The Transcriptome of the Baculovirus Autographa californica Multiple Nucleopolyhedrovirus in Trichoplusia ni Cells

Yun-Ru Chen, Silin Zhong, Zhangjun Fei, Yoshifumi Hashimoto, Jenny Z. Xiang, Shiying Zhang, Gary W. Blissard

Baculoviruses are important insect pathogens that have been developed as protein expression vectors in insect cells and as transcription vectors for mammalian cells. They have large double-stranded DNA genomes containing approximately 156 tightly spaced genes, and they present significant challenges for transcriptome analysis. In this study, we report the first comprehensive analysis of AcMNPV transcription over the course of infection in Trichoplusia ni cells, by a combination of strand-specific RNA sequencing (RNA-Seq) and deep sequencing of 5′ capped transcription start sites and 3′ polyadenylation sites. We identified four clusters of genes associated with distinctive patterns of mRNA accumulation through the AcMNPV infection cycle. A total of 218 transcription start sites (TSS) and 120 polyadenylation sites (PAS) were mapped. Only 29 TSS were associated with a canonical TATA box, and 14 initiated within or near the previously identified CAGT initiator motif. The majority of viral transcripts (126) initiated within the baculovirus late promoter motif (TAAG), and late transcripts initiated precisely at the second position of the motif. Analysis of 3′ ends showed that 92 (77%) of the 3′ PAS were located within 30 nucleotides (nt) downstream of a consensus termination signal (AAUAAA or AUUAAA). A conserved U-rich region was found approximately 2 to 10 nt downstream of the PAS for 58 transcripts. Twelve splicing events and an unexpectedly large number of antisense RNAs were identified, revealing new details of possible regulatory mechanisms controlling AcMNPV gene expression. Combined, these data provide an emerging global picture of the organization and regulation of AcMNPV transcription through the infection cycle.

Baculoviruses are invertebrate viruses that have large circular, double-stranded DNA genomes (1). They are used as biopesticides in agriculture, for protein production in research and industry, and as a gene delivery system for mammalian cell transduction (2–10). Because baculoviruses replicate only in insect cells, and infection results in exceptionally high levels of expression of certain viral late genes, they are widely used as eukaryotic expression vectors for recombinant proteins. In the pharmaceutical industry, baculovirus-infected insect cells are used for production of the virus-like-particle (VLP) vaccine Cervarix (GlaxoSmithKline) and numerous subunit- and VLP-based vaccine candidates. The type species, Autographa californica multiple nucleopolyhedrovirus (AcMNPV), is the most intensively studied model baculovirus and is the most widely used protein expression vector. The genome of AcMNPV is 133.9 kbp and contains at least 156 open reading frames (ORFs). The ORFs are densely packed on the genome, and most are found in close proximity to adjacent ORFs (11). The average distance between AcMNPV ORFs is 53 bp, and 48 (approximately 30%) of the 156 ORFs have some overlap with an adjacent ORF. Because of the close proximity of ORFs, transcripts frequently overlap adjacent ORFs and their transcripts. The extent of transcript overlap in the genome was not previously known, as no global transcription map is available. Because of the close arrangement of transcription units on the large genome, it has been difficult to interpret the results of studies utilizing microarrays, quantitative PCR (qPCR), and traditional RNA sequencing (RNA-Seq) strategies, especially for uncharacterized regions of the genome.

During the infection process, viral genes are regulated by host and viral transcription factors, and the replication cycle is divided into two general phases: an early phase and a late phase. Prior studies suggested that the early phase may be subdivided into immediate early and delayed early phases, and the late phase can be subdivided into late and very late stages (9, 12). Early genes are defined as those recognized and transcribed by the host cell RNA polymerase II, and early transcription was initially described as alpha amanitin sensitive, in contrast to late transcription, which is alpha amanitin resistant (13). The core elements of baculovirus early promoters are those recognized by host RNA polymerase II, and sometimes they include the TATA box motif and an initiator sequence (CAGT) (9, 12, 14). However, in many cases early promoters contain no known or recognizable motifs (15–17). Viral late transcription appears to begin after DNA replication, since inhibitors of DNA replication also inhibit late gene transcription. It was also demonstrated that late gene transcription can be reconstituted by transient expression of approximately 19 viral early genes, collectively referred to as late expression factor (lef) genes (18–20). In transient-expression studies, it was found that approximately 9 of the viral LEF proteins are necessary to reconstitute DNA replication (21, 22). Similarly, an additional 9 or 10 LEF proteins are necessary to reconstitute late transcription and therefore appear to comprise the baculovirus late RNA polymerase and regulatory proteins (19, 23–27). The viral RNA polymerase recognizes what appears to be an invariant late promoter motif, 5′-TAAG-3′ (2, 9, 28–30). In many studies in which late transcription initiation sites were mapped, the mapped initiation sites were
variable but were typically found within and very near the TAAG motif.

Although the AcMNPV replication cycle has been intensively studied, the compact nature of gene organization and the overlapping nature of viral mRNAs have made global analysis of viral gene expression difficult. Several prior studies of AcMNPV transcriptomes have been performed with microarrays and by real-time PCR analysis, although data are often difficult to interpret in light of the inherent problems associated with the tightly packed AcMNPV genome and overlapping transcripts (31–33). Recently, the Bombyx mori NPV (BmNPV) transcriptome was examined using RNA-Seq, and Illumina reads were mapped to a reference transcriptome database (34). However, the issue of extensive overlapping transcripts and antisense transcripts was not specifically addressed and therefore remains largely unresolved. In the present study, we integrated a combination of deep-sequencing methods to characterize the transcriptome of AcMNPV. These methods included 5′ capped RNA-Seq, 3′ RNA-Seq, and strand-specific RNA-Seq. Using these methods, we globally mapped 5′ ends and 3′ polyadenylation sites of mRNAs isolated from AcMNPV-infected Trichoplusia ni cells. Through analysis of mRNAs isolated at various times postinfection, we characterized the rapidly changing landscape of AcMNPV transcription in T. ni cells. Using this very sensitive methodology, our analyses show that (i) certain highly expressed late transcripts can be readily detected in infected cells within the first hour after infection, suggesting the likelihood that some abundant late mRNAs may be packaged and transported into cells via the virion; (ii) few early transcripts initiate from canonical TATA box and CAGT motifs, although TATA box motifs are found upstream of 29 transcription start sites, and CAGT motifs are associated with 14 start sites; (iii) late promoter initiation is highly precise, initiating at the second position of the conserved TAAG motif; (iv) viral late genes comprise the vast majority of viral mRNAs, and viral transcript levels from the genes polyhedrin, p10, and p6.9 are exceptionally high but with dramatically different profiles of temporal abundance; and (v) late in infection, viral transcripts comprise >80% of the total cellular mRNA population. This study represents the first comprehensive single nucleotide resolution analysis of the AcMNPV transcriptome, integrating genomic mapping of transcription start sites, polyadenylation sites, and mRNA abundance. Thus, these results will serve as a resource for research on baculovirus biology and for biotechnological applications of baculoviruses. The sequencing methods described here should also be broadly applicable to the analysis of other viruses and host cells.

MATERIALS AND METHODS
Cells, viruses, and infections. The cell line Tnms42 was subcloned from High Five (BTI-Tn5B14) cells. Tnms42 cells were cultured in TNNF medium (35) supplemented with 10% FBS (Gibco, BRL) at 28°C. Our extensive testing has shown that unlike High Five cells, Tnms42 cells contain no detectable alphaviruses. A detailed characterization of the Tnms42 cell line will be provided in a separate publication (our unpublished results). The wild-type AcMNPV strain E2 was used for cell infections, and virus titer was determined by measurements of 50% tissue culture infective dose (TCID50), as described previously (10). To generate each biological replicate of infected-cell mRNA, 3 × 10^6 T. ni (Tnms42) cells were infected with AcMNPV (multiplicity of infection [MOI] = 10) in a T25 flask. After a 1-h incubation period, the inoculum was removed, and the cells were rinsed with Grace’s medium. This time point was labeled 0 h postinfection (h p.i.), and cells were then cultured in TNMFH medium supplemented with 10% fetal bovine serum (FBS) at 28°C. Total RNA was purified from infected cells at 0, 6, 12, 18, 24, 36, and 48 h p.i. using a Qiagen RNeasy minikit.

Illumina strand-specific RNA sequencing. Illumina sequencing libraries were constructed by following a modified strand-specific RNA-Seq protocol (for details, see reference 36). Briefly, polyadenylated RNA was isolated from 2 μg total RNA using Dynabeads oligo(dT)25 (Invitrogen). The poly(A)-enriched sample was simultaneously eluted and fragmented in 2X SuperScript III buffer at 94°C for 7 min in the presence of hexamer and oligo(dT)(10)VN (5′-p- TTTTTTTTTTTT 3′; V = A, G, and C; Integrated DNA Technologies, Inc.). Hexamer- and oligo(dT)-primed first-strand cDNA synthesis was carried out using SuperScriptIII (Invitrogen) in the presence of deoxyxynucleoside triphosphate (dNTP). The RNA/cDNA hybrid was purified with AMPure RNA Clean XP (Agencourt). Second-strand cDNA was synthesized using RNase H (NEB) and the Klenow fragment of DNA polymerase I (NEB) with a dUTP mix at 16°C for 2.5 h. The double-stranded cDNA fragments were end rephosphorylated, dA tailed, and then ligated with the TruSeq Y-shape adapter. The dUTP-containing strand was then removed by digestion with uracil DNA glycosylase (NEB) (36). The library was then PCR amplified with TruSeq-indexed PCR primers and sequenced on the Illumina HiSeq2000 platform (Well Cornell Medical College).

