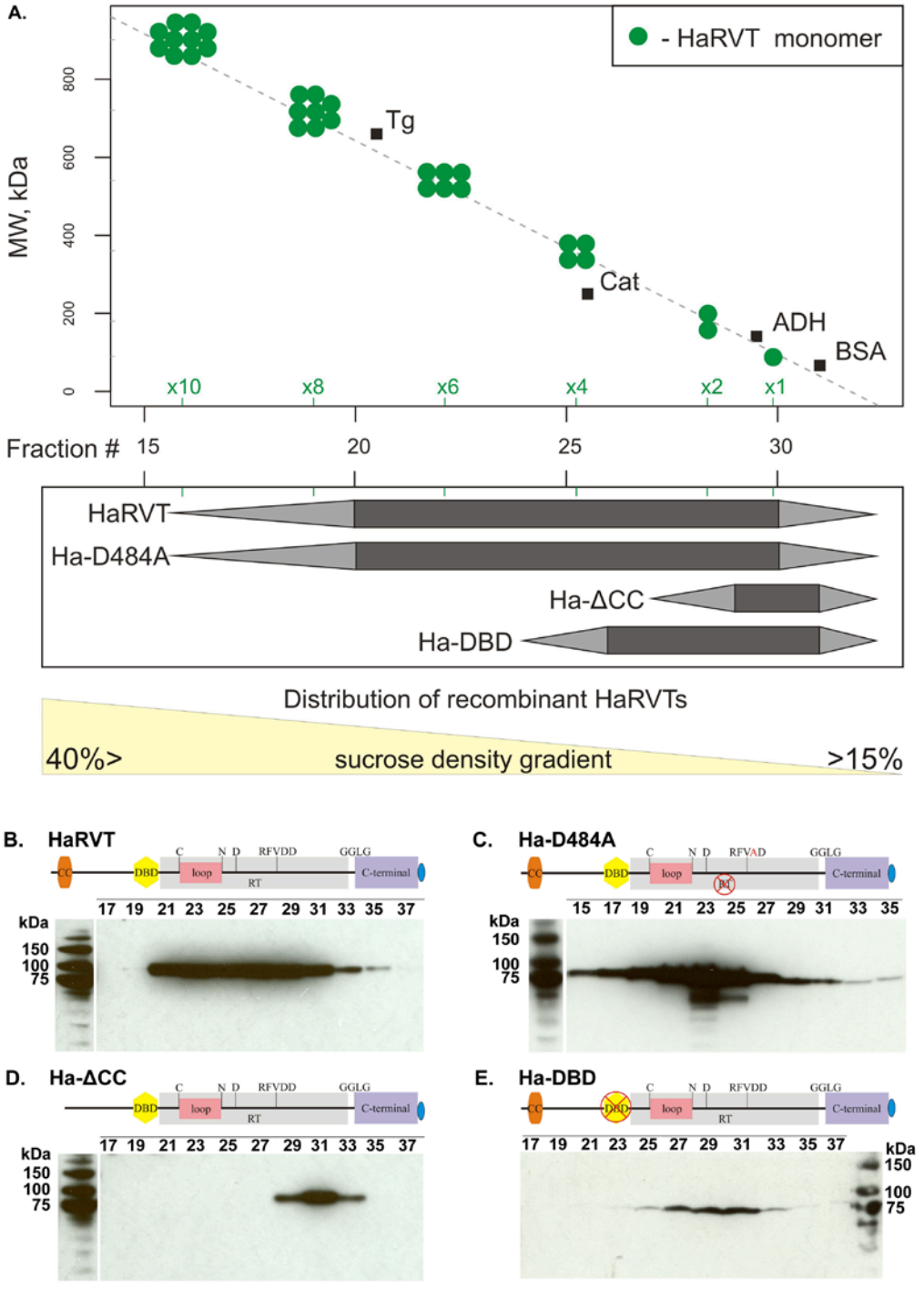

Fig. 4. Nucleotidyltransferase activity of HaRVT and NcRVT. Proteins were pre-incubated with [α-^{32}P]dCTP in the presence of 3 mM Mn^{2+}. Reactions were chased with NTPs ("N"), dNTPs ("dN") or nuclease-free water as a control ("–"). After proteinase K treatment, reactions were separated on 12% denaturing PAGE. M, marker. Asterisks denote [α-^{32}P]dCTP bound to amino acid(s). Arrows show extended products.

