doi:10.1016/j.jmb.2010.10.049 J. Mol. Biol. (2011) 405, 125–142

Contents lists available at www.sciencedirect.com

Journal of Molecular Biology

journal homepage: http://ees.elsevier.com.jmb

Temporal Regulation of Gene Expression of the *Thermus thermophilus* Bacteriophage P23-45

Zhanna Berdygulova[1,2], Lars F. Westblade[3], Laurence Florens[4], Eugene V. Koonin[5], Brian T. Chait[3], Erlan Ramanculov[2], Michael P. Washburn[4,6], Seth A. Darst[3], Konstantin Severinov[1,7,8]* and Leonid Minakhin[1]*

[1]*Waksman Institute for Microbiology, Piscataway, NJ 08854, USA*
[2]*National Center for Biotechnology of Republic of Kazakhstan, Kazakhstan*
[3]*The Rockefeller University, New York, NY 10065, USA*
[4]*Stowers Institute for Medical Research, Kansas City, MO 64110, USA*
[5]*NCBI, NLM, NIH, Bethesda, MD 20894, USA*
[6]*Department of Pathology and Laboratory Medicine, The University of Kansas Medical Center, Kansas City, KS 66160, USA*
[7]*Department of Molecular Biology and Biochemistry, Rutgers, the State University of New Jersey, Piscataway, NJ 08854, USA*
[8]*Institute of Molecular Genetics, Russian Academy of Sciences, Moscow 123182, Russia*

Received 30 August 2010; received in revised form 25 October 2010; accepted 27 October 2010
Available online 2 November 2010

Edited by M. Gottesman

Keywords:
Thermus thermophilus; thermophage; phage promoters; RNA polymerase; RNAP-binding proteins

Regulation of gene expression during infection of the thermophilic bacterium *Thermus thermophilus* HB8 with the bacteriophage P23-45 was investigated. Macroarray analysis revealed host transcription shut-off and identified three temporal classes of phage genes; early, middle and late. Primer extension experiments revealed that the 5′ ends of P23-45 early transcripts are preceded by a common sequence motif that likely defines early viral promoters. *T. thermophilus* HB8 RNA polymerase (RNAP) recognizes middle and late phage promoters *in vitro* but does not recognize early promoters. *In vivo* experiments revealed the presence of rifampicin-resistant RNA polymerizing activity in infected cells responsible for early transcription. The product of the P23-45 early gene 64 shows a distant sequence similarity with the largest, catalytic subunits of multisubunit RNAPs and contains the conserved metal-binding motif that is diagnostic of these proteins. We hypothesize that ORF64 encodes rifampicin-resistant phage RNAP that recognizes early phage promoters. Affinity isolation of *T. thermophilus* HB8 RNAP from P23-45-infected cells identified two phage-encoded proteins, gp39 and gp76, that bind the host RNAP and inhibit *in vitro* transcription from host promoters, but not from middle or late phage promoters, and may thus control the shift from host to viral gene expression during infection. To our knowledge, gp39 and gp76 are the first characterized bacterial RNAP-binding proteins encoded by a thermophilic phage.

**Corresponding authors.* E-mail addresses: severik@waksman.rutgers.edu; minakhin@waksman.rutgers.edu.
Abbreviations used: RNAP, RNA polymerase; DPBB, double-φ β-barrel.

Introduction

Transcription is the first step and primary regulatory determinant of gene expression. Multisubunit DNA-dependent RNA polymerases (RNAPs) are complex, highly regulated molecular machines. RNAP, alone or in complex with regulatory factors, is central to the transcription process. Every stage of transcription is regulated via RNAP interactions with various transcription factors. Bacteriophages (phages) have evolved highly effective mechanisms to modify bacterial RNAP to serve the needs of the phage.[1] Recent studies indicate that phages are the most abundant life form in the biosphere,[2,3] and that the phage gene pool is the largest source of natural gene diversity. Consequently, while many phages use common strategies to subjugate their hosts, the number of phage-encoded regulatory mechanisms is virtually unlimited.[4] At the time of writing, more than 580 complete phage genome sequences (NCBI, last modified July 2010) had been determined. Comparative genomic analysis provides important insights into the diversity and evolution of phages and their hosts. However, our understanding of gene expression regulation strategies utilized by phages is relatively limited. Classical studies of gene regulation mechanisms in a handful of model *Escherichia coli* phages (e.g., λ, T4 and T7) have led to key discoveries in molecular biology. More recently, transcription profiling and bioinformatic predictions have been successfully applied to aid our understanding of the molecular features of phage regulatory networks.[5–8] The ultimate understanding of transcription regulatory mechanisms is based on structure–function analysis of RNAP alone, and RNAP–regulator complexes. However, a high-resolution structure of *E. coli* RNAP is absent; therefore, almost no structural information on the action of transcription factors encoded by model *E. coli* phages in complex with *E. coli* RNAP has been obtained. Currently, only the structures of bacterial RNAP from the thermophilic eubacteria *aquaticus*[9–11] and *Thermus thermophilus*[12,13] have been determined. Thus, at present, the most reasonable way to structurally investigate phage-driven prokaryotic transcription regulation is to study phages that infect thermophilic eubacteria (i.e., thermophages). Despite the advances in our understanding of the structures of thermophilic eubacterial RNAPs, there is insufficient knowledge about thermophages, in particular their biology and the gene regulation mechanisms they employ during host bacterial infection. The genomes of several thermophages infecting *Thermus* species have been sequenced completely,[14–16] and the gene expression strategy of one of these phage, ϕYS40, has been investigated in detail.[8]

In this work, we studied the temporal regulation of transcription of another *T. thermophilus* HB8 phage, P23-45, whose genome has recently been sequenced.[15] We identified three temporal classes of P23-45 genes and their corresponding promoters by a combination of gene macroarray, *in vivo* primer extension and *in vitro* transcription experiments. P23-45 middle and late promoters have consensus elements that differ from those of *T. thermophilus* housekeeping promoters and ϕYS40 thermophage promoters.[8] Yet, P23-45 middle and late promoters are recognized by unmodified host *T. thermophilus* RNAP *in vitro* and *in vivo*. In contrast, P23-45 early promoters are not recognized by host *T. thermophilus* RNAP either *in vitro* or *in vivo*. The early promoters are defined by an unusual 11 bp conserved sequence motif, which is likely recognized by a unique phage-encoded RNAP. Affinity isolation of *T. thermophilus* RNAP from P23-45-infected cells identified two phage-encoded *T. thermophilus* RNAP-binding proteins, gp76 and gp39, which are the products of an early and a middle gene, respectively. These proteins bind the host RNAP *in vitro* and efficiently inhibit transcription from host bacterial promoters but not from middle or late phage promoters. Thus, these proteins might be responsible for the shut-off of host transcription and therefore the switch from host to viral gene expression. To our knowledge, this is the first description of thermophage-encoded thermophilic bacterial RNAP-binding proteins.

Results

Macroarray analysis of gene expression during P23-45 infection

Almost half of the P23-45 genome (ORFs 1–78) is transcribed in the same direction (leftward in Fig. 1a). These genes form a single cluster in the left arm of the genome. The remaining P23-45 ORFs (ORFs 79–117) are transcribed in the rightward direction. These ORFs form a single cluster at the right arm of the genome. The only exception is ORF5, which is transcribed in the rightward direction but is located in the left arm of the genome (Fig. 1a). To characterize the temporal profile of P23-45 gene expression, a macroarray of P23-45 phage genes was prepared. The array contained spots with equal amounts of PCR-amplified DNA fragments of 20 P23-45 genes and one non-coding region of the P23-45 genome (marked by black dots in Fig. 1a and b). The genes chosen for the array encode proteins of different functional classes. One group of spots represented the abundance of mRNA from a cluster of small genes with unknown functions from the left arm (genes 57, 68, 69, 76, 78 and the non-coding region of triplex-forming mirror repeats).[15] The second group of spots represented left arm genes involved in DNA replication, recombination and nucleotide metabolism (genes 4, 5, 11, 14, 24, 39 and 46). The third group of spots on the array

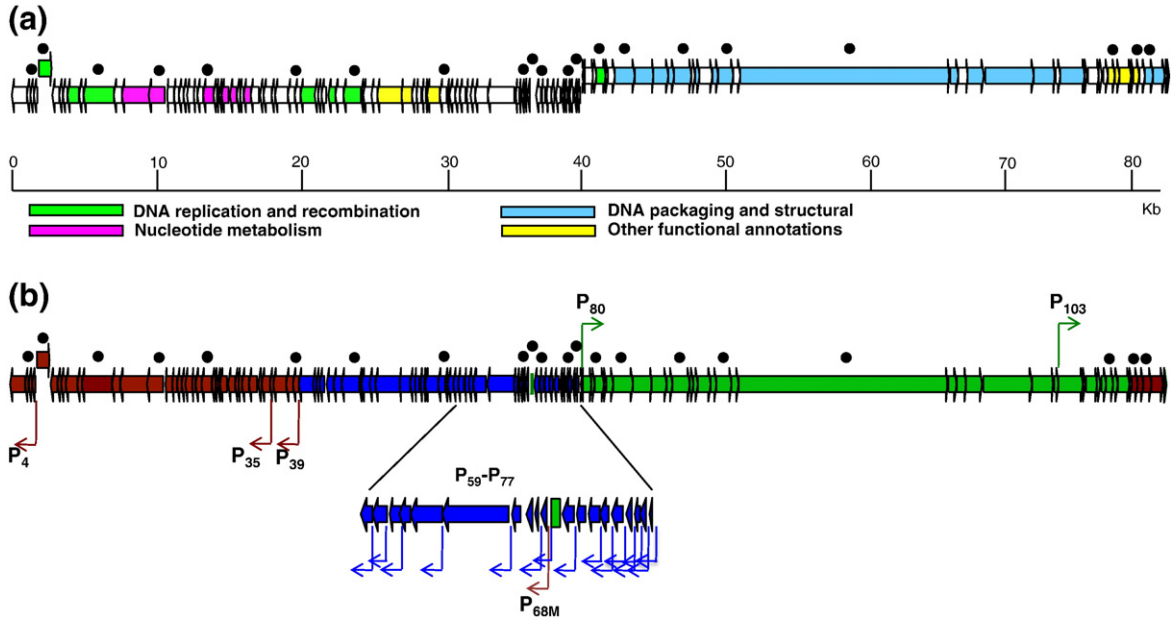

Fig. 1. Genome and transcription map of the *T. thermophilus* HB8 bacteriophage P23-45. (a) Functional annotation of the predicted P23-45 ORFs. The different colors of the arrows represent the different functions assigned for the proteins encoded by P23-45. Upward pointing arrows denote a rightward orientation of the genes and downward pointing arrows denote a leftward orientation of the genes. (b) The different colors of the transcription map indicate different temporal classes of genes: blue, early; brown, middle; green, late. Promoters are colored corresponding to their temporal class and are depicted as bent arrows. A section of the early cluster is shown in more detail. In both (a) and (b) black dots indicate genes used in gene macroarray analysis of the temporal transcription program.

represented right arm genes that encode the P23-45 virion structural proteins (genes 82, 89, 94, 96 and 114), and predicted DNA packaging (gene 85) and lysis proteins (genes 108 and 112). Closely spaced or partially overlapping genes are usually transcribed from the same promoter and some spots on the array are likely to report the abundance of transcripts of multiple P23-45 genes.

In order to determine whether P23-45 shuts off host transcription, PCR-amplified DNA fragments of seven housekeeping *T. thermophilus* HB8 genes were spotted on the membrane; *rpoC* (encoding the RNAP β′ subunit), *sigA* (encoding the primary sigma factor σ^A), *dnaK* (encoding a chaperone), *TTHA0466* (encoding alcohol dehydrogenase), *infB* and *infC* (encoding translation initiation factors IF2 and IF3, respectively) and *rpsA* (encoding the ribosomal protein S1). The array also contained spots with total *T. thermophilus* HB8 genomic DNA, P23-45 genomic DNA, and a PCR-amplified DNA fragment of the *Drosophila melanogaster zfrp8* gene that was used as a normalizing and loading control.