5′ rapid amplification of cDNA end (RACE) sequencing. To identify 5′ ends of mRNAs, poly(A) RNA was isolated as described above. Free (noncapped) 5′ ends were dephosphorylated with alkaline phosphatase by incubating poly(A) mRNA isolated from 50 μg total RNA in 2 U alkaline phosphatase (Fermentas) in a 10-μl reaction mixture containing 10 mM Tris-HCl (pH 8.0 at 37°C), 5 mM MgCl2, 0.1 M KCl, 0.02% Triton X-100, and 0.1 mg/ml bovine serum albumin (BSA) for 20 min at 37°C. The RNA was extracted once with phenol-chloroform and ethanol precipitated. The 5′ cap structure was removed by incubation of the poly(A) mRNA with 5 U tobacco acid pyrophosphatase (Epiprec) at 37°C for 30 min in a 20-μl reaction mixture containing 0.05 M sodium acetate (pH 6.0), 1 mM EDTA, 0.1% β-mercaptoethanol, and 0.01% Triton X-100. The reaction was stopped by phenol-chloroform extraction, and the RNA was ethanol precipitated. The decapped RNAs (containing a 5′ phosphate) were then ligated to an RNA linker (5′-GUUCAGAGUUCUACAGUCGCCGAGCUAC-3′; IDT) by incubation with 10 U of T4 RNA ligase 1 (NEB) in the presence of 20% polyethylene glycol (PEG) 8000 at 22°C overnight. Reverse transcription was carried out with SuperScript III using a tagged-heptamer (5′-GTGCCTCTCCGCTATCNNTNNNNNN-3′) under these conditions: 65°C for 30 s, 25°C for 10 min, 50°C for 50 min, and 70°C for 15 min. The first-strand cDNA was then purified as indicated above. Second-strand cDNA was synthesized using RNase H (NEB) and the Klenow fragment of DNA polymerase I (5′-exo-; NEB) with dUTP at 37°C for 2 h. The double-stranded cDNA was then separated in a 1% agarose gel, and cDNAs in the 250- to 450-bp range were excised and purified. The DNA was then digested with uracil DNA glycosylase (NEB), PCR amplified with Illumina PCR oligonucleotides, and sequenced.

Identification of transcription start sites (TSS). The 5′ RACE Illumina sequencing reads were first filtered by alignment to adapter, rRNA, and RNA sequences using Bowtie (37, 38) to remove those sequences and then mapped to the AcMNPV genome (NCBI accession no. NC_001623) using TopHat, allowing 1 mismatch per segment. TSS were defined on the following criteria. (i) Each TSS site identified was supported by ≥100 reads mapping to a site. (ii) The read count of the TSS site (peak) was ≥20% of the total read counts in the 1-kb region upstream of the downstream start codon. (iii) In general, when multiple TSS were identified clustered around a single highly abundant TSS (usually within ±5 nt), only the single abundant TSS site was mapped (39). The TSS data set was manually curated, and RACE-PCR and Sanger sequencing were performed to validate the TSS of 156 genes.

Genome sequences ≥50 nt from the TSS were analyzed and known core promoter motifs, TATAA, CAGT, TAAG, were first identified. The consensus motif search was carried out using multiple expectation max-
imization for motif elicitation (MEME) (40) from genes in each of the groups (parameters, zero or one motif per sequence and an E value of <0.001) (41, 42).

Identification of 3′ polyadenylation sites (PAS). The 3′ RACE analysis was performed in parallel with the strand-specific RNA-Seq library construction by priming the RNA with hexamer and oligo(dT)(10)VN (V = A, G, and C) simultaneously. Conventional Illumina RNA-Seq protocols rely on random hexamers to prime reverse transcription and tend to bias against both 5′ and 3′ ends of the transcripts (36). By using phosphorylated oligo(dT)(10)VN for the reverse transcription step, we primed cDNA synthesis precisely at the poly(A) track. After sequencing, RNA-Seq reads starting with 10 thymidines at their 5′ ends were recovered from the unmappable data as raw 3′ RACE-Seq reads. After trimming of the 10 thymidines, the reads were aligned back to the viral genome using TopHat to identify polyadenylation sites. All the candidates were further filtered to remove sites resulting from mispriming, by removing all ends that were followed by ≥3 adenines in the genomic DNA sequence unless (i) a consensus polyadenylation signal was identified ≥20 nt upstream, (ii) the RNA expression pattern from RNA-Seq data provided strong evidence for transcription termination at the mapped site, or (iii) both.

RNA-Seq read processing, alignment, and analysis. Strand-specific RNA-Seq reads were first filtered by alignment to adapter, rRNA, and tRNA sequences using Bowtie (37, 38) to remove those sequences. For analysis of the viral transcriptome, the AcMNPV genome (NCBI accession no. NC_001623), with the exon model defined as a total of 156 putative ORFs, was used. The filtered reads were mapped to the AcMNPV genome using TopHat (43), allowing 1 mismatch. The number of mapped reads derived from the sense RNA strand was calculated and then normalized to reads per kilobase of exon model per million mapped reads (RPKM) (44). The threshold for all subsequent analysis was set to 5 RPKM, which is roughly equivalent to a sensitivity of one mRNA copy per cell in eukaryotes. The cluster analysis was performed using hierarchical clustering by the complete-linkage clustering method in the Cluster 3.0 package (45).

Sequencing data accession number. Illumina RNA-Seq data were deposited in the Sequence Read Archive (SRA) under accession number SRA057390.

RESULTS

Genome-wide identification of transcription start sites and polyadenylation sites. Analysis of the baculovirus transcriptome presents unique difficulties because of a relatively large proportion of overlapping transcripts. In addition, the majority of transcription start sites and polyadenylation sites of the AcMNPV genome were not previously mapped. To identify and define AcMNPV transcripts throughout the genome, we therefore profiled the transcription start sites (TSS) and the polyadenylation sites (PAS) for AcMNPV transcripts at 7 time points in the infection cycle.

(i) 5′ RACE analysis of transcription start sites (TSS). We first performed genome-wide transcription start site (TSS) mapping using the Illumina platform to determine sequences at the 5′ cap sites of infected cell mRNAs (see Materials and Methods). We also then individually confirmed TSS data by Sanger sequencing of 5′ RACE PCR products for each gene or ORF. By Illumina sequencing, a total of 25.3 million 5′ RNA-Seq reads were generated for analysis of transcription start sites (TSS). Of those, 13.4 million reads (>52%) were aligned to the AcMNPV genome (see Table S1 in the supplemental material). The mapped genome positions for TSS associated with individual AcMNPV ORFs are shown in Table S2 in the supplemental material. Quantitative data (raw read numbers) for specific TSS genome locations at various times postinfection are shown in Table S3A in the supplemental material. TSS reads were filtered by including sites validated by >100 reads in most cases and by ignoring relatively weak TSS positions that were clustered at low levels around a single abundant TSS (usually within ±5 nt) (see the example in Fig. S2C in the supplemental material). Using this approach, we identified 218 TSS for 156 AcMNPV ORFs (Fig. 1). Validation experiments confirmed 186 TSS by 5′ RACE-PCR and Sanger sequencing. A total of 132 TSS (associated with 130 ORFs) were mapped to identical positions by both methods (see Table S2). The mapped 5′ end read numbers are provided in Table S3A. Of the 156 annotated ORFs, we identified 46 that contain multiple transcription start sites. In addition, in a number of cases a significant TSS was identified within the annotated ORF (see Table S3b), potentially encoding an internal protein product, such as that originally described for the OpMNPV p34 gene (46). We also examined the TSS of 72 genes for which TSS were previously published (see Table S2). Of those, 48 TSS were identical to previously mapped sites or differed by less than 5 nt. While many of the previously published TSS data were generated from AcMNPV-infected S. frugiperda cells, our results suggest that different hosts have little impact on the TSS usage of AcMNPV, as expected.