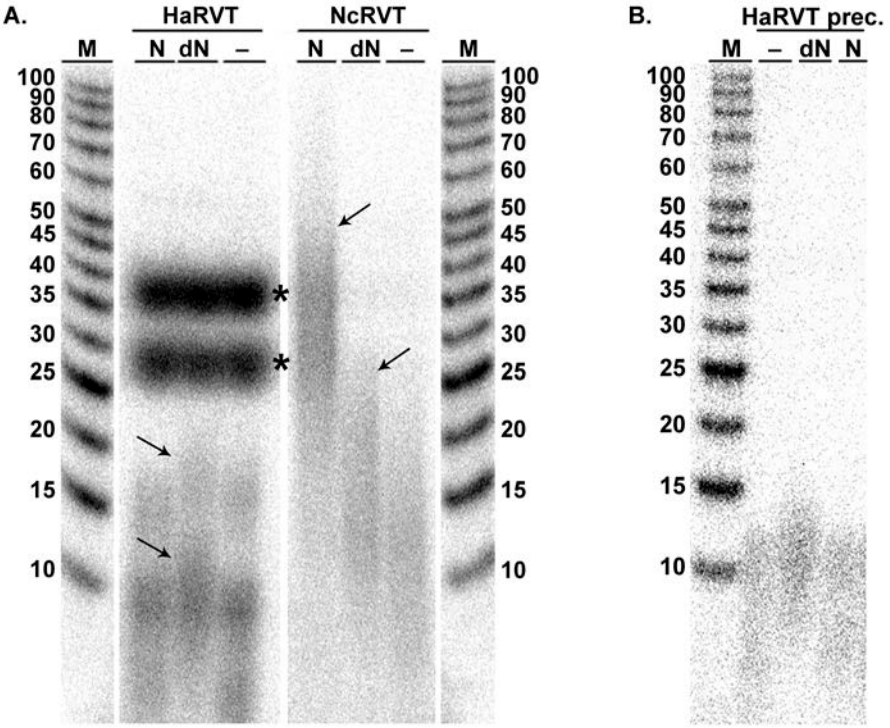

Fig. 5. Formation of protein-nucleotide linkage by purified recombinant HaRVT. Purified HaRVT (2.5 μg) was incubated with [α-^{32}P]dCTP in the presence of 3 mM Mg^{2+}, and the reaction was separated on 10% SDS-PAGE. M, marker. Phosphor image (left), same gel after staining (middle), and Western blot with His-tag specific antibody (right) are shown.

