

1 **Structure and function of virion RNA polymerase of crAss-like phage**

2 Arina V. Drobysheva^{1#}, Sofia A. Panafidina^{1,2#}, Matvei V. Kolesnik¹, Evgeny I. Klimuk^{1,2}, Leonid
3 Minakhin³, Maria V. Yakunina⁴, Sergei Borukhov⁵, Emelie Nilsson⁶, Karin Holmfeldt⁶, Natalya
4 Yutin⁷, Kira S. Makarova⁷, Eugene V. Koonin⁷, Konstantin V. Severinov^{1,2,3*}, Petr G. Leiman^{8*}
5 and Maria L. Sokolova^{1,8*}

6 ¹Center of Life Sciences, Skolkovo Institute of Science and Technology, Moscow, 121205,
7 Russia

8 ²Institute of Molecular Genetics, Russian Academy of Sciences, 123182 Moscow, Russia

9 ³Waksman Institute for Microbiology, Rutgers, The State University of New Jersey, Piscataway,
10 NJ 08854, USA

11 ⁴Peter the Great St.Petersburg Polytechnic University, St. Petersburg, 195251, Russia

12 ⁵Department of Cell Biology, Rowan University School of Osteopathic Medicine at Stratford,
13 Stratford, NJ 08084-1489, USA

14 ⁶Linnaeus University, Faculty of Health and Life Sciences, Department of Biology and
15 Environmental Science, Kalmar, 39231, Sweden

16 ⁷National Center for Biotechnology Information, National Library of Medicine, National Institutes
17 of Health, Bethesda, MD 20894, USA

18 ⁸Department of Biochemistry and Molecular Biology, Sealy Center for Structural Biology and
19 Molecular Biophysics, University of Texas Medical Branch, Galveston, TX 77555-0647, USA
20

21 # Contributed equally

22 * Corresponding authors

23 E-mail: maria.sokolova@skolkovotech.ru, pgleiman@utmb.edu, severik@waksman.rutgers.edu

24

25 **Abstract**

26 CrAss-like phages are a recently described family-level group of viruses that includes the most
27 abundant virus in the human gut^{1,2}. Genomes of all crAss-like phages encode a large virion-
28 packaged protein^{2,3} that contains a DFDxD sequence motif, which forms the catalytic site in
29 cellular multisubunit RNA polymerases (RNAPs)⁴. Using *Cellulophaga baltica* crAss-like phage
30 phi14:2 as a model system, we show that this protein is a novel DNA-dependent RNAP that is
31 translocated into the host cell along with the phage DNA and transcribes early phage genes. We
32 determined the crystal structure of this 2,180-residue enzyme in a self-inhibited, likely pre-virion-
33 packaged state. This conformation is attained with the help of a Cleft-blocking domain that
34 interacts with the active site motif and occupies the RNA-DNA hybrid binding groove. Structurally,
35 phi14:2 RNAP is most similar to eukaryotic RNAPs involved in RNA interference^{5,6}, although most
36 of phi14:2 RNAP structure (nearly 1,600 residues) maps to a new region of protein folding space.
37 Considering the structural similarity, we propose that eukaryal RNA interference polymerases
38 take their origin in a phage, which parallels the emergence of the mitochondrial transcription
39 apparatus⁷.

40

41 Transcription of bacterial, archaeal, and nuclear eukaryal genes is performed by multisubunit
42 DNA-dependent RNA polymerases (RNAPs), complex molecular machines that have a common
43 ancestor^{4,8-10}. Their active site is located at the interface of two double-psi β -barrel (DPBB)
44 domains that belong to two different polypeptide chains. One of the DPBB domains carries the
45 universally conserved amino acid motif DFDGD, where the three aspartates coordinate Mg^{2+} ions
46 required for catalysis^{11,12}. Gene g066 of *Cellulophaga baltica* crAss-like phage phi14:2 encodes
47 a 2,180-residue protein that shows a limited sequence similarity to one of the two DPBB domains
48 of cellular RNAPs and contains a motif (¹³⁶¹DFDID¹³⁶⁵) that is conserved in orthologs of this protein
49 across the crAss-like phage family². Gp66 protein has been identified as a component of the
50 phage particle³. We hypothesized that gp66 is an evolutionarily divergent virion-packaged RNAP
51 of phi14:2 that is delivered into the host cell early in the infection process where it transcribes the
52 early phi14:2 genes. To test this hypothesis, we examined the *in vitro* and *in vivo* activity of gp66
53 and solved its crystal structure.

54 **RNAP gp66 transcribes single-stranded and denatured double-stranded DNA *in vitro***

55 We expressed recombinant gp66 in *Escherichia coli*, purified it (**Extended data Fig. 1**), and tested
56 its RNA synthesis activity in a diverse set of assays.

57 First, we tested whether gp66 could extend the RNA primer of an 8-nucleotide long RNA-DNA
58 hybrid in the presence of ribonucleoside triphosphates (rNTPs). This hybrid molecule mimics the
59 nucleic acid structure in the transcription elongation complex¹³. Gp66 was inactive in this assay
60 whereas both *E. coli* and T7 RNAPs extended the RNA primer (**Extended data Fig. 2**).

61 Next, we examined whether gp66 can initiate transcription of double-stranded and single-
62 stranded DNA templates. Gp66 did not transcribe the genomic DNA of phage M13 in a double-
63 stranded form and showed weak transcription of the phi14:2 genome (**Fig. 1a**). In contrast, single-
64 stranded M13 genome and denatured phi14:2 DNA were transcribed very efficiently (**Fig. 1a, 1b**).
65 The reaction products were resistant to DNase RQ1 treatment and sensitive to RNase T1
66 indicating that these high-molecular weight nucleic acids comprised entirely of newly synthesized
67 RNA (**Fig. 1c**).

68 All experimentally characterized RNAPs require Mg^{2+} ions for template-dependent polymerization
69 of rNTPs, and all three aspartates of the DFDGD motif must be present to form a Mg^{2+} -binding
70 site. Gp66 had no activity in the absence of Mg^{2+} or one of rNTPs (rATP, **Fig. 1d**). Furthermore,
71 RNA synthesis activity of gp66 was abolished if any of the aspartates of the ¹³⁶¹DFDID¹³⁶⁵ motif
72 was replaced with an alanine (**Fig. 1e**).

73 **Transcription of phi14:2 genome during infection is organized in three temporal stages**

74 Genes of phi14:2 can be divided into three classes according to the timing of transcript
75 accumulation throughout the infection (**Fig. 2a,b, Supplementary Table 2**). These classes

76 generally correspond to three functional modules – replicative, gene expression, and capsid
77 genes – that have been identified by comparative genomics of crAss-like phages². The early class
78 includes the entire replicative gene module and is transcribed in the rightward direction (**Fig. 2a**).
79 The middle and late classes are transcribed in the leftward direction and contain the gene
80 expression and capsid modules, respectively (**Fig. 2a**). The gene expression module as well as
81 other upstream middle genes are also actively transcribed late in infection because they encode
82 putative virion proteins, namely, tail genes g071 and g072, the virion-packaged predicted RNAP
83 gene (g066), and two neighboring genes (g065 and g067), whose products are also present in
84 the phage phi14:2 particle³ (**Fig. 2a**).

85 The transcript of the major capsid protein gene (g091) was the most abundant (**Fig. 2a**).
86 Remarkably, two long intergenic regions (g004-g005 and g005-g006 junctions) were transcribed
87 at a higher level than most protein coding genes (**Fig. 2a, Supplementary Table 2,3**). These
88 intergenic regions are transcribed in the rightward direction and are present at the earliest time
89 point sampled (40 min post-infection), but in contrast to all other early genes transcripts, their
90 abundance does not drop later in infection (**Fig. 2a**). The functions of these long non-coding RNAs
91 remain to be determined.

92 **Virion-packaged gp66 transcribes early genes of phi14:2**

93 *In vitro*, gp66-dependent transcription was resistant to rifampicin, an inhibitor of bacterial
94 RNAPs¹⁴, whereas *C. baltica* RNAP-dependent transcription was sensitive (**Fig. 2c**). This finding
95 made it possible to examine the role of gp66 and *C. baltica* RNAP in the transcription of phi14:2
96 genome *in vivo*. Addition of rifampicin to a *C. baltica* culture infected with phi14:2 increased the
97 relative abundance of phi14:2 transcripts reads in libraries obtained for every time point sampled
98 and, conversely, reduced the abundance of *C. baltica* reads (**Extended data Fig. 3**). Rifampicin
99 severely inhibited the transcription of the middle and late genes of phi14:2, whereas the
100 transcription of the early genes was only moderately affected (**Fig. 2d** in comparison with **Fig.**
101 **2b**). Thus, the early genes of phi14:2 are transcribed by gp66, a rifampicin-resistant RNAP, which
102 must be translocated into the host alongside phi14:2 DNA.

103 **Middle and late genes of phi14:2 are transcribed by the host RNAP**

104 In order to delineate the 5' ends of phi14:2 transcripts and identify promoters, we performed
105 primer extension analysis of RNA purified from infected cells (**Extended data Fig. 4**). Early
106 transcripts that are synthesized by gp66 RNAP did not contain detectable common upstream
107 motifs. By contrast, an extended tripartite motif was present upstream of middle genes transcripts
108 (**Extended data Fig. 4, Fig. 2e**). Two blocks of this motif resembled the '-35' and '-10' promoter
109 consensus elements recognized by bacterial RNAPs containing primary σ -factors¹⁵. Indeed, *E.*
110 *coli* σ^{70} -RNAP holoenzyme transcribed PCR fragments with such promoters *in vitro* (**Extended**

111 **data Fig. 4d**). The 5' ends of these transcripts matched those of RNAs purified from infected *C.*
112 *baltica* cells (**Extended data Fig. 4**).

113 The genomes of crAss-like phages in the candidate genus VI of the beta-crassvirinae subfamily¹⁶
114 possess motifs that are similar to phi14:2 middle promoters (**Supplementary file 1, Fig. 2e**).
115 These putative promoters are located upstream of homologs of phi14:2 middle and late genes
116 (**Supplementary file 1**). Thus, middle and late genes in other crAss-like phages are likely also
117 transcribed by the respective host RNAPs.

118 **Crystal structure of gp66 reveals a unique active site conformation**

119 To better characterize gp66 RNAP, we crystallized it and solved its structure to a resolution of 3.5
120 Å. Two different crystal forms were produced (monoclinic and orthorhombic) and both contained
121 two RNAP molecules in the asymmetric unit (4,388 amino acids including affinity tags). The phase
122 information was obtained with the help of Ta₆Br₁₂ and SeMet derivatives by the single wavelength
123 anomalous diffraction technique. The atomic model comprising 2,166 amino acids was refined to
124 R/R_{free} values of 0.19/0.24 and contained 0.02% Ramachandran outliers (**Table 1**).

125 The structure of gp66 is most similar to that of *Neurospora crassa* single-subunit DNA/RNA-
126 dependent RNAP QDE-1. QDE-1 and its homologs in other eukaryotes synthesize short RNAs
127 involved in RNA interference^{5,6}. Both gp66 and QDE-1 contain two DPBB domains that belong to
128 two different subunits in multisubunit RNAPs (β and β' in bacterial RNAP) within a single chain.
129 Furthermore, in both gp66 and QDE-1 the two DPBB domains are connected by a similar ~140-
130 residue long Connector domain (residues 1095 – 1238 and 793 – 919 in gp66 and QDE-1,
131 respectively) (**Fig. 3**).

