Molecular Microbiology (2005) **55**(3), 764–777 doi:10.1111/j.1365-2958.2004.04442.x

The tale of two RNA polymerases: transcription profiling and gene expression strategy of bacteriophage Xp10

Ekaterina Semenova,[1] Marko Djordjevic,[2] Boris Shraiman[3] and Konstantin Severinov[1,4]*
[1]*Waksman Institute, Rutgers University, Piscataway, NJ 08854, USA.*
[2]*Department of Physics, Columbia University, New York, NY 10027, USA.*
[3]*Department of Physics and* [4]*Molecular Biology and Biochemistry, Rutgers University, Piscataway, NJ 08854, USA.*

Summary

Bacteriophage Xp10 infects rice pathogen *Xanthomonas oryzae*. Xp10 encodes its own single-subunit RNA polymerase (RNAP), similar to that found in phages of the T7 family. On the other hand, most of Xp10 genes are organized in a manner typical of lambdoid phages that are known to rely only on host RNAP for their development. To better understand the temporal pattern of viral transcription during Xp10 development, we performed global transcription profiling, primer extension, chemical kinetic modelling and bioinformatic analyses of Xp10 gene expression. Our results indicate that true to its mosaic nature, Xp10 relies on both host and viral RNAPs for expression of genes coding for virion components and host lysis. The joint transcription of the same set of genes by two types of RNA polymerases is unprecedented for a bacteriophage. Curiously, such a situation is realized in chloroplasts.

Introduction

Bacteriophage Xp10 infects *Xanthomonas oryzae*, a rice pathogen. The genomic sequence of Xp10 has been determined (Yuzenkova *et al*., 2003). Bioinformatic analysis of the genome revealed that Xp10 is a highly unusual phage. Half of the Xp10 genome contains genes coding for structural proteins and host lysis proteins in an arrangement typical for many lambdoid phages; the other half contains genes coding for proteins involved in host shut-off, enzymes of viral genome replication and a T7-like RNA polymerase (RNAP). The two groups of genes are divergently transcribed and are separated by a regulatory intergenic region which contains divergent promoters recognized by the host RNAP *in vitro* (Fig. 1; see also Yuzenkova *et al*., 2003). Because λ and T7 have entirely different transcriptional control strategies, analysis of Xp10 gene expression could potentially reveal new paradigms of viral transcription regulation.

Like T7, Xp10 executes host transcription shut-off (Liao and Kuo, 1986). A 73-amino-acid-long Xp10 protein p7 binds to and inhibits *X. oryzae* host RNAP and therefore may be responsible for host shut-off (Nechaev *et al*., 2002). P7 is a novel protein, with no similarity to proteins in public databases. P7 interacts with the *X. oryzae* RNAP β′ subunit; the p7 binding site on RNAP must be evolutionarily variable, because p7 binds to *X. oryzae* RNAP but not to the closely related *Escherichia coli* enzyme. Xp10 p7 inhibits the ability of *X. oryzae* RNAP to recognize the major, –10/–35, class promoters but has no effect on the recognition of the minor, extended –10, class promoters. Leftward-oriented *X. oryzae* RNAP promoters identified in the Xp10 regulatory region are effectively inhibited by p7 *in vitro* (Yuzenkova *et al*., 2003). Thus, p7 may be involved in a switch from early, leftward-oriented transcription to rightward-oriented transcription of late genes coding for structural components of the Xp10 virion. The structural genes are separated by recognizable intrinsic terminators, suggesting that transcription antitermination may be required for their efficient expression. Indeed, Xp10 p7 prevents transcription termination by host RNAP at all intrinsic terminators tested irrespective of a promoter from which transcription is initiated or of transcribed sequence upstream of the terminator (Nechaev *et al*., 2002). Thus, p7 could act as an antiterminator that allows late gene expression. In fact, a rightward-oriented Xp10 promoter P3 was found to be resistant to p7 *in vitro* (Yuzenkova *et al*., 2003), suggesting that this promoter can function as a late promoter and be analogous to λ pR′ promoter, which is used to transcribe late λ genes located downstream of transcription terminators in the presence of phage-encoded antiterminator protein Q (Roberts *et al*., 1999). The problem with such a view is that it leaves no role for Xp10-encoded RNAP, whose gene occupies more than 8% of Xp10 genome. Moreover, it has been shown that late stages of Xp10 development are

Accepted 18 October, 2004. *For correspondence. E-mail severik@waksman.rutgers.edu; Tel. (+1) 732 445 6095; Fax (+1) 732 445 5735.

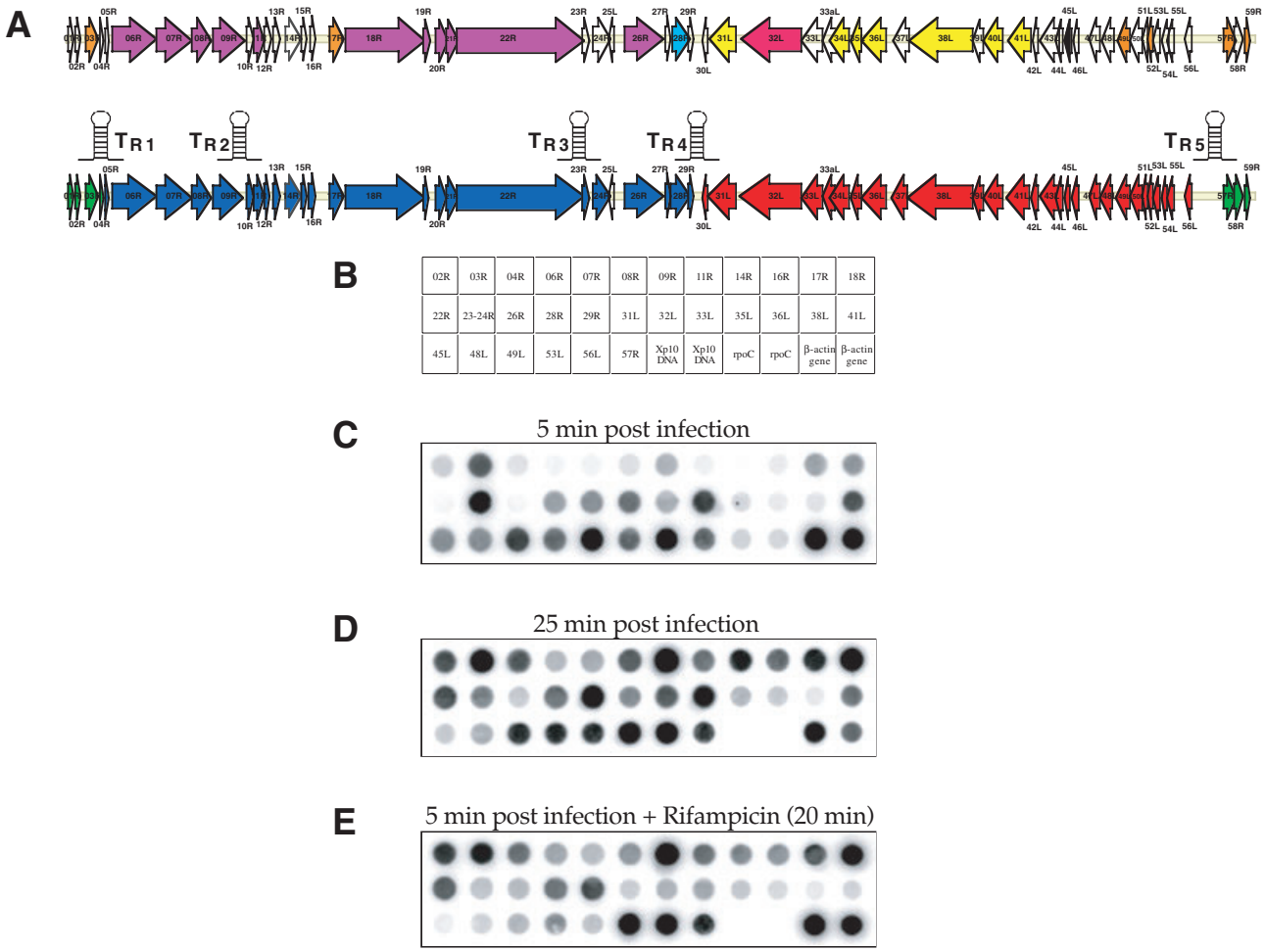

Fig. 1. Macroarray analysis of bacteriophage Xp10 gene expression.
A. The 44373 bp Xp10 genome is schematically shown. Genes are numbered from left to right end; Letters L and R identify genes which are transcribed in the leftward or rightward direction, correspondingly. At the top, the colour code indicates Xp10 genes belonging to different functional classes: structural genes, purple; genes involved in viral DNA replication, yellow, H-N-H endonuclease genes, orange, host lysis genes, cyan. Xp10 RNAP gene (32L) is coloured red-orange, the p7 gene (45L) is black. Xp10 genes for whose products no function can be predicted are indicated in white. At the bottom, the colour code corresponds to three temporal classes of Xp10 genes. L genes are red, early R genes are green, late R genes are blue. The positions of predicted intrinsic terminators (Table 2) are schematically shown.
B. The layout of a nylon macroarray containing PCR fragments corresponding to 30 Xp10 genes is schematically shown. The nomenclature of genes is the same as in panel A. See text for the explanation of control spots.
C–E. Representative macroarray data obtained with RNA prepared from cells collected 5 and 25 min post infection (C, D) or from cells that were treated with Rif 5 min post infection and were collected 20 min later (E) are shown.