T. thermophilus HB8 cells were infected with P23-45 at a multiplicity of infection of 10 and total RNA was extracted 5, 20, and 40 min post infection. As a control, RNA was extracted from *T. thermophilus* HB8 cells immediately before P23-45 infection. Equal amounts of total RNA were used to generate radioactively labeled cDNA by random priming/reverse transcription followed by hybridization to the macroarray membrane. The amount of radioactivity hybridized to different spots of the macroarray reflected the abundance of transcripts of each corresponding gene. Three independent macroarray experiments were performed for each time point. To quantitatively analyze the macroarray data, the radioactive signal from each spot was corrected for background and normalized according to the relative strength of the signal from the *D. melanogaster zfrp8* spot. The mean amount of radioactivity for each macroarray spot was plotted as a function of time post infection (Fig. 2). As expected, the total amount of P23-45 transcripts increased with time post infection relative to the control *zfrp8* spots (Fig. 2a). In contrast, the total amount of *T. thermophilus* HB8 transcripts decreased throughout the same period (Fig. 2a), indicating that P23-45 either executes host transcription shut-off or increases the rate of host RNA decay. The behavior of individual host transcripts during P23-45 infection was complex; while the abundance of most transcripts decreased throughout the infection cycle, the rates at which the transcripts decreased varied and in one case (*sigA*) the transcript abundance was unchanged (data not shown). The reasons for the observed differences in the abundance of the host

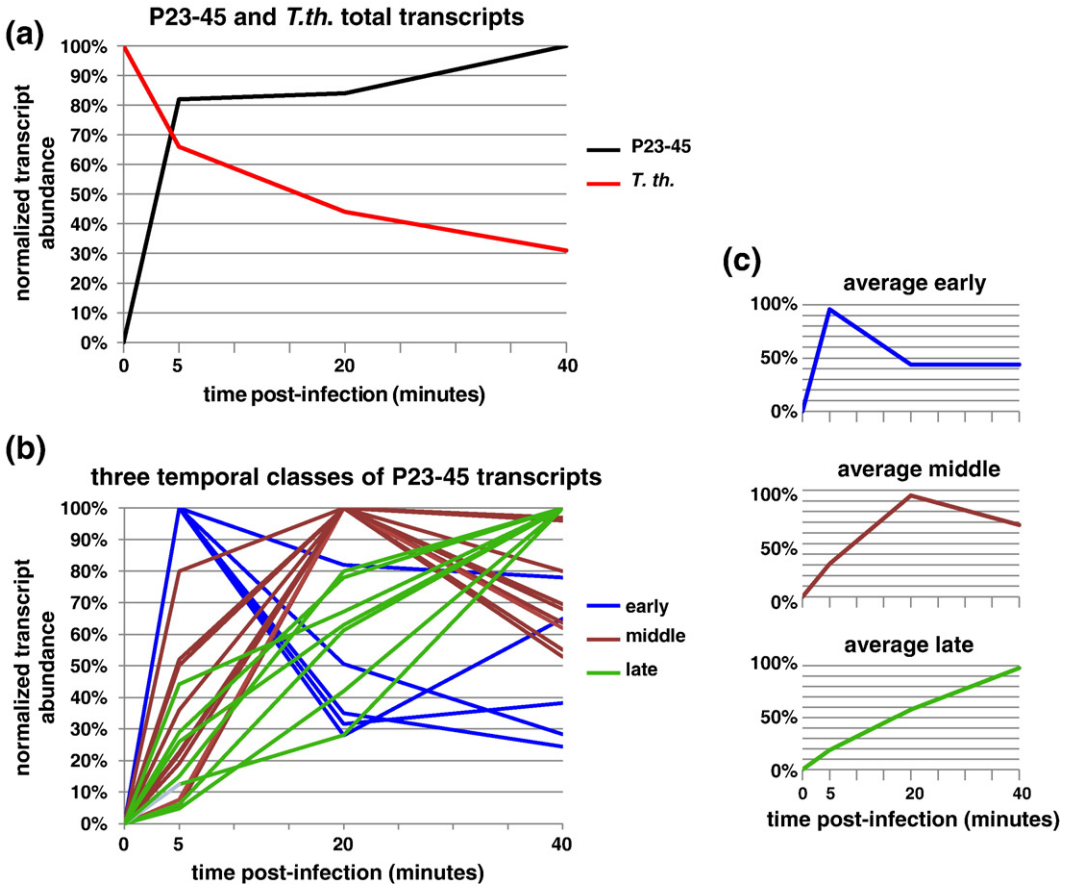

Fig. 2. Macroarray data analysis. (a) Normalized abundances of total *T. thermophilus* HB8-encoded transcripts and total P23-45-encoded transcripts are shown as red and black lines, respectively. (b) Normalized abundances for individual P23-45 transcripts are presented. Transcripts that belong to different temporal classes are colored: blue, early; brown, middle; green, late. The expression curves represent the average of three independent experiments. (c) Average normalized abundances of P23-45 transcripts of different temporal classes.

transcripts were not investigated; however, one explanation could be the interplay between the rates of host transcript synthesis and stability in the infected cell.

To compare the behavior of individual P23-45 transcripts, plots of normalized spot signal intensities as a function of time post-infection were scaled to equalize the mean abundance of each transcript. The accumulation of individual transcripts peaked at different times during the infection cycle, indicating the presence of different temporal classes of phage genes (Fig. 2b). Phage genes were clustered on the basis of the time when their transcripts became most abundant.[17] Analysis of transcript abundance patterns revealed three distinct temporal classes of genes; early, middle and late (Fig. 2b). The early class gene transcripts peaked at 5 min post infection, while middle class gene transcripts peaked at 20 min post infection. Finally, the abundance of transcripts of the late class of genes increased dramatically by the end of infection (under our conditions the eclipse period of P23-45 was 35 min and lysis of *T. thermophilus* HB8 by the phage was complete by ~60 min post infection). The average value of the scaled abundances calculated for each of the three temporal classes are shown as separate panels in Fig. 2c.

P23-45 genes used in the macroarray analysis and the temporal classes they belong to are indicated in Fig. 1 by black dots and different colors, respectively. Most centrally located genes with unknown functions belong to the early class; however, gene 78 and the non-coding triplex-forming region located in this part of the P23-45 genome belong to the middle and the late classes, respectively. Some genes encoding DNA replication and recombination components also belong to the early temporal class. Other genes from this functional group, as well as genes encoding nucleotide metabolism enzymes, comprise the middle class. As expected, genes of the late class were found exclusively in the right arm of the genome. They encode the P23-45 virion

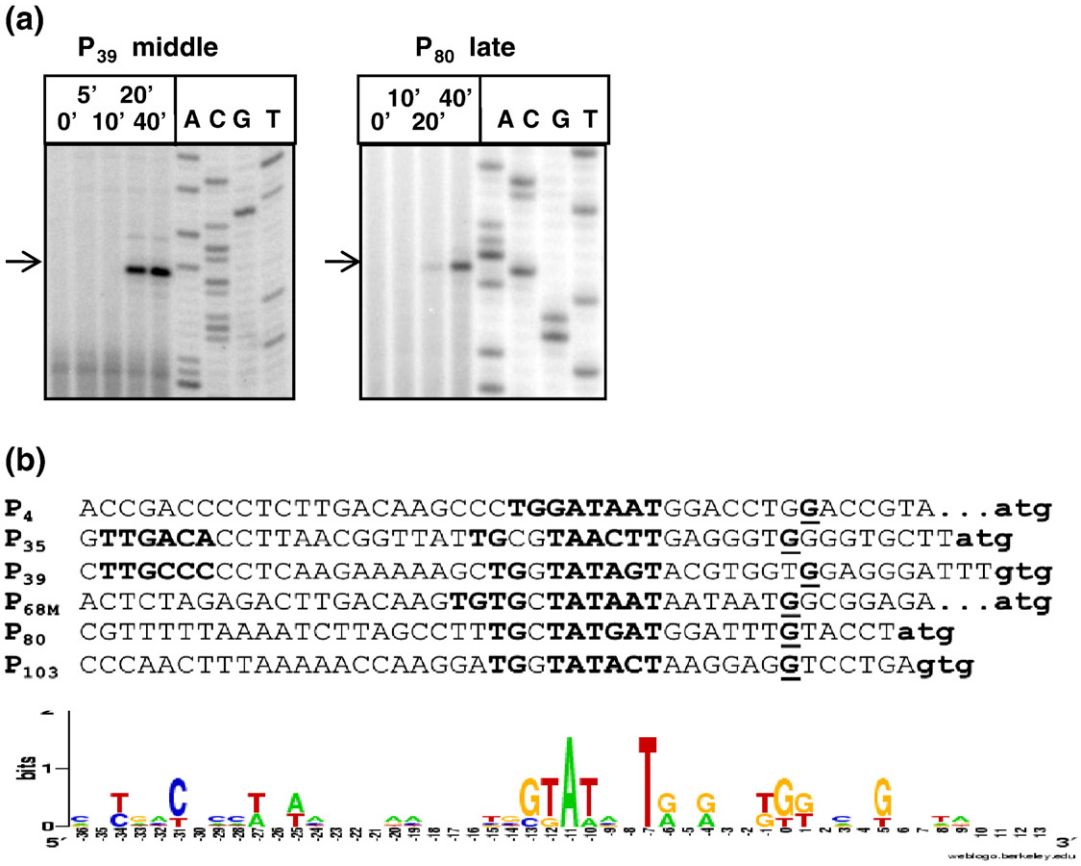

Fig. 3. Characterization of middle and late P23-45 promoters. (a) The kinetics of accumulation of *in vivo* primer extension products obtained with P23-45 transcripts from representative middle (P_{39}) and late (P_{80}) promoters during infection; the DNA products corresponding to 5′ RNA ends are labeled with arrows. (b) An alignment of P23-45 middle and late promoter sequences is shown. The putative –35 (where identified), –10 and TG/TGTG promoter elements are highlighted in uppercase bold. Experimentally identified transcription start sites are in bold and underlined. The annotated translation initiation codons are shown in lowercase bold. The corresponding sequence logos for P23-45 middle and late promoters are depicted below the alignment. The size of the letters indicates the degree of conservation; positions are defined with respect to identified transcription start sites.

structural proteins, DNA packaging proteins and analysis proteins.

Mapping P23-45 promoters

The previous automated annotation of P23-45 genes suggested that P23-45 does not encode its own RNAP and must therefore rely on host RNAP to transcribe its early, middle, and late genes throughout the infection.[15] Such temporal regulation can be achieved either by using specific sequences defining promoters of different temporal classes and/or by modification of RNAP promoter specificity by phage-encoded transcription factors. In the following section, we report our analysis of the P23-45 middle and late promoters followed by a discussion of the analysis of the P23-45 early promoters.

P23-45 middle and late promoters

Non-coding P23-45 regions upstream of middle and late phage genes were examined by primer extension. Five primer extension products corresponding to the 5′ RNA ends of three middle (P_4, P_{35} and P_{39}) and two late (P_{80} and P_{103}) putative promoters were detected. Additionally, one middle promoter, P_{68M}, located in the early gene cluster was identified by primer extension. The kinetics of primer extension product accumulation for these promoters during P23-45 infection matched the macroarray data for middle and late genes. Representative primer extension experiments with primers specific to middle and late genes are shown in Fig. 3a.

To confirm that the *in vivo* identified 5′ RNA end points are transcription start points, we tested the

ability of P23-45 genomic DNA fragments containing putative middle and late P23-45 promoters to serve as templates for *in vitro* abortive transcription initiation with purified *T. thermophilus* HB8 σ^A-associated holoenzyme. For each promoter tested, combinations of nucleotide substrates that should have permitted transcription initiation (based on the results of *in vivo* primer extension analysis) were used. In all cases, robust transcription was detected†; therefore, we conclude that *in vivo* primer extension products correspond to transcription start points and that unmodified host RNAP-σ^A holoenzyme can recognize P23-45 middle and late promoters *in vitro*.