Because of the highly conserved nature of AcMNPV late promoters and the prior observations that late transcription almost invariably initiates within or near the TAAG motif, we separated TSS sequences into two classes, those initiating at the late promoter motif (the TAAG motif recognized by the viral RNA polymerase) and those not initiating within or near a TAAG motif (non-TAAG sequences, which are likely recognized by host cell RNA polymerase II) (Fig. 2A and D). Of the 218 TSS identified in the AcMNPV genome, approximately 92 were non-TAAG TSS. Of those 92, 89 TSS were located immediately upstream of 77 AcMNPV ORFs. In contrast, 126 TSS were identified at the consensus baculovirus late promoter or TAAG motif, upstream of 101 ORFs (Fig. 3A and B). In addition, for 21 genes (ldf2, orf3, orf23, pkp5, v-jff, pp31, odv-66, orf74, orf79, orf82, p15, cg30, helicase, he65, orf114, pk-2, gp64, alk-exo, p35, me33, and ieo), we identified upstream TSS from multiple sites that included both early (non-TAAG) and late (TAAG) motifs. In addition, two genes (gp16 and p35) each contained a TSS located appropriately downstream of a TATA box, yet initiation was at the +2 position of a TAAG motif, suggesting overlapping early and late motifs for both genes. Based on the TSS positions relative to downstream ORFs, the average length of the 5′ untranslated region (UTR) was approximately 162 nt for all AcMNPV genes (see Table S2 in the supplemental material). Separately however, the average 5′ UTR lengths of non-TAAG and TAAG motif-containing ORFs are 155 and 169 nt, respectively (Fig. 2D).

(ii) TSS at non-TAAG sites. Approximately 84% of the TSS at non-TAAG sites initiated at an A residue, consistent with a strong bias for initiation at A residues for eukaryotic RNA polymerase II transcripts (47, 48). We also noted a slight bias for G/C at position −1 and a strong bias for A/T at the second nucleotide position in the transcript (+2) (Fig. 2C; also, see Fig. S1A, panel a, in the supplemental material). To identify motifs associated with the non-TAAG (early) baculovirus promoters, we analyzed sequences 50 nucleotides (nt) upstream and 50 nt downstream of each non-TAAG TSS. Of the 92 transcripts initiating at non-TAAG sequences, 29 contain canonical TATA boxes or TATA-like sequences located at 24 to 40 bp upstream of the TSS (Fig. 2A and D). Of the TSS with upstream TATA sequences, only 8 TSS initiated at the previously described start site sequence motif, CAGT

June 2013 Volume 87 Number 11 jvi.asm.org 6393

Downloaded from http://jvi.asm.org on November 13, 2017 by guest
A visual inspection of the 92 non-TAAG TSS shows that the following sequences are most abundant at the TSS sites. Approximately 52% of the non-TAAG TSS contain a C preceding the A at the TSS. For approximately 24% of the non-TAAG TSS, a T precedes the A at the TSS. In addition, the following sequences are also found at the TSS at the following approximate frequencies: CAGT, 15%; CATT, 9%; CAAT, 8.7%). Using the MEME algorithm (40) for motif analysis, a degenerate motif (5′-AACK-YGTYGRYSAWCRYGWT-3′; degenerate positions are underlined) was found associated with 21 of the non-TAAG TSS sites (Fig. 2D, M1), although the position relative to the TSS was not conserved and the motif was identified downstream of the TSS in 3 cases. The identification of functionally significant sequence motifs among (non-TAAG) promoters that are presumably recognized by host cell RNA polymerase II is problematic because of the relatively small size of the AcMNPV genome and the resulting difficulty in establishing statistical significance of motifs identified. As noted earlier, the start site or initiator motif (5′-CAGT-3′) identified from baculovirus early promoters (49, 51–53) is similar to the highly conserved TCAGTY motif identified from studies of Drosophila melanogaster RNA polymerase II promoters (54–56). Previous functional studies of the CAGT initiator motif established its functional role in baculovirus early gene transcription (49–51). The functional significance of other, less well defined motifs is more difficult to assess in the absence of experimental evidence.

(iii) TSS at TAAG motifs. Based on prior studies of baculovirus early and late promoters, a functional late promoter motif was generally defined as a promoter in which the TSS was found within a few nucleotides of the TAAG motif. For TSS identified by 5′ RNA-Seq in this study, we found that transcription initiation mapped precisely to an A in the second position (TAAG) in 100% of the TAAG motifs (126 TSS) (see Table S2 in the supplemental material). This specific initiation of transcription was also confirmed by Sanger sequencing of 5′ RACE products (see Table S2). A number of genes (at least 21 of the 156 ORFs) have multiple TSS that map to multiple TAAG motifs located immediately upstream of the ORF, and thus, at least 13% of the AcMNPV genes have
We found that the average AT content (61%) in the region 30 nt upstream and downstream of the TAAG of the AcMNPV TSS was somewhat higher than the average AT content of the AcMNPV genome (59%) (see Fig. S1A, panel b, in the supplemental material), and certain positions (−1 to −6 and +1 to +7) have a very high AT content of approximately 75% (Fig. 2C; also, see Fig. S1A, panel b). We also aligned all TAAG motifs from TSS and asked whether there was any specific nucleotide sequence conservation in the 30 nt flanking the core TAAG motif TSS sites. We identified no additional conserved motifs flanking the TAAG motifs. We also aligned several groups of late promoters that showed (i) moderate to strong promoter activity or (ii) similar patterns of gene expression through the infection cycle (see clusters below). For promoters with moderate to strong expression, we separately analyzed 4 late promoters from genes with RPKM values of >40,000, 22 late promoters from genes with RPKM values of 28,000 to 40,000, and 22 late promoters from genes with RPKM values of 10,000 to 28,000. From these alignments, we were unable to identify any motifs correlated with high levels of late transcripts.

(iv) 5′ UTR sequence content. It was previously observed that eukaryotic mRNAs typically have low AT content in the 5′ UTR (58, 59). In contrast, we observed that the AT content of the 30 nt downstream of TSS of non-TAAG-initiated transcripts was 58%, while the same region from transcripts initiated within TAAG motifs was 65%. The higher AT content of sequence downstream of the 5′ TSS of baculovirus late transcripts could result from a role of that sequence in late promoter regulation, since a regulatory role for downstream sequences has been demonstrated for two late genes (28, 29). In the case of early transcripts, it is also possible that these sequences play a regulatory role, but alternatively, the AT content of the 5′ UTR may simply reflect the high AT content of the AcMNPV genome.

(v) Summary of TSS analysis. Based on the genomic and transcriptomic analysis of AcMNPV ORFs and TSS at non-TAAG and TAAG sites, the following picture of AcMNPV early, late, or early/late genes emerges. (i) We identified 92 TSS that do not initiate at TAAG motifs and thus likely represent early transcripts mediated by the host cell RNA polymerase. Those 92 TSS are associated with 77 ORFs, approximately 49% of the annotated AcMNPV genes. Of those 92 TSS, we identified canonical TATA box sequences (Fig. S1A, panel b) that were somewhat higher than the average AT content of the genome (59%) (see Fig. S1A, panel b, in the supplemental material). Numbers above and below regions or arrows represent average distances (in nucleotides). The position M1 represents the position of a degenerate sequence motif (AACGTYGRSYAWCRTYGWT; degenerate positions are underlined) identified from TSS with no consensus TATA or TAAG motif at the TSS. Positions of TSS are indicated by a thick arrow above each gene region.

We also identified 21 ORFs that contain both early (TATAA or CAGT) and late (TAAG) motifs. We also aligned several groups of late promoters that showed (i) moderate to strong promoter activity or (ii) similar patterns of gene expression through the infection cycle (see clusters below). For promoters with moderate to strong expression, we separately analyzed 4 late promoters from genes with RPKM values of >40,000, 22 late promoters from genes with RPKM values of 28,000 to 40,000, and 22 late promoters from genes with RPKM values of 10,000 to 28,000. From these alignments, we were unable to identify any motifs correlated with high levels of late transcripts.

We also identified 21 ORFs that contain both early (TATAA or CAGT) and late (TAAG) motifs. We also aligned several groups of late promoters that showed (i) moderate to strong promoter activity or (ii) similar patterns of gene expression through the infection cycle (see clusters below). For promoters with moderate to strong expression, we separately analyzed 4 late promoters from genes with RPKM values of >40,000, 22 late promoters from genes with RPKM values of 28,000 to 40,000, and 22 late promoters from genes with RPKM values of 10,000 to 28,000. From these alignments, we were unable to identify any motifs correlated with high levels of late transcripts. (iv) 5′ UTR sequence content. It was previously observed that eukaryotic mRNAs typically have low AT content in the 5′ UTR (58, 59). In contrast, we observed that the AT content of the 30 nt downstream of TSS of non-TAAG-initiated transcripts was 58%, while the same region from transcripts initiated within TAAG motifs was 65%. The higher AT content of sequence downstream of the 5′ TSS of baculovirus late transcripts could result from a role of that sequence in late promoter regulation, since a regulatory role for downstream sequences has been demonstrated for two late genes (28, 29). In the case of early transcripts, it is also possible that these sequences play a regulatory role, but alternatively, the AT content of the 5′ UTR may simply reflect the high AT content of the AcMNPV genome.