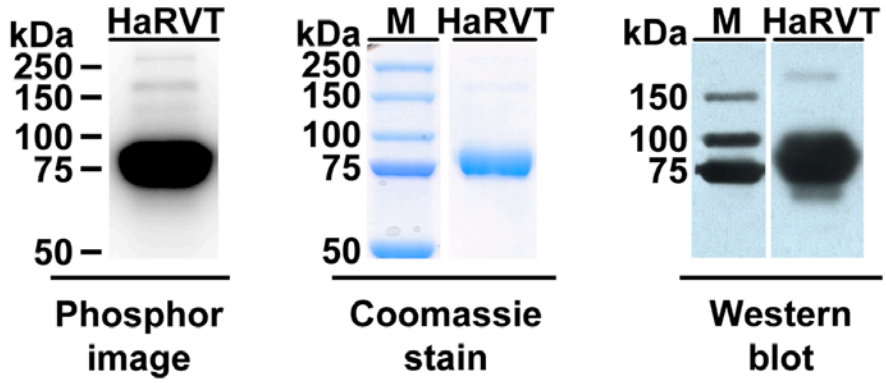

Fig. 6. Protein priming activity of recombinant HaRVT variants. Identical sucrose gradient fractions (#23 in Fig. S1A and S1B) for HaRVT (left) and Ha-D484A (right), enriched with recombinant His-tagged protein, were treated as denoted on the top, and separated on 10% SDS-PAGE. (**A**) Phosphor image. (**B**) Stained gel shown in A. (**C**) Phosphor image of the same gel as in A and B for HaRVT after alkali treatment. After staining and imaging, the gel was dried, incubated in 1 M KOH at 55ºC for 2 h, neutralized with four changes of 10% acetic acid, 10% isopropanol, and dried for phosphor imaging. (**D**) Western blot with His-tag specific antibody of sucrose gradient fractions used in HaRVT and Ha-D484A assays, showing the presence of His-tagged proteins of expected size; M, marker.