Rifampicin-resistant, indicating that Xp10 RNAP transcribes late genes, at least in the presence of Rifampicin (Liao *et al.*, 1987).

To better understand transcriptional regulation of Xp10, we analysed the whole-genome pattern of Xp10 gene expression during the course of infection. Our analysis allows us to cluster Xp10 genes into three different classes, and to determine which Xp10 genes are transcribed by Xp10 RNAP and which genes are transcribed exclusively by the host RNAP. We also identify promoter sequences recognized by Xp10 RNAP, a diverse member of the T7 RNAP family.

Results

Experimental setup

Our goal was to understand the strategy of Xp10 transcription, particularly late in infection, because we hypothesized that the viral RNAP should be most active at this stage, transcribing genes coding for Xp10 virion components, i.e. the R genes. To this end, a macroarray of Xp10 genes was created by spotting equal amounts of PCR-amplified fragments corresponding to 30 representative Xp10 genes on a nylon membrane. The genes that were chosen represent a subgroup of the total of 59 annotated

Xp10 genes (Yuzenkova *et al*., 2003). The array contained 18 spots corresponding to 31 annotated Xp10 R genes. The number of spots is less than the total number of genes because some of the R genes either partially overlap or are so closely spaced (less than 14 bp separating the translation termination codon of the upstream gene and translation initiation codon of the downstream gene), that they are likely to be under common transcription control. Therefore, some of the spots on the array reported the amounts of transcripts from several different genes. For example, the spot labelled 11R reports transcript abundance from partially overlapping genes 10R, 11R, 12R and 13R. The array contained 12 spots corresponding to 28 annotated Xp10 L genes. Some of the spots reported the amounts of transcripts from several L genes which are likely to be under common transcriptional control (i.e. spot 38L reports transcript abundance of overlapping genes 37L, 38L, 39L and 40L). A spot corresponding to 25L, a gene of unknown function and the only leftward-oriented gene in the R cluster was not present in the array. In addition, spots for 42L, 43L, 46L, 51L and 55L, left-cluster genes of unknown function were also absent.

Xanthomonas oryzae cells were infected with Xp10 at the multiplicity of infection of 10. Cells were either immediately harvested or infections were allowed to proceed at 30°C for various times (from 1 to 60 min), followed by the preparation of the total cellular RNA (note that cells started to lyse at about 60 min). Equal amounts of total RNA from each time point were used to generate radioactively labelled cDNA by random priming/reverse transcription, and the samples were then hybridized to the Xp10 macroarray. In addition to PCR fragments corresponding to individual Xp10 genes, the array contained the following spots as controls (see Fig. 1): (i) two spots containing different fragments of the *X. oryzae rpoC* gene (to follow the fate of a representative cellular mRNA), (ii) two spots with different amounts of total Xp10 DNA (to quantify the total amount of Xp10-specific transcripts and to ensure that the differences in expression of individual Xp10 genes are within the linear range and (iii) two spots containing different amounts of a fragment of human β-actin gene (to control for hybridization efficiency and to allow comparisons between different arrays). Before hybridization, a random-primed/reverse transcribed product of total human RNA was added to the reaction to control for hybridization efficiency. After hybridization and washing, membranes were subjected to phosphoimager analysis; several representative membranes are shown in Fig. 1. Visual inspection of the array revealed several noteworthy features. First, the amounts of transcripts corresponding to different genes varied dramatically (compare, for example, spots corresponding to genes 07R and 56L). Second, the amounts of transcript corresponding to the same gene changed during the course of infection (compare, for example, spots corresponding to gene 09R at 5 and 25 min post infection – increase in the abundance of transcript during the infection – and spots corresponding to gene 45L – decrease in transcript abundance).

Macroarray data analysis reveals three classes of Xp10 genes

To quantitatively analyse macroarray data, the radioactive signal from each spot was corrected for background and normalized based on the relative strength of the β-actin spot signal. Next, the amount of radioactivity in each spot was plotted as a function of time post infection. The total amount of Xp10 RNA remained low till about 3 min post infection, increased rapidly between 3 and 10 min post infection, and remained relatively stable afterwards (Fig. 2A). In contrast, the amount of the *rpoC* transcript remained constant till about 7 min post infection and then decreased rapidly, and no transcript could be detected 25 min post infection (Fig. 2A), indicating that Xp10 executes host transcription shut-off, as expected.

Data analysis of individual Xp10 gene spots on the array was also performed by plotting normalized spot signal intensity (which corresponds to transcript abundance) as a function of time post infection. Abundance versus time plots show that L transcripts appear to behave according to the same pattern, characterized by rapid increase of abundance early in infection, and decrease of abundance at later times. As two typical examples, the pattern of abundance change for the host RNAP regulator p7 transcript (gene 45L) and for the Xp10 RNAP transcript (gene 32L) are shown in Fig. S1. In contrast, the R genes transcripts appeared to show two different abundance versus time patterns. For example, the abundance of the 03R transcript rapidly increased early in infection (even faster than the calculated average abundance of L transcripts) and remained at high level later in infection. The abundance of the 07R transcript, however, remained low and increased very slowly early in infection (significantly slower than the calculated average for all R genes) but then started to increase rapidly later (see Fig. S1).

To systematically cluster all phage transcripts corresponding to different spots on the array into classes the following procedure was used. First, for each transcript, the mean abundance during the first 15 min of infection was determined by calculating the area below the time versus abundance curve corresponding to times less than 15 min post infection and dividing this value by 15 min. The 15 min time point was chosen as a reference because at this time the abundance of L transcripts started to decrease. The mean abundance of each transcript during the entire infection period (60 min) was determined in a similar way. The ratio μ of the two numbers was next

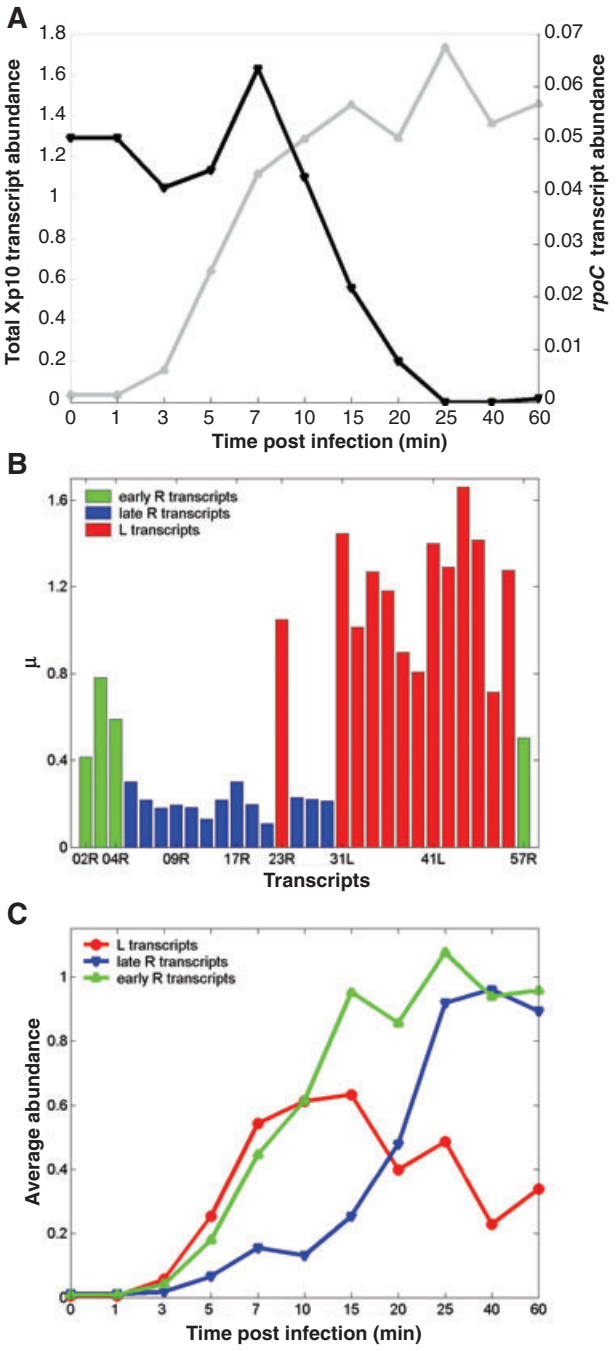

Fig. 2. Three classes of Xp10 genes.
A. The normalized (relative to β-actin spot) abundance of total Xp10-encoded RNA (grey line, determined as the amount of radioactivity hybridized to array spots containing total Xp10 genomic DNA), and *X. oryzae rpoC* mRNA abundance during the course of the Xp10 infection (black line) is presented.
B. Clustering of Xp10 genes into three classes was done by determining parameter μ (see *Supplementary material*). The value of μ calculated for all macroarray spots containing Xp10 genes is shown. Order of transcripts on the horizontal axis corresponds to the spatial position of corresponding genes in the genome.
C. Average abundance of Xp10 genes' transcripts belonging to different expression classes during the course of the Xp10 infection is presented.

calculated (see *Supplementary material*). From the definition of μ it follows that transcripts with an expression pattern like that of L transcripts should exhibit high μ values, transcripts with a pattern similar to 03R should exhibit intermediate values, and those similar to 07R should exhibit low μ values.

