Available online at www.sciencedirect.com

SCIENCE DIRECT

Virology 333 (2005) 190–199

VIROLOGY

www.elsevier.com/locate/yviro

Comparison of the complete genome sequence between C1 and G4 isolates of the *Helicoverpa armigera* single nucleocapsid nucleopolyhedrovirus ☆

Chuan-Xi Zhang*, Xiu-Cui Ma, Zhong-Jian Guo

Institute of Applied Entomology, Zhejiang University, Kaixuan Road 268, Hangzhou, Zhejiang 310029, China

Received 14 November 2004; returned to author for revision 30 November 2004; accepted 22 December 2004
Available online 27 January 2005

Abstract

The complete nucleotide sequence of *Helicoverpa armigera* single nucleocapsid nucleopolyhedrovirus isolate C1 (HearSNPV-C1) was determined and analyzed by comparing with the genome of HearSNPV-G4 isolate. C1 and G4 isolates occurred in the same host species and geographic location but showed different virulence. The HearSNPV-C1 genome consisted of 130,759 bp and 137 putative open reading frames larger than 150 nucleotides were identified. The two genomes shared 98.1% nucleotide sequence identity, with a total number of 555 bp substitutions, 1354 bp deletions, and 710 bp insertions in HearSNPV-C1. Comparison of ORFs and homologous repeat (*hr*) regions of the two genomes showed that there were four highly variable regions *hr*1, *hr*4, *hr*5, and *bro-b*, all in repeat regions. These results suggest that baculovirus strain heterogeneity may be often caused by SNPs and changes in the *hr*s and *bro* genes.

Keywords: Baculovirus; *Helicoverpa armigera*; Nucleocapsid nucleopolyhedrovirus; Complete genome; Isolate comparison

Introduction

The Baculoviridae is a family of invertebrate viruses with large, circular, and double-stranded DNA genomes ranging in size from 81.7 (NeleNPV) to 178.7 kb (XecnGV). They are pathogenic to arthropods, mainly insects of the orders Lepidoptera, Hymenoptera, and Diptera (Adams and McClintock, 1991). The family is subdivided into two genera, *Nucleopolyhedrovirus* (NPV) and *Granulovirus* (GV), based on the morphology of occlusion bodies (OBs). The NPVs are designated as viruses forming polyhedral OBs, each of which contains many virions, whereas the GVs typically produce ovoid OBs with a single virion (Blissard et al., 2000). A further phenotypic distinction for NPVs is their recognition as either single nucleocapsid NPVs (SNPVs) or multiple nucleocapsid NPVs (MNPVs) depending on the number of nucleocapsids packaged into each virion (Blissard et al., 2000). This, however, is not correlated to genetic relatedness and appears to have no phylogenetic trait (Murphy et al., 1995).

NPVs are pathogenic to a number of lepidopteran insects and are attractive biological agents for the control of agriculturally important insect pests. The current interest in the molecular biology of these viruses is fostered by their potential as modified virus pesticides with increased toxicity (Stewart et al., 1991) and as gene therapy vectors in medical science (Huser and Hofmann, 2003; Tani et al., 2003), also by their successful use as vectors for the expression of foreign proteins in the baculovirus-insect system (Smith et al., 1983).

So far, the genomic nucleotide sequences of 25 baculoviruses have been completely sequenced. These include 15 lepidopteran NPVs: *Autographa californica* (Ac) MNPV (Ayres et al., 1994), *Bombyx mori* (Bm) NPV (Gomi et al., 1999), *Orgyia pseudotsugata* (Op) MNPV (Ahrens et al.,

☆ The GenBank accession number of the HaSNPV-C1 genomic sequence reported in this paper is AF303045.

* Corresponding author. Fax: +86 571 86971697.
E-mail address: chxzhang@zju.edu.cn (C.-X. Zhang).

0042-6822/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.virol.2004.12.028