Comparisons of sequences upstream of the transcription start points of P23-45 middle and late transcripts revealed a motif that was common to promoters of both classes (Fig. 3b). The middle and late P23-45 promoters are characterized by a –10-like element (consensus sequence 5′-GTATanT-3′) with the highest conservation at positions –11 (A) and –7 (T) relative to the experimentally determined transcription start point (Fig. 3b). In addition, an extended –10 "TG/TGTG" motif is present 0–2 bp upstream of the –10 element. In two cases, appropriately positioned motifs similar to the consensus *T. thermophilus* –35 promoter element were identified in two middle phage promoters, P_{35} and P_{39} (Fig. 3b; see Fig. 4 for the consensus *T. thermophilus* –35 promoter element). Our failure to differentiate between the middle and late P23-45 promoter sequences might be due to the small number of promoter sequences examined. Alternatively, the distinct temporal patterns of activity from these promoters might not be caused by differences in basal promoter elements or by binding of transcription factors to specific regulatory sites, but by the phage-dependent modification of host *T. thermophilus* HB8 RNAP and/or differences in intrinsic promoter strengths.

P23-45 early promoters

On the basis of earlier studies of other phages, one would expect that transcription of early P23-45 genes should be driven by strong promoters recognized by the housekeeping form of the host RNAP holoenzyme, the *T. thermophilus* HB8 σ^A-associated holoenzyme. Early phage promoters need to be strong to compete efficiently with host promoters for host RNAP. Thus, early phage promoters would be expected to have a good match to the host σ^A-associated holoenzyme consensus promoter sequence. Indeed, such an expectation was fulfilled by the phage ϕYS40, a *T. thermophilus* HB8 phage that has been studied in our laboratory.[8]

† For example, see Fig. 5b, lanes 14 and 15.

We utilized a *T. thermophilus* RNAP-σ^A holoenzyme promoter bioinformatic profile[8] to search P23-45 DNA upstream of early P23-45 genes. To our surprise, with the exception of a likely σ^A-dependent promoter upstream of gene 68 located in the middle of the early cluster (P_{68E}), no high-matching candidate sequence was found. Visual inspection of non-coding regions separating P23-45 early genes revealed the presence of a common 11 bp sequence motif (5′-TTATTCcTTTA-3′) located immediately upstream of annotated start codons (Fig. 4). Copies of this motif were identified upstream of early genes 59–61, 63, 64, 67–69 and 71–77. No additional copy of the 11 bp motif is present in any other region of the P23-45 genome. The logo of the motif is shown below the alignment in Fig. 4 and compared to the *T. thermophilus* promoter consensus logo. As can be seen, the two logos are clearly distinct from each other.

We hypothesized that this 11 bp motif defines early P23-45 promoters. To test this hypothesis, RNA samples used in macroarray experiments were subjected to primer extension analysis using primers specific to eight genes from the early gene cluster (see the alignment in Fig. 4, where the experimentally identified 5′ ends are underlined). The result was identical in all cases; a primer extension product whose end corresponds to the last nucleotide of the 11 bp motif. Moreover, the kinetics of primer extension product accumulation for early P23-45 genes was in agreement with the macroarray data. Representative primer extension experiments with primers specific to genes 64 and 68 are shown in Fig. 5a.

The results presented above suggest that the 11 bp motif defines the early P23-45 promoters that are obviously distinct from known host or phage promoters. Alternatively, the 11 bp motif could be the site of post-transcriptional processing of an early P23-45 polycistronic transcript(s) that initiates elsewhere upstream. The following experiments were done to distinguish between these possibilities. First, putative P23-45 early promoters whose transcription start sites had been identified *in vivo* were tested in an *in vitro* abortive transcription initiation assay using purified *T. thermophilus* HB8 RNAP-σ^A holoenzyme. For each putative promoter tested, combinations of nucleotide substrates that should have permitted transcription initiation (based on the results of *in vivo* primer extension analysis) were used. Unexpectedly, *T. thermophilus* RNAP-σ^A holoenzyme was either unable to transcribe or yielded only small amounts of product for almost every transcription template tested (Fig. 5b, lanes 1–3 and 6–13); the only exception was P_{68E} (lanes 4 and 5). However, robust transcription from this promoter could be explained by the fact that the 11 bp motif is embedded in a recognizable σ^A-dependent promoter (as revealed by bioinformatic analysis; Fig. 4 alignment, the –35-like and –10-like

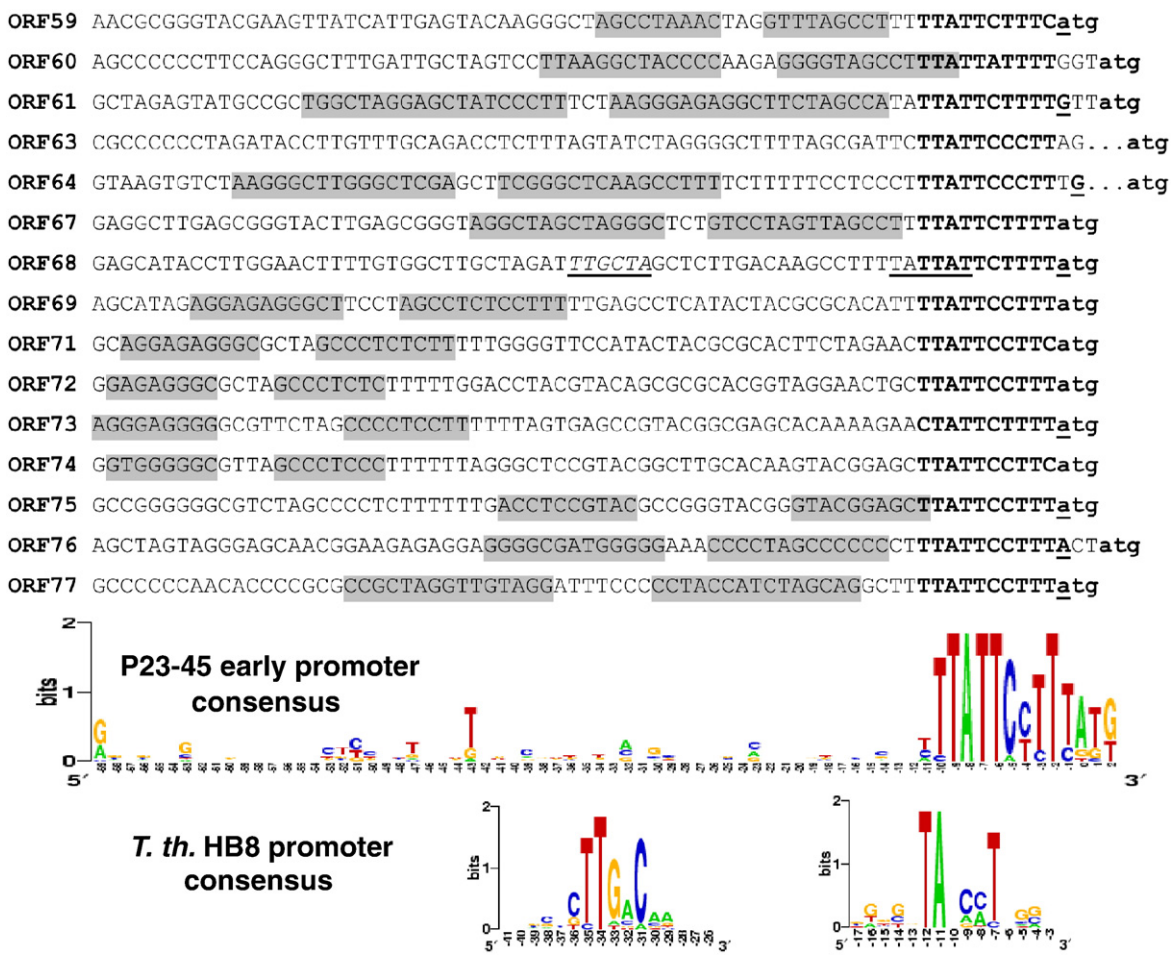

Fig. 4. Alignment of DNA sequences upstream of the annotated translation initiation start codons of P23-45 early genes. The novel promoter elements are in uppercase bold. Experimentally identified 5′ RNA ends are in bold and underlined. The annotated translation initiation start codons are in lowercase bold. The gray boxes indicate stem-forming regions of putative *rho*-independent terminators. The sequence logos for the early P23-45 promoters and the independently aligned –35 and –10 regions of *T. thermophilus* HB8 –10/–35 promoters are depicted below the alignment. The size of the letters indicates the degree of conservation; positions are defined with respect to putative or identified transcription start sites.

elements are underlined). For comparison, *T. thermophilus* HB8 RNAP-σ^A holoenzyme actively transcribed from P23-45 middle and late promoters (P_{68M} and P_{103}; see Fig. 5b, lanes 14 and 15). Therefore, we conclude that unmodified host *T. thermophilus* RNAP-σ^A holoenzyme is unable to initiate transcription from DNA fragments containing the 11 bp motif *in vitro*.

If processing of an early polycistronic precursor transcript was responsible for the appearance of primer extension products whose 5′ ends are located just downstream of the 11 bp motif, then transcription initiation of this precursor transcript is most likely to originate in the non-coding region of the phage genome that separates the divergently transcribed early and late gene clusters (Fig. 1). This region, between genes 77 (an early gene) and 80 (a late gene), contains an 11 bp motif upstream of gene 77. The following experiment was done in an attempt to identify a hypothetical early P23-45 promoter located upstream of gene 77 and responsible for early viral transcription. The intergenic region between P23-45 genes 77 and 80 was cloned into the *T. thermophilus–E. coli* shuttle plasmid pMKE1 to generate pMKE77-80, a plasmid containing the validated late promoter P_{80} and the putative divergent early promoter P_{77} (Fig. 6a).[18] T. *thermophilus* HB8 was transformed with pMKE77-80 and the resulting strain was infected with P23-45 (Fig. 6a). Total RNA was extracted from infected cells at various times post infection and an *in vivo* primer extension analysis was done with primers complementary to plasmid sequences upstream and downstream of the P23-45 insert. As a control, primer extension reactions with P23-45 phage genome-specific primers that reported the transcriptional

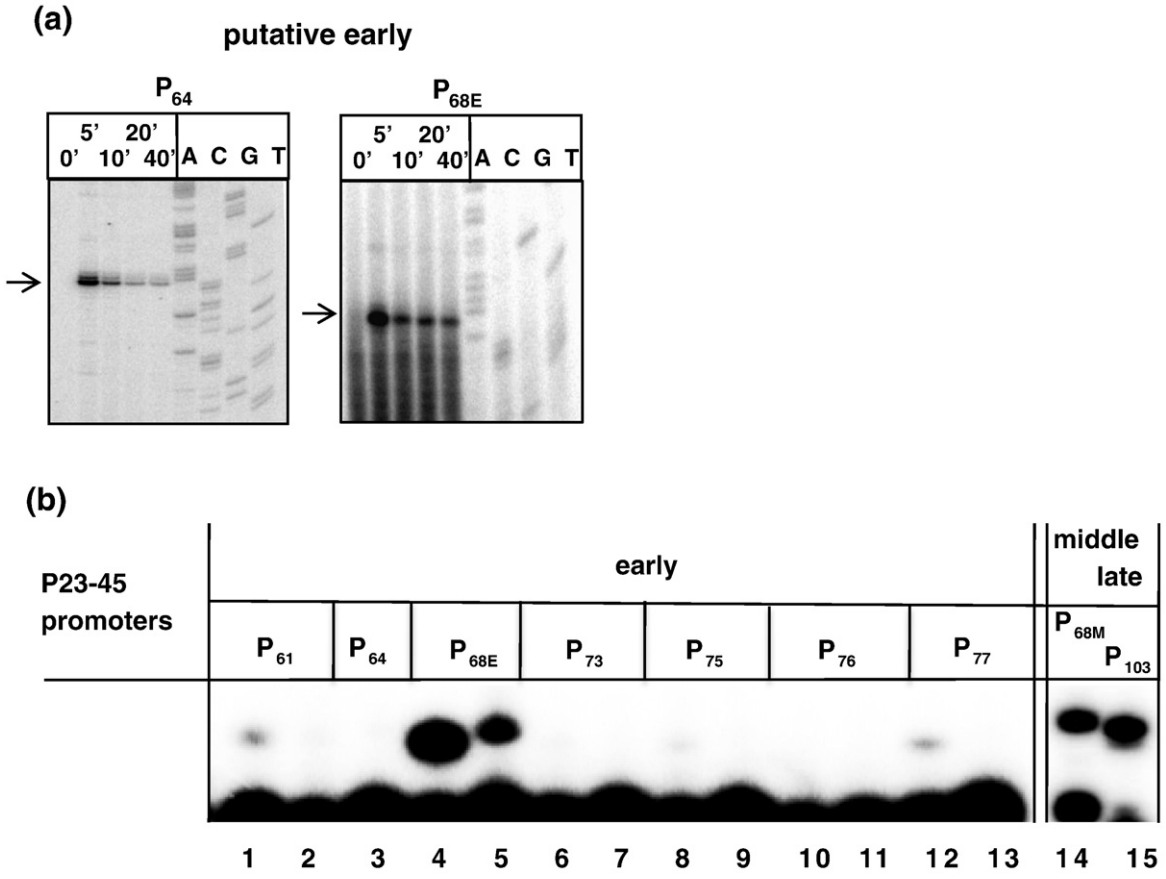

Fig. 5. Characterization of putative early P23-45 promoters. (a) The kinetics of accumulation of representative *in vivo* primer extension products obtained with P23-45 transcripts from putative early promoters P_{64} and P_{68E} during infection; the DNA products corresponding to 5′ RNA ends are labeled with arrows. (b) The results of abortive transcription initiation by *T. thermophilus* HB8 RNAP-σ^A holoenzyme at representative P23-45 early promoters (lanes 1–13) are shown. Dinucleotide RNA substrates corresponding to the –2/–1 and/or –1/+1 positions with respect to the putative transcription start sites (see Fig. 4) were used as primers. Radioactively labeled NTP corresponding to the putative +1 or +2 positions was used to form the trinucleotide RNA product. Abortive transcription at representative P23-45 middle and late promoters is shown as a positive control (lanes 14 and 15).

