Mass identification of transcriptional units expressed from the *Bombyx mori* nucleopolyhedrovirus genome

Susumu Katsuma,1 WonKyung Kang,2 Tadasu Shin-i,3 Kazuko Ohishi,3 Koji Kadota,4 Yuji Kohara3 and Toru Shimada1,4

Correspondence
Susumu Katsuma
katsuma@ss.ab.u-tokyo.ac.jp

1Department of Agricultural and Environmental Biology, University of Tokyo, Yayoi 1-1-1, Bunkyo-ku, Tokyo 113-8657, Japan
2Molecular Entomology Laboratory, RIKEN ASI, 2-1 Hirosawa, Wako 351-0198, Japan
3Center for Genetic Resource Information, National Institute of Genetics, Mishima 411-8540, Japan
4Agricultural Bioinformatics Research Unit, Graduate School of Agricultural and Life Sciences, University of Tokyo, Yayoi 1-1-1, Bunkyo-ku, Tokyo 113-8657, Japan

In order to identify the transcriptional units expressed from an entire nucleopolyhedrovirus (NPV) genome during infection, we constructed a full-length-enriched cDNA library from *Bombyx mori* NPV (BmNPV)-infected BmN cells. We randomly sequenced 11 520 clones from both ends to obtain a total of 4679 BmNPV-derived transcriptional units. The data revealed a number of novel transcripts, including putative non-coding RNAs, most of which are expressed from recognized baculovirus early or late promoter motifs. These findings provide new insights into the complex transcriptional regulation of an NPV genome and suggest roles for as-yet-uncharacterized transcripts.

The nucleopolyhedroviruses (NPVs) are members of the genus *Alphabaculovirus* of the *Baculoviridae* and possess a large circular, dsDNA genome encapsidated within a rod-shaped virion. Gene expression of NPVs is regulated mainly at the transcription level and is marked by early, late and very late phases. Early transcripts are synthesized by host RNA polymerase II, while late transcription is mediated by a virus-encoded RNA polymerase. Mapping analysis of late transcripts has revealed a TAAG sequence motif that is required for late and very late promoter activity, and two genes, *polyhedrin* (*polh*) and *p10*, are selectively and highly expressed at the late stage of infection.

To understand how transcription regulation operates in NPV-infected host cells, several groups have performed genome-wide expression analyses of NPV genes using DNA microarrays (Yamagishi et al., 2003; Iwanaga et al., 2004; Jiang et al., 2006). This technology, however, cannot discriminate between the variable-length baculovirus transcripts that commonly overlap more than one ORF (Smith, 2007). To obtain a precise overview of the population of transcriptional units expressed from an NPV genome, we generated and characterized a full-length-enriched cDNA library from *Bombyx mori* NPV (BmNPV)-infected BmN cells.

BmN cells were infected with BmNPV at an m.o.i. of 5. At 20 h post-infection (p.i.), total RNA was prepared using Trizol (Invitrogen). A full-length-enriched cDNA library was constructed by the vector-capping method using the pGCAP10 vector (Hitachi High-Tech Manufacturing & Service Corporation). Using this method, approximately 95% of cDNA clones are full-length when cultured cells or fresh tissues dissected from an animal organ are used (Kato et al., 2005). The DNA of each clone was directly amplified from bacterial cultures by the RCA method (Dean et al., 2001) using a TempliPhi HT DNA amplification kit (GE Healthcare). Single-pass sequencing of both ends of 11 520 clones was performed using an ABI3730XL DNA sequencer (Applied Biosystems) with forward (5'-TGGTGGTG-GCAAATCAAAGAA-3'; the annealing site is located 135 bases upstream of the cloning site) and reverse anchored oligo-dT, dNT([17]) primers. Note that, in some cases, 3'-expressed sequence tags (ESTs) are not identical to 3'-end sequences of cDNA clones, depending on the annealing sites of an anchored oligo-dT primer and/or cDNA length.

We obtained a total of 20 635 ESTs (10 281 5'-tags and 10 354 3'-tags), and all sequences were submitted to the DNA Database of Japan, under accession numbers FY043077–FY063711. We performed BLASTN searches and...
selected BmNPV-derived clones, both ends of which matched sequences in the genome of BmNPV isolate T3 (accession number L33180). After eliminating clones whose 5’ or 3’ ends were of low quality and trimming the vector sequences from the remainder, we used 4679 clones for further analyses (Fig. 1a). The genomic location and length of each transcriptional unit are listed in Supplementary Table S1, available in JGV Online. As shown in Fig. 1(b), the distribution of sense (transcripts in the same orientation as that of polh) and antisense transcripts was very similar: the numbers of sense and antisense transcripts were 2071 and 2608, their mean lengths were 1003 and 962 bases and the percentages of long transcripts (>2.0 kb) in each direction were 4.4 and 5.9%, respectively. The transcripts cover approximately 96% of the genome (Fig. 1c), showing that almost all of the 128 kbp BmNPV genome is transcriptionally active in BmNPV-infected BmN cells. In addition, we found that transcriptional units exist with high frequency in the sense orientation where the corresponding antisense sequences are not transcribed or are only transcribed with low frequency at this time point (Fig. 1c).