1997), *Lymantria dispar* (Ld) MNPV (Kuzio et al., 1999), *Spodoptera exigua* (Se) MNPV (Ijkel et al., 1999), *Helicoverpa armigera* (Hear) SNPV-G4 (Chen et al., 2001), *Helicoverpa zea* (Heze) SNPV (Chen et al., 2002), *Adoxophyes honmai* (Adho) NPV (Nakai et al., 2003), *Spodoptera litura* (Spli) MNPV (Pang et al., 2001), *Epiphyas postvittana* (Eppo) NPV (Hyink et al., 2002), *Mamestra configurata* (Maco) NPV-90/2 (Li et al., 2002a, 2002b), MacoNPV-96B (Li et al., 2002a, 2002b), *Rachiplusia ou* (Raou) MNPV (Harrison and Bonning, 2003), *Choristoneura fumiferana* (Cf) MNPV (GenBank NC_004778), and *Choristoneura fumiferana* defective (CfDEF) NPV (GenBank NC_005137). The genomes of these NPVs range in size from 113.2 kb (AdhoNPV) to 161.0 kb (LdMNPV). In contrast to NPVs, only seven of GVs, *Xestia cnigrum* (Xecn) GV (Hayakawa et al., 1999), *Plutella xylostalla* (Plxy) GV (Hashimoto et al., 2000), *Cydia pomonella* (Cypo) GV (Luque et al., 2001), *Phthorimaea operculella* (Phop) GV (NC_004062), *Adoxophyes orana* (Ador) GV (Wormleaton et al., 2003), *Cryptophlebia leucotreta* (Crle) GV (Lange and Jehle, 2003), and *Agrotis segetum* (Agse) GV (NC_005839) have been determined, with the genomes ranging in size from 99.7 kb (AdorGV) to 178.7 kb (XecnGV). A baculovirus from dipteran insect host, *Culex nigripalpus* (Cuni) baculovirus (Afonso et al., 2001) and two NPVs from hymenopteran hosts, *Neodiprion lecontei* (Nele) NPV (Lauzon et al., 2004) and *Neodiprion sertifer* (Nese) NPV (Garcia-Maruniak et al., 2004), have also been determined, with genome sizes of 108.3, 86.5, and 81.8 kb, respectively. All these genomic sequences give us a better understanding of the distinctive features, evolution, and extent of diversity of baculoviruses. However, data concerning the strain polymorphism at the complete genomic sequence level are limited.

H. armigera is one of the most serious pests in China. As an economically polyphagous pest, it has caused considerable economic loss to many vegetable and field crops such as cotton, corn, baccy, tomato, and wheat. As an important pathogen to *H. armigera*, HearSNPV was the first commercial baculovirus pesticide used to control the *H. armigera* in China, and also has been extensively used for the control of the pests in cotton and vegetable crops (Zhang, 1994). HearSNPV strains with different virulence or molecular characteristics have been isolated (Jia et al., 2003; Sun and Zhang, 1994). Bioassay showed that the virulence of HearSNPV isolate C1 was higher than that of HearSNPV isolate G4. The median lethal doses (LD_{50}) for C1 and G4 against the third instar *H. armigera* were 568 (95% confidence interval 424–740) and 1584 (95% confidence interval 1065–2221) PIBs/larva, respectively. The genomic sequence of HearSNPV isolate G4 has already been determined (Chen et al., 2001). During the present study the genome of HearSNPV-C1 was completely sequenced and compared with the genome of HearSNPV-G4 to further understand the strain heterogeneity and the possible reasons for variation of the virulence in this virus and to provide clues for baculovirus evolution.

Results and discussion

Analysis of the HearSNPV-C1 genome

The complete nucleotide sequence of the HearSNPV-C1 genome has been determined. The sequence data were assembled into a contiguous sequence of 130,759 bp (Tables 1), which was in good agreement with a previous estimate of 130 kb based on restriction enzyme analysis and physical mapping of DNA fragments (Zhang and Wu, 2001). According to the adopted convention (Hayakawa et al., 1999; Ijkel et al., 1999; Vlak and Smith, 1982), *polyhedrin* was designated as the first gene (ORF1) and the adenine residue at the translation-initiation codon of the *polyhedrin* gene was designated as the start point of the circular HearSNPV map.

Using computer-assisted analysis and the criteria of selecting ORFs starting with methionine-initiated codons (ATG) and at least 50 aa having minimal overlap with other ORFs, 137 putative ORFs and five homologous repeat (*hr*) regions were identified for further detailed analysis in the HearSNPV-C1 genome. The location, orientation, and size of the predicted ORFs are shown in Table 1. The number of 137 ORFs is proportional to that of other completely sequenced baculoviruses ranging from NeleNPV (89) to MacoNPV-90/2 (169), especially similar to that of HezeSNPV (139), EppoNPV (135), SeMNPV (139), SpliMNPV (141), and BmNPV (143). The HearSNPV-C1 ORFs have average length of 843 bp with ORF84 (*Helicase*) being the largest (3762 bp) and ORF40 being the smallest (153 bp). The 137 predicted ORFs encoded 38,362 aa. The total coding sequence and the intergenic regions were 114,394 and 10,041 bp and represent 87.5% versus 7.7% of the genome, respectively. The five *hr*s were distributed along the genome with sizes varying from 297 to 2253 bp and the total sequence was 6324 bp accounting for 4.8% of the genome. Twenty-three ORFs overlapped with adjacent ORFs in lengths ranging from 4 to 161 bp, totaling 1103 bp. Of the 137 HearSNPV-C1 ORFs identified, 135 (99.3%) had homologues in HezeSNPV, a possible variant of HearSNPV isolated from *Helicoverpa zea* (Chen et al., 2002).