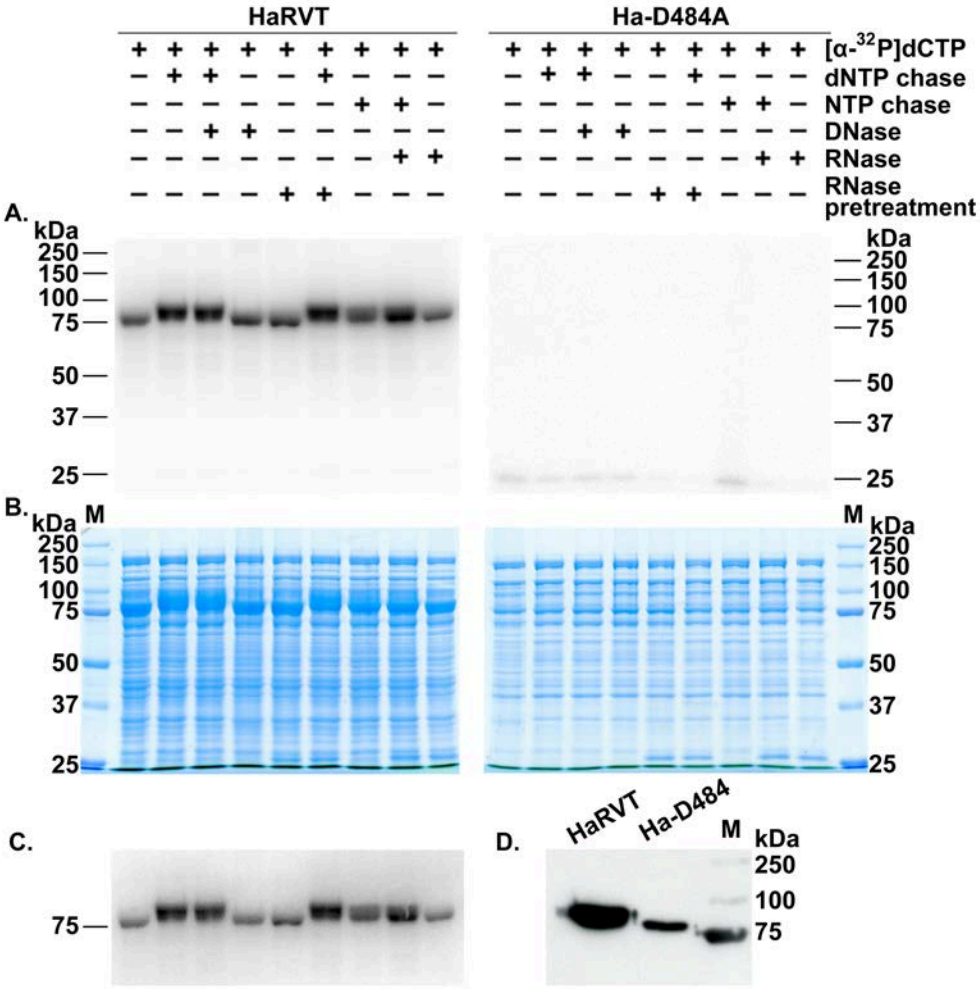

Fig. 7. Effect of divalent cations on HaRVT protein priming activity. The purified HaRVT protein was incubated with [α-^{32}P]dCTP ("+[α-^{32}P]dCTP") without divalent ions ("-") and in the presence of 3 mM Mg^{2+} ("Mg") or 3 mM Mn^{2+} ("Mn"); the protein without addition of [α-^{32}P]dCTP is shown as a control. Reactions were separated on 10% SDS-PAGE, stained with Coomassie stain (**A**) and analyzed by phosphor imaging (**B**). M, marker.

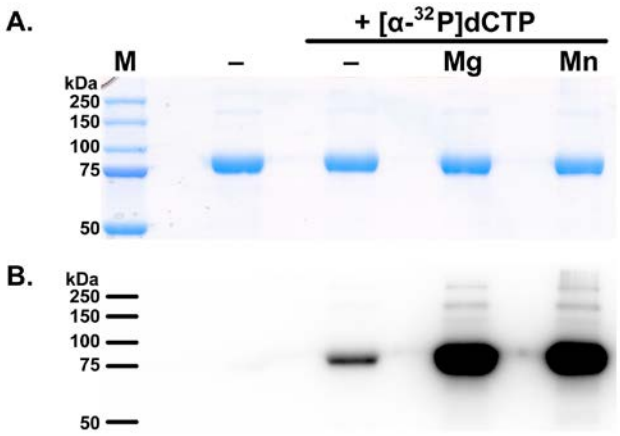

Fig. 8. Protein priming assays of recombinant HaRVT. Purified proteins were incubated with [α-^{32}P]dCTP in the presence of 3 mM Mg^{2+}, and reactions were separated on 10% SDS-PAGE. Arrows point at the HaRVT band. M, marker. (**A**) Analysis of HaRVT with Y-to-F substitutions. Residues conserved between HaRVT and NcRVT were subjected to mutagenesis as follows: Ha-Y335F; Ha-Y626F; Ha-YF (Y741F, Y749F). Phosphor image (top), same gel after staining (middle), and Western blot with His-tag specific antibody (bottom) are shown. (**B**) Assay for *trans*-complementation of protein priming activity in Ha-YF and Ha-D484A mutants compared with HaRVT. Phosphor image (top) and the same gel after staining (bottom) are shown. (**C**) Western blot with His-tag specific antibody for HaRVT catalytic mutant (Ha-D484A), Ha-YF (Y741F, Y749F) and wild type HaRVT.