132 Gp66 RNAP shares two conserved structural elements and, possibly, corresponding functional
133 features with multisubunit cellular RNAPs. Specifically, gp66 contains a trigger loop (residues
134 1598 – 1636), which loads rNTPs into the active site in multisubunit RNAPs^{17,18}, and a bridge
135 helix (residues 1529 – 1559) that is essential for RNAP translocation along the template¹⁹⁻²¹ (**Fig.**
136 **3**). Both, the trigger loop and the bridge helix of gp66 are more similar to those of QDE-1 than to
137 the corresponding elements of cellular RNAPs. All strictly conserved residues of gp66 are located
138 around the ¹³⁶¹DFDID¹³⁶⁵ motif and most of them have counterparts in QDE-1 and/or multisubunit
139 RNAPs (**Extended data Table 1**).

140 The overall structural similarity of gp66 to QDE-1 and multisubunit RNAPs is, however, low.
141 Automatic superposition²² of gp66 onto QDE-1 identifies 479 equivalent residues that display
142 8.6% sequence identity and whose C α atoms superimpose with a root mean square deviation
143 (RMSD) of 3.5 Å. Superposition of gp66 onto *T. thermophilus* RNAP contains 489 residues with
144 8.2% sequence identity and an RMSD of 4.2 Å.

145 The structure of gp66 presents multiple unique features. Besides the two DPBB domains,
146 Connector, and two structural elements involved in catalysis (the trigger loop and the bridge helix),
147 the rest of gp66 domains comprising nearly 1,600 residues have no homologs that could be
148 identified with existing tools²³. The functions of these domains remain to be determined.
149 Furthermore, the ¹³⁶¹DFDID¹³⁶⁵ catalytic motif of gp66 is in a conformation that is incompatible
150 with catalysis (**Fig. 4**). In all previously studied RNAPs, the fourth position in this motif is occupied
151 by a glycine, and the three aspartate side chains point roughly towards the same point where they
152 coordinate a Mg²⁺ ion required for catalysis. As shown above, Mg²⁺ and each of the aspartates of
153 the catalytic motif are required for gp66 RNA synthesis activity, so this motif must be responsible
154 for catalysis despite its unusual conformation in the crystal structure. Thus, in an actively
155 transcribing gp66 RNAP, the catalytic motif apparently refolds to allow Mg²⁺ ion coordination by
156 the three aspartate side chains. One way to accomplish this is for the isoleucine to adopt a left-
157 handed turn conformation. RNAPs in all crAss-like phages contain an isoleucine or valine in the
158 fourth position of the catalytic motif, so their active sites conformations and properties are likely
159 to be similar to those of gp66.

160 **Regulation of activity of virion RNAPs of crAss-like phages**

161 A notable feature of the gp66 structure is that its RNA-DNA hybrid binding cavity is occupied by
162 a Cleft-blocking domain (residues 196 – 233) (**Fig. 4**). Besides forming a number of interactions
163 with the cavity 'walls', it interacts with the catalytic ¹³⁶¹DFDID¹³⁶⁵ motif (there is a nearly ideal
164 hydrogen bond between G218 and the side chain of D1365) (**Fig. 4**). This interaction stabilizes
165 the unusual conformation of the catalytic motif.

166 We hypothesize that the crystal structure represents a self-inhibited, pre-virion-packaged form of
167 the enzyme as is likely required by the virion assembly pathway. At late stages of infection, newly
168 synthesized copies of gp66 have to be available for packaging into the virus particle and thus
169 have to be excluded from transcription of the phage genome. Gp66 attains its fully active
170 conformation upon translocation into the cell during infection. In the active conformation, the Cleft-
171 blocking domain and, possibly, the domain upstream of it, refold or are cleaved to free RNA-DNA
172 hybrid binding groove. Notably, recombinant gp66 shows a strong *in vitro* single-stranded DNA
173 transcription activity (**Fig. 1a,b**). Most likely, a single-stranded DNA template fits into the
174 remaining space in the cleft and is able to displace the Cleft-blocking domain from the cavity for
175 transcription to take place.

176 The self-inhibited conformation of virion-packaged gp66 RNAP parallels the assembly-coupled
177 maturation in other viruses. This process is typically accompanied by a large-scale conformational
178 change of the virus particle and involves proteolysis²⁴⁻²⁶. Whether activation of gp66 RNAP
179 requires proteolysis or is accomplished by a novel and unique mechanism, remains to be
180 determined. Notably, orthologs of gp66 RNAP and two proteins encoded by the adjacent genes
181 (g065 and g067 in phi14:2) are present in the virions of phi14:2³ and other crAss-like phages²⁷.

182 In some phages, the three proteins are fused into a single huge polyprotein², which is likely
183 cleaved into individual components (one of which is the RNAP). One of these proteins (gp65 of
184 phi14:2 and its orthologs) contains a Zincin-like metal-dependent protease domain² that might be
185 involved in the activation of RNAP and/or functions to digest the host peptidoglycan layer.

186 **Eukaryotic RNAPs involved in RNA interference and crAss-like phage RNAP share a**
187 **common ancestor**

188 QDE-1 and its orthologs comprise a family of RNAPs that is widespread in eukaryotes and is
189 likely to have been present in the Last Eukaryotic Common Ancestor (LECA)^{28,29}. These proteins
190 were originally characterized as RNA-dependent RNAPs and were directly implicated in the
191 production and/or amplification of small interfering RNAs⁵. However, it has been subsequently
192 shown that *in vitro* they transcribe single-stranded DNA much more robustly than RNA^{30,31}.
193 Moreover, Replication Protein A (a single-stranded DNA-binding complex) and DNA helicase
194 QDE-3 are required for RNA synthesis activity of QDE-1 on single-stranded DNA templates and
195 for RNA silencing³¹. Thus, synthesis of small interfering RNAs by QDE-1 and related enzymes
196 likely begins from transcription of a DNA template.

197 Structural similarity of crAss-like phage and QDE-1 RNAPs and the critical role of DNA-binding
198 proteins in the function of the latter³¹, strongly suggests that the RNA interference RNAP was
199 acquired by the LECA (or an earlier organism) from a phage, which infected a protomitochondrial
200 endosymbiont²⁹. This evolutionary scenario mimics the accepted view of the emergence of the
201 mitochondrial transcription apparatus that takes its origin in an unrelated single-subunit RNAP of
202 a T7-like phage⁷.

203

204 **Materials and Methods**

205 **Bacterial and phage growth conditions, biological properties of phi14:2**

206 *Cellulophaga* phage phi14:2 and its host *Cellulophaga baltica* strain #14 were previously
207 isolated³². *C. baltica* strain #14³² was grown at room temperature (RT) on agar plates (12 g sea
208 salt (Sigma), 1 g yeast extract (Helicon), 5 g Bacto Peptone (Helicon), and 15 g of agar (Helicon)
209 per liter). Bacterial colonies were visible after 2-3 days of incubation. A single colony was
210 inoculated into MLB liquid media (12 g sea salt (Sigma), 0.5 g yeast extract (Helicon), 0.5 g Bacto
211 Peptone (Helicon), 0.5 g casamino acids (Difco), 3 mL glycerol (Sigma) per liter) and grown
212 without agitation overnight. A high titer phi14:2 lysate was prepared using the top-agar plating
213 technique as follows: 100 µL of the phi14:2 phage lysate diluted in MSM buffer (450 mM NaCl
214 (Helicon), 50 mM MgSO₄ (Panreac), 50 mM Tris-HCl (Sigma), pH 8.0, 0.01% gelatin (Dr.Oetker))
215 was mixed with 300 µL of bacterial overnight culture and 5 mL of molten soft agar (MSM buffer
216 containing 0.4% TopVision Low Melting Point Agarose (ThermoFisher Scientific)) cooled to 32°C;
217 the suspension was dispersed on agar plates. Plates were incubated at RT in the dark overnight.
218 Further, 4 mL of MSM buffer was added to fully lysed plates, the top-agar surface was shredded
219 and the plates were shaken for 30 min at RT, the liquid suspension was collected and centrifuged
220 (4°C, 10,000 g, 10 min). The supernatant was 0.22-µm filtered (PES membrane filters, BIOFIL).
221 The resultant phage stock (~ 10¹⁰ - 10¹¹ PFU/mL) was stored at 4°C.

222 To plot the growth curves of the *C. baltica* during infection by phi14:2, *C. baltica* cultures
223 (n=3) were infected at OD₆₀₀~0.11 with phi14:2 at different multiplicity of infection (MOI) levels
224 (0.01, 0.1, 1 and 10). The growth was monitored using EnSpire Multimode Plate Reader
225 (PerkinElmer) by measuring OD every 30 min during 48 hours. At MOI of 10, culture lysis was
226 observed 3.5 hours post-infection (**Extended data Fig. 5a**)

227 To perform single-burst experiment, the *C. baltica* cultures (n=3) were infected at
228 OD₆₀₀~0.15 with phi14:2 at a MOI of 0.5 and immediately split into two flasks; one of the cultures
229 was supplemented with rifampicin (10 µg/mL). Aliquots of infected cultures were withdrawn every
230 hour. The number of plaque forming units (PFU) was determined by the top-agar plating
231 technique. The latent period was 3 hours (**Extended data Fig. 5b**). During the next 3 hours, a
232 gradual, ~ 20-fold, increase of the number of plaque forming units in the culture was observed
233 (**Extended data Fig. 5b**). Addition of rifampicin – an inhibitor of bacterial RNA polymerase
234 (RNAP) – prevented the production of phage progeny (**Extended data Fig. 5b**).

235 ***C. baltica* genome sequencing and assembly**

236 Genomic DNA of *C. baltica* strain #14³² was extracted from 2 mL of overnight culture by
237 Genomic DNA Purification Kit (Thermo Fisher Scientific) according to manufacturer's protocol for
238 Gram-negative bacteria. DNA libraries were generated by the Skoltech Genomics Core Facility
239 using NEBNext Ultra II DNA Library Prep Kit for Illumina (NEB) following the manufacturer's
240 instructions and sequenced on Miseq (Illumina) instrument using Miseq reagents v.3, 600 cycles.

241 Sequence reads were quality-checked using FastQC v0.11.8. The adapters and low-quality
242 sequences were eliminated using Trimmomatic v0.38. Reads were assembled by SPAdes
243 v3.13.0 with standard parameters.

244 **Sample collection and RNA purification for RNA-Sequencing**

245 *C. baltica* strain #14 culture was grown to OD₆₀₀ of 0.14, split into two flasks and one of
246 the two cultures was supplemented with rifampicin (10 µg/mL). The cultures were infected with
247 phi14:2 at a MOI of 10. To synchronize the infection, 40 min after the infection, the two cultures
248 were centrifuged (RT, 5000g, 15min) and the pellets were resuspended in the same amount of
249 fresh MLB medium with and without rifampicin correspondingly. At various time points (40, 90,
250 140, 190 min post-infection), 20-mL aliquots of infected cultures were withdrawn, collected by
251 centrifugation and kept at -20°C. Efficiency of infection was measured by comparing colony-
252 forming units (CFU) before the infection with CFU determined 90 min post-infection. Total RNA
253 was purified from the cell pellets using GeneJET RNA Purification Kit (Thermo Fisher Scientific)
254 following the manufacturer's instruction (Bacteria Total RNA Purification Protocol) with an
255 additional step: after resuspension in the Lysis Buffer the cells were disrupted by sonication (two
256 rounds of exposure for 10 seconds with a 50 seconds interval at an amplitude 20% (Q500
257 Sonicator by Qsonica)). RNA samples (5 µg of each) were treated with RNase-free DNase I
258 (Thermo Fisher Scientific) in the presence of RiboLock (Thermo Fisher Scientific) for 1 h at 37°C
259 and RNA was subsequently purified by GeneJET RNA Purification Kit (Thermo Fisher Scientific)
260 according to the manufacturer's instructions. RNA concentrations were determined with a
261 NanoDrop spectrophotometer. The overall levels of rRNA did not change throughout the infection,
262 as determined by visual inspection of agarose gel lanes.