Of the 137 ORFs identified in HearSNPV-C1, only 27 (19.7%) possess a consensus early promoter motif (TATA box followed by a 20–25 bp downstream CAC/GT motif) within 180 bp of the initiation codon, 53 (38.7%) contain a late promoter motif ((A/T/G) TAAG) within 120 bp of the initiation codon, and 17 (12.4%) have both an early and late promoter motif, which may allow transcription of these genes during both early and late stages of infection, as has been reported for Spli19 (*p10*) and Spli57 (*fp*) (Kool and

Table 1
Comparison of ORFs between HearSNPV isolate C1 and G4

ORF	Name	Position	Length (aa)	Predicted MW (kDa)	C1 compared with G4									
					G4ORF	Nucleotide				Amino acid residue				
						Substitution	Insertion	Deletion	Substitution	Insertion	Deletion	G4 length	Identity %	Note
1	*polyhedrin*	1 > 741	246	28.9	1	5[5]	3	0	5	1	0	245	97.6	d/i
2	*p78/83*	738 < 1982	414	46.1	2	14[8]	3	0	6	1	0	413	98.3	d/i
3	*pk*	1997 > 2800	267	31.6	3	2[1]	0	0	1	0	0	267	99.6	S
4	*hoar*	2923 < 5190	755	85.5	4	53 [15](3)	33	36	34	11	12	756	94.7	d/i
5b		5447 < 5638	63	7.4		2(2)	1	4	–	–	–	–	–	–
6		5733 > 6590	285	34.4	6	10 [9] (1)	0	0(1)	2	0	0	285	98.6	S
7		6790 < 6966	58	6.7	7	0	6	0	0	2	0	56	94.8	d/i
8	*ie-0*	6954 > 7811	285	33.2	8	2 [2]	0	0	0	0	0	285	100	I
9	*p49*	7828 > 9234	468	55.2	9	5[5] (1)	0	0	1	0	0	468	99.8	S
10	*odv-e18*	9245 > 9490	81	8.8	10	2[2]	0	0	0	0	0	81	100	I
11	*odv-ec27*	9505 > 10359	284	33.3	11	9[9]	0	0	0	0	0	284	100	I
12		10404 > 10682	92	10.8	12	1[1]	0	0	0	0	0	92	100	I
13	*ep23*	10709 < 11314	201	22.6	13	8[5]	0	6	3	0	2	203	97.5	T
14	*ie-1*	11356 > 13341	661	76.5	14	18[14]	18	0	4	6	0	655	98.5	d/i
15	*odv-e56*	13395 < 14459	354	38.8	15	8[8]	0	0	0	0	0	354	100	I
16	*me-53*	14620 > 15699	359	42.7	16–17	0	0	1	0	0	0	284	100[a]	E
18		15702 > 15869	55	6.4	18	0	0	0	0	0	0	55	100	I
19		15922 < 16203	93	11.1	19	0	0	0	0	0	0	93	100	I
20	*p74*	16224 > 18290	688	78.4	20	3	0	0	2	0	0	688	99.7	S
21	*p10*	18344 < 18607	87	9.3	21	0	0	0	0	0	0	87	100	I
22	*p26*	18690 < 19493	267	30.5	22	0	0	0	0	0	0	267	100	I
23		19607 > 19810	68	8.3	23	0	0	0	0	0	0	67	100	I
24	*lef6*	19886 < 20449	187	22.2	24	0	0	0	0	0	0	187	100	I
25	*dbp*	20463 < 21434	323	37.6	25	0	0	0	0	0	0	323	100	I
26		21578 > 22054	158	18.1	26	0	0	1	0	0	0	133	100[a]	E
repeat	*hr1*	22129–24097			*hr*1	8	0	106						
27		24225 < 24992	255	29.5	27	1[b]	0	0	1	0	0	255	99.6	S
28	*ubiquitin*	24832 > 25083	83	9.2	28	1[1]	0	0	0	0	0	83	100	I
29		25147 > 25653	168	20.4	29	0	0	0	0	0	0	168	100	I
30	*e125*	25673 > 26251	192	22.6	30	1	6	0	1	2	0	190	98.4	d/i
31	*39K/pp31*	26310 < 27248	312	35.3	31	1	3	0	1	1	0	311	99.3	d/i
32	*lef11*	27214 < 27597	127	14.6	32	0	0	0	0	0	0	127	100	I
33	*bv-e31*	27566 < 28282	238	28.4	33	0	0	0	0	0	0	238	100	I
34		28513 > 29592	359	41.2	34	1	0	0	1	0	0	359	99.7	S
35	*p47*	29667 < 30905	412	48.1	35	0	1	0	0	0	0	333	100[a]	E
36	*lef12*	30978 > 31649	223	25.8	36	0	0	0	0	0	0	223	100	I
37		31735 > 31977	80	9.5	37	0	0	0	0	0	0	80	100	I
38	*lef8*	31974 < 34679	901	104.9	38	2[2]	0	0	0	0	0	901	100	I
39		34732 > 35316	194	22.5	39	0	0	0	0	0	0	194	100	I
40		35457 > 35609	50	6.3	40	0	0	0	0	0	0	50	100	I
41	*chitinase*	35617 < 37329	570	65.5	41	0	0	0	0	0	0	570	100	I
42		37408 < 37950	180	21.3	42	0	0	0	0	0	0	180	100	I
43		38067 > 38477	136	16.4	43	1(1)	0	0	1	0	0	136	99.3	S