The μ values for every Xp10 gene spot on the array are presented in Fig. 2B in the order that corresponds to the linear order of genes in the genome. As can be seen, spots corresponding to the L gene cluster behaved as a uniform group and exhibited μ values of 0.7 or significantly higher. Most spots corresponding to the R gene cluster exhibited μ values of 0.3 or less. The obvious exception is spot 23–24R. The unusual behaviour of this spot is likely resulting from leftward transcription from 25L, the only left-oriented gene in the R gene cluster (see Fig. 1). Indeed, a potential leftward-oriented host RNAP promoter containing a consensus –35 element sequence TTGACA, a plausible –10 element TAc*T*cT (consensus sequence TATAAT), and an optimal 17 bp spacer is present in the intergenic segment between 25L and 26R.

Four spots corresponding to R genes located close to genome ends exhibited μ values close to 0.6. Because Xp10 genome contains *cos* sites, genes located proximal to genome ends should become close to each other when the genome assumes circular form and may likely be under common transcription control.

Although no information about spatial positions of genes entered the calculation of μ, transcripts with similar μ values showed clear separation according to their order in the genome (Fig. 2B). We note that clustering of Xp10 genes into three classes was robust with respect to a particular 'clustering method' (M. Djordjevic *et al.,* in preparation).

Once all transcripts were systematically clustered according to their μ values, abundance averages for all three classes were calculated for each time point, and the results are plotted in Fig. 2C. As can be seen, on average, transcripts corresponding to L genes started to accumulate approximately 1 min post infection, reached peak levels at about 10–15 min post infection and then declined. Bioinformatic analysis suggests that as a group, the L genes encode proteins involved in host shut-off, phage DNA replication and control of phage gene expression.

Rapid accumulation of transcripts of the R genes with low μ values started approximately 15 min post infection and their abundance continued to increase until late in infection. We call these genes 'late' R genes. The late R genes include all the structural, host lysis as well as the viral DNA packaging genes.

Accumulation of transcripts of R genes with intermediate μ values began together with the L transcripts accumulation but, unlike the L transcripts, their abundance did not decrease late in infection and they continued to accu-

mulate until 25 min post infection. We call these genes 'early' R genes. Of the seven early R genes, three encode endonucleases, while the rest are of unknown function. The product of the early R gene 04R encodes a protein that is highly similar to the N-terminal portion of Xp10 terminase, a product of late R gene 06R. Thus, 04R may be a pseudogene.

Macroarray analysis in the presence of Rifampicin reveals genes transcribed exclusively by host RNAP and genes transcribed by both viral RNAP and host RNAP

Xp10 encodes a T7-like single-subunit RNAP that likely participates in the later stages of Xp10 transcription. To determine Xp10 genes that are transcribed by the viral RNAP, we performed Xp10 infections followed by the addition of Rifampicin (Rif), a drug that inhibits the host RNAP, at 0, 1, 3, 5, 7, 10, 15, 20, 25, 40 and 60 min post infection and an additional 20 min incubation in the presence of the drug. Macroarrays obtained with RNA from cells that have been incubated with Rif were compared with corresponding arrays obtained using RNA prepared from cells immediately before Rif addition. Normalized signals for individual spots were determined and a difference (Δ) between corresponding spots' signals was calculated and plotted as a function of time of Rif addition (Fig. 3). The value of Δ equals the amount of transcript that was generated during the period when infected cells are incubated with Rif less the amount that decayed during the same period. Because transcript decay is expected to be directly proportional to transcript concentration, the contribution of decay to Δ-value shall become significant only at later stages of infection, when sufficient amounts of transcript are produced. Conversely, early in infection, the value of Δ is dominated by the amount of transcript that has been synthesized during the period of drug administration, provided that this amount is not zero. In the absence of significant decay (i.e. early in infection), for a gene transcribed by Rif-resistant Xp10 RNAP, Δ should assume a positive value or, if Rif was added before Xp10 RNAP was produced, equal 0. For a gene transcribed only by Rif-sensitive host RNAP Δ should assume a non-positive value at all times during infection. As can be seen from Fig. 3, Δ was uniformly negative for the L genes, uniformly positive for the late R genes, and initially assumed a positive value but eventually became negative for the early R genes. In all cases, the absolute value of Δ decreased at later stages, presumably because of decreased decay for leftward-transcribed genes (because their transcript abundances decreased at later stages) and increased decay for rightward-transcribed genes (because their transcript abundances increased at later stages).

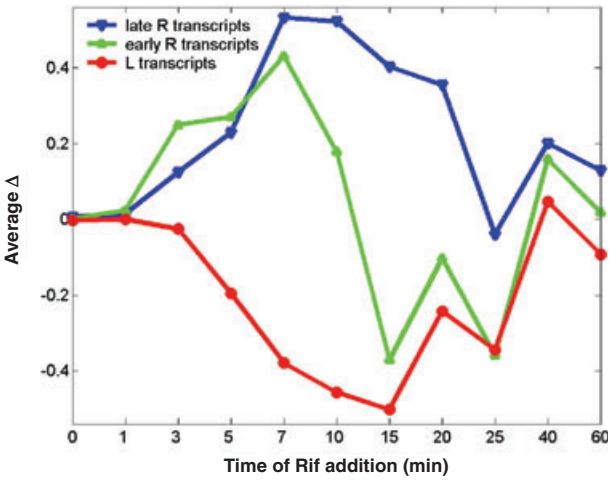

Fig. 3. mRNA of R, but not L genes continues to accumulate after the treatment with Rifampicin. Macroarray data for individual Xp10 gene spots were compared by determining the Δ parameter, a difference between normalized mRNA abundance in cells that were subjected to a 20 min Rifampicin treatment at *t* and cells collected at time point *t*. Average Δ-values for the three classes of Xp10 genes (Fig. 2C) are shown.

The results in Fig. 3 clearly show that the host RNAP is involved in L gene transcription. In fact, the Δ parameter values for these genes throughout the course of infection are in agreement with a model that assumes zero generation during the time of Rif administration (M. Djordjevic *et al*., in preparation). Thus, the L genes should be exclusively transcribed by host RNAP without any contribution from Xp10 RNAP. For all L transcripts, half-life estimates obtained using the decay model are much shorter than the time of the infection (M. Djordjevic *et al*., in preparation). Analysis of L gene transcript abundance in the absence of Rif, using decay rates extracted from data in the presence of Rif, allowed us to predict how host RNAP transcription activity for L genes (defined as amount of L gene transcripts generated per unit of time) changes in the course of infection (M. Djordjevic *et al*., in preparation). The results indicate that the activity of host RNAP on L genes reaches maximal value 5 min post infection, decreases rapidly afterwards and is not significantly different from zero for times greater than 15 min post infection.

The data presented in Fig. 3 indicate that viral RNAP has to be involved in both early and late R gene transcription. The fact that Δ became positive for R genes at 3 min post infection indicates that at least some amounts of Xp10 RNAP are produced at this early point of infection. The fact that Δ becomes negative for early R genes 15–25 min post infection and remains positive for late R genes can be explained by the fact that transcript abundances of early R genes approach their maximal values at this time, while the late R transcript abundances are still low. This should result in much larger amount of early R tran-

scripts decaying during Rif administration (as compared to late R transcripts), which leads to negative Δ-values.

To determine the contribution of host RNAP to R gene transcription, macroarray data were further analysed. For each R gene, α, a difference between the measured transcript abundance 20 min after Rif addition at time *t* and measured (or interpolated from measurements) transcript abundance at time *t* + 20 min in the absence of Rif was determined (see *Supplementary material*). α therefore reflects a change in transcript abundance at time *t* + 20 min caused by incubation with Rif during time interval from *t* to *t* + 20 min If α assumes a value of zero, host RNAP does not transcribe the corresponding genes during the time interval from *t* to *t* + 20. The negative value of α can be resulting from (i) inhibition of host RNAP by Rif (a direct effect) or (ii) decrease in the amount of (and hence transcription by) viral RNAP because of inhibition of viral RNAP gene transcription by host RNAP (an indirect effect). A combination of the two effects is also possible. Because host RNAP ceases to transcribe L genes at 15 min post infection, the indirect effect of Rif addition can be discounted at times later that 15 min.