Next, we assessed whether known early and late genes of BmNPV exist in this library. Although late transcripts were completely included in the present list, Bm21, one of the BmNPV early genes (Huang et al., 2008), was missing, suggesting that our library generated from BmNPV-infected cells at 20 h p.i. does not include a proportion of the early transcripts. We also observed that 55.4% (214/386 sites) of baculoviral late promoter motifs (TAAGs) in the genome are actually used in BmNPV-infected cells.

Many studies of baculovirus temporal gene expression have revealed the existence of transcripts spanning two or more

---

**Fig. 1.** Generation and characterization of a full-length-enriched cDNA library from BmNPV-infected BmN cells. (a) Scheme for construction and data processing of a full-length-enriched cDNA library from BmNPV-infected BmN cells. (b) Size distribution of BmNPV-derived transcriptional units. (c) Mapping of 4679 transcriptional units onto the BmNPV genome.
ORFs (see Ooi & Miller, 1990). Our current study clearly shows that overlapping transcripts are distributed throughout the genome (see Supplementary Table S1). As an illustration, transcripts mapping in the p26–adv-ec27 region are shown in Fig. 2. In this region, an extremely long, 8332-base transcript is located, and many kinds of sense and antisense transcriptional units overlap each other. According to previous reports (Happ et al., 1991; Smith, 2007), multiple proteins are probably not translated from this type of mRNA. Such expression profiles cannot be discovered by the strategy used

Fig. 2. Mapping of transcriptional units located in the p26–adv-ec27 region. Arrows show each transcriptional unit. The bold arrow indicates an extremely long transcript located in this region.

Fig. 3. Identification of putative regulatory ncRNAs. Bold arrows indicate putative regulatory ncRNAs found in the polh (a) and pif-1 (b) loci. Late (TAAG) and early (CAGT/CAGA) motifs are shown at the 5' end of each ncRNA. Arrows show all transcriptional units identified in this region.
in previous transcriptome studies (Yamagishi et al., 2003; Iwanaga et al., 2004; Jiang et al., 2006).

We also observed a number of transcripts that are incapable of encoding known baculoviral proteins. As summarized in Supplementary Fig. S1, we identified 56 kinds of putative non-coding RNAs (ncRNA) that were represented at least twice in the library. Most of these ncRNAs are initiated from baculoviral late (TAAG; 84%) or early (CAGT/CAGA; 7%) promoter motifs (Fig. 3 and Supplementary Fig. S1), suggesting strongly that they are not incomplete cDNAs generated during library construction. Intriguingly, about a quarter of these ncRNAs exist in the antisense orientation near the 5′ ends of known ORFs (Fig. 3). In the polh region, two antisense ncRNAs are transcribed from a TAAG motif at different positions within the polh coding region and have a common 3′ end (Fig. 3a). Similarly, the pif-1 gene, a gene required for per os infectivity of NPVs (Kikhno et al., 2002), encompasses five different antisense ncRNAs whose 3′ termini are located close to its 5′ end (Fig. 3b). Three ncRNAs are transcribed from a TAAG motif at different positions within the pif-1 coding region, whereas two are initiated from putative early motifs (CAGT/CAGA). Two groups have observed an RNA traversing the polh gene in the antisense direction in NPV-infected cells (Ooi & Miller, 1990; Yang et al., 2007). Interestingly, a mutant virus that does not express polh, as a result of a point mutation at the essential RNA initiation site of polh, exhibited higher levels of this antisense transcript than did wild-type virus (Ooi & Miller, 1990), suggesting that the polh gene regulates the level of this RNA. Taking these observations together with our current study, we speculate that some of the baculoviral transcripts that overlap coding RNAs, and are in the antisense orientation, are regulatory ncRNAs.

In summary, we have performed a large sequencing of cDNA clones derived from a full-length-enriched cDNA library from BmNPV-infected BmN cells and shown, for the first time, a precise overview of the population of transcriptional units expressed from an NPV genome. We have also discovered a number of novel transcripts, including putative regulatory ncRNAs. To date, researchers have focused almost exclusively on the protein-coding genes (ORFs) predicted from genome information. However, our results suggest that more attention should be paid in future to all of the transcriptional units produced from the viral genome. To understand better the roles of these putative ncRNAs, we are currently generating NPVs whose ncRNA promoters are inactivated.

Acknowledgements

We would like to thank Ian Smith for critical reading of the manuscript and Yukiko Yamazaki for EST data deposition. This work was supported by grants from MEXT (nos 19688004 and 22380033 to S. Katsuma, no. 17018007 to T. Shimada and the Professional Program for Agricultural Bioinformatics) and NIAS/MAFF (Agrigenome Research Program). cDNA clones were supplied by the National Bio-Resource Project ‘Silkworm’ of MEXT.

References