activity of P_{77} and P_{80} located on the P23-45 genome were also done (Fig. 6b). Control reactions revealed the expected late accumulation of the P_{80} transcript and the early accumulation of a transcript whose 5′ end coincided with the last nucleotide of the 11 bp motif; i.e. P_{77} (Fig. 6b, two lower panels). A primer extension product corresponding to plasmid-located P_{77} was absent from uninfected cells, peaked 5 min post infection and decreased steadily afterwards (Fig. 6b, upper left panel). A primer extension product corresponding to plasmid-located late P_{80} was detectable in the absence of infection and increased continuously throughout the infection (Fig. 6b, upper right panel). The appearance of the P_{80} transcript in the absence of the phage shows that this promoter is recognized by host RNAP in uninfected cells *in vivo* and is consistent with our findings that unmodified host *T. thermophilus* RNAP initiates transcription from P_{80} (and other P23-45 middle and late promoters) *in vitro*. The absence of this transcript immediately post infection might be due to phage-dependent modification(s) of the host RNAP that regulates the coordinated temporal transcription program of P23-45 and prohibits premature recognition of late promoters via unknown mechanisms. No primer extension product upstream of P_{77} was detected in uninfected cells. Our attempts to identify a hypothetical promoter located upstream of P_{77} with several additional primers, both P23-45 phage genome-specific and pMKE77-80 plasmid-specific, were also unsuccessful (data not shown). These data led us to conclude that it is very unlikely that there is a strong viral promoter from which early P23-45 genes are transcribed by host RNAP to produce a single precursor transcript.

An "internal" putative early promoter P_{59} was also cloned into pMKE1 and the activity of this promoter in P23-45 infected and uninfected cells was monitored. A result identical with that described for P_{77} (no activity in uninfected cells and an

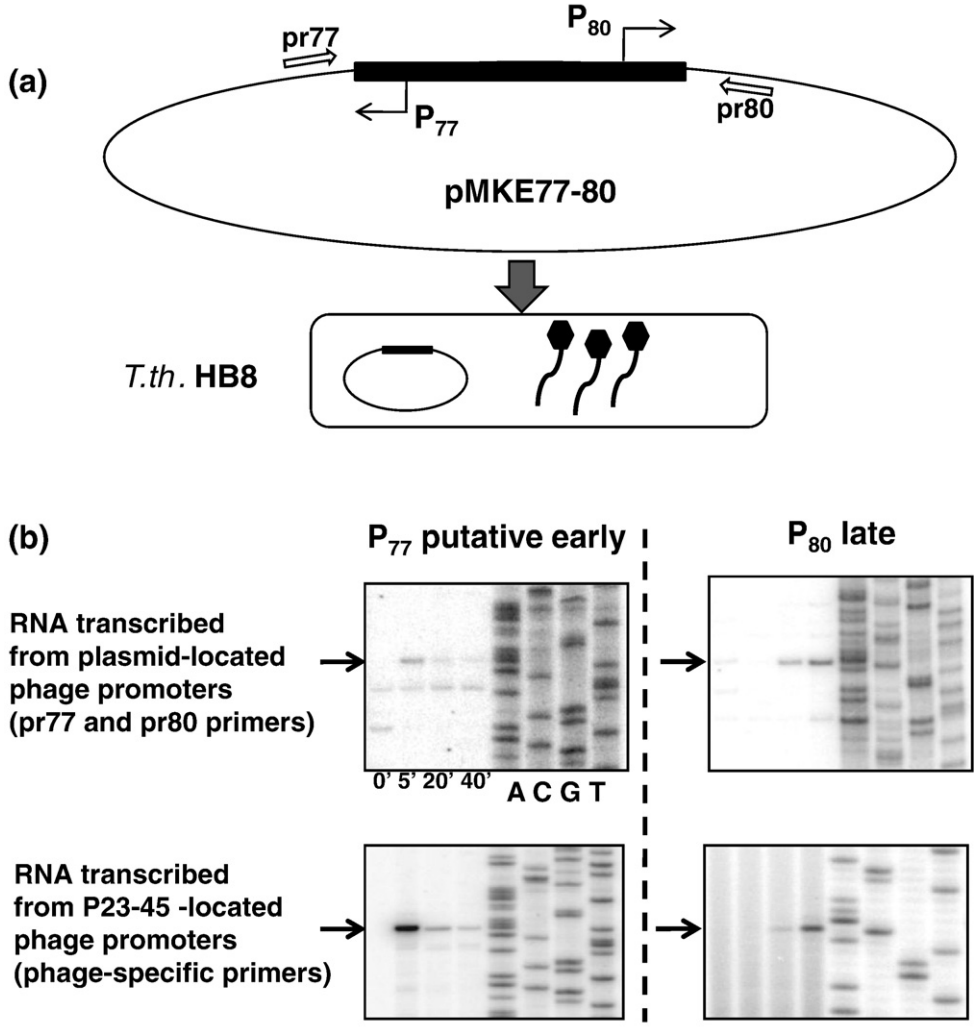

Fig. 6. Putative P23-45 early promoters are functional *in vivo*. (a) A diagram illustrating an experiment to probe the functionality of P23-45 promoters *in vivo*. *T. thermophilus* HB8 harboring the plasmid pMKE77-80 was infected with P23-45 and total RNA was extracted throughout the infection at the time points indicated. (b) The results of *in vivo* primer extension analysis of RNA transcribed from the divergent plasmid-located P23-45 early (P_{77}) and late (P_{80}) promoters (upper panels) and from the genome-located P_{77} and P_{80} promoters (lower panels). The DNA products corresponding to 5′ RNA ends are labeled with arrows.

"early" pattern of activity during infection) was observed (data not shown). Taken together, these data suggest that despite the lack of *in vitro* activity, P_{59} and P_{77} (and, likely, other genomic sites defined by the presence of the 11 bp motifs) might function as early phage promoters that are recognized immediately after infection by either a phage-encoded, but unidentified RNAP, or a phage-modified host RNAP.

Transcription of P23-45 early genes is rifampicin-resistant

Rifampicin (Rif), a strong inhibitor of bacterial RNAPs, including the *T. thermophilus* RNAP, binds to a pocket formed by the RNAP β subunit and efficiently blocks the synthesis of transcripts longer than 2 or 3 nt.[19] Conversely, all phage-encoded RNAPs studied thus far are resistant to Rif.[6,20] Therefore, if early P23-45 genes were transcribed by host RNAP, then the appearance of these transcripts shall be suppressed by the addition of Rif. On the other hand, if a phage RNAP was responsible for early P23-45 transcription, the appearance of these transcripts shall not be affected by Rif.

We performed P23-45 infection followed by the addition of Rif at different time points of infection followed by additional 10 min incubation and primer extension analysis of selected phage transcripts. Since middle and late P23-45 mRNAs are transcribed by host RNAP, we expected that Rif will inhibit accumulation of these transcripts. This

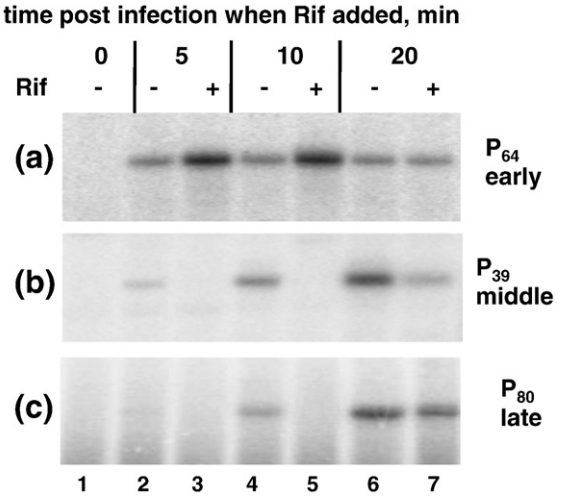

Fig. 7. Transcription from P23-45 early promoters is Rif-resistant and does not depend on host RNAP. The results of *in vivo* primer extension analysis of RNA extracted from phage-infected cells in the absence of Rif (–) or in the presence of Rif (+) added at different time points post-infection. (a–c) The kinetics of accumulation of transcripts from representative early, middle and late P23-45 promoters, respectively.

expectation was fulfilled (Fig. 7b and c): the addition of at Rif 5, 10 and 20 min post infection inhibited accumulation of a middle (P_{39}) and late (P_{80}) promoter transcripts compared to a control infection that was not treated with Rif. The effect was less pronounced when Rif was added 20 min post infection, because by this time P_{39} and P_{80}-originated transcripts started to accumulate before Rif addition. At the 5 and 10 min time points, Rif addition led to complete disappearance of primer extension bands corresponding to either P_{39} or P_{80}. In contrast, the addition of Rif had a completely different effect on the abundance of primer extension product corresponding to the early P_{64} phage transcript (Fig. 7a). The addition of Rif at 5 or 10 min post infection led to increase of this transcript abundance compared to untreated control cells. Addition of Rif 20 min post infection, when early transcription ceases based on the kinetics of early transcript accumulation, had no effect on the P_{64} transcript abundance. The difference in the abundances of the P_{64}-originated transcript in total RNA extracted from Rif-treated and Rif-untreated cells might be explained by an increased proportion of the P_{64} transcript in Rif-treated cells, where host RNAP-dependent transcripts are not accumulating. We conclude that the synthesis of early P23-45 transcripts is Rif-independent and might therefore be due to activity of Rif-resistant RNAP that is distinct from Rif-sensitive host RNAP.

Identification of P23-45-encoded *T. thermophilus* RNAP-binding proteins

Based on our macroarray data, P23-45 executes shut-off of *T. thermophilus* HB8 transcription, expresses its genes in a coordinated manner and must rely on the host transcription machinery for the expression of middle and late genes. Therefore, we hypothesized that P23-45 encodes transcription factors to alter *T. thermophilus* HB8 RNAP promoter specificity and activity. To identify such proteins, we affinity-isolated *T. thermophilus* HB8 RNAP genomically tagged with a protein A (4PrA) tag from *T. thermophilus* HB8 cells infected with P23-45 and identified RNAP co-isolating proteins in the sample using the MudPIT technique (Fig. 8).[21] As a control, analysis of proteins affinity-isolated from P23-45-infected wild type *T. thermophilus* HB8 cells (untagged RNAP) was done. Proteins present in both the *T. thermophilus* HB8/P23-45-infected (untagged) and the *T. thermophilus* HB8 RNAP-4PrA/P23-45-infected (tagged) samples were filtered out of the data set obtained for the RNAP-4PrA tagged strain.