263 **RNA-Seq library preparations and sequencing**

264 cDNA libraries were constructed by the Skoltech Genomics Core Facility as follows.
265 Ribosomal RNA was depleted from the total-RNA samples using Ribo-Zero rRNA Removal Kit
266 (Illumina), according to the manufacturer's protocol. Subsequently, strand-specific cDNA libraries
267 were generated by NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (NEB) following
268 the manufacturer's instructions, with exception of fragmentation time (10 minutes instead of 15).
269 Eight libraries were created (40 min Rif-, 90 min Rif-, 140 min Rif-, 190 min Rif-, 40 min Rif+, 90
270 min Rif+, 140 min Rif+, 190 min Rif+). The single-end strand-specific sequencing with 84 bp length
271 of the reads was performed on an Illumina Nextseq500. In total, 13 million to 25 million sequence
272 reads were obtained for each cDNA library.

273 **RNA-Seq data analysis**

274 The raw reads were subjected to quality filtering and adaptor trimming using Trimmomatic
275 v0.38³³ with the following parameters: SE -phred33 Illuminaclip:TruSeq3-se:2:30:10 leading:3
276 trailing:3 slidingwindow:4:15 minlen:36. The quality before and after processing was examined

277 using FastQC tool. Processed reads were mapped to the reference sequences (phi14:2 genome
278 (NC_021806.1) and the *C. baltica* strain #14 genome (BioProject ID PRJNA552277) using
279 bowtie2 v2.3.4.3 with default settings. Overall, 88 – 99 % of reads from each library aligned with
280 the reference genomes of *C. baltica* and phi14:2 in a strand-specific manner. Ratio of phage and
281 host transcripts abundances is shown in **Extended data Fig. 3**. The quantification of reads by
282 phage genes was performed using featureCounts function from the Rsubread package v1.34.3
283 in a strand-specific mode and allowed multiple overlapping of reads with features; other
284 parameters were set to default. RPKM (Reads Per Kilobase of transcript, per Million mapped
285 reads) values were calculated with normalization on a total number of mapped reads
286 (**Supplementary Tables 2,3**). These RPKM values were used to create the abundance curves
287 and heat maps (**Fig. 2**).

288 **Criteria for classification of phi14:2 genes**

289 Each gene was assigned to one of three temporal classes – Early, Middle, or Late –
290 according to its transcript abundance within a certain period post infection. The dynamics of
291 transcript abundance was quantified with the help of a Log-Fold Change parameter (LogFC) that
292 was calculated as follows: $\text{LogFC}_{XvsY} = \log_{10}A(Y) - \log_{10}A(X)$, where $A(X)$ and $A(Y)$ are
293 normalized transcript abundances of the gene at time points X and Y post infection
294 (**Supplementary Table 2**).

295 The maximum value of transcript abundance of the Early class genes was within the first
296 90 min post infection, so their LogFC values obeyed the following criterion: $\text{LogFC}_{90vs140} < 0$
297 and $\text{LogFC}_{140vs190} < 0$. The maximum value of transcript abundance of the Middle class genes
298 was in the 90-190 min post infection period and the increase of abundance within that period did
299 not exceed 10 times: $\text{LogFC}_{90vs190} > 0$ and $\text{LogFC}_{90vs190} \leq 1$. The transcript abundances
300 of the Late class genes increased by more than 10 times in the 90-190 min post infection period:
301 $\text{LogFC}_{90vs190} > 1$.

302 **RT-qPCR**

303 Results obtained by RNA-Seq for both Rif- and Rif+ cultures were validated by reverse
304 transcription-quantitative PCR (RT-qPCR) with primers specific to randomly chosen Early, Middle
305 and Late phage genes (**Supplementary Table 5** and **Extended data Fig. 6**).

306 Total RNA was purified as described in the sample collection and RNA purification section.
307 First-strand cDNA synthesis was performed with Maxima reverse transcriptase (Thermo Fisher
308 Scientific) and random hexamer primers (Thermo Fisher Scientific) with 150 ng of total RNA
309 according to the manufacturer's instructions. The subsequent qPCR analysis was performed
310 using iTaq Universal SYBR Green Supermix (Bio-Rad), on Applied Biosystems QuantStudio 3
311 amplifier with primers listed in **Supplementary Table 5**. The cycle threshold (Ct) values of the
312 16S RNA were used to normalize the Ct values of selected phi14:2 transcripts ($\Delta\text{Ct} = (\text{mean Ct}$

313 gene) – (mean Ct 16S rRNA)). To follow the relative differences in amplicon concentrations for
314 different samples, a $2^{(-\Delta Ct)}$ value was used.

315 **Primer extension and sequencing reactions**

316 Gene-specific primers (**Supplementary Table 4**) were labeled with [γ - ^{32}P]ATP by phage
317 T4 polynucleotide kinase (New England Biolabs), as recommended by the manufacturer. Primer
318 extension reactions were performed with 1 pmol of [γ - ^{32}P]ATP end-labeled primers and 5 μ g of
319 total RNA using Maxima reverse transcriptase (Thermo Fisher Scientific) according to the
320 manufacturer's instructions. Reactions were terminated by the addition of an equal volume of
321 denaturing loading buffer (95% formamide, 18 mM EDTA, 0.25% SDS, 0.025% xylene cyanol,
322 0.025% bromphenol blue). Sequencing reactions were performed with the same primers as the
323 ones used for the primer extension reactions and with PCR fragments (amplified from phi14:2
324 genomic DNA) using the USB Thermo Sequenase Cycle Sequencing Kit (Thermo Fisher
325 Scientific) according to manufacturer's instructions. The reactions were terminated as above. The
326 reaction products were resolved on 6–8% (w/v) denaturing polyacrylamide gels and visualized
327 with Typhoon FLA scanner (GE Healthcare). In total, 23 phi14:2 genome regions, which could
328 contain promoters were analyzed and primer extension products for ten of them were detected
329 (**Supplementary Table 4**).

330 **Search for nucleotide sequence motifs**

331 To identify motifs similar to the phi14:2 Middle promoter motif in genomes of other crAss-
332 like phages, 36 previously analyzed representative genomes²
333 (ftp://ftp.ncbi.nih.gov/pub/yutin/crassphage_2017/), the 242 genomes from a subsequent study¹⁶
334 and the genome of phicrAss001²⁷ were scanned. First, we searched for occurrences of the
335 phi14:2 Middle promoter motif by using the program FIMO³⁴ (Supplementary file 1). Thirteen
336 genomes that contained at least four unique hits with a score greater than 1 and genome of IAS
337 phage that contained three unique hits were used to create new consensus motifs, which were
338 then used as new templates to search motifs in the same 14 genomes. The new searches resulted
339 in 137 hits of which 17 corresponded to coding regions and the rest to intergenic regions. All but
340 four intergenic hits were in the sense direction.

341 Identification of coding regions required annotation of the following twelve phage
342 genomes: cs_ms_27, err843924_ms_3, ERR844029_ms, ERR844058_ms_2,
343 ERR844065_ms_1, SRR4295173_s_14, SRR4295175_s_4, eld298-t0_s_3, ERR844030_ms_2,
344 cs_ms_22, Fferm_ms_11, and HvCF_E4_ms_5. HMM profiles of conserved protein families of
345 crAss-like phages from Yutin et al² were generated from multiple sequence alignments published
346 by Yutin et al² using hmmbuild tool from the HMMER v3.1b2 package (<http://hmmer.org/>) with
347 default settings. tRNA and tmRNA genes were predicted using ARAGORN v1.2.38³⁵. ORFs were
348 predicted with Prodigal v2.6.3³⁶. Amino acid sequences of predicted ORFs were scanned against
349 Pfam-A v32.0 supplemented with aforementioned HMM profiles using hmmscan tool from

350 HMMER v3.1b2 package and hits with an e-value of less than 10^{-6} were considered a match.
351 Homologs of phi14:2 gp65 were found with the help of the jackhammer tool from the HMMER
352 v3.1b2 package. Matching sequences had an e-value of less than 10^{-6} . Two phage genomes
353 (IAS² and phicrAss001²⁷) have been annotated previously. The putative promoter motifs were
354 found more frequently upstream of phage genes encoding homologs of the phi14:2 middle
355 proteins gp069 (function unknown), gp66 (RNAP), and gp074 (integration host factor IHF
356 subunit), and late proteins gp092 (a structural protein of unknown function) and gp093 (portal)
357 (**Supplementary file 1**). In 12 out of 14 phages the motif was at least once located upstream of
358 a gene coding for tRNA. The DNA Logos of the motifs were constructed using WebLogo³⁷.

359 **Purification of phi14:2 gp66 and *C. baltica* RNAP**

360 The gene coding for the predicted phi14:2 RNAP catalytic subunit (g066 in this work;
361 GeneID 16797463 in NCBI Reference Sequence NC_021806.1) was PCR amplified from phi14:2
362 genomic DNA and cloned into pETDuet-1 between BamHI and SacI restriction sites. This plasmid
363 was used as a template to create mutant versions of g066 by site directed mutagenesis (list of
364 corresponding primers is in **Supplementary Table 6**). Resulting plasmids were transformed into
365 BL21 Star (DE3) chemically competent *E. coli* cells. The culture (3 L) was grown at 37°C to A₆₀₀
366 ~0.7 in LB medium supplemented with ampicillin at a concentration of 100 ug/mL and
367 recombinant protein over-expression was induced with 1 mM IPTG for 3 hours at 20°C. Cells
368 containing over-expressed recombinant protein were harvested by centrifugation and disrupted
369 by sonication in buffer A (40mM Tris-HCl pH 8, 300mM NaCl, 3mM β-mercaptoetanol) followed
370 by centrifugation at 15,000 g for 30 min. Cleared lysate was loaded onto a 5 mL HisTrap
371 sepharose HP column (GE Healthcare) equilibrated with buffer A. The column was washed with
372 buffer A supplemented with 20 mM Imidazole. The protein was eluted with a linear 0-0.5 M
373 Imidazole gradient in buffer A. Fractions containing gp66 were combined and diluted with buffer
374 B (40mM Tris HCl pH 8, 5% Glycerol, 0.5 mM EDTA, 1mM DTT) to the 50 mM NaCl final
375 concentration and loaded on equilibrated 5 mL HiTrap Heparin HP sepharose column (GE
376 Healthcare). The protein was eluted with a linear 0-1 M NaCl gradient in buffer B. Fractions
377 containing gp66 were pooled and concentrated (Amicon Ultra-4 Centrifugal Filter Unit with
378 Ultracel-30 membrane, EMD Millipore) to a final concentration 4 mg/ml, then glycerol was added
379 up to 50% to the sample for storage at -20°C (the sample was used for transcription assays). For
380 crystallization, fractions were diluted with buffer C (20 mM Tris HCl pH 8, 0.5 mM EDTA, 1mM
381 DTT) to the 100 mM NaCl and loaded onto MonoQ 10/100 GL column (GE Healthcare). Bound
382 proteins were eluted with a linear 0.1– 1 M NaCl gradient in buffer C. The fractions containing
383 gp66 were pooled, diluted with buffer C to the 100 mM NaCl final concentration and concentrated
384 to a final concentration 15 mg/mL and used for crystallization immediately.