(v) Summary of TSS analysis. Based on the genomic and transcriptomic analysis of AcMNPV ORFs and TSS at non-TAAG and TAAG sites, the following picture of AcMNPV early, late, or early/late genes emerges. (i) We identified 92 TSS that do not initiate at TAAG motifs and thus likely represent early transcripts mediated by the host cell RNA polymerase. Those 92 TSS are associated with 77 ORFs, approximately 49% of the annotated AcMNPV genes. Of those 92 TSS, we identified canonical TATA box sequences (Fig. S1A, panel b). (ii) We mapped 126 TSS that initiate at TAAG motifs and therefore represent typical late transcription. Those 126 TSS are located upstream of 101 ORFs, approximately 65% of the annotated AcMNPV genes. Of those late genes, 80 contain only one mapped late promoter TSS and 21 contain two or more late promoters. (iii) We also identified 21 ORFs that contain upstream TSS that initiate from both non-TAAG and TAAG sequences (Fig. 3A), suggesting that these genes are regulated by both early and late promoters. Of those, 12 genes (orf13, orf23, pkip, v-ffg, pp31, adv-e66, orf82, he65, gp64, p35, me53, and ieto) contain both early (TATAA or CAGT) and late (TAAG) motifs. Thus, based on analysis of transcription from early and late classes of promoter motifs, we estimate that the AcMNPV genome contains 77 early genes and 101 late genes (including 24 genes that are both early and late) (Fig. 3A). This distribution (49% early, 65% late, and 15% early/late) is in contrast to poxviruses, in which the distribution was reported as approximately 56% early, 25% intermediate, and 18% late (60). Of course, it must be recognized that these types of early/late analyses examine the general timing of gene activation and must be viewed in the context of other factors,
such as the timing of expression within early or late phases, steady-state levels of the mRNAs, and other types of regulation, such as translation efficiency and protein turnover.

3′ polyadenylation sites (PAS). Viral early genes are transcribed by the host RNA polymerase II, and 3′ processing of transcripts is thought to be mediated by the host enzyme complex (20). In contrast, the mechanism of 3′ processing for late transcripts is not understood. Coterminal 3′ polyadenylation sites (PAS) for several genes were previously reported. In one of the best-studied cases, multiple tandem ORFs are present on transcripts that each initiate from distinct 5′ TSS located immediately upstream of each ORF, and all these transcripts were reported to share a 3′ PAS (61, 62). We analyzed the 3′ PAS of AcMNPV transcripts using a 3′ RACE Illumina sequencing protocol, as described in Materials and Methods. The protocol for identifying poly(A) sites was validated by comparing our results with a variety of previously mapped polyadenylation sites (63–68), and the PAS mapped by Illumina sequencing (see Table S4 in the supplemental material) corresponded closely to sites previously mapped by S1 nuclease protection and/or 3′ RACE. For this analysis, we analyzed mRNAs isolated at times ranging from 6 to 48 h p.i. A total of 120 PAS were identified associated with 135 ORFs by 3′ RNA-Seq analysis. A total of 84 ORFs were associated with a single and unique PAS located immediately downstream of the ORF, while 14 ORFs were associated with multiple PAS. We identified 16 genome regions in which multiple overlapping mRNAs (encoding multiple ORFs) appear to terminate at a common 3′ PAS (for examples, see Fig. S3A in the supplemental material). In two of

![Diagram](http://jvi.asm.org/jvi.asm.org)
these cases, multiple PAS were identified in the downstream region. Overall, a single PAS was identified downstream of 84 genes and multiple PAS were found downstream of 14 genes.

Analysis of all the PAS mapped in the AcMNPV genome revealed that the average distance from the stop codon of the upstream ORF to the downstream PAS was 338 nt. It should be recognized that in cases where a PAS was not identified downstream of an ORF, the absence of detection could result from low abundance of some transcripts and/or inefficient 3′ RACE. It is also important to note that our analysis examined only polyadenylated transcripts, since nonpolyadenylated transcripts from AcMNPV infections have not been described. Nevertheless, any nonpolyadenylated transcripts, should they exist, would not be detected by our methods. To better understand how transcripts are terminated and polyadenylated, we also analyzed the sequence 50 nt upstream and downstream from the PAS and found that the AT content of the 3′ UTR was higher (71%) than the average AT content of the AcMNPV genome (59%), an observation consistent with prior anecdotal observations (69). The AT content in the 40 nt region immediately surrounding the PAS (±20 nt) was even higher at approximately 80% (see Fig. S1A, panel d, in the supplemental material).

A number of AcMNPV PAS were previously mapped downstream of a canonical eukaryotic polyadenylation site: AAUAAA or AUUAAA. In eukaryotes, this signal is recognized by a multicomponent cellular complex called the cleavage and polyadenylation specificity factor, which determines the site of RNA cleavage (70). A second typically conserved sequence element (a G/U- or U-rich sequence) is located up to 30 nt downstream of the polyadenylation signal, and this sequence is bound by another factor called cleavage-stimulating factor (CsrF), which serves to increase the efficiency of mRNA processing. Additional sequences may also be recognized by alternative mechanisms where canonical polyadenylation signals are absent. The assembly and positioning of enzymatic machinery on the mRNA are then followed by endonucleolytic cleavage and subsequent addition of adenine residues by a nuclear poly(A) polymerase. In a prior study of baculovirus polyadenylation, transcripts (generated from a heterologous globin gene construct) by a purified late RNA polymerase in a virus polyadenylation, transcripts (generated from a heterologous donor nucleolytic cleavage and subsequent addition of adenine residues) were analyzed at each time point. The poly(A) RNA fractions were purified using a modified library construction strategy that processes RNAs detected in prior studies (61, 71). To analyze temporal patterns of AcMNPV gene expression inferred from temporal transcriptome data. To analyze temporal patterns of AcMNPV gene expression, Tnms42 cells were infected with AcMNPV at an MOI of 10, and poly(A) RNAs were isolated at seven times postinfection (0, 6, 12, 18, 24, 36, and 48 h p.i.). Three biological replicates were analyzed at each time point. The poly(A) RNA fractions were processed using a modified library construction strategy that provides RNA orientation information (strand-specific RNA-Seq)
and share a 3′-TAAG-5′ motif. In addition, 35 of the 40 transcripts detected are among the most abundant transcripts at 6 h p.i., representing all but 32 viral genes were detected at RPKM values of ≥8, indicating that by 6 h p.i., the majority of the genome (approximately 80% of viral genes) is transcriptionally active. Between 6 and 12 h p.i., expression increased dramatically for many viral genes. At 6 h p.i., RPKM values for AcMNPV mRNAs ranged from undetectable to approximately 2,300. In contrast, by 12 h p.i., RPKM values ranged from 14 to 84,000 (see Table S5 in the supplemental material). A series of graphs illustrating the 20 most abundant mRNAs at each time sampled (0, 6, 12, 18, 24, 36, and 48 h p.i.) is provided in Fig. S3B in the supplemental material. In addition, comparative expression profiles for all individual genes are shown as a series of graphs in Fig. S3C. Genes with RPKM values of ≥8 were considered positive for transcript detection. An RPKM value of 5 is roughly equivalent to a sensitivity of one mRNA copy per cell in mammals (44). For time zero samples, mRNA was isolated from cells after a 1-h virus absorption and incubation period. At time zero, we detected very low levels of transcripts (RPKM values ranging from 8 to 160) from 40 of the 156 AcMNPV genes. Of the transcripts detected at time zero, the 5 most abundant transcripts were p6.9, orf27, orf79, and orf59. Several factors suggest that these may represent mRNAs delivered to the cells from the viral inoculum and not from transcription within the first 60 min after addition of inoculum. Each of the 5 most abundant mRNAs is transcribed from only a late (5′-TAAG-3′) motif. In addition, 35 of the 40 transcripts detected are among the most abundant transcripts at 48 h p.i. Thus, carryover of abundant late transcripts could account for most of the transcripts detected at a very low level at time zero. We can only speculate that abundant mRNAs may be trapped in budded virions as they assemble or bud from the infected cell. Of the most abundant mRNAs detected at 6 h p.i., the majority were early mRNAs initiating at non-TAAG sites, thus representing transcripts from early AcMNPV promoters (see Fig. S3B and Table S2 in the supplemental material). Represented among the 20 most abundant AcMNPV mRNAs detected at 6 h p.i. are transcripts from 5 lef genes (lef-2, lef-3, lef-6, pp31, and pp35) (27), 4 genes that encode proteins known to bind viral DNA (lef-2, lef-3, dbp, and pp31) (72–76), and the gene encoding the major budded virus envelope glycoprotein (gp64) (see Fig. S3B). Based on the relative levels of 5′-TSS at this time (see Table S3A in the supplemental material), a number of the genes that are most highly expressed at 6 h p.i. have active early (non-TAAG) and late (TAAG) promoters. In some cases (such as orf82), transcription from the early promoter appears to predominate. In other cases (gp64 and pp31), a moderately strong early promoter activity appears to be accompanied by even stronger late promoter activity at 6 h p.i. (see Tables S2 and S3A). The abundant detection of a number of late mRNAs (initiating at TAAG sites) at 6 h p.i. was somewhat surprising but may result partially from the improved dynamic range of RNA-Seq. It is also possible that detection of late gene transcripts could be explained by an accelerated infection cycle in the Tnms42 cell line. However, when we compared viral DNA accumulation in Tnms42 and Sf9 cells, the results suggested that the pace of the infection cycle is generally similar in the two lines (Fig. 4). At 6 h p.i., transcripts from 42 genes (representing approximately 27% of the genome) had RPKM values above 200 (see Table S5 in the supplemental material). Interestingly, transcripts representing all but 32 viral genes were detected at RPKM values of >8, indicating that by 6 h p.i., the majority of the genome (approximately 80% of viral genes) is transcriptionally active.