Analysis of the material affinity isolated from the *T. thermophilus* HB8 RNAP-4PrA/P23-45-infected cells revealed the presence of the core RNAP subunits ($\alpha\beta\beta'\omega$) and the primary σ factor, σ^A. The RNAP subunits and σ^A were present at stoichiometric levels, as estimated by normalized spectral counts (Fig. 8), and another protein present at stoichiometric levels was CarD (TTHA0168). In a recent study, it was demonstrated that *T. thermophilus* HB8 CarD could interact with the N-terminus of the RNAP β subunit in a bacterial two-hybrid assay.[22,23] Several known tanscription elongation and antitermination factors (NusA, NusG, and

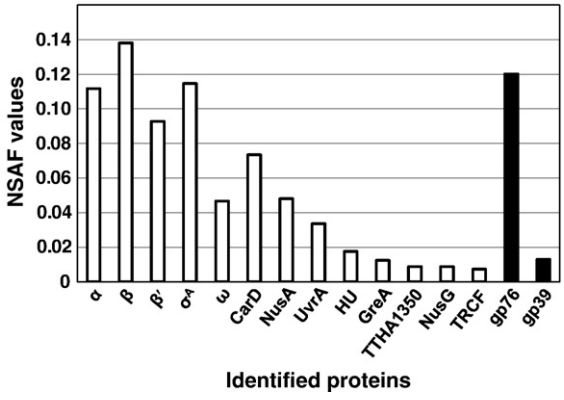

Fig. 8. MudPIT analysis of proteins that co-isolate with *T. thermophilus* HB8 RNAP affinity-isolated from P23-45 infected cells. Normalized spectral abundance factor values for RNAP subunits and co-isolated *T. thermophilus* HB8- and P23-45-encoded proteins are shown. *T. thermophilus* HB8-encoded proteins are shown as open bars and P23-45-encoded proteins are shown as filled bars.

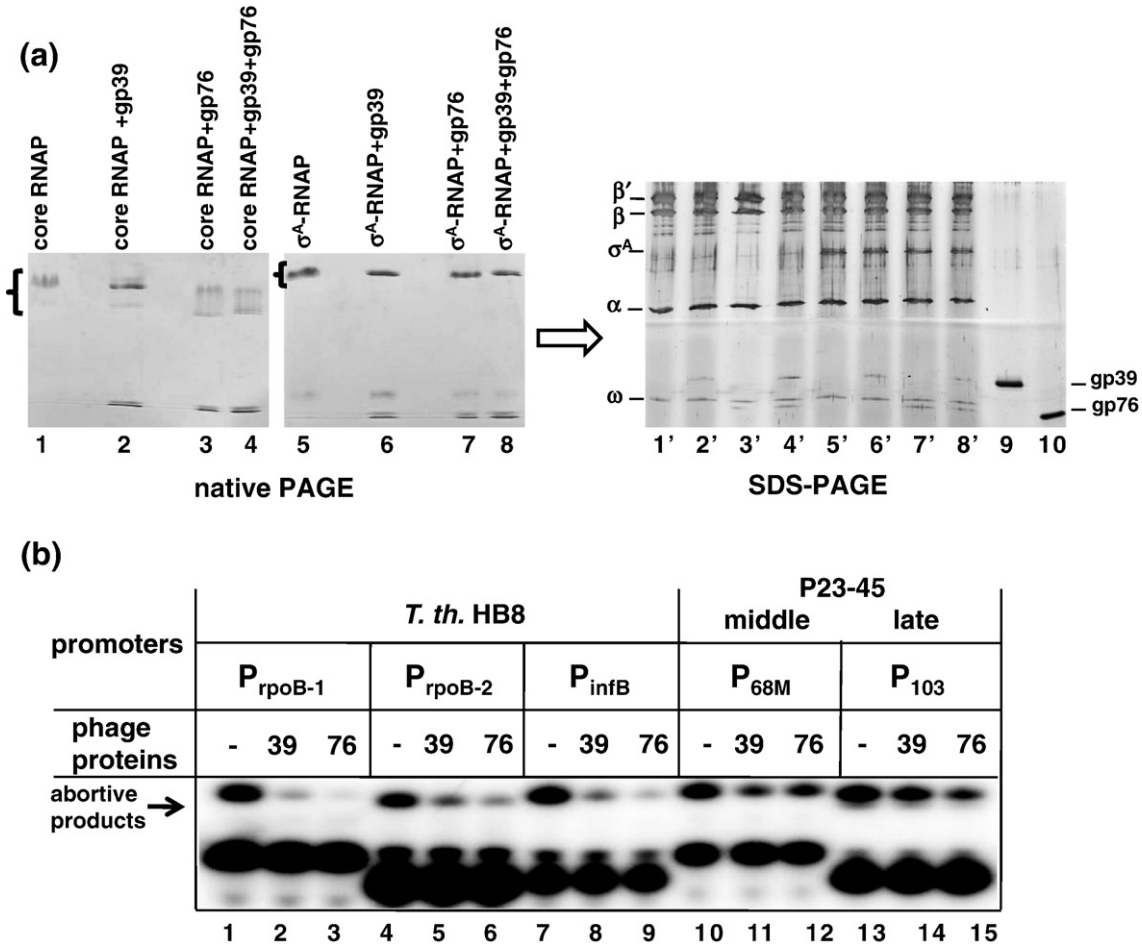

Fig. 9. The P23-45 phage proteins gp39 and gp76 bind to *T. thermophilus* HB8 RNAP and inhibit transcription initiation from host promoters but not from middle or late phage promoters. (a) gp39 and gp76 bind to both core RNAP and RNAP-σ^A holoenzyme. The proteins alone or together were incubated with RNAP core (lanes 2–4) or with RNAP-σ^A holoenzyme (lanes 6–8) and analyzed by gradient PAGE. Next, bands that were stained with Coomassie brilliant blue (indicated by brackets in a) were excised from the gels and their composition was determined by denaturing SDS-PAGE (lanes 1′–8′). gp39 (lane 9) and gp76 (lane 10) were loaded as markers. (b) The results of abortive transcription initiation by *T. thermophilus* HB8 RNAP-σ^A holoenzyme at several representative host *T. thermophilus* HB8 (lanes 1–9) and middle/late phage (lanes 10–15) promoters in the presence or in the absence of either gp39 or gp76 are shown.

GreA), the transcription-repair coupling factor (TRCF), the nucleoid-associated protein HU and the exonuclease ABC subunit A (UvrA) were also detected, although these proteins were present at lower levels. An uncharacterized protein, TTHA1350, was identified; although TTHA1350 was detected at low levels, this result indicates that TTHA1350 might play a role in the *T. thermophilus* HB8 transcription cycle. In addition to the host bacterial proteins, two P23-45-encoded proteins were identified (Fig. 8): gp39 (16.2 kDa; detected at a level substoichiometric to the core RNAP subunits) and gp76 (5.8 kDa; detected at a level stoichiometric to the core RNAP subunits). To corroborate the affinity isolation results, the DNA encoding these two proteins was cloned into an *E. coli* pET28-derived expression vector[24] and recombinant gp39 and gp76 proteins were purified to homogeneity and assayed for their ability to bind to host *T. thermophilus* HB8 RNAP core and σ^A-associated holoenzymes (Fig. 9a). The phage proteins were incubated with *T. thermophilus* RNAP core or σ^A-holoenzymes and the mixtures were resolved by native gradient PAGE (Fig. 9a, left-hand panel). The results of subsequent SDS-PAGE indicated that gp39 and gp76 can interact with both the core and the σ^A-holoenzymes (Fig. 9a, for gp39 in lanes 2, 6, 2′ and 6′; gp76 in lanes 3, 7, 3′ and 7′). Thus, we conclude that gp39 and gp76 are able to bind *T. thermophilus* HB8 RNAP *in vivo* and *in vitro*. As can be seen from Fig. 9, they can interact with RNAP simultaneously (Fig. 9a, lanes 4, 8, 4′ and 8′) but do not interact with each other (Z.B. and L.M., unpublished results), suggesting that they bind to

distinct sites on RNAP. To our knowledge, this is the first documentation of thermophage-encoded thermophilic host bacterial RNAP-binding proteins. gp39 (a middle phage protein) and gp76 (an early phage protein) have no recognizable conserved motif or similarity to other proteins in the public databases.[15] Thus, they are novel bacterial RNAP-binding proteins.

In order to elucidate the possible role(s) of gp39 and gp76 in P23-45 infection, we tested their ability to influence transcription by the host RNAP. *In vitro* abortive transcription initiation reactions using DNA fragments containing three different *T. thermophilus* HB8 σ^A-dependent promoters (P_{rpoB-1}, P_{rpoB-2} and P_{infB}) and two P23-45 promoters, a middle promoter (P_{68M}) and a late promoter (P_{103}), were performed in the presence or in the absence of either gp39 or gp76 (Fig. 9b). As can be seen, both proteins efficiently inhibited transcription from the host bacterial promoters that belong to the –10/–35 promoter class (Fig. 9b, lanes 1–9). When added together, gp39 and gp76 demonstrated a strong additive effect in transcription inhibition at these promoters (data not shown). In contrast, both gp39 and gp76 were much less efficient at inhibiting transcription from the phage middle and late promoters that belong to the extended –10 class of promoters that lack the –35 promoter element (Fig. 9b, lanes 10–15, see also P23-45 middle and late promoter alignment in Fig. 3b). Thus, the binding of gp39 and gp76 to *T. thermophilus* HB8 RNAP leads to promoter-specific transcription inhibition. Both proteins were purified as polyhistidine-tagged versions. To check if the tags could introduce non-physiological activity, we compared His-tagged and untagged (with the His tag removed by thrombin) proteins and found that the tag does not interfere with their RNAP binding or transcription inhibition *in vitro.* Our data suggests that gp39 and gp76 might be responsible for host transcription shut-off during P23-45 infection and are likely to act by interfering with –35 promoter element–RNAP interactions. However, the molecular mechanisms underlying host transcription inhibition remain to be fully elucidated.

Discussion

In this work, we investigated the transcription strategy of P23-45, a lytic thermophilic siphovirus infecting the thermophilic eubacterium *T. thermophilus* HB8. , We used a combination of bioinformatic and biochemical methods to study the P23-45 gene expression pattern, an approach that was successfully used by us to study host and viral gene expression during infection of *T. thermophilus* HB8 by an unrelated thermophage, the large myovirus ϕYS40.[8] In the case of P23-45, macroarray analysis and *in vivo* primer extension revealed that (i) host bacterial transcription is shut off during the P23-45 infection cycle and (ii) three temporal classes of viral genes exist; early, middle and late. Most of the known phages do not encode their own RNAP for expression of early genes and must therefore rely on host RNAP to transcribe their genes. Early promoters of such phages tend to be very strong, with a good match to the host promoter consensus, to successfully compete with host promoters for host RNAP at the onset of infection. Initially, automated bioinformatic analysis of the P23-45 genome did not reveal a recognizable RNAP gene,[15] leading us to hypothesize that P23-45 early promoters are similar to host promoters and are recognized by the host *T. thermophilus* RNAP-σ^A holoenzyme. Contrary to this expectation, no early P23-45 promoter with homology to host σ^A-dependent promoters was identified. Instead, we determined that many early phage genes are preceded by a common sequence motif. This highly conserved 11 bp motif has the consensus sequence 5′-TTATTCcTTTA-3′, with the highest conservation at positions –9 (T), –8 (A), –7 (T), –6 (T) and –2 (T) relative to the experimentally determined 5′ end of the transcript. In 13 of the 15 early P23-45 promoters, the 11 bp motif is located immediately upstream of the annotated translation start codon. Thus, many early phage transcripts appear to be leaderless. A similar situation was observed during the analysis of late and middle transcripts of the ϕYS40 thermophage.[8]

Comparison of putative P23-45 early promoters with the *T. thermophilus* –10/–35 consensus promoter elements used in bioinformatic searches indicated that they are clearly different (Fig. 4, compare logos). Thus, it was not at all surprising that the *T. thermophilus* RNAP-σ^A holoenzyme did not recognize these sequences as promoters *in vitro*, or in the absence of P23-45 infection *in vivo*. Nevertheless, DNA fragments containing the 11 bp motifs, when positioned on a *T. thermophilus* plasmid, led to the appearance of "correctly" initiated RNAs that behaved as early transcripts. We take these results as a strong indication that the 11 bp conserved sequences define early phage promoters. Experiments done in the presence of host RNAP inhibitor Rif clearly show that early P23-45 transcription is resistant to Rif, while middle and late transcription, which is carried out by *T. thermophilus* RNAP, is Rif-sensitive. The data suggest that P23-45 encodes a Rif-resistant RNAP that transcribes its early genes from promoters defined by the 11 bp consensus motif.