From quantitative analysis it follows that in the absence of indirect effects of Rif addition (that is for times greater than 15 min post infection), α is equal to the increase of transcript abundance resulting from host RNAP activity from *t* to *t* + 20 min (see *Supplementary material*). Average α values for different times post infection were calculated separately for early R transcripts (green line) and for late R transcripts (blue line) (Fig. 4). As can be seen, the α value curves for both the early and the late R transcripts matched, except for very early times when absolute values of α for early R genes were significantly higher than absolute α values for late R genes. The fact that incubation with Rif during the time interval between 15 min and 35 min post infection leads to an α value that is clearly less than zero strongly indicates that host RNAP contributes to R genes transcription at least until 15 min post infection. For both early R and late R transcripts the absolute values of α decreased during infection, indicating that for times greater than 15 min post infection, the relative involvement of host RNAP in transcription of R genes decreases.

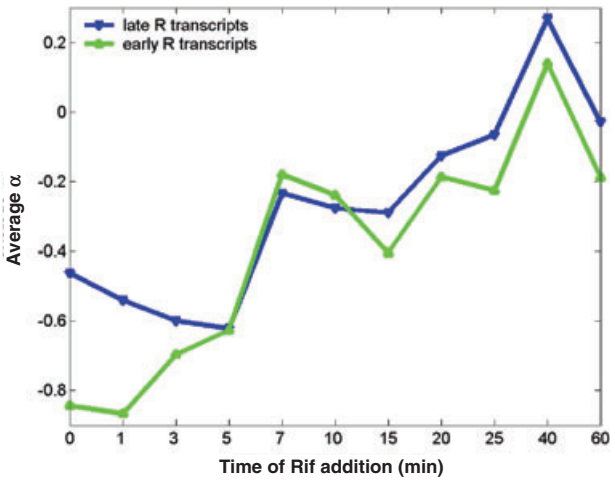

Fig. 4. Analysis of R genes transcription by host RNAP during Xp10 infection. Macroarray data for Xp10 R gene spots were compared by determining the α parameter, a normalized difference between mRNA abundance in cells collected at time point *t* + 20 min with Rifampicin and the same time without Rifampicin. Average α values for the two classes of R genes are shown.

Identification of rightward Xp10 promoters

Preliminary experiments showed that purified recombinant Xp10 RNAP was inactive on Xp10 DNA and did not recognize T7 RNAP promoters (data not shown). Furthermore, a bioinformatic search of Xp10 genome for T7 RNAP promoter-like sequences, using a weight matrix constructed on the basis of known or predicted T7-like RNAP promoters (see *Experimental procedures*) obtained by QPMEME algorithm (Djordjevic *et al*., 2003) found no statistically significant matches. This suggested strongly that Xp10 RNAP promoters are different from other known T7-like RNAP promoters.

To identify viral RNAP promoters, we combined data obtained from the macroarray experiment with a bioinformatic approach. The macroarray data lead to a conclusion that R genes have to be transcribed by Xp10 RNAP. Therefore, we carried out a search for statistically overrepresented motifs focusing on the intergenic regions upstream of R-transcribed genes. We looked for overrepresented motifs, because in T7-like phages viral RNAP promoters are present in multiple copies (Molineux, 2005). Two independent 'unsupervised' searches were carried out using Gibbs sampling algorithm (Lawrence *et al*., 1993), and MEME algorithm (Beily and Elkan, 1994). Both algorithms converged to the same result and revealed two motifs, designated as Motif 1 and Motif 2 (Table 1). We have used the QPMEME algorithm to construct weight matrices for Motif 1 and Motif 2 (see *Experimental procedures*) on the basis of sequences presented in Table 1, and used them to search the entire Xp10 genome for additional sites. However, no additional sequences similar to either Motif 1 or Motif 2 were revealed.

Two perfect direct 43 bp repeats are present inside the long (1085 bp) intergenic region which separates the R and L genes (first copy at 42335–42377, second copy at 42599–42641). These repeats contain copies of both Motif 1 and Motif 2. The 'unsupervised' search converges to Motif 1 even if one 43 bp repeat is omitted from the analysis, which shows that this motif is not entirely based on the repeat. On the other hand we do not have conver-

Table 1. Over-represented sequences inside Xp10 intergenic regions revealed by 'unsupervised' search.

Xp10 motif	Positions	Sequence	Score[a]
Motif 1	42335–42357	tgaggcacctatagagaagaact	5.51
	42599–42621	tgaggcacctatagagaagaact	5.51
	42643–42665	gtaggcacattgagagcagggca	5.25
	42675–42697	tgacagctagtaacagcagacga	5.16
	42941–42963	gtaaagcttataacagaaaagca	5.22
	42972–42994	tgaagcattataacagcaaagca	5.66
	44300–44322	tgaaagatagagacagaagaggg	5.39
Motif 2	42305–42327	atacctgcaagatagccgacgtt	6.26
	42354–42376	aactctgtattttagacgaagtt	7.04
	42618–42640	aactctgtattttagacgaagtt	7.04
	44337–44359	aagtcggcatgatcgccgaactt	6.43

a. Scores were obtained by using weight matrices given in *Experimental procedures*.

gence to Motif 2 if we omit one repeat copy, which is a consequence of a small number of sequences that belong to Motif 2 (Table 1).

To experimentally locate R transcript start points, primer extension reactions were performed using primers that were designed to reveal transcription initiation events in Motif 1 and 2 sequences presented in Table 1. In addition, primer extension reactions were performed with a set of primers annealing downstream of intergenic regions of the R genes. As a template for primer extension we used RNA prepared from cells that were collected 15 min post infection, a condition when the R transcripts are present in significant amounts and continue to accumulate.

Several distinct primer extension products inside the Xp10 intergenic regions were revealed. It should be noted that in addition to transcription start points, primer extension reveals sites of RNA processing and that these two classes of primer extension products can not be distinguished. Secondary-structure analysis of RNA sequences in the immediate vicinity of primer extension end-points was performed using MFOLD program (Mathews *et al*., 1999). Almost all primer extension end-points occurred in regions that had a potential to form extensive secondary structures, which are presented in Table S1. The only primer extension product for which no secondary structure was found is located at Xp10 genome position 42588. Neither the primer extension sites listed in Table S1 nor the 42588 site matched potential promoter sequences for either host of viral RNAP. We therefore suggest that these primer extension products correspond to processing events and do not reflect transcription initiation events within the late gene cluster.

Primer extension analysis with primers annealing downstream of the long intergenic region separating the L and R genes revealed RNA 5′ ends originating in Motif 1 and Motif 2 sequences of each 43 bp repeat. For Repeat 1, these primer extension products are labelled as R1M1 and R1M2 (for Repeat 1, Motif 1 and Motif 2 respectively). R2M1 and R2M2 denote corresponding primer extension products in Repeat 2 (Fig. 5). In addition, a primer extension product corresponding to the previously identified P3 host RNAP promoter (Yuzenkova *et al*., 2003) was present. Finally, a transcript that originated upstream of both 43 bp repeats and preceded by appropriately positioned −10 and −35 bacterial promoter elements was identified and is referred to as P_{UP} in Figs 5 and 6.

Identification of Xp10 RNAP promoters

Because primer extension products originating from the 43 bp repeats corresponded to predicted Xp10 RNAP promoter sequences, we determined whether these transcripts (i) accumulate late in infection and (ii) accumulate in the presence of Rif. The kinetics of accumulation of transcripts that originated in the intergenic region in cells collected at different times after infection is shown in Fig. 6. In this experiment, two primers were used. The primer used in experiments presented in Fig. 6A and C annealed approximately 80 bases downstream of Repeat 1 and allowed to follow the accumulation of R1M1, R1M2 and P_{UP} primer extension products. The primer used in experiments presented in Fig. 6B and D annealed 125 bases downstream of P3 and allowed to follow the accumulation of P3 and both the Repeat 1 and Repeat 2 primer extension products. However, Motif 1 and 2 transcripts originating within the Repeats could not be resolved in this case. As can be seen, the P3 and P_{UP} products accumulated early in infection; they became detectable 3 min post infection and reached their maximum 5 min (P_{UP}) and 10 min (P3) post infection (Fig. 6A and B, compare lanes 2, 3 and 5 respectively). In contrast, the amount of products from Repeats 1 and 2 was low in the initial stages of the infection and reached significant levels 20 min post infection (Fig. 6A and B, compare, for example, lanes 4 and 6).