FIG 4 AcMNPV DNA replication and mRNA abundance in Tnms42 cells. The accumulation of AcMNPV genomic DNA and mRNA in Tnms42 cells over a 48-h period are indicated by dashed and solid lines, respectively. The x axis indicates times after infection with AcMNPV (MOI, 10). The right y axis indicates viral DNA copy numbers relative to the host cell haploid genome. The left y axis indicates viral mRNA reads as a percentage of total mRNA reads from the cell (bar = SD; n = 3). For comparison of the infection cycle, viral genomic DNA was similarly measured from Sf9 cells infected in parallel with the same virus (MOI, 10). Relative levels of AcMNPV DNA from infected Tnms42 or Sf9 cell lines were measured by qPCR. The viral DNA copy number was normalized to the host cell genome based on the hsp90 copy number. hsp90 was previously reported as a single copy gene per haploid genome in Spodoptera frugiperda and Bombyx mori (98). qPCR was performed using the following gene-specific primer sets: Q_S9_hsp90_F, CTG ACT TAG CGA TGG TAC CAA GG, and Q_S9_hsp90_R, GGA CAA GAT CCG TTA TGA GTG; hsp90-3_F, CAG ACT TTG CAC TCG TAA GAA GG, and hsp90-3_R, GGA CAA GAT CCG TCA TGG GTG; Q_AC_vp39_F, AAT TGA TCG CCA GCA GCA CCG CC, and Q_AC_vp39_R, TAC CGA AGG TTT CAC AAG CAA C. To illustrate the various expression levels of viral genes, a series of graphs showing the 20 most abundant mRNAs at each time sampled (0, 6, 12, 18, 24, 36, and 48 h p.i.) is provided in Fig. S3B in the supplemental material. In addition, comparative expression profiles for all individual genes are shown as a series of graphs in Fig. S3C. Genes with RPKM values of ≥8 were considered positive for transcript detection. An RPKM value of 5 is roughly equivalent to a sensitivity of one mRNA copy per cell in mammals (44). For time zero samples, mRNA was isolated from cells after a 1-h virus absorption and incubation period. At time zero, we detected very low levels of transcripts (RPKM values ranging from 8 to 160) from 40 of the 156 AcMNPV genes. Of the transcripts detected at time zero, the 5 most abundant transcripts were p6.9, orf27, orf79, and orf59. Several factors suggest that these may represent mRNAs delivered to the cells from the viral inoculum and not from transcription within the first 60 min after addition of inoculum. Each of the 5 most abundant mRNAs is transcribed from only a late (5′-TAAG-3′) motif. In addition, 35 of the 40 transcripts detected are among the most abundant transcripts at 48 h p.i. Thus, carryover of abundant late transcripts could account for most of the transcripts detected at a very low level at time zero. We can only speculate that abundant mRNAs may be trapped in budded virions as they assemble or bud from the infected cell. Of the most abundant mRNAs detected at 6 h p.i., the majority were early mRNAs initiating at non-TAAG sites, thus representing transcripts from early AcMNPV promoters (see Fig. S3B and Table S2 in the supplemental material). Represented among the 20 most abundant AcMNPV mRNAs detected at 6 h p.i. are transcripts from 5 lef genes (lef-2, lef-3, lef-6, pp31, and pp35) (27), 4 genes that encode proteins known to bind viral DNA (lef-2, lef-3, dbp, and pp31) (72–76), and the gene encoding the major budded virus envelope glycoprotein (gp64) (see Fig. S3B). Based on the relative levels of 5′-TSS at this time (see Table S3A in the supplemental material), a number of the genes that are most highly expressed at 6 h p.i. have active early (non-TAAG) and late (TAAG) promoters. In some cases (such as orf82), transcription from the early promoter appears to predominate. In other cases (gp64 and pp31), a moderately strong early promoter activity appears to be accompanied by even stronger late promoter activity at 6 h p.i. (see Tables S2 and S3A). The abundant detection of a number of late mRNAs (initiating at TAAG sites) at 6 h p.i. was somewhat surprising but may result partially from the improved dynamic range of RNA-Seq. It is also possible that detection of late gene transcripts could be explained by an accelerated infection cycle in the Tnms42 cell line. However, when we compared viral DNA accumulation in Tnms42 and Sf9 cells, the results suggested that the pace of the infection cycle is generally similar in the two lines (Fig. 4). At 6 h p.i., transcripts from 42 genes (representing approximately 27% of the genome) had RPKM values above 200 (see Table S5 in the supplemental material). Interestingly, transcripts representing all but 32 viral genes were detected at RPKM values of >8, indicating that by 6 h p.i., the majority of the genome (approximately 80% of viral genes) is transcriptionally active. Between 6 and 12 h p.i., expression increased dramatically for many viral genes. At 6 h p.i., RPKM values for AcMNPV mRNAs ranged from undetectable to approximately 2,300. In contrast, by 12 h p.i., RPKM values ranged from 14 to 84,000 (see Table S5 in the supplemental material). The left y axis indicates viral DNA copy numbers relative to the host cell haploid genome. The left y axis indicates viral mRNA reads as a percentage of total mRNA reads from the cell (bar = SD; n = 3). For comparison of the infection cycle, viral genomic DNA was similarly measured from Sf9 cells infected in parallel with the same virus (MOI, 10). Relative levels of AcMNPV DNA from infected Tnms42 or Sf9 cell lines were measured by qPCR. The viral DNA copy number was normalized to the host cell genome based on the hsp90 copy number. hsp90 was previously reported as a single copy gene per haploid genome in Spodoptera frugiperda and Bombyx mori (98). qPCR was performed using the following gene-specific primer sets: Q_S9_hsp90_F, CTG ACT TAG CGA TGG TAC CAA GG, and Q_S9_hsp90_R, GGA CAA GAT CCG TTA TGA GTG; hsp90-3_F, CAG ACT TTG CAC TCG TAA GAA GG, and hsp90-3_R, GGA CAA GAT CCG TCA TGG GTG; Q_AC_vp39_F, AAT TGA TCG CCA GCA GCA CCG CC, and Q_AC_vp39_R, TAC CGA AGG TTT CAC AAG CAA C.
the supplemental material). As a percentage of total cellular mRNA, AcMNPV mRNA increased from an estimated 3% at 6 h p.i., to approximately 38% by 12 h p.i. (Fig. 4). This dramatic increase in viral transcription also corresponds to the period in which viral DNA replication begins or accelerates (Fig. 4) and therefore reflects both increased promoter activity as well as increased template DNA available for transcription. For 30 genes, we detected a >200-fold increase in transcript levels between 6 and 12 h p.i. It is of special note that the levels of three transcripts encoding structural proteins (p6.9, odv-ec27, and odv-e18) were extremely high (RPKM values of approximately 50,000 to 75,000), and the levels for these transcripts were 2- to 3-fold higher than those of any other viral transcripts at 12 h p.i. (see Fig. S3B and Table S5 in the supplemental material). A group of 23 additional genes (odv-e25, orf75, orf73, orf82, orf81, orf74, cg30, pp31, vp39, alkexo, orf132, bv/odv-e42, orf76, Ac-bro, cxx, orf102, orf93, pp34, 49k, gp64, orf78, orf79, and gp16) also had high transcript levels (RPKM values of 7,000 to 20,000) at 12 h p.i. Approximately 109 genes had transcript levels between 200 and 7000 at that time. At 12 and 18 h p.i., the p6.9, odv-e18, and odv-ec27 mRNAs represented the most highly expressed transcripts. However, by 24 h p.i., the levels of polyhedrin mRNA exceeded all others, and by 36 to 48 h p.i., polyhedrin and p10 were the most abundant viral mRNAs. By 48 h p.i., the levels of polyhedrin and p10 mRNAs were approximately 3- to 4-fold higher than those of p6.9, odv-e18, and odv-ec27, the next most abundant group of mRNAs. At 48 h p.i., polyhedrin and p10 mRNAs represented 24% and 7.5% of total cellular RNA, respectively.

Total transcript abundance through the infection cycle. Baculoviruses are highly productive viruses and expression vectors largely because they commandeer the cell’s resources and express enormous quantities of structural proteins associated with the occluded form of the virus. Proteins associated with the structure or formation of occlusion bodies are not generally necessary for viral replication or propagation in cultured cells. The promoters of the two most highly expressed occlusion-related baculovirus genes (polyhedrin and p10) (9) are typically used for recombinant protein production. However, for some applications, it may be desirable to vary the timing of recombinant protein expression in the baculovirus infection cycle while still achieving high levels of protein expression. Thus, using the comparative expression data provided here, specific promoters such as those from p6.9, odv-e18, and odv-ec27 might be selected in order to shift protein production to an earlier portion of the infection cycle, for optimizing protein production, activity, or processing. Alternative promoters have been used for protein expression previously (77). However, the present analysis provides comparative temporal and quantitative information on many promoters and should permit a wide variety of choices for expression levels and timing of protein production.