Given that biochemical data strongly suggest the existence of P23-45-encoded RNAP, a more thorough bioinformatic analysis of the P23-45 genome was undertaken. All known RNAPs are divided into two unrelated families based on sequence, structure and subunit composition. One family includes large multisubunit RNAPs of bacteria, archaea and

eukaryotes.[25–27] The other family consists of small single-subunit RNAPs related to phage T7 RNAP and found in some bacteriophages, in mitochondria and in chloroplasts.[28,29] The principal enzymatic activities are performed in both families of RNAPs via the same two-metal catalytic mechanism. All multisubunit RNAPs share a universal metal-binding signature motif, NADFDGD, in their largest subunits that together with additional conserved domains forms the active site. All T7-related RNAPs have other catalytic motifs and conserved domains/amino acids involved in the active center formation. In principle, it would seem more likely that P23-45 would encode a single-subunit RNAP that is present in many other bacteriophages. However, despite careful searches, we did not identify any sequence similarity to single-subunit RNAPs in the P23-45 genome. In contrast, a BLASTP search with the ORF64 sequence used as the query retrieved as the best hit (after the closely related ortholog from bacteriophage P74-26) the A subunit of RNAP I from the stramenopile *Thalassiosira pseudonana*.[30]

The detected region of similarity included a 65 amino acid segment that aligned with 32% amino acid identity and 47% similarity; the similarity was not statistically significant (expected value 1.4). However, it was notable that the alignment encompassed a portion of the catalytic double-φ β-barrel (DPBB) domain of RNAP, the most conserved portion of both the β′ and β subunits of all multisubunit RNAPs that is directly involved in nucleotide polymerization.[26,27,31] Moreover, the RNAP amino acid signature that includes the three invariant and essential aspartates required for the coordination of the two magnesium ions that participate in catalysis was fully conserved in ORF64 ([NA]AD[FY]DGD, the magnesium-chelated aspartates are underlined). When the ORF64 sequence was combined with the *T. pseudonana.* RNAP sequence to generate a position-specific scoring matrix, the second iteration of the PSI-BLAST search readily retrieved numerous RNAP sequences.[30] Using the PHI-BLAST program,[32] we found that when the ORF64 sequence was compared to the non-redundant protein sequence database at the NCBI under the additional requirement that the signature Mg^{2+}-binding motif was matched, the only sequences retrieved were those of RNAP subunits. Using the HHPred program,[33] we found that the ORF64 sequence produced the best hit with the DPBB domain of the *T. thermophilus* RNAP β′ subunit, with the *E*-value 0.091 when the Interpro collection of hidden Markov models was searched. When the HHPred search was initiated with the isolated sequence of the predicted DPBB domain of ORF64, the same best hit was obtained, with a statistically significant *E*-value of 0.0031.

A multiple alignment of the putative DPBB domain of ORF64 and a representative set of eukaryotic, archaeal and bacterial RNAPs is shown in Fig. 10. The alignment suggests that ORF64 contains a short version of the DPBB domain similar to the highly diverged β′ homologs of RNAPs from baculoviruses and fungal mitochondrial plasmids.[31] The DPBB domain consists of six β-strands.[27,31] In addition to the (NA)D(FY)DGD signature motif, ORF64 retains several other invariant and essential amino acid residues of the DPBB domain, such as the arginine at the end of S3 (corresponding to β′ R704 in *T. aquaticus* RNAP) and the proline and leucine in the loop downstream of S3 (β′ P706 and L708, respectively, in *T. aquaticus* RNAP; Fig. 10).

We propose that ORF64 is a distant member of the multisubunit family of RNAPs. However, the putative P23-45 RNAP contains no counterpart to several other essential residues of the β′ RNAP subunit, suggesting that this enzyme could be mechanistically distinct from the known RNAPs. In our earlier work, highly sensitive mass spectrometric analysis of pure P23-45 virions did not detect ORF64.[15] More recently, Western blot analysis revealed no ORF64 trace among P23-45 virion proteins using anti ORF64 polyclonal antibodies (Minakhin *et al.*, unpublished results). It remains to be determined whether ORF64 functions on its own or requires additional host or phage-encoded factors for activity.

Macroarray and *in vivo* primer extension analysis also revealed P23-45 middle and late genes and the corresponding P23-45 promoters upstream of these genes; at present we cannot distinguish between the middle and the late promoters on the basis of their consensus element sequences. Similar to ϕYS40, the P23-45 middle/late promoters are characterized by a –10 consensus element supplemented with a "TG/TGTG" motif (Fig. 3b). Middle and late promoters of P23-45 are recognized by unmodified *T. thermophilus* HB8 RNAP *in vitro*; as these promoters are inactive early in infection, it follows that their activity is somehow repressed until later stages of the infection cycle.

A combined view that emerges from our work indicates that the P23-45 transcription strategy is a simplified version of the transcription strategy employed by *E. coli* phage N4. In the case of N4 infection, early genes are transcribed by the Rif-resistant phage-encoded RNAP that is encapsulated in the virion and injected into the infected cell with the N4 genome.[20,29] The middle genes are transcribed by another phage-encoded RNAP, a product of early N4 genes.[34] Late N4 genes are transcribed by phage-modified host RNAP. In the case of P23-45, early genes are likely to be transcribed by the Rif-resistant RNAP encoded by ORF64, whereas the middle and late genes are transcribed by host RNAP. Earlier analysis did not reveal the presence of ORF64 in P23-45 virions, which could have been caused by a low copy number of the putative phage

```
Secondary structure ORF64                    eee                 eeeee
P23_ORF64_YP_001467917.1   243-341  GPNTGFGG-ILLSPKILPFLGLHGLEDGGVLAYTRRWKP
P74_ORF62_YP_001468032.1   270-368  GPNTGFGG-ILLSPKILPFLGLHGLEDGGLLAYTRRWRP
RPA1_XP_002290924.1 Thaps  466-631  GKRVNFACRSVISPD--PYI--------GTNEIGLPLYF  75
RPA1_EEY66296.1     Phyin  466-637  GKRVNYAARSVISPD--PYI--------STSQIGVPLRF  84
RPA1_XP_794863.2    Strpu  430-597  GKRVNYAARSVISPD--PYI--------NTDEIGIPQVI  82
RPA1_EAW99467.1     Homsa  434-604  GKRVDYAARSVICPD--MYI--------NTNEIGIPMVF  82
RPA1_EEH52215.1     Micpu  443-631  GKRVNYAARSVIMPD--PYL--------KTSEIGVPPVF 100
RPA1_ACN85301.1     Oryco  430-605  GKRVNYACRSVISPD--PYL--------AVNEIGIPPVF  87
RPA'_2WAQ_A         Sulsh  315-472  GKRVDFSSRTVISPD--PNI--------SIDEVGVPEII  79
RPA'_ZP_06213371.1  Metsp  318-482  GKRVNFSSRTVISPD--PCL--------SINEVGVPEVV  82
RPb'_AAB59112.1     Bacsu  333-445  GKRVDYSGRSVIVVG--PHL--------KMYQCGLPKEM  57
RPb'_2GHO_D         Theaq  329-444  GKRVDYSGRSVIVVG--PQL--------KLHQCGLPKRM  57
                                    *   :::   ::       :
Secondary structure 2WAQ                  eeeeeeeee               eeeee
Secondary structure 2GHO                 eeeeeeeee                 eeeee
DPBB consensus structure:                  eeeS1eee                eeS2ee


Secondary structure ORF64     eeeee          eeeeee   hhhhhhhh                eeeeee
P23_ORF64_YP_001467917.1   GERVIFNRRPDLPTGQSAVELTYVGLSPIADSVIAHEHDIAPTGADYDGD-IGYVFPTPEMG
P74_ORF62_YP_001468032.1   GERVIFNRRPDLPTGQSAVELTYLGLSPIADSVIAHEGDIASTGADYDGD-IGYLFPTPEKG
RPA1_XP_002290924.1 Thaps  GDMVLMNRQPTLHKPGIMAHRVRVLFSPTQNTLRMHYANCNTYNADYDGDEMNCHFPQSYLA
RPA1_EEY66296.1     Phyin  GDVVLMNRQPTLHKPSIMAHTARVLTNPKMQTIRMHYANCNTFNADFDGDEMNMHFPQNELA
RPA1_XP_794863.2    Strpu  GDIVLLNRQPTLHKPGIMAHKVRVL--PGEKTLRLHYSACKTYNADFDGDEMNVHFPQNELG
RPA1_EAW99467.1     Homsa  GDILLLNRQPTLHRPSIQAHRARIL--PEEKVLRLHYANCKAYNADFDGDEMNAHFPQSELG
RPA1_EEH52215.1     Micpu  GDVLLVNRQPTLHKPGIMAHTAKVL--PGQRTIRMHYANCSTYNADFDGDEMNLHFPQDHLA
RPA1_ACN85301.1     Oryco  GDIVLVNRQPTLHKPSMMAHFVRVL--PGEKTIRMHYANCSTYNADFDGDEMNVHFPQDEIS
RPA'_2WAQ           Sulsh  GDVVLFNRQPSLHRISMMAHRVRVL--KG-LTFRLNLLVCPPYNADFDGDEMNLHVPQSEEA
RPA'_ZP_06213371.1  Metsp  GDIVLYNRQPSLHRMSIMAHRVRVL--PY-RTFRHNLCVCPPYNADFDGDEMNLHVPQSEEA
RPb'_AAB59112.1     Bacsu  EHPVLLNRAPTLHRLGIQAFEPTLV--EG-RAIRLHPLVCTAYNADFDGDQMAVHVPLSAEA
RPb'_2GHO           Theaq  GKVVLLNRAPTLHRLGIQAFQPVLV--EG-QSIQLHPLVCEAFNADFDGDQMAVHVPLSSFA
                              :::** * *      :    :        .  :       :**:*** :   :*
Secondary structure 2WAQ   eeeeeeee       eeeeeeeeee      eeeeee            eeeeeee
Secondary structure 2GHO     eeeeee       eeeeeeeee      eeee                eeeeee
DPBB consensus structure:    eeS3e           eeeS4ee       eS5                eeS6ee
```

Fig. 10. Alignment of the double-φ β-barrel (DPBB) domains of ORF64 from P23-45 and ORF62 from closely related P74-26 and β′ orthologs from selected multisubunit RNAPs. The alignment was generated using the MUSCLE program,[41] and the N-terminal part was manually modified on the basis of secondary structure predictions. The range of amino acid residues of the aligned segments in the respective protein precedes each sequence. Inserts in RNAP subunit sequences are shown as the number of amino acids. The residues conserved in all aligned sequences are denoted by asterisks; similar residues are denoted by colons. The most conserved segments of the alignment including the NADFDGD motif are shown as gray boxes. The predicted secondary structure of ORF64 (consensus of the PredictProtein and JPred predictions) is shown above the alignment. The secondary structure from two crystal structures, those of the RNAPs of the archaeon *Sulfolobus shibatae* (PDB code 2WAQ)[44] and the bacterium *T. aquaticus* (2GHO)[45] are shown underneath the alignment along with the consensus structure of the DPBB domain.[31] In the secondary structure lines: e, extended conformation (β-strand); h, α-helix; S1–S6 numbering is from Ref. 31. Thaps, *Thalassiosira pseudonana*; Phyin, *Phytophthora infestans*; Strpu, *Strongylocentrotus purpuratus*; Homsa, *Homo sapiens*; Micpu, *Micromonas pusilla*; Oryco, *Oryza coarctata*; Sulsh, *Sulfolobus shibatae*; Metsp, *Methanocaldococcus* sp.; Bacsu, *Bacillus subtilis*; Theaq, *Thermus aquaticus*.