385 To produce a Se-methionine (SeMet) derivative of gp66, the cells were first grown in the
386 2xTY medium until OD₆₀₀ of 0.35, then pelleted by centrifugation at 4000 g for 10 min at 4°C and

387 transferred to the SelenoMet Medium (Molecular Dimensions, Newmarket, Suffolk, UK) prepared
388 according to the manufacturer's instructions and supplemented with ampicillin at a concentration
389 of 100 ug/mL. All the subsequent steps including the expression at low temperature and protein
390 purification were the same as for the native protein.

391 For purification of *C. baltica* RNAP, 3 g of pelleted *C. baltica* cells were disrupted by
392 sonication in 15 mL of buffer B (40 mM Tris-HCl pH 8.0, 0.5 mM EDTA, 1 mM DTT, 5% glycerol),
393 containing 50 mM NaCl followed by centrifugation at 15,000 g for 30 min. Polyethylenimine P (pH
394 8.0) solution was added with stirring to the cleared lysate to the final concentration of 0.8 %. The
395 resulting suspension was incubated on ice for 30 min and centrifuged at 10,000 g for 15 min. The
396 pellet was washed by resuspension in buffer B with 0.3 M NaCl following centrifugation as
397 previously. For elution, the pellet was resuspended in buffer B with 0.6 M NaCl. Eluted proteins
398 were precipitated by adding ammonium sulfate to 67% saturation and centrifuged. The pellet was
399 dissolved in 10 mL of buffer B and loaded onto a 1 mL HiTrap Heparin HP sepharose column (GE
400 Healthcare) equilibrated with buffer B supplemented with 0.1 M NaCl. The column was washed
401 with buffer B with 0.3 M NaCl, and RNAP was eluted with buffer B with 0.6 M NaCl. The fraction
402 was concentrated by ultrafiltration (Amicon Ultra-4 Centrifugal Filter Unit with Ultracel-30
403 membrane, EMD Millipore) and loaded onto a Superdex 200 Increase 10/300 gel filtration column
404 (GE Healthcare) equilibrated with buffer B containing 0.2 M NaCl. The fractions containing RNAP
405 were pooled and concentrated up to 1 mg/mL, then glycerol was added up to 50% to the sample
406 for storage at -20°C.

407 **DNA templates for transcription assay**

408 For phi14:2 RNAP transcription assay genomic DNA of phi14:2 was purified using the
409 Phage DNA Isolation Kit (Norgen Biotek Corp) according to the manufacturer's instructions.
410 Commercial genomic DNA of M13 bacteriophage (double- and single-stranded forms, New
411 England Biolabs) were used.

412 For transcription by *C. baltica* RNAP, the PCR fragment containing T7 A1 promoter was
413 used (5' to 3' sequence:
414 tccagatcccgaaaatttatcaaaaaagagtattgacttaaagtctaacctataggatacttacagcCatcgagagggccacggcgaa
415 cagccaaccaatcgaacaggcctgctgtaatcgaggcctttttatttgatccccgggta).

416 **In vitro transcription**

417 Transcription reactions were performed in 5 µl of transcription buffer (20 mM Tris-HCl
418 pH=8, 40 mM KCl, 10 mM MgCl₂, 0.5 mM DTT and 100 µg/mL bovine serum albumin, RNase
419 inhibitor) and contained 100 nM gp66 and 50 ng genomic DNA. Where indicated, genomic DNA
420 were denatured by heating to 100°C for 5 minutes following rapid cooling at 0°C. The reactions
421 were incubated for 10 min at 22°C (or 10°C and 30°C where indicated), followed by the addition
422 of 100 µM each of ATP, CTP, and GTP; 10 µM UTP and 3 µCi [α-³²P]UTP (3000 Ci/mmol).
423 Reactions proceeded for 30 min at 22°C (or 10°C and 30°C where indicated) and were terminated

424 by the addition of an equal volume of denaturing loading buffer (95% formamide, 18 mM EDTA,
425 0.25% SDS, 0.025% xylene cyanol, 0.025% bromophenol blue). Where indicated, rifampicin was
426 added to the final concentration of 50 µg/mL. Treatment with RNase T1 (Thermo Fisher Scientific)
427 and DNase RQ1 (Promega) were performed as follows: after the 30 min incubation of
428 transcription reactions at 22°C, corresponding enzyme was added to 5 µl reactions; reactions
429 were incubated for additional 15 min at 37°C and were terminated by the addition of an equal
430 volume of denaturing loading buffer.

431 The reaction products were resolved by electrophoresis on 5 % (w/v) denaturing 8 M urea
432 polyacrylamide gel. Since high-molecular weight RNA was expected to be synthesized from
433 genomic DNA templates, the electrophoresis was run for 2 hours. Transcription reaction products
434 by *C. baltica* RNAP were loaded on the gel with a delay to observe both, high-molecular weight
435 RNA synthesized by gp66 from genomic DNA and 67 nucleotides RNA synthesized by *C. baltica*
436 RNAP from PCR fragment. Results were visualized by Typhoon FLA scanner (GE Healthcare).

437 Transcription reactions from RNA-DNA scaffolds were set at the same buffer as above
438 transcription reactions and contained 15 nM RNA-DNA scaffold and 15 nM of gp66, T7 RNAP or
439 *E. coli* RNAP core (New England Biolabs). Reactions were incubated for 10 min at 22°C, followed
440 by the addition of 1mM each of ATP, CTP, GTP and UTP. Reactions proceeded for 30 min at
441 22°C and were terminated by the addition of an equal volume of denaturing loading buffer; the
442 products were resolved by electrophoresis on 16 % (w/v) denaturing 8 M urea polyacrylamide
443 gel. Results were visualized by Typhoon FLA scanner (GE Healthcare).

444 All transcription experiments were repeated at least three times.

445 **Crystallization and structure determination of phi14:2 RNAP (gp66)**

446 The initial crystallization screening was carried out by the sitting drop method in 96 well ARI
447 Intelliwell-2 LR plates using Jena Bioscience crystallization screens at 19°C. PHOENIX pipetting
448 robot (Art Robbins Instruments, USA) was employed for preparing crystallization plates and
449 setting up drops each containing 200 nl of the protein and the same volume of the well solution.
450 Optimization of crystallization conditions was performed in 24 well VDX plates and thin siliconized
451 cover slides (both from Hampton Research) by hanging drop vapor diffusion. Crystallization drops
452 of the 24-well plate setup contained 1.5 µl of the protein solution in 20 mM Tris-HCl pH 8.0, 100
453 mM NaCl, 1 mM DTT, 0.5 mM EDTA mixed with an equal volume of the well solution. Best crystals
454 of both native and SeMet gp66 were obtained with the protein having the initial concentration of
455 15 mg/ml and equilibrated against 700 µl of the well solution containing 100 mM Tris-HCl pH 8.5,
456 200 mM NaOAc, 11% PEG 4000, 2 mM TCEP. Ta₆Br₁₂ derivatized crystals of gp66 were
457 produced by soaking native crystals in a pre-equilibrated crystallization solution that contained a
458 freshly prepared Ta₆Br₁₂ compound at a 1-2 mM concentration. Upon soaking for 1-3 days,
459 Ta₆Br₁₂ derivatized crystals acquired an emerald green color. For data collection, the crystals
460 were dipped for 15 seconds into cryo solutions containing 30% of glycerol in addition to the well

461 solution components and flash frozen in liquid nitrogen. X-ray diffraction data and fluorescent
462 spectra were collected in a nitrogen stream at 100 K.

463 X-ray fluorescence emission spectra of both Ta₆Br₁₂ and SeMet derivative crystals displayed
464 a strong “white line” at the L_{III} and K absorption edges of Ta and Se, respectively. The
465 corresponding excitation wavelengths – 1.25478 Å for Ta₆Br₁₂ crystals and 0.97872 Å for SeMet
466 – were then used for data collection. Diffraction data were collected on two different beamlines of
467 the Life Sciences Collaborative Access Team at Advanced Photon Source, Chicago: Ta₆Br₁₂ on
468 21-ID-D (Dectris Eiger 9M area detector), and SeMet on 21-ID-F (Rayonix MX300 area detector).
469 All datasets comprised a full 360° swath that was cut into 0.125° frames on the Eiger detector
470 (2880 frames) or 0.5° frames on the Rayonix detector (720 frames). The datasets were indexed,
471 integrated, and reduced with the help of the XDS suite. The heavy atom substructure of Ta₆Br₁₂
472 datasets that had an anomalous signal greater than 1.37 (as defined by XDS) could be easily
473 solved. On the other hand, all attempts at *ab initio* solution of the Se atom substructure in SeMet
474 datasets with an anomalous signal as great as 1.28 failed. The Se substructure was expected to
475 consist of around 80 atoms because the asymmetric unit contained two molecules of gp66 with
476 39 methionines each.

477 An interpretable electron density was obtained as follows. First, we solved the heavy atom
478 substructure of one of the Ta₆Br₁₂ soaked dataset with the help of HKL2MAP and SHELXD suite.
479 These phases were improved by two-fold non-crystallographic averaging. A large fraction of the
480 polypeptide chain could be traced in this electron density, but it had a resolution of 3.75 Å and it
481 was discontinuous and disordered in places. The partial model was then used in a molecular
482 replacement procedure to solve the best SeMet dataset. The height of the peaks in the Bijvoet
483 difference Fourier synthesis map of the SeMet dataset decreased gradually, and the exact
484 number of ordered Se sites could not be established. For this reason, the first 71 peaks that were
485 somewhat higher than the rest were input into PHASER to find all Se sites and obtain new phases.
486 These phases were improved by two-fold non-crystallographic averaging with the help of the
487 program PARROT. The resulting 3.5 Å resolution electron density could be traced with relative
488 ease and interpreted in nearly all 2,180 residues comprising gp66 barring for a few residues at
489 both termini.

490 The atomic model was refined with the help of the programs PHENIX and COOT. Molprobity
491 was used in the validation procedure. The final model has 94.76% of residues in the favorable
492 region of the Ramachandran plot and 0.07% outliers.

493

494 **Data availability**

495 Genome of *C. baltica* strain 14 has been deposited in the NCBI BioProject and is accessible
496 through BioProject ID PRJNA552277.

497 The RNA-Sequencing data have been deposited in the NCBI Gene Expression Omnibus³⁸ and
498 are accessible through GEO Series GenBank accession no. GSE133609.

499 The refined atomic model of phi14:2 gp66 and the associated experimental data have been
500 deposited to the Protein Data Bank under the accession number 6VR4.