We also examined expression patterns associated with genes transcribed only from early or late promoters. Promoter distinctions were based on the following criteria. (i) Early promoters were defined by the absence of the canonical late-promoter motif and often by the presence of TATA box and/or CAGT start site motifs. (ii) Late promoters were defined by transcription initiation within the late-promoter (TAAG) motif (Fig. 2D). We analyzed the expression patterns of AcMNPV genes associated with different promoter motifs. Of the 20 most abundant viral mRNAs detected at 6 h p.i. (see Fig. S3B in the supplemental material), all but one represent transcripts generated from early promoters, and approximately half of the early transcripts are transcribed from promoters containing TATA boxes or TATA-like sequences. Several (pp31, gp64, and he65) of the 20 early genes that were most highly expressed at 6 h contain a functional late promoter motif in addition to an early promoter. Transcript levels of genes transcribed from promoters containing one or both early motifs (TATAA and CAGT) were typically highest at 6 to 12 h p.i. (Fig. 3C and D). In contrast, genes that contain a functional late promoter (initiating at a TAAG motif) typically showed a delayed peak of maximal transcript levels, with most peaking at 12 to 18 h p.i. or later. When multiple motifs are present, the overall expression pattern appears to be determined primarily by the late promoter, since all genes with TAAG-containing promoters peaked at 24 h p.i. or later (Fig. 3C and D). Late genes display large expression level differences that range from tens to thousands in RPKM values. As described above, we also compared the sequences flanking TAAG motifs. For this analysis, we clustered genes expressed in similar patterns (Fig. 5A and B), selected from each cluster those with only one late promoter motif, and then performed a motif analysis using the MEME algorithm (40). The flanking sequences showed little conservation, and motifs generated had very low reliability scores, perhaps due to the limited sample size. Thus, experimental analysis of isolated promoter sequences will likely be necessary to identify promoter motifs that regulate expression patterns.

Genome-wide characterization of promoters in model eukaryotic systems such as human and Drosophila genomes shows that the majority of genes do not have a universal promoter sequence or structure. However, approximately 20 to 40% may be directed by an upstream TATA box, which is the most highly conserved feature of promoters recognized by RNA polymerase II. Promoters from these model systems may be classified as either dispersed or focused, based on the precision of transcription start site initiation (47, 78). Mammalian promoters are predominantly of the dispersed type, initiating transcription at multiple nucleotide positions over a wide area of perhaps 50 to 100 bp. In contrast, Drosophila and perhaps most insects have mostly focused promoters. For focused promoters, transcription typically initiates at a single nucleotide or within a few nucleotides. The TSS usage we observed for baculovirus early (non-TAAG motif) genes was highly stringent and consistent with focused promoters of RNA polymerase II. Late promoters were also highly precise, initiating at position 2 (the first A) of the 5’-TAAG-3’ motif, as described above (see Fig. S2C in the supplemental material).

ORFs with unresolved gene expression patterns. The expression of 5 overlapping transcripts could not be clearly resolved. The genes orf44, orf56, orf113, and alkexo are located in tandem with adjacent genes, and they share 3’ termination sites with overlapping mRNAs, such that their expression levels could not be readily separated from those of the overlapping genes. In addition, transcripts from lef-10 overlap adjacent mRNAs to such a large extent that lef-10 transcripts could not be separately resolved. In these cases, mRNA levels might be measured only by techniques that can discriminate between overlapping transcripts based on transcript size differences.

Clustering analysis of viral mRNA expression patterns. Baculovirus genes are often assigned to four groups based on the timing of transcriptional activation: immediate early, delayed early, late, and very late. While the timing of transcription activa-
tion is an important feature of regulation, the abundance of transcripts at any given time likely plays a more important role in terms of protein expression. Transcript abundance for a given gene is determined by a combination of factors, including the timing and rate of transcription activation, the rate of transcript elongation and processing, and mRNA stability or half-life. To generate a better overview of regulation of the AcMNPV transcriptome over time, we performed cluster analysis with RNA-Seq data (RPKM values) using the complete-linkage hierarchical clustering method (45), to identify AcMNPV gene expression patterns based on transcript abundance. The cluster analysis revealed four major groups (Fig. 5; also, see Table S5 in the supplemental material), based on the temporal abundance and the pattern of changes in transcript accumulation through the infection cycle. These groups are arbitrarily named groups 1 to 4 (G1 to G4).

G1 and G2 showed patterns of higher transcript abundance at the earliest times with G1 transcripts present at higher levels initially and reaching higher overall levels of the two groups (Fig. 5A, G1 versus G2). In both cases, transcript levels increased dramatically between 0 and 6 h.p.i. (Fig. 5B). Peak transcript levels for G1 and G2 were observed at 12 to 18 h.p.i. and levels either were maintained or decreased slightly through 48 h.p.i. In general, early genes and regulatory genes are the most abundant gene categories in G1 and G2 (Fig. 5B, pie charts). Early genes comprised approximately 54 and 40% of the genes in G1 and G2, respectively, and late genes represent approximately 27% and 20% of genes in G1 and G2, respectively. Thus, the timing of transcription activation is strongly but not absolutely correlated with mRNA abundance. Also, regulatory genes represented approximately 25 and 32% of the genes in G1 and G2, respectively, while structural protein genes represent a relatively minor fraction in these clusters.

Transcripts in G3 and G4 did not generally increase rapidly between 0 and 6 h.p.i. but increased dramatically between 6 and 12 h.p.i. and on average peaked between 12 to 18 h.p.i. (G3) or 18 to 24 h.p.i. (G4). Late genes and genes encoding structural proteins make up the largest identified categories of genes in G3 and G4. Late genes account for 57% of the genes in G3 and 86% of the genes in G4. Transcripts encoding structural proteins make up approximately 23% of G3 and 32% of G4. While most genes in G4 appeared to peak around 18 to 24 h.p.i., several genes (very late genes) continued to increase substantially beyond 24 h.p.i. Individual genes that comprise each of the clusters are identified in Table S5 in the supplemental material, and individual gene profiles are illustrated in Fig. S3B and C in the supplemental material.

Thus, the analysis of transcript abundance through the infection cycle shows that functional classes of genes (such as the reg-

![FIG 5 Cluster analysis of AcMNPV gene expression patterns. (A) The schematic shows a heat map of normalized viral gene expression levels from 0 to 48 h.p.i. Each gene is represented as a single horizontal line, and times postinfection are indicated above the heat map. Gray indicates no detected expression. (B) Based on gene expression levels and patterns, annotated genes were clustered into four groups, G1 to G4. In each group, individual gene patterns are shown as thin colored lines, and an average pattern for the cluster is indicated by the thick black line. The x axis shows hours postinfection, and the y axis shows a relative measure of normalized gene expression levels. On the right, the percentage of genes corresponding the transcription activation phases (early, late, very late, and unknown) and functions (regulatory, structure, accessory, and unknown) are shown in pie charts for each group (G1 to G4).]
ulotary or structural genes) are not strictly regulated in parallel, but the identified regulatory protein genes were found predominately in clusters that are expressed early in infection. Transcripts encoding structural proteins were found primarily in clusters of mRNAs that accumulate late in the infection cycle. While these trends are not unexpected for this virus, these data provide a detailed comparative and comprehensive view of AcMNPV transcripts and their abundance throughout the infection cycle.

**Antisense transcripts.** Because of the compact nature of the AcMNPV genome, we observed antisense RNAs overlapping many ORFs. A frequent form of antisense RNA appears to result from overlapping mRNAs transcribed in a head-to-head or tail-to-tail orientation. An example of head-to-head overlap is the mRNA of vp39, in which three vp39 TSS are located within the coding region of lef-4, which is immediately upstream and in the opposite orientation (head to head), resulting in 40- to 304-nt overlaps of the vp39 mRNA 5’ ends with the ORF of lef-4. Such antisense overlaps in transcripts may result in downregulation of some genes, and this type of regulation has been reported previously (79). We found 50 genes for which there was more antisense than sense RNA at one or more time points in the infection cycle (see Table S5B and Fig. S4A in the supplemental material). Approximately 40 of those 50 genes appear to be similarly downregulated. For example, the level of p74 mRNA declines rapidly after 18 h.p.i., coinciding with a rapid increase of antisense RNA from the adjacent p10 gene (see data for orf112 and orf113 in Tables S5 and S6; also, see Fig. S4B). Thus, these correlations might suggest that antisense RNAs may play a significant role in regulating baculovirus gene expression. The RNAs that are antisense to three ORFs (orf119, 49K, and odv-e56) contain small ORFs of their own and potentially encode proteins of 84, 105, and 119 amino acids. Thus, some of the antisense RNAs detected in this study may represent new AcMNPV genes that are not yet annotated (see Table S5B).