The experiment was next repeated with RNA prepared from cells collected at 5 and 25 min post infection (Fig. 6C and D, lanes 1 and 3, correspondingly), and RNA prepared from cells that were subjected to a 20 min treatment with Rif 5 min post infection (Fig. 6C and D, lane 2). As can be seen, the amount of P3 and P_{UP} products in RNA prepared from cells treated with Rif was lower than in RNA prepared from cells either at 5 or at 25 min post infection. This is consistent with the idea that these products are synthesized by Rif-sensitive host RNAP, and that upon addition of Rif some (P3) or all (P_{UP}) of these RNAs become processed or degraded. In contrast, products that originated within Repeats 1 and 2 accumulated even in the presence of Rif, such that the final levels at 25 min post infection were similar with or without Rif treatment (Fig. 6C and D, compare lanes 2 and 3, correspondingly). We therefore conclude, that RNAs originating from

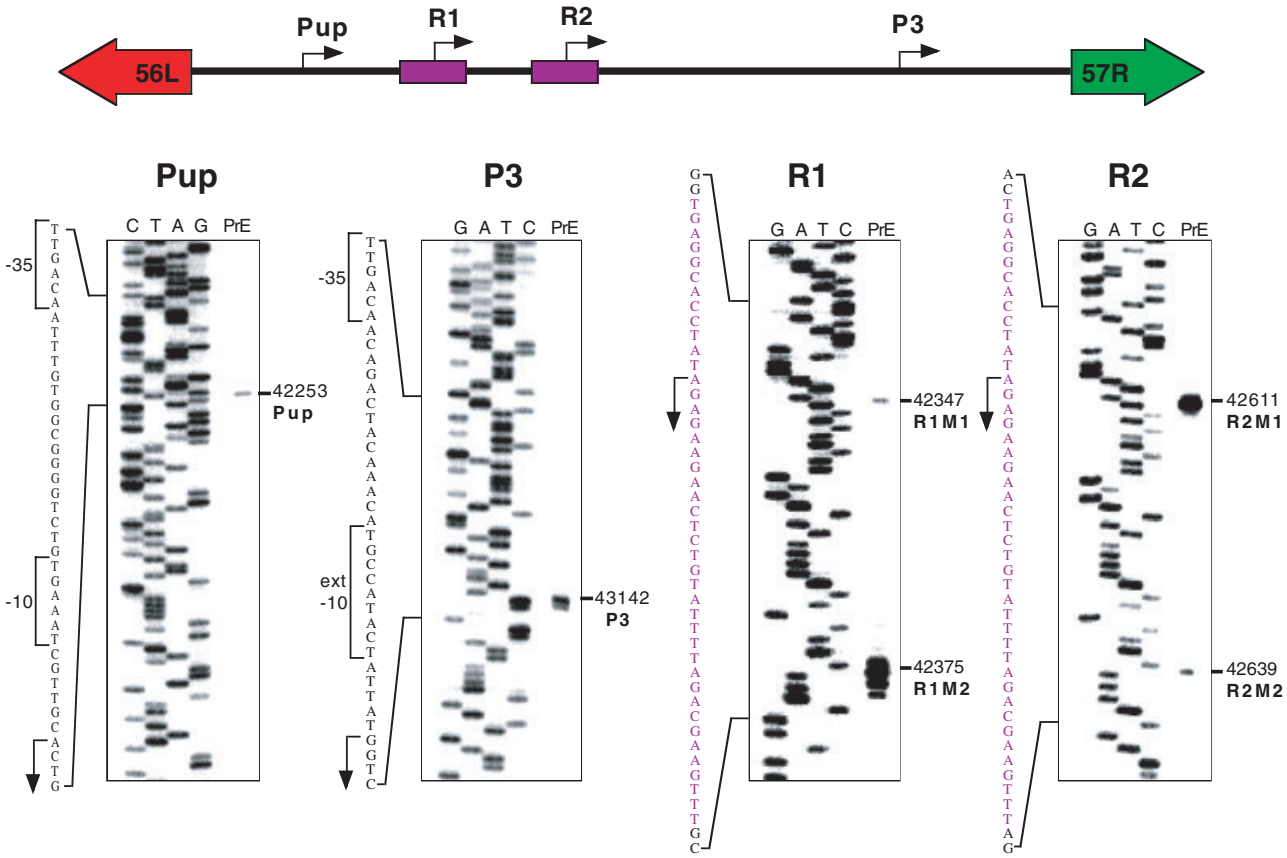

Fig. 5. Primer extension analysis of rightward-transcribed Xp10 RNAs originating in the long intergenic region. At the top of the figure, the intergenic region (genome positions 42098–43181) and the flanking divergently transcribed Xp10 genes are shown. The locations of experimentally determined primer extension products are indicated by rightward-directed arrows. The positions of two perfect 43 bp long repeats are indicated. Below, the results of primer extension with several primers annealing within the intergenic region are shown. A sequencing reaction run with the same primer is shown alongside the products of primer extension reaction and the primary sequence is written on the left of each panel. The primer extension product positions are indicated. Potential host RNAP promoter elements are also indicated. The sequences of 43 bp repeats are indicated in purple colour. Only the primer extension reactions whose products matched with predicted or experimentally determined *X. oryzae* or Xp10 RNAP promoters are shown; the complete list of all primer extension products is presented in Table S1.

Repeats 1 and 2 must be transcribed by Rif-resistant Xp10 RNAP and that the viral RNAP is relatively inactive early in infection but becomes highly active at later stages. In addition, we conclude that both the host and viral RNAPs transcribe R genes; the host RNAP is most active early in infection; however, it continues to transcribe R genes at least until 5 min post infection. This must be so, for if host RNAP became inactivated after 5 min post infection, we would expect to see little or no host RNAP transcripts by 25 min post infection (a situation observed when Rif is added at 5 min post infection). Instead, we see substantial amounts of host RNAP transcripts. In fact, the amount of the P3 transcript does not decrease until 25 min post infection (Fig. 6B, lanes 1–8), indicating that host RNAP must be active in transcription from the P3 promoter at least to this time. On the other hand, the amount of P_{UP} transcript starts to decrease significantly earlier (at around 5 min post infection, Fig. 6A, lanes 1–8). This is probably a consequence of the fact that P3 belongs to extended −10 class (and is therefore resistant to p7, Yuzenkova *et al.*, 2003) while P_{UP} is a −10/−35 class promoter and is therefore sensitive to p7.

The primer extension experiment does not allow to determine which of the two motifs in the 43 bp repeats, Motif 1 or Motif 2, are recognized by Xp10 RNAP. To experimentally locate the 5′ ends of Xp10 RNAP-initiated transcripts, RNA ligase-mediated RT-PCR (Bensing *et al.*, 1996) was performed using RNA samples prepared from cells that were infected for 10 min and treated with Rif for additional 20 min. In this method, one half of the RNA sample is treated with TAP (tobacco acid pyrophosphatase), an enzyme that converts 5′ triphosphates of RNA into monophosphates that can be ligated to exogenously added RNA oligonucleotide by RNA ligase. Another half of the RNA sample is left untreated and is used as a control. After the ligation, reverse transcription and PCR amplification are used to identify DNA fragments whose abundance *increases* after TAP treatment (such

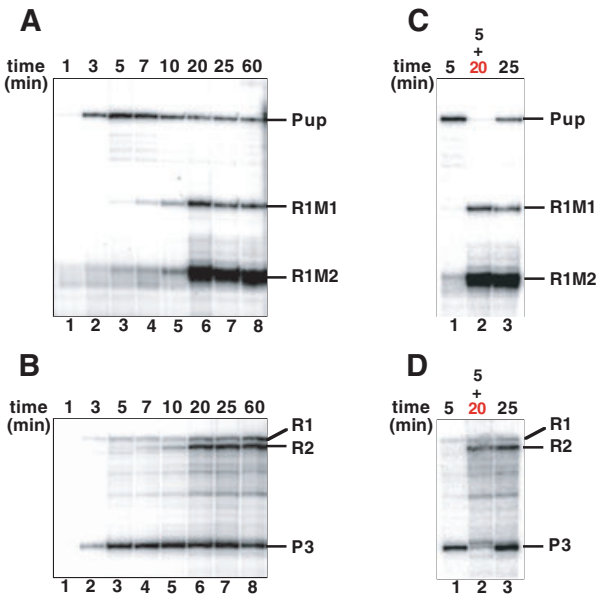

Fig. 6. Transcripts originating from 43 bp repeats accumulate late in infection in Rifampicin-resistant manner. Results of primer extension with two primers annealing downstream of 43 bp Repeat 1 (A and C) and host RNAP P3 promoter (B and D) are shown. As a template for primer extension, RNA prepared from cells collected at the indicated times post infection was used. In C and D, RNA prepared from cells that were treated for 20 min with Rifampicin 5 min post infection was used for primer extension in lanes 2.

fragments must be generated from RNA molecules that contained 5′ triphosphates, and must therefore correspond to transcription initiation start points). The results of an experiment performed with a set of oligonucleotides that should reveal transcription initiation events in both Repeat 1 and Repeat 2 are presented in Fig. 7. As can be seen, for both Repeats, two PCR fragments were observed, as expected from the results of the primer extension experiments, Fig. 6. DNA sequencing of PCR fragments revealed that they originated from a product of ligation between the RNA oligonucleotide and Xp10 RNAs that corresponded exactly to primer extension products in Motifs 1 and 2. The appearance of longer PCR fragments was stimulated by TAP treatment (Fig. 7, compare lanes 1 and 2, and 5 and 6). We therefore conclude the 5′ ends of R1M1 and R2M1 primer extension products correspond to transcripts initiated by Xp10 RNAP and that Motif 1 sequences are therefore part of the Xp10 RNAP promoter. The R1M2 and R2M2 primer extension products must arise from RNA processing at Motif 2.