501 Additional data are available from the corresponding authors upon request.

502

503 **Acknowledgments**

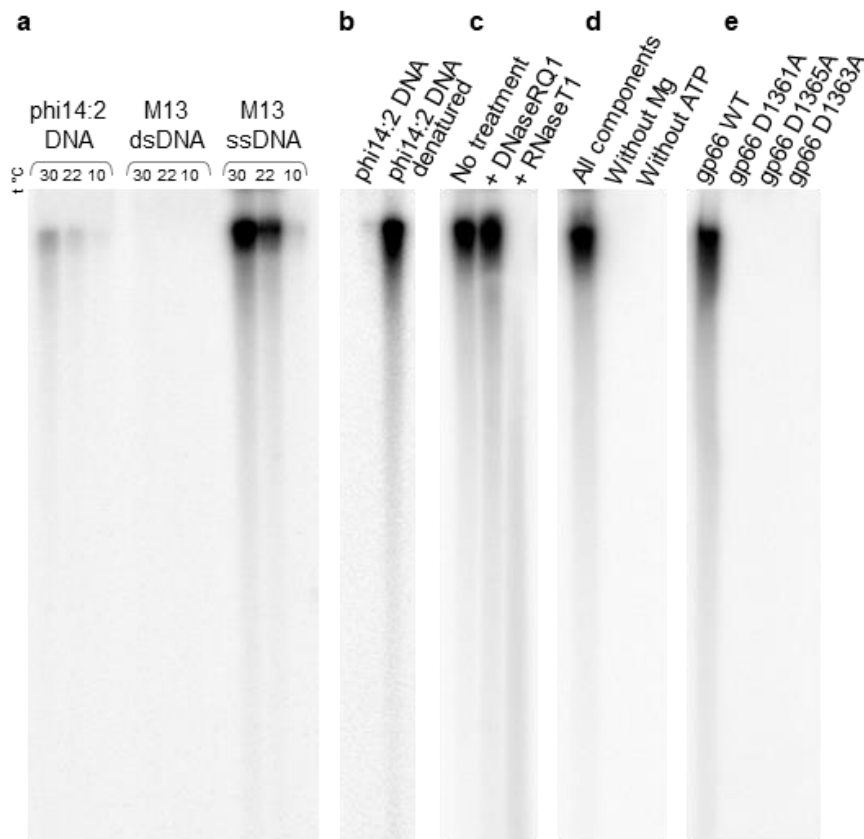
504 We would like to thank Sofia Medvedeva (Skolkovo Institute of Science and Technology, Moscow,
505 Russia) for help with promoter search. The study was carried out using resources of the Skoltech
506 Genomics Core Facility. The work was supported by the Russian Science Foundation (grant no
507 19-74-00011 to M. L. Sokolova).

508

509 **Author contributions**

510 **K.V.S., M.L.S.** and **E.V.K** conceived the study. **K.H.** and **E.N.** provided *C. baltica* cells, phi14:2
511 phage and phi14:2 DNA. **A.V.D.** cultivated *C. baltica* and phi14:2, prepared RNA for RNA-Seq
512 and primer extension experiments (PE), performed RT-qPCR. **S.P.** purified phi14:2 RNAP and its
513 mutants, performed all *in vitro* transcription assays and some of the PE. **M.K.** processed and
514 analyzed RNA-Seq data, annotated crAss-like phage genomes. **E.I.K.** performed mutagenesis of
515 phi14:2 RNAP. **L.M.** performed PE. **M.V.Y.** purified *C. baltica* RNAP. **M.L.S.** performed search for
516 promoters, prepared crystals. **P.G.L.** solved crystal structure. **M.L.S., P.G.L., S.B.** analyzed the
517 structure. **M.L.S., P.G.L., K.V.S.** wrote the manuscript, which was read, edited and approved by
518 all authors.

519



520

521 **Fig. 1. *In vitro* transcription activity of the phi14:2 RNAP gp66.**

522 **a**, Transcription by gp66 of genomic DNA of phages phi14:2 and M13 (double- and single-
523 stranded forms) at 30, 22, and 10°C; the reaction products were resolved by electrophoresis in 5
524 % (w/v) denaturing 8 M urea polyacrylamide gel and revealed by autoradiography.

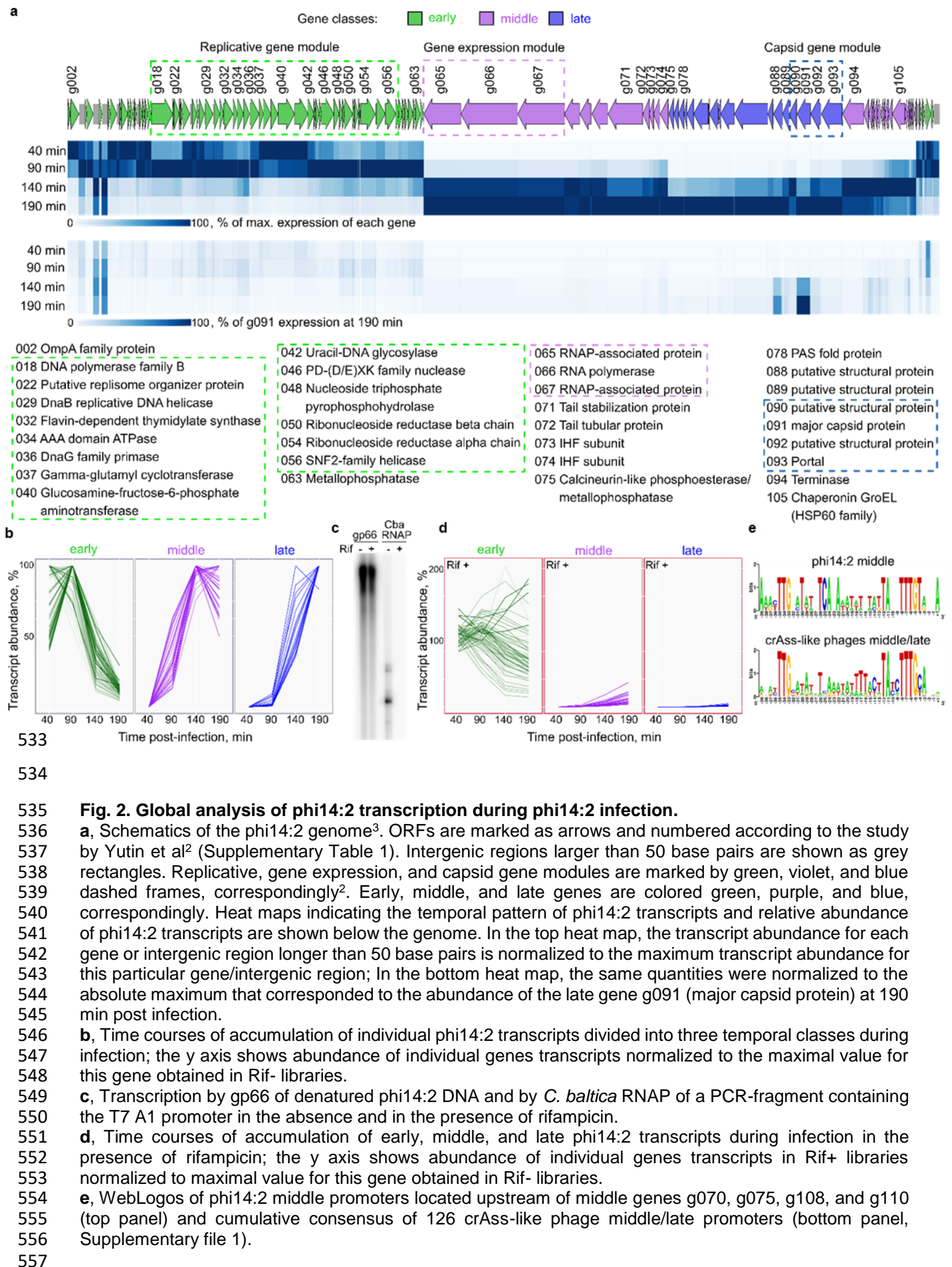
525 **b**, Transcription by gp66 of native and denatured genomic DNA of phage phi14:2.

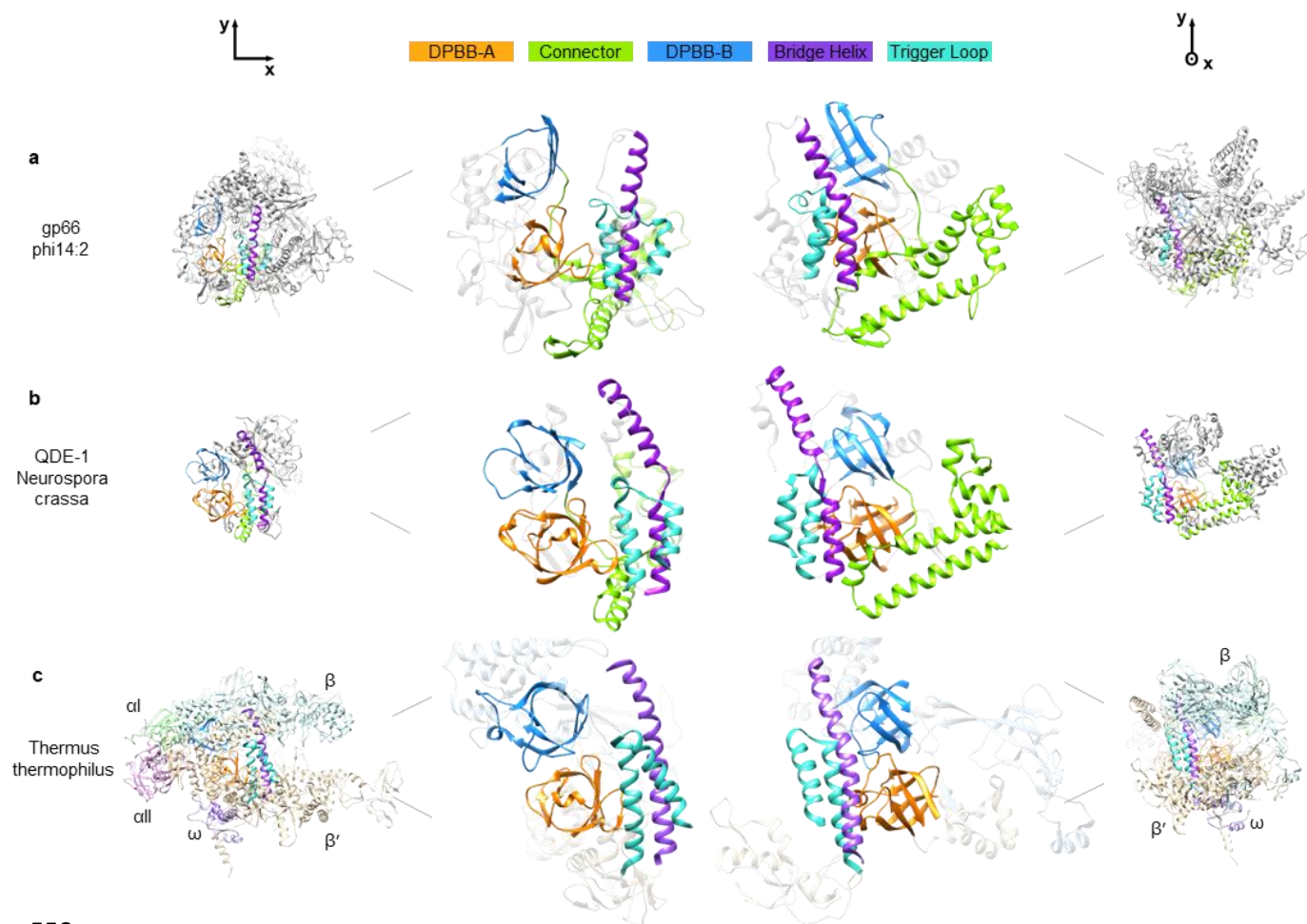
526 **c**, Completed transcription reactions of phi14:2 genomic DNA were treated with DNase RQ1 or
527 RNase T1 prior to loading on the gel.

528 **d**, Activity of gp66 requires Mg ions and ATP. Denatured genomic DNA of phi14:2 phage has
529 been used as a template.