**Splicing in the AcMNPV transcriptome.** Although RNA splicing is common in eukaryotes and many viruses, splicing is thought to be rare in baculoviruses (9). Only one spliced gene, the immediate early gene ie-0, was previously identified in AcMNPV, and splicing of that gene is conserved in other related baculoviruses (see Fig. S5E in the supplemental material) (80–84). In the closely related virus Orgyia pseudotsugata MNPV (OpMNPV), an additional spliced gene product was identified at late times postinfection (85). In that case, a splice site in the 3′ UTR of gene op18 results in an extension of the length of the 3′ UTR of op18. In the present study, we identified 11 additional AcMNPV mRNA splice sites for which the resulting spliced RNAs were present at a frequency of approximately 1% or more of the target gene mRNA (Table 1). Spliced mRNAs were detected at early, intermediate, and late time points. Among the total of 12 identified splice sites, 8 are consistent with the canonical GT/AG splice site signature, 3 are of the rare AT/AC class (86, 87), and one utilizes a GC/AG donor/acceptor sequence (Table 1). Most of the identified splice sites (10 of 12) were detected initially at 12 h p.i., and two were detected earlier, at 6 h p.i. Only 3 of the identified splice sites modify the protein coding regions of AcMNPV genes. The IE0 and IE1 proteins are splice variants of a highly active and extensively studied transcription factor. By analyzing RNA-Seq reads at the ie0/exon0 splice junction and measuring the ratio of spliced (ie0) to unspliced (exon0) transcripts at that junction, we found that the spliced form (ie0) represents approximately 84% of the ie0+exon0 transcripts at 6 h.p.i., but the spliced form decreases dramatically to 15, 7, 3, 4, and 2% of the total ie0+exon0 transcripts at 12, 18, 24, 36, and 48 h.p.i., respectively. The measured ratios of spliced versus nonspliced forms of ie-0/exon0 were consistent with prior reports of a decrease in ie-0 splicing early in the infection cycle (80, 81, 83, 88). The other two splices that affect predicted protein-coding regions remove small introns (130 to 498 nt) and shift the reading frames in p47 and exon0 genes, generating severely truncated proteins. However, the latter two splices represent less than 2% of the total target mRNA populations at the times of detection (Table 1, transcripts 6 and 11), and thus their relative importance remains unclear. Of the remaining noncoding region splice sites, some appear to represent substantial proportions of the mRNA populations, including those from the 3′ UTR of Ac-sod and from the antisense and 3′ UTR regions of orf114 (Table 1, transcripts 5 and 7 to 9). Splicing in the 3′ UTR region could possibly regulate gene expression through effects on mRNA half-life or on translation efficiency (58, 59). However, the significance of multiple spliced products of the sense and anti-sense mRNAs of the orf114 region is unknown. Based on the decrease in ie-0 splicing as infection progresses, it was proposed previously that the cellular splicing machinery might be impaired or inactivated during viral infection (62). However, the relative abundance of some spliced mRNAs at late times and the relatively constant ratios of spliced to nonspliced mRNA of the 11 newly identified spliced transcripts suggest that the splicing machinery remains viable though late times in the viral infection cycle.

**DISCUSSION**

We used deep sequencing to measure AcMNPV transcript levels over the course of the infection cycle. Transcript levels result from a combination of factors, but the primary determinants include timing of promoter activation, transcription initiation rates, polymerase processivity, transcript termination and polyadenylation, and transcript stability or turnover. An advantage of the current method of analysis is the simultaneous and comparative examination of transcripts representing the entire genome. A difficulty in analyzing baculovirus transcription in this manner, however, is the close proximity of viral genes and the high degree of transcript overlap. To address these issues, we combined global mapping of 5′ and 3′ ends of AcMNPV transcripts with strand-specific RNA-Seq data. From the measurements of the relative abundance of transcripts at various times postinfection, several general concepts emerge regarding AcMNPV gene transcription throughout the infection cycle. First, the levels of transcription for different viral genes vary over a large dynamic range. Measured RPKM values ranged from less than 100 to over 300,000, a >3,000-fold difference. Second, because of the high sensitivity of our methods, we were able to detect very low levels of 40 different viral transcripts in host cells after the 1-h inoculation period. It should be noted that these viral transcripts represented only approximately 0.1% of the RNA-Seq reads detected at that time. The most abundant of these low-level viral mRNAs represent late viral mRNAs that are highly abundant when the virus assembles. Thus, these data combined with prior proteomic analyses (89, 90) may suggest that upon infection, budded AcMNPV virions may deliver a fairly extensive array of viral mRNAs and proteins (although at a very low level) in addition to the viral genome. Alternatively, nucleocapsids could carry a small quantity of the viral RNA polymerase that mediates transcription of these very low levels of late transcripts. Whether these mRNAs and proteins have biological significance is...
not yet known, but it is intriguing to imagine that they may provide a potential benefit to the virus at early stages of the infection. Third, the levels of canonical early transcripts detected were low in comparison to transcripts initiated from late promoter motifs (late genes). At 6 h p.i. the most abundant transcripts detected were transcribed predominantly from (non-TAAG) early promoters. Peak RPKM values for the most abundant viral transcripts at 6 h p.i. were 72,600, and viral transcripts represented only 3.2% of the total cellular transcripts (Fig. 4, 6 h p.i.; also, see Table S1 in the supplemental material). In contrast, by 12 h p.i., late transcripts from p6.9, odv-e18, and odv-ec27 had RPKM values exceeding 40,000, and viral transcripts represented approximately 38% of the total transcripts detected. This increase in viral transcripts is due to both the activity of the late RNA polymerase and the amplification of the viral genome template (Fig. 4). We cannot rule out the additional possibility that late transcripts may be more stable than early transcripts.

We documented a wide variety of expression patterns that can now be viewed as a whole for direct gene-to-gene comparisons or for inferred promoter-to-promoter comparisons. Transcripts from genes such as lef-3, dbp, pp31, p35, and egt are among the most abundant viral transcripts at 6 h p.i., although their individual patterns of expression differ substantially through the remainder of the infection cycle (see Fig. S3B and C in the supplemental material).

### Table 1: Spliced AcMNPV Transcripts

<table>
<thead>
<tr>
<th>Transcript</th>
<th>Strand</th>
<th>Intron position</th>
<th>Intron donor site</th>
<th>Intron acceptor site</th>
<th>Intron length</th>
<th>Associated gene or sequence</th>
<th>Hours postinfection</th>
<th>% spliced reads (total no. of spliced reads)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>+</td>
<td>264–354</td>
<td>5′-GC</td>
<td>AG-3′</td>
<td>90</td>
<td>Ac-tp5 5′ UTR</td>
<td>12</td>
<td>3.40 (46)</td>
</tr>
<tr>
<td>2</td>
<td>−</td>
<td>12979–11354</td>
<td>5′-GT</td>
<td>AG-3′</td>
<td>1,625</td>
<td>lef-1 5′ UTR</td>
<td>12</td>
<td>1.32 (10)</td>
</tr>
<tr>
<td>3</td>
<td>−</td>
<td>12979–12718</td>
<td>5′-GT</td>
<td>AG-3′</td>
<td>261</td>
<td>Antisense Ac-egt</td>
<td>12</td>
<td>1.71 (14)</td>
</tr>
<tr>
<td>4</td>
<td>+</td>
<td>26431–26502</td>
<td>5′-AT</td>
<td>AC-3′</td>
<td>71</td>
<td>Ac-sod 3′ UTR</td>
<td>12</td>
<td>2.58 (21)</td>
</tr>
<tr>
<td>5</td>
<td>+</td>
<td>26447–26518</td>
<td>5′-GT</td>
<td>AT-3′</td>
<td>71</td>
<td>Ac-sod 3′ UTR</td>
<td>12</td>
<td>2.84 (23)</td>
</tr>
<tr>
<td>6</td>
<td>−</td>
<td>33305–33175</td>
<td>5′-GT</td>
<td>AG-3′</td>
<td>130</td>
<td>p47 coding region</td>
<td>12</td>
<td>2.1 (13)</td>
</tr>
<tr>
<td>7</td>
<td>−</td>
<td>97872–97631</td>
<td>5′-GT</td>
<td>AG-3′</td>
<td>241</td>
<td>orf114 3′ UTR</td>
<td>12</td>
<td>56.41 (198)</td>
</tr>
<tr>
<td>8</td>
<td>+</td>
<td>97641–97713</td>
<td>5′-AT</td>
<td>AC-3′</td>
<td>72</td>
<td>Antisense orf114</td>
<td>12</td>
<td>72.60 (1,486)</td>
</tr>
<tr>
<td>9</td>
<td>−</td>
<td>97835–97747</td>
<td>5′-AT</td>
<td>AC-3′</td>
<td>88</td>
<td>orf114 3′ UTR</td>
<td>12</td>
<td>41.47 (423)</td>
</tr>
<tr>
<td>10</td>
<td>+</td>
<td>99610–100001</td>
<td>5′-GT</td>
<td>AG-3′</td>
<td>391</td>
<td>Antisense orf115</td>
<td>6</td>
<td>8.82 (20)</td>
</tr>
<tr>
<td>11</td>
<td>+</td>
<td>122946–123444</td>
<td>5′-GT</td>
<td>AG-3′</td>
<td>498</td>
<td>exon0</td>
<td>12</td>
<td>0.49 (19)</td>
</tr>
<tr>
<td>12</td>
<td>+</td>
<td>122946–127149</td>
<td>5′-GT</td>
<td>AG-3′</td>
<td>4,203</td>
<td>ie-0</td>
<td>6</td>
<td>83.69 (209)</td>
</tr>
</tbody>
</table>
of small RNAs revealed hotspots of small RNAs that mapped non-uniformly across the genome and to both strands (93). Hotspots were located mostly in late genes identified as encoding structural proteins and genes involved in viral dispersal. Notably, hotspots of siRNA in the HaSNPV genome were within the ORFs of the very highly expressed genes polyhedrin, p6.9, and p10, as well as polh, env, p49, odv-ec27, the cathepsin gene, p74, the chitinase gene, p26, the gene encoding envelope protein, and odv-e25. Those authors (93) suggested that an RNAi response to these genes may in some way regulate viral replication. Small RNAs have not been mapped across the AcMNPV genome, and it is not known if there are viral siRNAs that correspond to those identified in HaSNPV. Although the highest-level antisense RNAs do not correlate well with the ORFs identified in the HaSNPV study, the level of antisense RNA may not be a reliable indicator of RNAi response. It will be of interest in future studies to determine whether AcMNPV antisense RNAs play a role in an RNAi response or whether they have another role in regulating baculovirus gene expression.