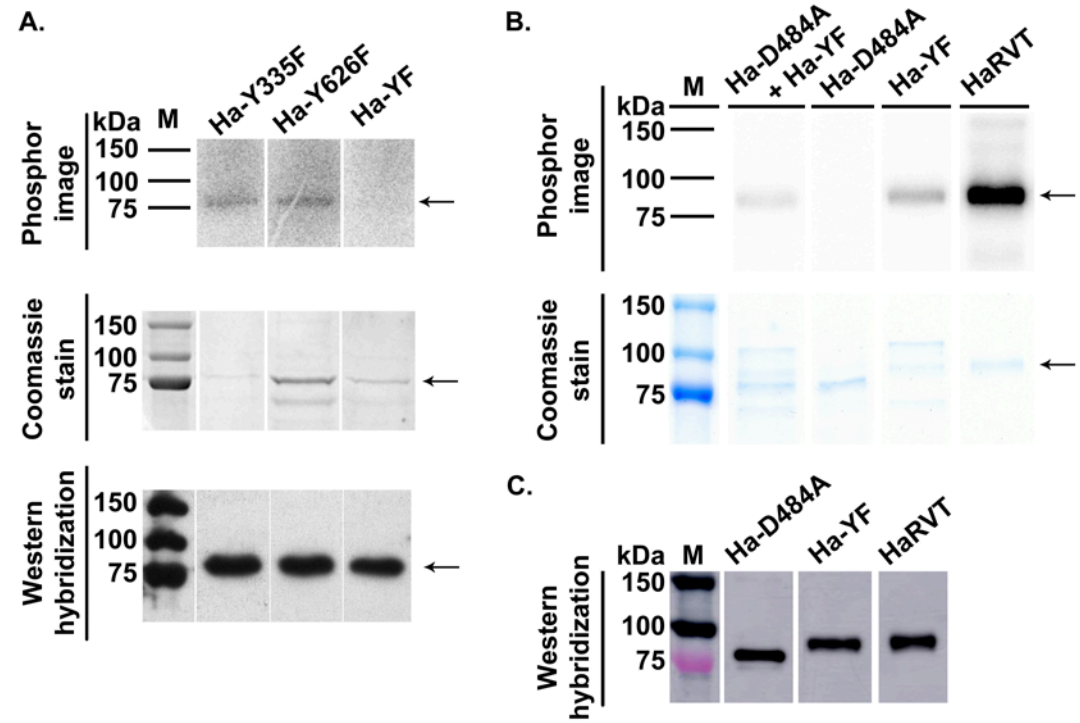

Fig. 9. Nucleotide preferences of HaRVT. Protein was pre-incubated with [α-^{32}P]dCTP in the presence of 3 mM Mn^{2+}. Reactions were chased with NTPs ("+N"), ATP ("+A"), CTP ("+C"), GTP ("+G"), UTP ("+U"), dNTPs ("+dN"), dATP ("+dA"), dCTP ("+dC"), dGTP ("+dG"), dTTP ("+dT") or nuclease-free water ("control"). After proteinase K treatment, reactions were separated on 20% denaturing PAGE. Arrow denotes the band of notably increased intensity, corresponding to dCTP incorporation.

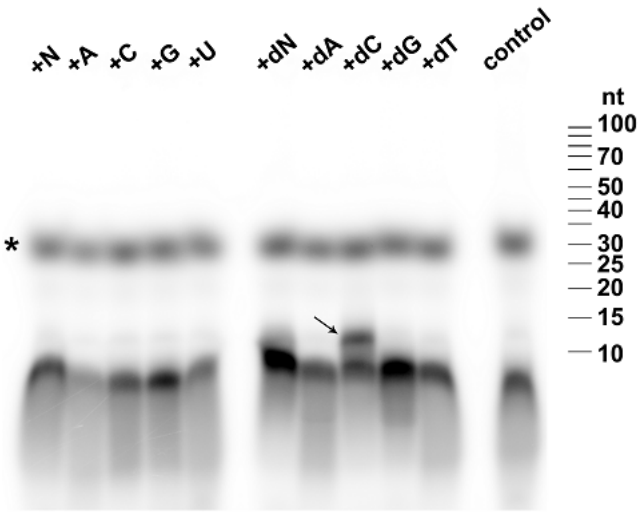

Table 1. HaRVT recombinant proteins.

Protein_ID	Description of mutants	Length, aa	MW, kDa	pI
HaRVT-N	HaRVT, N-terminal 6xHis-Tag (wild-type)	799	90.7	5.97
HaRVT	HaRVT, C-terminal 6xHis-Tag (wild-type)	792	90.1	5.93
Ha-ΔCC	HaRVT, E16_D40del (coiled-coil mutant)	767	87.3	6.06
Ha-D484A	HaRVT, D484A (catalytic mutant)	792	90.1	5.97
Ha-DBD	HaRVT, M211_A216delins GPGPGP (DBD mutant)	792	89.8	5.88
Ha-Y335F	HaRVT, Y335F (tyrosine mutant, loop)	792	90.1	5.93
Ha-Y626F	HaRVT, Y626F (tyrosine mutant, thumb)	792	90.1	5.93
Ha-YF	HaRVT [Y741F;Y749F] (tyrosine mutant, C-term.)	792	90.1	5.93