Bioinformatic analysis of Xp10 transcription terminators

The existence of two classes of R genes whose transcription initiates from a common area in the intergenic region separating the L and the R genes suggests that there should be a transcription terminator(s) that separates R genes belonging to the two classes. To identify putative transcription terminators we have scanned both strands of the Xp10 genome for the presence of stem–loop structures with uracil-rich tails, using the algorithm described by Ermolaeva *et al*. (2000). The algorithm assigns potential terminators with an 'energy score' that measures the stability of the stem–loop and a 'tail score' which characterizes the proximity and composition of the uracil-rich sequence. In Table 2, we list all candidates with 'energy score' greater than 7.4 and 'tail score' greater than 3.3 (these putative terminators are also schematically illustrated in Fig. 1A). As a reference, a highly efficient terminator TE found in the T7 phage genome has an 'energy score' of 11.4 and a 'tail score' of 4.0. As can be seen from Table 2, a putative terminator T_{R1} that separates the early R genes from late R genes exists, in the 3′ end of the 04R gene. Based on its position, we propose that the function of this putative terminator accounts for different kinetics of early and late R transcripts accumulation. The lower score of T_{R1} may indicate lower termination efficiency, which would be consistent with the fact that small but significant amounts of late R transcripts accumulate during the first 10 min post infection (Fig. 2C).

Most other putative terminators of R gene transcription are located in intergenic regions separating late R genes. They may have biologically significant functions, because they separate genes coding for viral head structural proteins from genes coding for head-tail joining proteins (T_{R2}), and genes coding for tail proteins from host lysis genes (T_{R3}). The strongest predicted terminator, T_{R4}, is located at the end of the string of R-genes and could function to prevent invasion of R gene transcription into the L gene cluster. One predicted terminator is located at genome

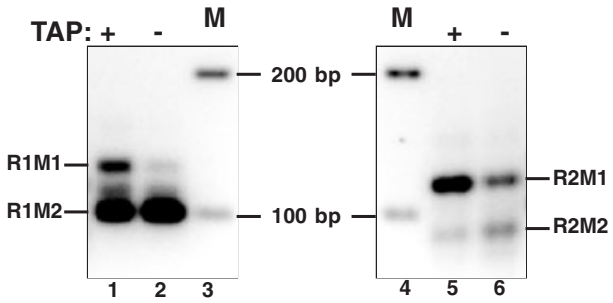

Fig. 7. Identification of Xp10 RNAP promoters. RNA prepared from Xp10-infected cells was treated with TAP (lanes 1 and 5). RNA in lanes 2 and 6 was not treated with TAP and used as a control. The samples were ligated to an exogenous RNA oligonucleotide and subjected to RT-PCR as described in *Experimental Procedures*. Primers used for PCR amplify transcripts originated in 43 bp Repeat 1 (lanes 1 and 2) or Repeat 2 (lanes 5 and 6). The PCR products were sequenced and the junction sites between the exogenous oligonucleotide sequence and Xp10 sequences identified the 5′ ends of Xp10 RNAs originating from Repeats 1 and 2. The 5′ ends corresponded to primer extension products from Fig. 5 and are labelled correspondingly. Lanes 3 and 6 are DNA molecular weight marker (M) lanes.

Table 2. Predicted host RNAP intrinsic terminators found in the Xp10 genome.

Terminator name[a]	Positions	Strand	Sequence	Energy score	Tail score
T_{R1}	1390–1411	R	ctgccctacttatgggcagttt[b]	7.4	3.5
T_{R2}	6573–6603	R	gggaggggctgggaaactggccccctctcttt	13.4	3.6
	10445–10429	L	gcgtcgttgacgctttt	9.1	4.1
T_{R3}	19285–19308	R	ggggcagggtttcctgccccattt	15.0	3.3
	21983–22201	R	gtctgtggcaacagactttt	7.6	3.6
T_{R4}	23718–23760	R	gggagggagctaagccttaatggcctagcccctccctttttt	15.5	5.9
	31618–31636	R	gagggccatgctctctttt	7.8	4.1
	36624–36662	R	ggcgccgtcgccagtagctactgccacgccgccttgttt	8.4	4.1
T_{R5}	42740–42759	R	ctgaacgatccgttcagttt	10.9	3.5
	44027–44011	L	cagggtcaaccctgttt	8.1	3.6

a. Only terminators with plausible biological functions (see text) are named.
b. Stretches of sequence forming potential stems of the stem–loop structure are underlined.

position 21983, deep inside the 26R gene, the last gene in the tail genes cluster, and may not be biologically significant. Two additional putative terminators of rightward transcription are located within the L genes cluster and are unlikely to play a role in transcription regulation. Finally, one putative rightward transcription terminator, T_{R5}, is located in the intergenic region that separates the L and the R genes and may therefore be involved in regulation of rightward transcription.

Our search did not reveal any potentially significant terminators of leftward transcription (the locations of two putative terminators that are listed in Table 2 make it highly unlikely that they are biologically significant). No putative terminator sequences were located downstream of 25L gene, which is located in the R gene cluster and is likely transcribed from its own promoter (see above). It is therefore possible that terminators of L gene transcription have different structure than the 'hairpin+poly-U' model which is assumed by our search.

Discussion

Our analysis indicates that three transcriptional classes of Xp10 genes exist. All L genes belong to the first class; they are characterized by rapid increase of transcript abundances in early stages of infection and decrease of abundances in later stages. A small number of R genes located proximal to the Xp10 genome ends belong to the second class, which we call early R, characterized by rapid transcript abundance increase in early stages and the absence of transcript abundance decrease in later stages. Most of the R genes belong to the third class, which we call late R, characterized by very slow transcript accumulation in early stages and rapid growth of abundance in later stages. The following view of Xp10 gene expression is compatible with our data. Host RNAP appears to be solely responsible for transcription of L genes. It also transcribes early R genes at the beginning of infection. Significant transcription of late R genes, in the early stages of infection, is likely prevented by a terminator that separates the early and late R genes. Two L gene products allow expression of the R genes at the later stages of infection. Transcription initiated by Xp10 RNAP, a product of 32L, proceeds to the end of the late R gene cluster in apparent disregard of transcription terminators. Host RNAP should also be able to contribute to late R gene cluster transcription through the antiterminator function of the p7 protein, the product of 45L. Although both RNAPs can contribute to R gene expression, the fact that Xp10 virions are produced when Rif is added to infected cells, indicates that phage development can occur even in the absence of R gene transcription by host RNAP. However, the yield of phage particles produced is lower when Rif is added early (Yuzenkova *et al.*, 2003) suggesting that joint transcription by both enzymes may make phage development more efficient.

The view of Xp10 gene expression that emerges from our study is consistent with the earlier work of Liao *et al.* (1987). These authors used Xp10 RNAP purified from the infected cells as well as purified *X. oryzae* RNAP to transcribe from the Xp10 genome *in vitro*. They then hybridized the *in vitro* transcribed RNA to Xp10 genomic DNA digested with several restriction enzymes to determine which regions of the genome are transcribed by which enzyme. Analysis of their low-resolution data suggests that in agreement with our results, the Xp10 RNAP strongly transcribes the region of the genome where the R genes are located, while the host RNAP is mostly involved in L gene transcription.

The pattern of transcript abundance versus time characteristic for early R genes is most likely insured by simultaneous transcription by both host and phage RNAP. That is, in the early stages of infection, when phage RNAP did not significantly accumulate, rapid early R transcript increase must be resulting from host RNAP activity. In the later stages of infection, high early R transcript levels are maintained resulting from phage RNAP activity. Thus, the simultaneous transcription by phage and host RNAP ensures high transcript levels of early R genes throughout the infection. The significance of this is not known, how-

ever, because the function of the early R gene products is not known.

Our data strongly suggest that R gene transcription by host RNAP continues at least until 15 min post infection. Viral RNAP is also transcribing R genes at this time. Thus, Xp10 evolved a strategy that combines the late gene expression strategy of lambdoid phages (host RNAP modified by an antiterminator protein) and the T7-like phages (viral RNAP which is a product of an early gene). It is not clear what physiological circumstance may favour such a strategy. We note, that in chloroplasts, genes that are simultaneously transcribed by both the plastid-encoded, bacterial-like, RNAP and nucleus-encoded, T7-like, RNAP exist (Allison *et al*., 1986).

What is the relative contribution of host and viral RNAPs to the early R and late R gene transcription? Figure 4 indicates that very early in the infection, host RNAP contributes significantly more to transcription of early R genes than to the late R genes transcription. This is probably resulting from the T_{R1} terminator located between the early R and late R genes. Later in the infection, there appears to be no significant difference between host RNAP contribution to early R and late R gene transcription. This is probably a consequence of accumulation of the p7 protein which has been shown to exhibit antitermination activity *in vitro* (Yuzenkova *et al*., 2003). Analysis of the α parameter (Fig. 4) suggests that the contribution of host RNAP to R gene transcription gradually decreases as the infection proceeds. However, more refined analysis (M. Djordjevic *et al*., in preparation) indicates that even at late stages of infection (25 or more min post infection) the contribution of host RNAP to the total R gene transcription is about 25%. Because our results suggest that phage transcripts decay fast compared to the length of infection (M. Djordjevic *et al*., in preparation), it is likely that sufficient late R transcript levels could not be achieved in later stages of infection without the involvement of viral polymerase.