A global view of the baculovirus transcriptome will undoubtedly lead to improvements in the application of baculoviruses as expression vectors. Among the most highly expressed genes, the timing of transcript abundance varied substantially, suggesting that the promoters regulating these genes differ substantially. This information can be used to optimize protein expression from heterologous genes in AcMNPV expression vectors. By selecting a specific expression pattern for expression of a heterologous protein, production could be optimized and improved. For example, if a posttranslational processing step requires the production of a specific processing enzyme prior to the overexpression of the target protein, it should be possible to express the processing enzyme under the control of an appropriate promoter (such as the p6.9, odv-ec27, or odv-e18 promoter) that is highly active at an earlier time. The target protein that requires processing could be expressed under the control of a promoter that is active later in the infection cycle (for example, the p10 or polyhedrin promoter). Similarly, if the assembly of a functional protein complex requires expression of one component prior to another or requires a specific stoichiometry for assembly, the selection of promoters based on expression timing and levels may facilitate or enhance this process. Modulating the levels and timing of expression might also be used to enhance protein folding or stability by expression of chaperones, protein disulfide isomerase, or specific components of the desired glycosylation pathway (7, 95, 96). Thus, the availability of a detailed AcMNPV transcriptome will provide for a vastly improved understanding of the complexities of regulation of the 156 or more viral genes, and in addition will aid in the use of baculoviruses for protein expression and for other more complex applications, such as the assembly of virus-like particles for use as vaccines, production of pseudotyped baculovirus virions for use in transduction and gene therapy, and use of baculoviruses for biological insect control.

ACKNOWLEDGMENTS

We thank Ying Shao for help and assistance with Illumina sequencing and Gao Shan, Maria Huang, and Zen Yi for assistance and advice on bioinformatic analysis and approaches. We thank Abigail McSweeney for excellent technical assistance and Sicong Dong for reading the manuscript. We also thank the reviewers for insightful comments that were helpful in improving the manuscript.

This work was supported by grants from DARPA and the NIH.
REFERENCES


28. Yamagishi J, Burnett ED, Harwood SH, Blissard GW. 2007. The AcMNPV p31 gene is not essential for productive AcMNPV replication or late gene transcription but appears to increase levels of most viral transcripts. Virology 365:34–47.


47. Blissard GW, Rohrmann GF. 1989. Location, sequence, transcriptional mapping, and temporal expression of the gpe4 envelope glycoprotein gene
of the Orgyia pseudotsugata multicapsid nuclear polyhedrosis virus. Virol-
ogy 170:537–555.


54. Fitzgerald PC, Sturgill D, Shyakhentko A, Oliver B, Vinson C. 2006. Com-
putational analysis of Drosophila and human core promoters. Ge-
nome Biol. 7:R53. doi:10.1186/gb-2006-7-7-r53.

core promoters in the Drosophila genome. Genome Biol. 3:RE-
SEARCH0087. doi:10.1186/gb-2002-3-3-research0087.

composition, conservation and condition-specificity of single and alter-
native transcriptional start sites in the Drosophila genome. Genome Biol.
10:R73. doi:10.1186/gb-2009-10-7-r73.

57. Rankin C, Ooi BG, Miller LK. 1988. Eight base pairs encompassing the
transcriptional start point are the major determinant for baculovirus poly-

Genome Biol. 3:REVIEWS0004. doi:10.1186/gb-2002-3-3-reviews0004.

positional properties of untranslated regions of human mRNAs. Gene

analysis of the 5′ and 3′ ends of vaccinia virus early mRNAs deline-
ates regulatory sequences of annotated and anomalous transcripts. J. Vi-
rol. 85:5897–5909.

61. Gombart AF, Blissard GW, Rohrmann GF. 2002. Identification and charac-


63. Durantel D, Croizier L, Ayres MD, Croizier G, Possee RD, Lopez-
Ferber M. 1998. The pkp/jnl gene (orf 86) of Autographa californica
nucleopolyhedrovirus is a non-essential, immediate early gene. J. Gen.

64. Lu A, Carstens EB. 1992. Transcription analysis of the EcoRI D region of the baculovirus Autographa californica nuclear polyhedrosis virus identi-

65. Possee RD, Sun TP, Howard SC, Ayres MD, Hill-Perrins M, Gearing


67. Jin J, Guarino LA. 2000. 3′-end formation of baculovirus late RNAs. J.

68. Danckwardt S, Hentze MW, Kulozik AE. 2008. 3 prime end mRNA
processing: molecular mechanisms and implications for health and dis-

from the genome of Autographa californica nuclear polyhedrosis virus: a

70. Broussard DR, Guarino LA, Jarvis DL. 1996. Mapping functional do-

71. Evans JT, Leisy DJ, Rohrmann GF. 1997. Characterization of the inter-
action between the baculovirus replication factor LEP-1 and LEP-2. J.

72. Hang X, Dong W, Guarino LA. 1995. The lef-3 gene of Autographa cal-
ifornica Nuclear Polyhedrosis Virus encodes a single-stranded DNA


74. Mikhailov VS, Rohrmann GF. 2002. Baculovirus replication factor LEP-1 is
a DNA primase. J. Virol. 76:2287–2297.

75. Hill-Perkins MS, Possee RD. 1990. A baculovirus expression vector de-
283:37–49.

76. Cooper SJ, Trinklein ND, Anton ED, Nguyen L, Myers RM. 2006. Comprehen-
sive analysis of transcriptional promoter and structure and func-
tion in 1% of the human genome. Genome Res. 16:1–10.

77. Ooi BG, Miller LK. 1990. Transcription of the baculovirus polyhedrin
gene reduces the levels of an antisense transcript initiated downstream. J.
Virol. 64:3126–3129.

78. Chisholm GE, Henner DJ. 1988. Multiple early transcripts and splicing of
the Autographa californica nuclear polyhedrosis virus IE-1 gene. J. Virol.

79. Kovacs GR, Guarino LA, Graham BL, Summers MD. 1991. Identifica-
tion of spliced baculovirus RNAs expressed late in infection. Virology

80. Pearson MN, Rohrmann GF. 1997. Splicing is required for transactiva-
tion by the immediate early gene 1 of the Lymantria dispar multinucleo-

239:357–373.

82. Nie Y, Theilmann DA. 2010. Deletion of AcMNPV AC16 and AC17 results in

83. Braungel SC, Russell WK, Rosas-Acosta G, Russell DH, Summers MD.
2003. Determination of the protein composition of the occlusion-derived


85. van Mierlo JT, van Rij RP. 2011. Defense and counter-

86. Luebhart H, Doerfler W. 1984. Transcription of overlapping sets of RNAs
from the genome of Autographa californica nuclear polyhedrosis virus: a

87. Broussard DR, Guarino LA, Jarvis DL. 1996. Mapping functional do-

88. Evans JT, Leisy DJ, Rohrmann GF. 1997. Characterization of the inter-
action between the baculovirus replication factor LEP-1 and LEP-2. J.

89. Hang X, Dong W, Guarino LA. 1995. The lef-3 gene of Autographa cal-
ifornica Nuclear Polyhedrosis Virus encodes a single-stranded DNA


92. Hill-Perkins MS, Possee RD. 1990. A baculovirus expression vector de-

93. Cooper SJ, Trinklein ND, Anton ED, Nguyen L, Myers RM. 2006. Comprehen-
sive analysis of transcriptional promoter and structure and func-
tion in 1% of the human genome. Genome Res. 16:1–10.

94. Ooi BG, Miller LK. 1990. Transcription of the baculovirus polyhedrin
gene reduces the levels of an antisense transcript initiated downstream. J.
Virol. 64:3126–3129.