RNAP in the virion.[15] Alternatively, ORF64 might be transcribed initially by the host RNAP from P_{68}, an upstream early promoter that contains an 11 bp motif embedded into a host RNAP-σ^A promoter (see alignment in Fig. 4, where the –10/–35 elements are underlined). These possibilities are being investigated in our laboratory.

Using mass spectrometric analysis (MudPIT) of the one-step affinity-isolated host *T. thermophilus* HB8 RNAP, we identified two P23-45-encoded *T. thermophilus* HB8 RNAP-binding proteins, gp76 and gp39. To the best of our knowledge, this is the first documentation of thermophage-encoded thermophilic bacterial RNAP-binding proteins. gp76 and gp39 are encoded by early and middle P23-45 genes, respectively, and both proteins efficiently inhibit *in vitro* transcription by host RNAP from host promoters, but are less effective at inhibiting transcription from P23-45 middle and late promoters (Fig. 9b). Taken together, these data suggest that gp76 and gp39 might be involved in the shut-off of host transcription during the P23-45 infection program. Further biochemical and structural analysis of these proteins, in complex with *T. thermophilus* RNAP, should make it possible to obtain a structure-based model of the action of a phage-encoded transcription regulator.

Experimental Procedures

Bacterial growth and phage infection

Bacteriophage P23-45 was generously provided by Dr Michael Slater (Promega Corporation, Madison, WI). To isolate individual P23-45 plaques, 150 μL of a *T. thermophilus* HB8 culture (A_{600} ~0.4) freshly grown in TB medium (0.8% (w/v) tryptone, 0.4% (w/v) yeast extract, 0.3% (w/v) NaCl, 1 mM $MgCl_2$, 0.5 mM $CaCl_2$) was combined with a 100 μL dilution of phage stock, incubated at 65°C for 10 min, plated in soft TB agar (0.75 %, w/v), and incubated at 65°C overnight. To prepare a phage stock suspension, an individual plaque was picked and subjected to two more rounds of plaque purification. The phage lysate was prepared as described.[15]

E. coli strains XL-1Blue (New England Biolabs) and BL21(DE3) (Novagen) were used for molecular cloning and recombinant protein expression, respectively.

Total DNA purification and molecular cloning

P23-45 genomic DNA was extracted with a Lambda Midi kit (Qiagen) using the procedure recommended by the manufacturer. *T. thermophilus* HB8 genomic DNA was purified by extraction with phenol/chloroform and subsequent precipitation in ethanol.

Plasmids encoding either polyhistidine-tagged gp39 or gp76 for recombinant protein production and purification were constructed as follows: the DNA encoding gp39 or gp76 was PCR-amplified using primers that appended NdeI and EcoRI sites at the 5′ and 3′ ends of each gene, respectively. The resultant PCR products were cleaved with NdeI and EcoRI and cloned between the NdeI and EcoRI sites of a pET28a-based plasmid, creating pSKB2-39_{HIS} and pSKB2-76_{HIS}. The plasmid pMKE77-80 was constructed as follows: a 504 bp DNA fragment of the P23-45 genome comprising the divergent P_{77} and P_{80} promoters together with the proximal parts of ORF77 and ORF80 coding regions was PCR-amplified using primers that appended NdeI and HindIII sites at the 5′ and 3′ end, respectively. The resultant PCR fragment was digested with NdeI and HindIII and cloned between the NdeI and HindIII sites of the *E. coli – T. thermophilus –E. coli* shuttle plasmid pMKE1.[18] The resultant plasmid pMKE77-80 was used in *in vivo* primer extension experiments.

Strain construction and affinity isolation of protein A-tagged *T. thermophilus* HB8 RNAP

A similar strategy to that used to construct a *T. thermophilus* HB8 strain encoding a polyhistidine tag (His_{10}) appended to the 3′ end of the *rpoC* gene (encoding the RNAP β′ subunit) was used to construct a *T. thermophilus* HB8 strain encoding a genomic β′-protein A (4PrA) fusion protein.[8] The resultant strain, *T. thermophilus* HB8*rpoC::4PrA*, demonstrated a slightly slower growth rate than wild type cells, but was infected with P23-45 as efficiently as wild type.

To prepare P23-45-infected biomass, wild type *T. thermophilus* HB8 or *T. thermophilus* HB8*rpoC::4PrA* cells were grown at 65°C in 4 L of TB medium until A_{600} ~0.35 and were infected with P23-45 at a multiplicity of infection of 10. Infection was stopped at 20 min post infection by rapidly cooling the samples in an ice-water bath. Cells were harvested by centrifugation and washed once with ice-cold 10% (v/v) glycerol. Next, 1 mL of lysis buffer (20 mM Hepes (pH 7.), 0.2 mg/mL PMSF, 4 mg/mL pepstatin) was added to every 10 g of cell pellet, and the cells were frozen in liquid nitrogen. Both tagged and untagged *T. thermophilus* HB8 cells were cryogenically lysed using the MM 301 Mixer Mill (Retsch)[24] and stored at –80°C. Affinity-isolation of RNAP-4PrA and co-isolating proteins from P23-45 infected *T. thermophilus* HB8 cells was done as described for *E. coli* RNAP and co-isolating proteins.[24,35]

MudPIT analysis

Elution of *T. thermophilus* HB8 RNAP-4PrA and co-isolating proteins from the IgG conjugated Dynabeads (Invitrogen) was achieved with 0.5 M ammonium hydroxide and 0.5 mM EDTA. The eluted proteins were frozen in liquid nitrogen and dried in a SpeedVac (Thermo Savant). The dried protein pellets were denatured, reduced, alkylated, and digested with endoproteinase LysC (Roche Applied Science) followed by digestion with trypsin (Promega). The peptide mixtures were pressure-loaded onto triphasic microcapillary columns, installed in-line with a Quaternary Agilent 1100 series HPLC pump coupled to a Deca-XP ion trap tandem mass spectrometer (ThermoElectron) and analyzed via ten-step chromatography.[36] The MS/MS data sets were searched using SEQUEST[37] against a database of 117 P23-45 predicted gene products, combined with 2238 protein sequences from *T. thermophilus* HB8 as described.[15] Lists of detected proteins were established and compared using DTASelect/CONTRAST[38] as described.[15] Protein levels in different samples were compared using normalized spectral abundance factor values.[39,40]

Proteins

Polyhistidine-tagged *T. thermophilus* HB8 core RNAP and σ^A were used in native PAGE protein–protein interaction and *in vitro* transcription assays. The proteins were purified essentially as described.[8] Either recombinant polyhistidine-tagged gp39 or gp76 were produced as follows; the expression plasmids either pSKB2-39_{HIS} or pSKB-76_{HIS} were transformed into *E. coli* BL21 (DE3) cells and transformants were selected in the presence of 50 μg/mL kanamycin. Cultures (4 L) were grown at 37°C to A_{600} ~0.8 and recombinant protein over-expression was induced with 1 mM IPTG for 4 h at 37°C. Cells containing over-expressed recombinant proteins were harvested by centrifugation and disrupted by sonication in buffer A (10 mM Tris–HCl (pH 8.0), 500 mM NaCl, 5 mM imidazole, 5% (v/v) glycerol, 0.2 mg/mL PMSF). Inclusion bodies were dissolved in buffer B (10 mM Tris-HCl (pH 8.0), 500 mM NaCl, 2 mM imidazole, 7 M urea); loaded onto a 5 mL nickel-chelated Hi-Trap Sepharose column (GE Healthcare) equilibrated in buffer B, and the column was washed with buffer B supplemented with 25 mM imidazole. The bound proteins, either gp39 or gp76, were eluted from the column with buffer B

supplemented with 200 mM imidazole, and dialyzed against buffer C (20 mM Tris–HCl (pH 8.0), 50 mM NaCl, 0.5 mM EDTA). The proteins were loaded onto a MonoQ column (GE Healthcare) equilibrated in TGE buffer (10 mM Tris–HCl (pH 8.0), 50 mM NaCl, 1 mM EDTA, 5% (v/v) glycerol), eluted with a linear 150 mM – 450 mM gradient of NaCl , dialyzed against buffer D (10 mM Tris–HCl (pH 8.0), 100 mM NaCl, 1 mM EDTA, 50% (v/v) glycerol) and stored at –80°C.

Macroarray membrane preparation and data analysis

DNA fragments corresponding to each of the selected P23-45 ORFs, *T. thermophilus* HB8 housekeeping genes and the *D. melanogaster zfrp8* gene (control) were PCR-amplified from the corresponding genomic DNA using gene-specific primer pairs (the sequences of the primers are available from the authors upon request). Membrane preparation, cDNA synthesis and macroarray hybridization were done as described.[7] After hybridization, the amount of radioactivity from each spot was quantified using ImageQuant software (Molecular Dynamics) and the background signal was subtracted from signals corresponding to every ORF spot. To allow comparison between the signals on different membranes, the background-corrected signals were normalized relative to the average of the two *D. melanogaster zfrp8* spot signals; the normalized spot signals were used for data analysis.

Primer extension

Primer extension reactions were done essentially as described.[8] Exponential phase *T. thermophilus* HB8 cells were infected with P23-45 and harvested at the same time points post infection as used for the macroarray experiments. In experiments utilizing Rif (Sigma-Aldrich), it was added to *T. thermophilus* HB8 cells infected with P23-45 at the designated time points to yield a final concentration of 2 mg/mL followed by incubation at 65°C for 10 min before RNA extraction. Total RNA was extracted using the RNeasy Mini Kit (Qiagen) according to the manufacturer's procedure. For each primer extension reaction, 10 g of total RNA was reverse-transcribed with 100 units of SuperScript III enzyme from the First-Strand Synthesis kit for RT-PCR (Invitrogen) in the presence of 10 pmol of (γ-^{32}P) end-labeled primer. The reactions were treated with RNase H, precipitated with ethanol and dissolved in formamide loading buffer. DNA sequencing reactions, accomplished using the *fmol* DNA Cycle Sequencing kit (Promega), containing both the corresponding PCR-amplified P23-45 genome fragments and end-labeled primers used for the primer extension reaction were used to identify the 5′ ends of the primer extension products. The reaction products were separated by PAGE (6–8 % (w/v) polyacrylamide gel) and visualized with a PhosphorImager (Molecular Dynamics).

Protein complex analysis

Either *T. thermophilus* HB8 core RNAP or σ^A-holoenzyme (reconstituted with 1 μM core RNAP and 1 μM σ^A) was incubated with either gp39 (~5 μM) or gp76 (~5 μM) in 10 μL of transcription buffer (30 mM Tris–HCl (pH 7.9), 40 mM KCl, 10 mM $MgCl_2$, 2 mM β-mercaptoethanol) at 65°C for 10 min. Subsequently, 4 μL of the reaction mixture was resolved in a native 4%–15% (w/v) Phast gradient polyacrylamide gel (GE Healthcare); bands of protein were visualized by staining with Coomassie brilliant blue. To examine the protein composition, the bands were excised from the native gel and placed into the wells of a 12%–16% (w/v) gradient polyacrylamide/SDS gel, followed by electrophoresis and staining with silver.

In vitro transcription

A typical abortive transcription reaction was done in a final volume of 10 μL and contained 200 nM *T. thermophilus* HB8 σ^A-holoenzyme and 20–40 nM of a PCR-amplified DNA fragment containing either a *T. thermophilus* HB8 or a P23-45 promoter in standard transcription buffer (30 mM Tris–HCl (pH 7.9), 40 mM KCl, 10 mM $MgCl_2$, 2 mM β-mercaptoethanol). Reactions were supplemented (where indicated) with either 15 μM gp39 or 15 μM gp76, incubated at 65°C for 10 min, followed by the addition of various RNA dinucleotides (100–500 μM), [α-^{32}P]NTPs (3000 Ci/mmol) and the corresponding cold NTPs (100 μM). The reactions were incubated at 65°C for a further 10 min before being terminated by the addition of an equal volume of urea/formamide loading buffer. The reaction products were resolved by SDS-PAGE (20% (w/v) polyacrylamide gel) and visualized using a PhosphorImager.