530 **e**, Transcription of phi14:2 denatured genomic DNA by wild-type gp66 and gp66 mutants carrying
531 single alanine substitutions of each aspartate in the DFDID motif.

532



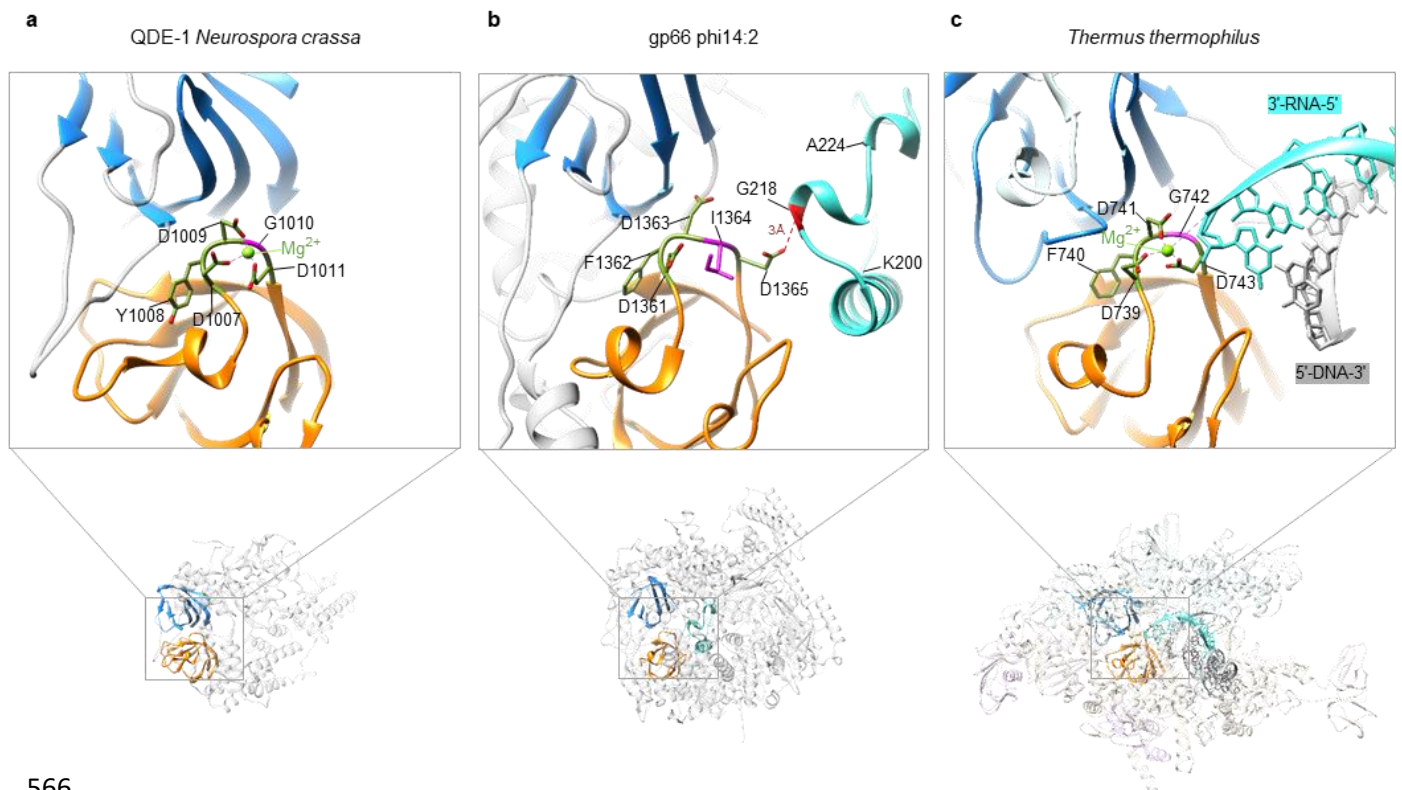


558

559 **Fig. 3. Phi14:2 RNAP gp66 is related to single- and multi-subunit RNAPs.**

560 **a**, **b**, and **c**, Crystal structures of phi14:2 gp66, QDE-1 from *N. crassa* (PDB 2J7N⁶), and *T.*
561 *thermophilus* RNAP (PDB ID 2O5J¹⁷) are shown as ribbon diagrams, respectively. Conserved
562 structural elements are colored according to the color code given above the top panels. Dissimilar
563 domains of QDE-1 and gp66 are shown in gray color. Each of the five subunits comprising the *T.*
564 *thermophilus* RNAP (αI , αII , β , β' , and ω) is rendered in a distinct color.

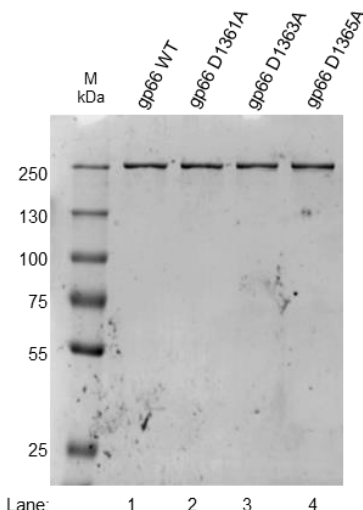
565



566

567 **Fig. 4. Cleft-blocking domain occupies the RNA-DNA hybrid binding site in phi14:2 RNAP**
568 **gp66.**

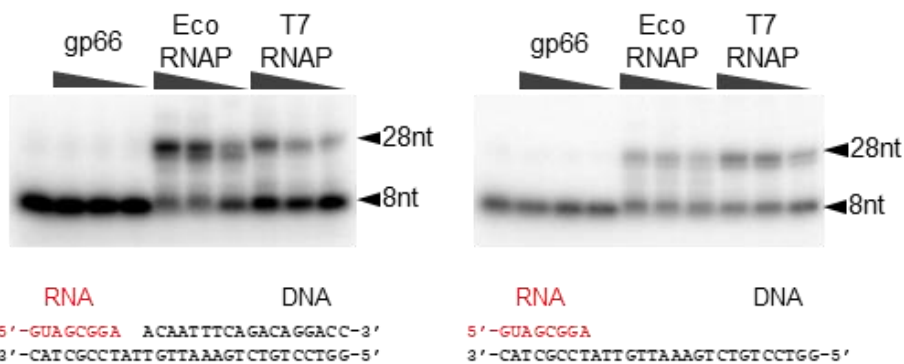
569 **a**, **b**, and **c**, The structure of the active site of QDE-1 from *N. crassa* (PDB 2J7N⁶), phi14:2
570 gp66, and *T. thermophilus* RNAP (PDB ID 2O5J¹⁷), respectively. The active site of phi14:2
571 RNAP gp66 (**b**) is in a conformation incompatible with Mg binding.
572



573
574
575
576

Extended data Fig. 1. SDS-PAGE analysis of wild-type gp66 and mutant gp66.

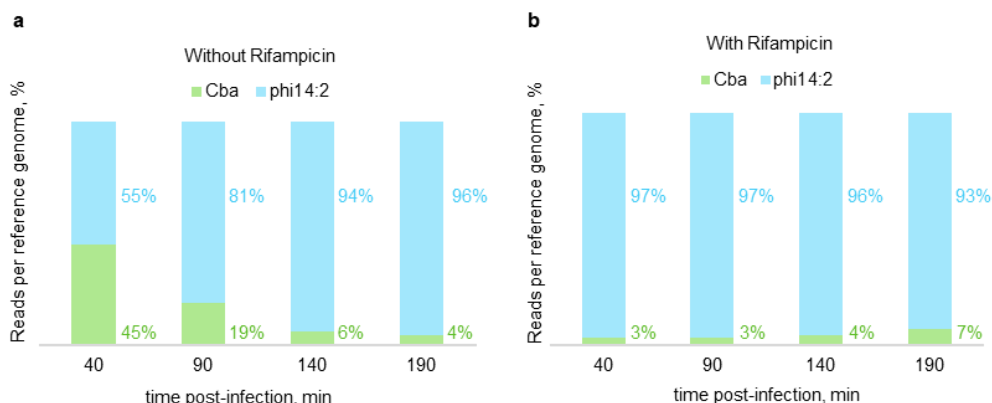
SDS-PAGE analysis of wild-type gp66 and gp66 mutants carrying single alanine substitutions of each aspartate in the DFDID motif purified by Heparin HP sepharose column chromatography.



577
578
579
580
581
582
583

Extended Data Fig. 2. Gp66 does not extend RNA primer in RNA-DNA scaffold.

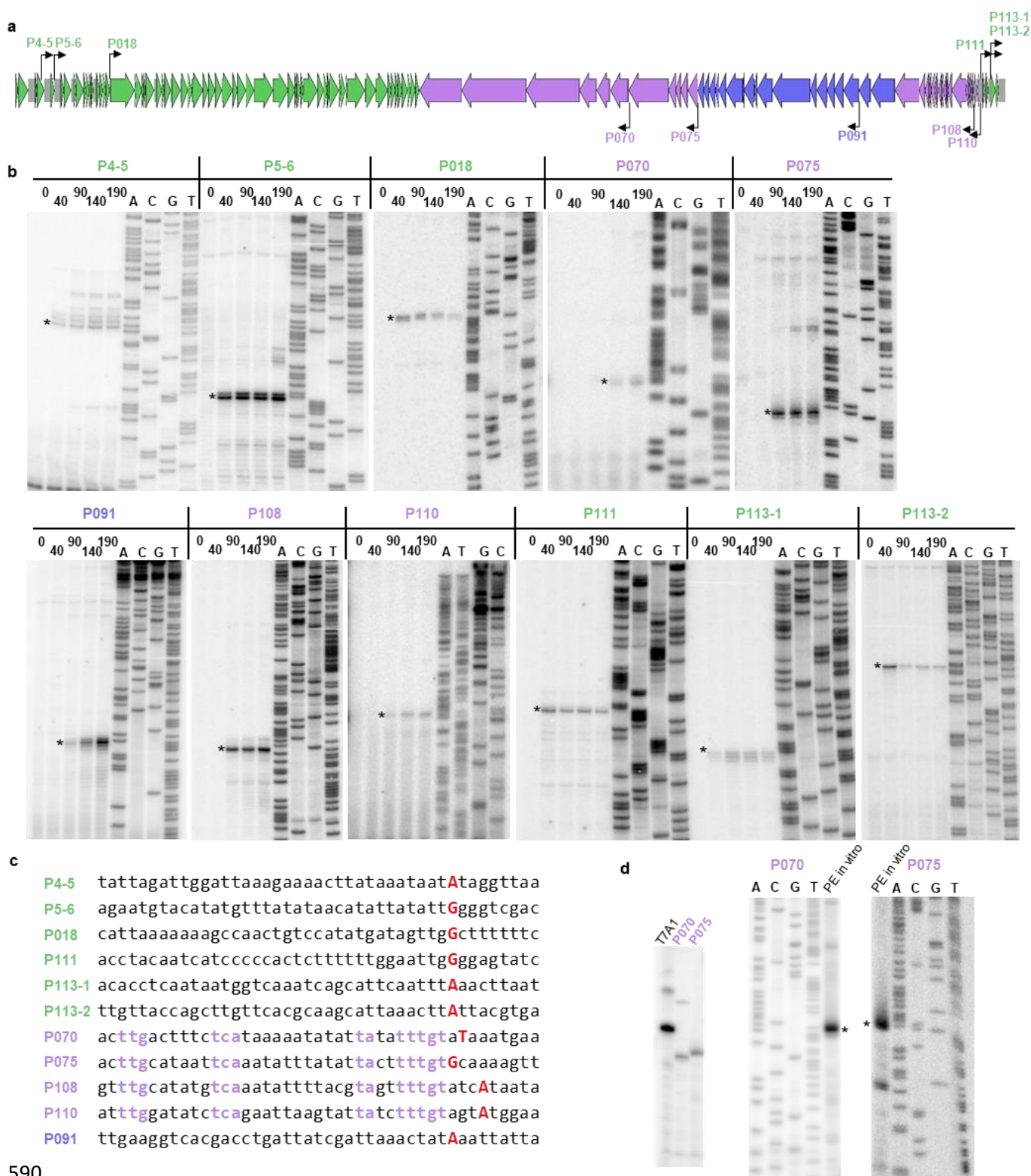
Extension of RNA primer in RNA-DNA scaffolds by gp66, *E. coli* (Eco) and T7 RNAPs as controls in the presence of ribonucleoside tri-phosphates. The sequences of RNA-DNA scaffolds used are shown under the gels; the RNA was radioactively labeled at the 5' end. The reaction products were resolved by electrophoresis in 16 % (w/v) denaturing 8 M urea polyacrylamide gel and revealed by autoradiography.