One of the goals of our work was to identify promoters recognized by Xp10 RNAP, because the enzyme appeared to be inactive *in vitro*. Our results demonstrate that Xp10 RNAP promoters coincide with bioinformatically identified Motif 1 sequences located in the 43 bp repeats in the Xp10 intergenic region. This Motif 1 sequence weakly resembles known T7 RNAP-like promoters: T7 consensus sequence is taatacgactcactata**G**ggaga, while DNA sequence proximal to Motif 1, aligned with T7 consensus is agtggtgaggcacctat**A**gagaa (transcription initiation sites are marked in bold, identical residues are underlined). Because Xp10 RNAP is inactive on Xp10 DNA *in vitro*, it is likely that an additional specificity factor is required for efficient *in vitro* transcription. Thus Xp10, a mosaic between a T7-like and a λ-like phage, may rely on a mechanism used by yet another unrelated phage N4 for expression of genes transcribed by viral RNAP (Carter *et al*., 2003).

Experimental procedures

Preparation of Xp10 macroarray

Macroarray membranes were designed on the basis of published Xp10 genome (GenBank/NCBI Accession No. AY299121). Thirty open reading frames (ORFs) were selected from the total of 59 potential Xp10 ORFs. When Xp10 ORFs partially overlapped or had no or a very short (<14 bp) intergenic region, one ORF of the set was chosen for macroarray.

DNA fragments (320 bp) corresponding to each of the selected Xp10 ORF were amplified from genomic Xp10 DNA using gene-specific primers. The sequences of the primers are available from the authors on request. PCR was performed using New England BioLabs Vent DNA polymerase under the following conditions: initial heating at 94°C for 2 min followed by 25 cycles of 94°C, 30 s; 50°C, 30 s; 72°C, 1 min, with final extension at 72°C for 10 min. PCR products were purified from agarose gels using QIAquick Gel Extraction Kit (QIAGEN). To increase the yield and purity of PCR products, a second round of amplification was performed with first-round products as templates. Second-round products were purified using QIAquick PCR Purification Kit (QIAGEN). The concentration of the products was determined by absorbance at 260 nm.

Two fragments of *X. oryzae rpoC* gene (823 and 911 bp) were amplified using gene specific primers and *X. oryzae rpoC* expression plasmid (Nechaev *et al*., 2002) as a template and were purified as above. A 353 bp fragment of human β-actin gene was obtained by RT-PCR from total human RNA provided in the SuperScript III first-strand synthesis system for RT-PCR (Invitrogen). Xp10 genomic DNA was purified from phage lysates as described previously (Yuzenkova *et al*., 2003).

Xp10 ORFs fragments and controls were spotted onto a positively charged nylon membrane Immobilon-Ny+ (Millipore) using S&S Minifold I (Filtration Manifold for Dot-Blot Assays, Schleicher and Shuell). DNA was denaturated by alkali/heat treatment (0.4 M NaOH and 10 mM EDTA, 10 min at 100°C). The amount of DNA per spot was 100 ng for Xp10 ORFs and *rpoC* gene, 25 ng and 10 ng for total phage DNA, 100 ng and 25 ng for β-actin gene. After applying the samples to the membrane, DNA was fixed with UV cross-linking. The array layout is shown in Fig. 1. Each membrane was used once for hybridization.

Bacterial and phage growth conditions

Xanthomonas oryzae strain XO604 (Liao and Kuo, 1986) was grown at 30°C with agitation in TGSC medium [10 g of Bactotryptone, 5 g of Bactosoytone, 5 g of NaCl, 2 g of glucose and 0.5 g of $Ca(NO_3)_2$ per litre]. *Xanthomonas oryzae* cells in the exponential phase of growth were infected by Xp10 phage at a multiplicity of 10. Two-millilitre aliquots of the infected cells were removed at various time points (0, 1, 3, 5, 7, 10, 15, 20, 25, 40 and 60 min) and added to an equal volume of cold stop-solution (Liao *et al*., 1987; 20 mM Tris-

HCl pH 8.0, 0.1 M $MgCl_2$, 0.1 mM EDTA pH 8.0, 0.1 mM DTT and 0.1 M sodium azide). To analyse an effect of Rif on Xp10 gene expression, additional 2 ml aliquots of the culture were withdrawn at times indicated above, combined with Rif at a final concentration of 100 μg ml^{-1} and incubated at 30°C with agitation for additional 20 min. The infection was stopped as above. Cells were collected by centrifugation at 5000 g for 5 min at 4°C and stored at −80°C.

RNA extraction and labelling

Total RNA from infected cells was isolated using RNeasy Mini Kit (QIAGEN) according to manufacturer's instructions and including the DNase I digestion step. The RNA samples were analysed by formaldehyde-agarose gel electrophoresis and quantified by absorbance at 260 nm.

The cDNA probes for array hybridization were synthesized from total RNA by reverse transcriptase using SuperScript III first-strand synthesis system for RT-PCR (Invitrogen) following manufacture's protocol. Briefly, 3 μg of RNA was combined with 50 ng of random hexamer primers, denaturated at 65°C for 5 min, and placed on ice. RNA was reverse transcribed with 200 units of Super Script III enzyme in the presence of 0.5 mM dCTP, dGTP, dTTP, 20 μCi ^{32}P-α-dATP (6000 Ci $mmol^{-1}$) and 40 units of RNaseOUT. The primers were annealed to RNA at 25°C for 10 min, cDNA synthesis was performed at 50°C for 50 min and terminated at 85°C for 5 min. RNA was removed by digestion with 2 units of RNase H at 37°C for 20 min. Labelled cDNA probes were purified by QIAquick PCR Purification Kit (QIAGEN) to remove the unincorporated radioisotopes and reduce background during hybridization.

The same procedure was used for synthesis of control cDNA probe from Human thymus total RNA (Ambion) with β-actin gene-specific antisense primer.

Hybridization and data analysis

The hybridization was performed in roller bottles in a hybridization oven. The membranes were incubated in 10 ml of a hybridization solution containing 5× SSC, 0.1% SDS, 5× Denhardt's and 100 μg ml^{-1} sheared, sonicated calf thymus DNA for 1–2 h at 65°C. Labelled Xp10 cDNA probes were added to 5 ml of the hybridization solution containing labelled β-actin cDNA probe which was used as a normalization control. cDNA was denaturated at 90–95°C for 10 min and loaded into the roller bottle containing the pre-hybridized membrane. Hybridization was performed overnight (12–18 h) at 65°C. After hybridization, the membranes were washed twice at room temperature in 2× SSC, 0.1% SDS for 5 min and three times at 65°C in 0.2 SSC, 0.1% SDS for 20 min. Membranes were air-dried for 5 min and analysed by PhosphorImager (Molecular Dynamics).

Quantification of gene expression signals was performed using PhosphorImager-generated image files using the ImageQuant (Molecular Dynamics) software. The signal intensities of each spot on membrane were determined, background signal was subtracted and the values obtained were normalized relative to the β-actin signals.

Primer extension

Total RNA for primer extension analysis was extracted from *X. oryzae* cells infected by Xp10 phage in the presence and absence of Rif and harvested at various time points throughout the course of infection. The RNA was isolated using TRIzol Max Bacterial RNA Isolation Kit (Invitrogen) according to manufacture's protocol. RNA concentration and purity were determined by measurements of absorbance at 260 nm and 280 nm and by electrophoresis in formaldehyde-agarose gel. For primer extension reactions, 2–10 μg of total RNA were reverse-transcribed with 100 units of Super Script III enzyme from first-strand synthesis system for RT-PCR (Invitrogen) in the presence of 1 pmol ^{32}P-end-labelled primer. Specific primers were annealed to RNA by heating for 20 min at the temperature equal to melting temperature for the primer and additional 10 min incubation at room temperature. Primer extension reactions were carried out for 50 min at 50°C and terminated by a 5 min incubation at 85°C, RNA was removed by RNase H treatment. After chloroform extraction, nucleic acids were precipitated with ethanol and dissolved in formamide-containing loading buffer. As a reference, sequencing reactions were performed on appropriate Xp10 PCR fragments using the same end-labelled primer as the one used in the primer extension reaction. The *fmol* DNA Cycle Sequencing System (Promega) was used for sequencing according to manufacturer's protocol. The products of the reactions were resolved on an 8% sequencing gel and revealed using PhosphorImager.