Sequence analysis

The sequences of the predicted proteins encoded in the P23-45 genome were searched against the non-redundant protein sequence database at the NCBI using the iterative PSI-BLAST[30] and the pattern-hit-initiated BLAST (PHI-BLAST)[32] program searches with P23-45 deduced ORFs. Additional searches were done with the HHPred program that implements pairwise comparison of hidden Markov models.[33] The multiple alignment of the putative DPBB domains of ORF64 from P23-45, ORF62 from P74-26, and a representative set of multisubunit RNAPs was constructed using the MUSCLE program.[41] The results of secondary structure prediction made using the PredictProtein[42] and JPred[43] programs were taken into consideration to manually refine the alignment.

Acknowledgements

Bacteriophage P23-45 was generously provided by Dr Michael Slater from Promega Corporation. We thank Dr E. Peter Geiduschek for critical reading of the manuscript and helpful advice. This work was supported by NIH grant R21 AI074769 and Federal Program "Scientific and scientific-pedagogical personnel of innovative Russia 2009-2013", state contract 02.740.11.5132 (to L.M.), by NIH grant R01 GM61898 (to S.A.D.), by NIH grants RR00862 and RR022220 (to B.T.C.), by the Stowers Institute for

Medical Research (to L.F. and M.P.W.) by NIH grant R01 GM59295, by Molecular and Cell Biology grant from the Presidium of Russian Academy of Sciences, Russian Foundation for Basis Research grant 07-04-00366-a, by Federal Program "Scientific and scientific-pedagogical personnel of innovative Russia 2009-2013", state contract 02.740.11.0771, and by a National Center for Biotechnology of the Republic of Kazakhstan grant (to K.S.); E.V.K. is supported by the US Department of Health and Human Resources (National Library of Medicine, National Institutes of Health).

References

1. Nechaev, S. & Severinov, K. (2003). Bacteriophage-induced modifications of host RNA polymerase. *Annu. Rev. Microbiol.* **57**, 301–322.
2. Hendrix, R. W., Smith, M. C., Burns, R. N., Ford, M. E. & Hatfull, G. F. (1999). Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. *Proc. Natl Acad. Sci. USA*, **96**, 2192–2197.
3. Wommack, K. E. & Colwell, R. R. (2000). Virioplankton: viruses in aquatic ecosystems. *Microbiol. Mol. Biol. Rev.* **64**, 69–114.
4. Pedulla, M. L., Ford, M. E., Houtz, J. M., Karthikeyan, T., Wadsworth, C., Lewis, J. A. *et al.* (2003). Origins of highly mosaic mycobacteriophage genomes. *Cell*, **113**, 171–182.
5. Ventura, M., Foley, S., Bruttin, A., Chennoufi, S. C., Canchaya, C. & Brussow, H. (2002). Transcription mapping as a tool in phage genomics: the case of the temperate Streptococcus thermophilus phage Sfi21. *Virology*, **296**, 62–76.
6. Semenova, E., Djordjevic, M., Shraiman, B. & Severinov, K. (2005). The tale of two RNA polymerases: transcription profiling and gene expression strategy of bacteriophage Xp10. *Mol. Microbiol.* **55**, 764–777.
7. Minakhin, L., Semenova, E., Liu, J., Vasilov, A., Severinova, E., Gabisonia, T. *et al.* (2005). Genome sequence and gene expression of Bacillus anthracis bacteriophage Fah. *J. Mol. Biol.* **354**, 1–15.
8. Sevostyanova, A., Djordjevic, M., Kuznedelov, K., Naryshkina, T., Gelfand, M. S., Severinov, K. & Minakhin, L. (2007). Temporal regulation of viral transcription during development of *Thermus thermophilus* bacteriophage phiYS40. *J. Mol. Biol.* **366**, 420–435.
9. Zhang, G., Campbell, E. A., Minakhin, L., Richter, C., Severinov, K. & Darst, S. A. (1999). Crystal structure of *Thermus aquaticus* core RNA polymerase at 3.3 Å resolution. *Cell*, **98**, 811–824.
10. Murakami, K. S., Masuda, S. & Darst, S. A. (2002). Structural basis of transcription initiation: RNA polymerase holoenzyme at 4 Å resolution. *Science*, **296**, 1280–1284.
11. Murakami, K. S., Masuda, S., Campbell, E. A., Muzzin, O. & Darst, S. A. (2002). Structural basis of transcription initiation: an RNA polymerase holoenzyme-DNA complex. *Science*, **296**, 1285–1290.
12. Vassylyev, D. G., Sekine, S., Laptenko, O., Lee, J., Vassylyeva, M. N., Borukhov, S. & Yokoyama, S. (2002). Crystal structure of a bacterial RNA polymerase holoenzyme at 2.6 Å resolution. *Nature*, **417**, 712–719.
13. Vassylyev, D. G., Vassylyeva, M. N., Perederina, A., Tahirov, T. H. & Artsimovitch, I. (2007). Structural basis for transcription elongation by bacterial RNA polymerase. *Nature*, **448**, 157–162.
14. Naryshkina, T., Liu, J., Florens, L., Swanson, S. K., Pavlov, A. R., Pavlova, N. V. *et al.* (2006). *Thermus thermophilus* bacteriophage phiYS40 genome and proteomic characterization of virions. *J. Mol. Biol.* **364**, 667–677.
15. Minakhin, L., Goel, M., Berdygulova, Z., Ramanculov, E., Florens, L., Glazko, G. *et al.* (2008). Genome comparison and proteomic characterization of *Thermus thermophilus* bacteriophages P23-45 and P74-26: siphoviruses with triplex-forming sequences and the longest known tails. *J. Mol. Biol.* **378**, 468–480.
16. Jalasvuori, M., Jaatinen, S. T., Laurinavicius, S., Ahola-Iivarinen, E., Kalkkinen, N., Bamford, D. H. & Bamford, J. K. (2009). The closest relatives of icosahedral viruses of thermophilic bacteria are among viruses and plasmids of the halophilic archaea. *J. Virol.* **83**, 9388–9397.
17. Paulose-Murphy, M., Ha, N. K., Xiang, C., Chen, Y., Gillim, L., Yarchoan, R. *et al.* (2001). Transcription program of human herpesvirus 8 (Kaposi's sarcoma-associated herpesvirus). *J. Virol.* **75**, 4843–4853.
18. Moreno, R., Zafra, O., Cava, F. & Berenguer, J. (2003). Development of a gene expression vector for *Thermus thermophilus* based on the promoter of the respiratory nitrate reductase. *Plasmid*, **49**, 2–8.
19. Campbell, E. A., Korzheva, N., Mustaev, A., Murakami, K., Nair, S., Goldfarb, A. & Darst, S. A. (2001). Structural mechanism for rifampicin inhibition of bacterial RNA polymerase. *Cell*, **104**, 901–912.
20. Falco, S. C., Zehring, W. & Rothman-Denes, L. B. (1980). DNA-dependent RNA polymerase from bacteriophage N4 virions. Purification and characterization. *J. Biol. Chem.* **255**, 4339–4347.
21. Washburn, M. P., Wolters, D. & Yates, J. R., III (2001). Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19**, 242–247.
22. Stallings, C. L., Stephanou, N. C., Chu, L., Hochschild, A., Nickels, B. E. & Glickman, M. S. (2009). CarD is an essential regulator of rRNA transcription required for *Mycobacterium tuberculosis* persistence. *Cell*, **138**, 146–159.
23. Westblade, L. F., Campbell, E. A., Pukhrambam, C., Padovan, J. C., Nickels, B. E., Lamour, V. & Darst, S. A. (2010). Structural basis for the bacterial transcription-repair coupling factor/RNA polymerase interaction. *Nucleic Acids Res*, **38**, 8357–8369.
24. Savalia, D., Westblade, L. F., Goel, M., Florens, L., Kemp, P., Akulenko, N. *et al.* (2008). Genomic and proteomic analysis of phiEco32, a novel Escherichia coli bacteriophage. *J. Mol. Biol.* **377**, 774–789.
25. Severinov, K. (2000). RNA polymerase structure-function: insights into points of transcriptional regulation. *Curr. Opin. Microbiol.* **3**, 118–125.

26. Lane, W. J. & Darst, S. A. (2009). Molecular evolution of multisubunit RNA polymerases: sequence analysis. *J. Mol. Biol.* **395**, 671–685.
27. Lane, W. J. & Darst, S. A. (2009). Molecular evolution of multisubunit RNA polymerases: structural analysis. *J. Mol. Biol.* **395**, 686–704.
28. Cermakian, N., Ikeda, T. M., Miramontes, P., Lang, B. F., Gray, M. W. & Cedergren, R. (1997). On the evolution of the single-subunit RNA polymerases. *J. Mol. Evol.* **45**, 671–681.
29. Kazmierczak, K. M., Davydova, E. K., Mustaev, A. A. & Rothman-Denes, L. B. (2002). The phage N4 virion RNA polymerase catalytic domain is related to single-subunit RNA polymerases. *EMBO J.* **21**, 5815–5823.
30. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.
31. Iyer, L. M., Koonin, E. V. & Aravind, L. (2003). Evolutionary connection between the catalytic subunits of DNA-dependent RNA polymerases and Eukaryotic RNA-dependent RNA polymerases and the origin of RNA polymerases. *BMC Struct. Biol.* **3**, 1–23.
32. Zhang, Z., Schaffer, A. A., Miller, W., Madden, T. L., Lipman, D. J., Koonin, E. V. & Altschul, S. F. (1998). Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Res.* **26**, 3986–3990.
33. Soding, J., Biegert, A. & Lupas, A. N. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* **33**, W244–W248.
34. Willis, S. H., Kazmierczak, K. M., Carter, R. H. & Rothman-Denes, L. B. (2002). N4 RNA polymerase II, a heterodimeric RNA polymerase with homology to the single-subunit family of RNA polymerases. *J. Bacteriol.* **184**, 4952–4961.
35. Westblade, L. F., Minakhin, L., Kuznedelov, K., Tackett, A. J., Chang, E. J., Mooney, R. A. *et al.* (2008). Rapid isolation and identification of bacteriophage T4-encoded modifications of *Escherichia coli* RNA polymerase: a generic method to study bacteriophage/host interactions. *J. Proteome Res.* **7**, 1244–1250.
36. Florens, L. & Washburn, M. P. (2006). Proteomic analysis by multidimensional protein identification technology. *Methods Mol. Biol.* **328**, 159–175.
37. Eng, J., McCormack, A. L. & Yates, J. R., III (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Mass Spectrom.* **5**, 976–989.
38. Tabb, D. L., McDonald, W. H. & Yates, J. R., III (2002). DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.* **1**, 21–26.
39. Zybailov, B., Mosley, A. L., Sardiu, M. E., Coleman, M. K., Florens, L. & Washburn, M. P. (2006). Statistical analysis of membrane proteome expression changes in *Saccharomyces cerevisiae*. *J. Proteome. Res.* **5**, 2339–2347.
40. Paoletti, A. C., Parmely, T. J., Tomomori-Sato, C., Sato, S., Zhu, D., Conaway, R. C. *et al.* (2006). Quantitative proteomic analysis of distinct mammalian Mediator complexes using normalized spectral abundance factors. *Proc. Natl Acad. Sci. USA*, **103**, 18928–18933.
41. Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797.
42. Rost, B., Yachdav, G. & Liu, J. (2004). The PredictProtein server. *Nucleic Acids Res.* **32**, W321–W326.
43. Cuff, J. A., Clamp, M. E., Siddiqui, A. S., Finlay, M. & Barton, G. J. (1998). JPred: a consensus secondary structure prediction server. *Bioinformatics*, **14**, 892–893.
44. Korkhin, Y., Unligil, U. M., Littlefield, O., Nelson, P. J., Stuart, D. I., Sigler, P. B. *et al.* (2009). Evolution of complex RNA polymerases: the complete archaeal RNA polymerase structure. *PLoS Biol.* **e102**, 7.
45. Kuznedelov, K., Lamour, V., Patikoglou, G., Chlenov, M., Darst, S. A. & Severinov, K. (2006). Recombinant *Thermus aquaticus* RNA polymerase for structural studies. *J. Mol. Biol.* **359**, 110–121.