584
585
586
587
588
589

Extended Data Fig. 3. Distribution of phage and host transcript abundances.

The total number of reads per reference sequence aligned with a corresponding genome (Cba – *C. baltica* strain 14, this study; phi14:2 – NC_021806) is shown for the Rif⁻ libraries (a) and for the Rif⁺ libraries as stacked bars (b). The percentages are indicated next to the bars.

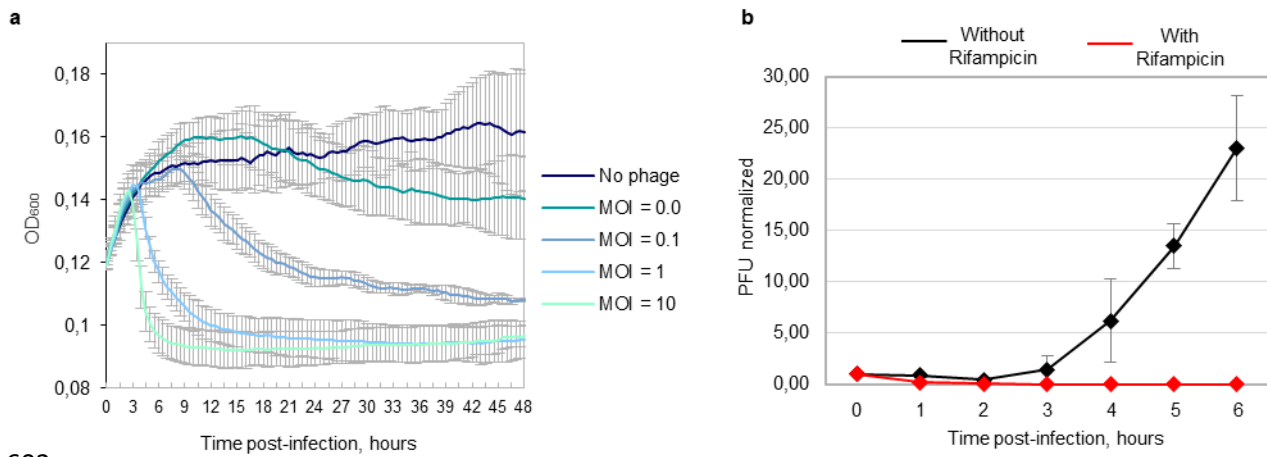


590

591

592 **Extended Data Fig. 4. Identification of putative promoters in phi14:2 genome.**

593 **a**, Schematic of the phi14:2 genome (see the legend of Fig. 2a for details) with putative promoters
594 marked by black arrows. **b**, Primer extension and sequencing reactions for eleven putative
595 promoters (Supplementary Table 4). Major primer extension products are marked with black
596 asterisks. **c**, Sequences flanking the primer extension endpoints are shown; nucleotides,
597 corresponding to primer extension endpoints are colored red. Conserved nucleotides of putative
598 middle promoters are shown in violet. **d**, Left panel: *In vitro* transcription of PCR-fragments
599 containing the T7 A1 promoter and predicted phi14:2 P070 and P075 promoters by *E. coli* RNAP;
600 Right panel: Primer extension reactions of RNA synthesized *in vitro* by *E. coli* RNAP from PCR-
601 fragments containing phi14:2 P070 and P075 promoters.



602

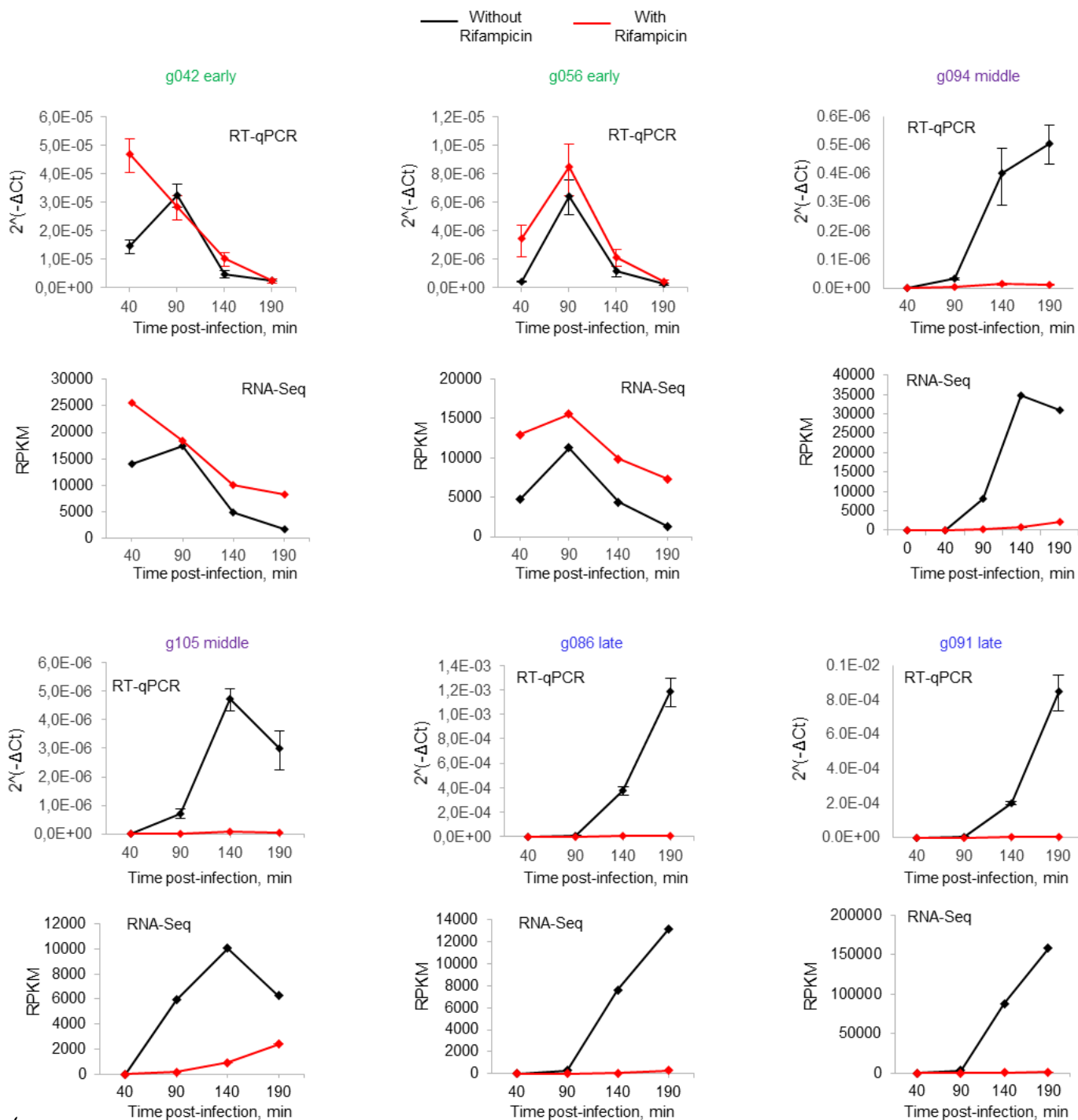
603

Extended Data Fig. 5. General parameters of phi14:2 infection.

604 **a**, Growth curves of *C. baltica* infected with phi14:2 at different MOIs in the log growth phase
605 (mean±SD of three biological replicates).

606 **b**, Single-burst curves of phi14:2 infecting *C. baltica* at MOI~0.5. Number of plaque forming units
607 (PFUs) normalized to the PFU immediately after the phage was added to the culture (0 time point)
608 (mean±SD of three biological replicas) are shown for cultures treated (red line) or not treated
609 (black line) with host RNAP inhibitor rifampicin (Rif) prior to infection.

610



612
 613 **Extended Data Fig. 6. Validation of RNA-Seq data by RT-qPCR.**
 614 Relative transcript abundances of six selected phi14:2 genes during the infection of *C. baltica*
 615 cells in the presence (red) and absence (black) of rifampicin were determined by RT-qPCR. The
 616 cycle threshold (Ct) values of the *C. baltica* 16S RNA were used to normalize the Ct values of
 617 selected phi14:2 transcripts as follows: $\Delta Ct = (\text{mean Ct gene}) - (\text{mean Ct 16S rRNA})$. The
 618 amplicon concentrations for different time points are plotted as $2^{(-\Delta Ct)}$ (mean \pm SD of three
 619 technical replicates). Corresponding RNA-Seq data are shown below the results of RT-qPCR.
 620

621 **Table 1. Data collection and refinement statistics.**

622

	Ta ₆ Br ₁₂ soaked	SeMet
Data collection		
Space group	P2 ₁ 2 ₁ 2	P2 ₁ 2 ₁ 2
Cell dimensions		
<i>a</i> , <i>b</i> , <i>c</i> (Å)	270.121, 299.343, 93.402	266.441, 297.181, 92.015
α, β, γ (°)	90.00, 90.00, 90.00	90.00, 90.00, 90.00
Resolution (Å)	50.0 – 3.75 (3.99 – 3.75) *	50.0 – 3.50 (3.71 – 3.50)
<i>R</i> _{merge}	0.151 (1.450)	0.176 (0.942)
<i>I</i> / σ <i>I</i>	13.04 (1.60)	11.90 (2.25)
Completeness (%)	98.7 (94.3)	99.7 (99.4)
Redundancy	13.85 (13.20)	7.90 (7.90)
Refinement		
Software (version)		Phenix.refine (1.41)
Resolution (Å)		50.0 – 3.50
No. reflections		177,613
<i>R</i> _{work} / <i>R</i> _{free}		0.191 / 0.244
No. atoms		
Protein		34,688
Ligand/ion		26
Water		0
<i>B</i> -factors		
Protein		88.81
Ligand/ion		72.90
R.m.s. deviations		
Bond lengths (Å)		0.003
Bond angles (°)		0.568

623

624

*Values in parentheses are for highest-resolution shell.