RNA ligase-mediated RT-PCR

RNA ligase-mediated (RLM) RT-PCR was carried out as originally described (Bensing *et al.*, 1996), with some modifications. Ten micrograms of total RNA from Xp10-infected cells was split into two halves (treated and untreated) and processed in parallel. Treated RNA was combined with 5 units of TAP (Epicentre) in a 25 μl reaction volume according to manufacturer's instructions, with the addition of 20 units RNasin RNase inhibitor (Promega). TAP was substituted by RNase-free water in the untreated sample. Following TAP treatment, the treated and untreated samples were processed identically. The RNA was phenol/chloroform extracted, ethanol precipitated and dried. The pellet was resuspended in a 50 μl ligation reaction containing 150 ng of RNA oligonucleotide (5′GGUAUUGCGGUACCCUUGUAC GC3′), and 100 units of T4 RNA Ligase (New England Biolabs) at conditions specified by the manufacturer. Following phenol/chloroform extraction and ethanol precipitation with 50 pmol cDNA primer, complementary to Xp10 genome positions 42722–42751, the RNA/primer pellet was resuspended in 20 μl of buffer for reverse transcription, heated at 60°C for 20 min, and incubated additional 10 min at room temperature. cDNA synthesis was carried out with 100 units of Super Script III enzyme from first-strand synthesis system for RT-PCR (Invitrogen) according to manufacturer's instructions for 50 min at 50°C. RNA was removed by RNase H treatment. cDNA was then purified by the Qiaquick PCR cleanup method (Qiagen) and PCR-amplified with Ampli *Taq* Gold DNA polymerase (Roche) for 40 cycles: 94°C, 30 s; 55°C, 30 s; 72°C, 1 min. As an upstream primer, a DNA oligonucleotide corresponding to the RNA oligonucleotide

was used. As downstream primers oligonucleoides complementary to Xp10 genome positions 42428–42453 (for Repeat 1) and 42682–42704 (for Repeat 2) were used. Reaction products were visualized on 2% agarose gel. PCR fargements were excised from the gel, eluted by Qiaquick gel extraction method (Qiagen), and sequenced. The site of transcription initiation was determined by the position of the sequence junction between the RNA oligonucleotide and Xp10 genomic sequence.

Weight matrices used in the 'supervised' searches of Xp10 genome

T7 RNAP-like promoter sequences weight matrix:

	A	T	C	G
1	–0.12	–0.20	0.10	0.20
2	–0.20	–0.07	0.20	0.06
3	–0.20	–0.22	0.20	0.20
4	–0.04	–0.39	0.20	0.20
5	–0.46	0.03	0.20	0.20
6	–0.14	0.05	0.07	0.02
7	0.031	0.20	–0.11	–0.10
8	0.031	–0.09	–0.11	0.17
9	0.20	0.20	–0.27	–0.10
10	–0.12	–0.31	0.20	0.20
11	0.20	0.20	–0.56	0.20
12	–0.62	0.20	0.20	0.20
13	0.20	0.20	–0.56	0.20
14	0.20	–0.62	0.20	0.20
15	–0.39	0.08	0.20	0.08
16	–0.07	–0.36	0.20	0.20
17	–0.15	0.06	0.10	–0.02
18	–0.06	0.20	0.20	–0.33
19	0.03	0.20	–0.02	–0.19
20	–0.12	–0.04	0.20	–0.05
21	–0.07	0.05	0.01	0.01

Motif 1 weight matrix:

	A	T	C	G
1	0.17	–0.24	0.17	–0.11
2	0.17	–0.14	0.17	–0.21
3	–0.55	0.17	0.17	0.17
4	–0.07	0.17	0.04	–0.14
5	–0.21	0.17	0.17	–0.14
6	0.17	0.17	–0.14	–0.18
7	–0.25	0.17	–0.11	0.17
8	0.17	–0.21	–0.14	0.17
9	–0.19	0.01	0.00	0.17
10	0.01	–0.17	0.17	–0.03
11	–0.10	–0.28	0.17	0.17
12	–0.13	–0.02	0.17	–0.03
13	–0.55	0.17	0.17	0.17
14	0.17	0.17	–0.18	–0.14
15	–0.55	0.17	0.17	0.17
16	0.17	0.17	0.17	–0.50
17	–0.26	0.17	–0.09	0.17
18	–0.55	0.17	0.17	0.17
19	0.01	0.17	0.17	–0.34
20	–0.40	0.17	0.17	0.04
21	–0.02	0.17	0.04	–0.19
22	0.17	0.17	–0.29	–0.03
23	–0.28	–0.02	0.17	0.10

Motif 2 weight matrix:

	A	T	C	G
1	–0.46	0.14	0.14	0.14
2	–0.24	0.07	0.14	0.14
3	–0.07	0.14	–0.04	–0.03
4	0.14	–0.24	–0.05	0.14
5	0.14	0.14	–0.41	0.14
6	0.14	–0.27	0.14	–0.03
7	0.14	0.14	0.14	–0.41
8	0.14	–0.06	–0.23	0.14
9	–0.46	0.14	0.14	0.14
10	–0.07	–0.24	0.14	0.14
11	0.14	–0.06	0.14	–0.23
12	–0.26	–0.06	0.14	0.14
13	0.14	–0.46	0.14	0.14
14	–0.27	0.14	–0.03	0.14
15	0.14	0.14	0.14	–0.41
16	–0.06	0.14	–0.23	0.14
17	0.14	0.14	–0.41	0.14
18	0.14	0.14	0.14	–0.41
19	–0.46	0.14	0.14	0.14
20	–0.24	0.14	–0.05	0.14
21	–0.14	–0.14	–0.03	–0.24
22	0.14	–0.46	0.14	0.14
23	0.14	–0.46	0.14	0.14

Acknowledgements

We are grateful to Drs E. Peter Geiduschek, Ian Molineux and Sergei Nechaev for critical reading of the manuscript. We are indebted to Dr Lucia Rothman-Denes for criticism and for suggesting the experiment of Fig. 7, and to Michael Dor Hammer for providing us with protocols for a TAP-RT PCR experiment. We thank Jing Liu for help with preparation of Fig. 1A. This work was supported by NIH Grants GM59295 (to K.S.) and GM67794 (to B.S.). E.S. was partially supported by Charles and Johanna Busch postdoctoral fellowship.

Supplementary material

The following material is available from
http://www.blackwellpublishing.com/products/journals/suppmat/mmi/mmi4442/mmi4442sm.htm

Appendix S1. Calculating μ.
Appendix S2. Interpretation of α.
Fig. S1. Abundances of representative Xp10 transcripts.
Table S1. Predicted RNA secondary structure of sequences at and around primer extension end-points.

References

Allison, L.A., Simon, L.D., and Maliga, P. (1986) Deletion of rpoB reveals a second distinct transcription system in plastids of higher plants. *EMBO J* **15:** 2802–2809.

Beily, T.L., and Elkan, C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Second Intl Conference on Intelligent Systems for Mol Biol* 28–36.

Bensing, B.A., Meyer, B.J., and Dunny, G.M. (1996) Sensitive detection of bacterial transcription initiation sites and differentiation from RNA processing sites in the pheromone-induced plasmid transfer system of *Enterococcus faecalis*. *Proc Natl Acad Sci USA* **93:** 7794–7799.

Carter, R.H., Demidenko, A.A., Hattingh-Willis, S., and Rothman-Denes, L.B. (2003) Phage N4 RNA polymerase II recruitment to DNA by a single-stranded DNA-binding protein. *Genes Dev* **17:** 2334–2345.

Djordjevic, M., Sengupta, A.M., and Shraiman, B.I. (2003) A biophysical approach to transcription factor binding site discovery. *Genome Res* **13:** 2381–2390.

Ermolaeva, M.D., Khalak, H.G., White, O., Smith, H.O., and Salzberg, S.L. (2000) Prediction of transcription terminators in bacterial genomes. *J Mol Biol* **301:** 27–33.

Lawrence, C.E., Altschul, S.F., Bogouski, M.S., Liu, J.S., Neuwald, A.F., and Wooten, J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262:** 208–214.

Liao, Y.D., and Kuo, T.T. (1986) Loss of sigma-factor of RNA polymerase of *Xanthomonas campestris* pv. *oryzae* during phage Xp10 infection. *J Biol Chem* **261:** 13714–13719.

Liao, Y.D., Tu, J., and Kuo, T.T. (1987) Regulation of transcription of the Xp10 genome in bacteriophage-infected *Xanthomonas campestris* pv. *oryzae*. *J Virol* **61:** 1695–1699.

Mathews, D.H., Sabina, J., Zuker, M., and Turner, D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* **288:** 911–940.

Molineux, I. (2005) The T7 group. In: *Bacteriophages.* Calendar, R. (ed.). Oxford: Oxford University Press (in press).

Nechaev, S., Yuzenkova, Y., Niedziela-Majka, A., Heyduk, T., and Severinov, K. (2002) A novel bacteriophage-encoded RNA polymerase binding protein inhibits transcription initiation and abolishes transcription termination by host RNA polymerase. *J Mol Biol* **320:** 11–22.

Roberts, J.W., Yarnell, W., Bartlett, E., Guo, J., Marr, M., Ko, D.C., *et al.* (1998) Antitermination by bacteriophage lambda Q protein. *Cold Spring Harb Symp Quant Biol* **63:** 319–325.

Yuzenkova, Y., Nechaev, S., Berlin, J., Rogulja, D., Kuznedelov, K., Schloss, M., *et al.* (2003) Genome of Xanthomonas oryzae bacteriophage Xp10: an odd T-odd phage. *J Mol Biol* **330:** 735–748.