625 **Extended data Table 1. Absolutely conserved amino acids of RNAPs of crAss-like phages**
 626 **and their analogs in other RNAPs based on structural alignments**

627 Residue numbers are given for gp66 of phi14:2, QDE-1 RNAP of *Neurospora crassa* and RNAP
 628 of *Thermus thermophilus* (*T. th*). Light green-colored cells describe amino acids conserved in all
 629 three types of RNAPs; dark green-colored cells indicate amino acids conserved in crAss-like
 630 phage and multisubunit RNAPs; dark blue-colored cells show amino acids conserved in crAss-
 631 like phage RNAPs and QDE-1 RNAP; light blue-colored cells contain amino acids unique to
 632 crAss-like phage RNAPs.

633

No	<i>phi14:2</i> <i>gp66</i>	<i>QDE-1</i>	<i>T. th</i>	Function in canonical DNA-dependent RNAPs according to analysis by Lane and Darst²⁰ Residue numbers are indicated for <i>T. th</i> RNAP
1	Lys893	-	-	-
2	Arg894	Arg671	β -Arg557	β -Arg557 interacts with the γ -phosphate of the rNTP;
3	Asp962	Asp709	β -Asp686	β -Asp686 interacts with: 1) β -Asp739/Phe740/Asp741 of the β '-NADFDGD motif; 2) the γ -phosphate of rNTP and MgII in the active site; 3) β -Arg879, which also interacts with the rNTP γ -phosphate;
4	Lys1012	-	β -Lys838	β -Lys838 and β -Lys846 interact with the backbone of the RNA transcript at the -1/-2 positions;
5	Lys1027	Lys743	β -Lys846	
6	Lys1065	Lys767	-	-
7	Gln1116	Gln797	-	-
8	Gly1235	-	-	-
9	Arg1322	Arg962	β '-Arg704	β '-Arg704 interacts simultaneously with the O4' of rNTP, the 2'-OH of the RNA transcript at -1 position, and β '-Asn737, β '-Ala738, and β '-Asp743 of the β '-NADFDGD motif;
10	Pro1324	Pro964	β '-Pro706	β '-Pro706 lines the path for the template DNA around the +1 position;
11	Ser1330	-	-	-
12	Gly1359	Gly1005	β '-Asn737	Within the β '-NADFDGD motif, β '-Asn737 interacts with O2' and O3' of the rNTP;
13	Asp1361	Asp1007	β '-Asp739	β '-Asp739 interacts with Mgl and MgII, and with absolutely conserved β -Asp686;
14	Asp1363	Asp1009	β '-Asp741	β '-Asp741 interacts with Mgl, the RNA transcript at the -1 position, and β -Asp686;
15	Asp1365	Asp1011	β '-Asp743	β '-Asp743 interacts with Mgl and the RNA transcript at the -1 position.
16	Asp1612	Asp1116	-	-
17	Lys1615	Lys1119	-	-

634

- 635 1 Dutilh, B. E. *et al.* A highly abundant bacteriophage discovered in the unknown sequences of
636 human faecal metagenomes. *Nat Commun* **5**, 4498, doi:10.1038/ncomms5498 (2014).
- 637 2 Yutin, N. *et al.* Discovery of an expansive bacteriophage family that includes the most abundant
638 viruses from the human gut. *Nat Microbiol* **3**, 38-46, doi:10.1038/s41564-017-0053-y (2018).
- 639 3 Holmfeldt, K. *et al.* Twelve previously unknown phage genera are ubiquitous in global oceans.
640 *Proc Natl Acad Sci U S A* **110**, 12798-12803, doi:10.1073/pnas.1305956110 (2013).
- 641 4 Werner, F. & Grohmann, D. Evolution of multisubunit RNA polymerases in the three domains of
642 life. *Nat Rev Microbiol* **9**, 85-98, doi:10.1038/nrmicro2507 (2011).
- 643 5 Cogoni, C. & Macino, G. Gene silencing in *Neurospora crassa* requires a protein homologous to
644 RNA-dependent RNA polymerase. *Nature* **399**, 166-169, doi:10.1038/20215 (1999).
- 645 6 Salgado, P. S. *et al.* The structure of an RNAi polymerase links RNA silencing and transcription.
646 *PLoS Biol* **4**, e434, doi:10.1371/journal.pbio.0040434 (2006).
- 647 7 Shutt, T. E. & Gray, M. W. Bacteriophage origins of mitochondrial replication and transcription
648 proteins. *Trends Genet* **22**, 90-95, doi:10.1016/j.tig.2005.11.007 (2006).
- 649 8 Griesenbeck, J., Tschochner, H. & Grohmann, D. Structure and Function of RNA Polymerases and
650 the Transcription Machineries. *Subcell Biochem* **83**, 225-270, doi:10.1007/978-3-319-46503-6_9
651 (2017).
- 652 9 Werner, F. Structural evolution of multisubunit RNA polymerases. *Trends Microbiol* **16**, 247-250,
653 doi:10.1016/j.tim.2008.03.008 (2008).
- 654 10 Sauguet, L. The Extended "Two-Barrel" Polymerases Superfamily: Structure, Function and
655 Evolution. *J Mol Biol*, doi:10.1016/j.jmb.2019.05.017 (2019).
- 656 11 Vassylyev, D. G. *et al.* Crystal structure of a bacterial RNA polymerase holoenzyme at 2.6 Å
657 resolution. *Nature* **417**, 712-719, doi:10.1038/nature752 (2002).
- 658 12 Zhang, G. *et al.* Crystal structure of *Thermus aquaticus* core RNA polymerase at 3.3 Å resolution.
659 *Cell* **98**, 811-824 (1999).
- 660 13 Sidorenkov, I., Komissarova, N. & Kashlev, M. Crucial role of the RNA:DNA hybrid in the
661 processivity of transcription. *Mol Cell* **2**, 55-64 (1998).
- 662 14 Campbell, E. A. *et al.* Structural mechanism for rifampicin inhibition of bacterial rna polymerase.
663 *Cell* **104**, 901-912, doi:10.1016/s0092-8674(01)00286-0 (2001).
- 664 15 Paget, M. S. Bacterial Sigma Factors and Anti-Sigma Factors: Structure, Function and
665 Distribution. *Biomolecules* **5**, 1245-1265, doi:10.3390/biom5031245 (2015).
- 666 16 Guerin, E. *et al.* Biology and Taxonomy of crAss-like Bacteriophages, the Most Abundant Virus in
667 the Human Gut. *Cell Host Microbe* **24**, 653-664 e656, doi:10.1016/j.chom.2018.10.002 (2018).
- 668 17 Vassylyev, D. G. *et al.* Structural basis for substrate loading in bacterial RNA polymerase. *Nature*
669 **448**, 163-168, doi:10.1038/nature05931 (2007).
- 670 18 Wang, D., Bushnell, D. A., Westover, K. D., Kaplan, C. D. & Kornberg, R. D. Structural basis of
671 transcription: role of the trigger loop in substrate specificity and catalysis. *Cell* **127**, 941-954,
672 doi:10.1016/j.cell.2006.11.023 (2006).
- 673 19 Cramer, P., Bushnell, D. A. & Kornberg, R. D. Structural basis of transcription: RNA polymerase II
674 at 2.8 angstrom resolution. *Science* **292**, 1863-1876, doi:10.1126/science.1059493 (2001).
- 675 20 Lane, W. J. & Darst, S. A. Molecular evolution of multisubunit RNA polymerases: structural
676 analysis. *J Mol Biol* **395**, 686-704, doi:10.1016/j.jmb.2009.10.063 (2010).
- 677 21 Weinzierl, R. O. The Bridge Helix of RNA polymerase acts as a central nanomechanical
678 switchboard for coordinating catalysis and substrate movement. *Archaea* **2011**, 608385,
679 doi:10.1155/2011/608385 (2011).
- 680 22 Krissinel, E. & Henrick, K. Secondary-structure matching (SSM), a new tool for fast protein
681 structure alignment in three dimensions. *Acta Crystallogr D Biol Crystallogr* **60**, 2256-2268,
682 doi:10.1107/S0907444904026460 (2004).
- 683 23 Holm, L. Benchmarking Fold Detection by DaliLite v.5. *Bioinformatics*,
684 doi:10.1093/bioinformatics/btz536 (2019).
- 685 24 Conway, J. F., Duda, R. L., Cheng, N., Hendrix, R. W. & Steven, A. C. Proteolytic and
686 conformational control of virus capsid maturation: the bacteriophage HK97 system. *J Mol Biol*
687 **253**, 86-99, doi:10.1006/jmbi.1995.0538 (1995).

- 688 25 Izaguirre, G. The Proteolytic Regulation of Virus Cell Entry by Furin and Other Proprotein
689 Convertases. *Viruses* **11**, doi:10.3390/v11090837 (2019).
- 690 26 Konvalinka, J., Krausslich, H. G. & Muller, B. Retroviral proteases and their roles in virion
691 maturation. *Virology* **479-480**, 403-417, doi:10.1016/j.virol.2015.03.021 (2015).
- 692 27 Shkoporov, A. N. *et al.* PhiCrAss001 represents the most abundant bacteriophage family in the
693 human gut and infects *Bacteroides intestinalis*. *Nat Commun* **9**, 4781, doi:10.1038/s41467-018-
694 07225-7 (2018).
- 695 28 Iyer, L. M., Koonin, E. V. & Aravind, L. Evolutionary connection between the catalytic subunits of
696 DNA-dependent RNA polymerases and eukaryotic RNA-dependent RNA polymerases and the
697 origin of RNA polymerases. *BMC Struct Biol* **3**, 1 (2003).
- 698 29 Shabalina, S. A. & Koonin, E. V. Origins and evolution of eukaryotic RNA interference. *Trends Ecol*
699 *Evol* **23**, 578-587, doi:10.1016/j.tree.2008.06.005 (2008).
- 700 30 Aalto, A. P., Poranen, M. M., Grimes, J. M., Stuart, D. I. & Bamford, D. H. In vitro activities of the
701 multifunctional RNA silencing polymerase QDE-1 of *Neurospora crassa*. *J Biol Chem* **285**, 29367-
702 29374, doi:10.1074/jbc.M110.139121 (2010).
- 703 31 Lee, H. C. *et al.* The DNA/RNA-dependent RNA polymerase QDE-1 generates aberrant RNA and
704 dsRNA for RNAi in a process requiring replication protein A and a DNA helicase. *PLoS Biol* **8**,
705 doi:10.1371/journal.pbio.1000496 (2010).
- 706 32 Holmfeldt, K., Middelboe, M., Nybroe, O. & Riemann, L. Large variabilities in host strain
707 susceptibility and phage host range govern interactions between lytic marine phages and their
708 *Flavobacterium* hosts. *Appl Environ Microbiol* **73**, 6730-6739, doi:10.1128/AEM.01399-07
709 (2007).
- 710 33 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence
711 data. *Bioinformatics* **30**, 2114-2120, doi:10.1093/bioinformatics/btu170 (2014).
- 712 34 Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif.
713 *Bioinformatics* **27**, 1017-1018, doi:10.1093/bioinformatics/btr064 (2011).
- 714 35 Laslett, D. & Canback, B. ARAGORN, a program to detect tRNA genes and tmRNA genes in
715 nucleotide sequences. *Nucleic Acids Res* **32**, 11-16, doi:10.1093/nar/gkh152 (2004).
- 716 36 Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site
717 identification. *BMC Bioinformatics* **11**, 119, doi:10.1186/1471-2105-11-119 (2010).
- 718 37 Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence logo generator.
719 *Genome Res* **14**, 1188-1190, doi:10.1101/gr.849004 (2004).
- 720 38 Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and
721 hybridization array data repository. *Nucleic Acids Research* **30**, 207-210,
722 doi:10.1093/nar/30.1.207 (2002).