44		38484 < 39620	378	42.8	44	3[3]	0	0	0	0	0	378	100	I
45		39628 < 39855	75	9.1	45	0	0	0	0	0	0	75	100	I
46	*lef10*	39815 > 40030	71	7.7	46	1[b]	0	0	1	0	0	71	98.6	S
47	*vp1054*	39903 > 40958	351	41.7	47	4[1]	0	0	3	0	0	351	99.1	S
48		41078 > 41284	68	8.0	48	1	0	0	0	0	0	68	100	I
49		41285 > 41479	64	7.4	49	0(1)	0	0	0	0	0	64	100	I
50		41765 > 42280	171	20.7	50	3[2]	0	0	1	0	0	171	99.4	S
51	*8.2kDa*	42331 < 42810	159	18.9	51	3	0	3	3	0	1	160	97.5	d/I
52		42822 < 43088	88	10.2	52	0	0	0(1)	0	0	0	88	100	I
53	*fp 25K*	43300 < 43953	217	25.4	53	0	0	0	0	0	0	217	100	I
54		44125 > 44310	61	7.3	54	0	0	0(2)	0	0	0	61	100	I
55	*lef9*	44420 > 45979	519	59.6	55	1	0	0	1	0	0	519	99.8	S
56	*cathepsin*	46063 < 47160	366	42.4	56	0	0	0	0	0	0	365	100	I
57		47201 < 47788	195	21.3	57	0	0	0	0	0	0	195	100	I
58	*gp37*	47859 < 48698	279	32.1	58	0	0	0	0	0	0	279	100	I
repeat	*hr2*	48813–49719			*hr*2	2(1)	1	0						
59	*bro-a*	49850 > 50584	244	28.3	59	1(14)	0	0	1	0	0	244	99.6	S
60	*bro-b*	50708 > 51781	357	40.2	60	70 [13] (1)	21	531	34	7	177	527	59.6	d/i
repeat	*hr3*	51949–52245			*hr*3	3(2)	0	0(1)						
61	*he65*	52536 > 53246	236	27.5	61	0	0	0(1)	0	0	0	236	100	I
62	*iap-2*	53322 < 54074	250	29.2	62	7[7]	0	0	0	0	0	250	100	I
63		54122 < 54946	274	31.6	63	0	0	0	0	0	0	274	100	I
64		54915 < 55316	133	15.6	64	0	0	0	0	0	0	133	100	I
65	*lef3*	55336 > 56475	379	44.0	65	7[5]	0	0	2	0	0	379	99.5	S
66	*93 kDa*	56584 < 58940	785	88.9	66	2[2]	0	0	0	0	0	785	100	I
67	*DNA pol*	58971 > 62033	1020	119.2	67	11[11]	0	0	0	0	0	1020	100	I
68	*30.5 kDa*	62114 < 62572	152	17.6	68	0	0	0	0	0	0	152	100	I
69		62634 < 63017	127	14.9	69	0	0	0	0	0	0	127	100	I
70		63023 < 63280	85	10.0	70	0	0	0	0	0	0	85	100	I
71	*vlf-1*	63321 < 64562	413	48.0	71	0	3	0	0	1	0	412	99.8	d/i
72		64575 < 64907	110	12.7	72	0	0	0	0	0	0	110	100	I
73	*gp41*	64976 < 65944	322	36.6	73	2[2]	0	0	0	0	0	322	100	I
74		65874 < 66599	241	27.7	74	1[1]	0	0	0	0	0	241	100	I
75		66472 < 67149	225	24.9	75	0	0	0	0	0	0	225	100	I
76	*vp91capsid*	67079 > 69529	816	93.5	76	0	0	0	0	0	0	816	100	I
77	*cg30*	69657 < 70508	283	32.3	77	1[1]	0	0	0	0	0	283	100	I
78	*p39*	70597 < 71478	293	33.4	78	0	0	0	0	0	0	293	100	I
79	*lef4*	71477 > 72862	461	50.0	79	1	0	0	1	0	0	461	99.8	S
80	*p33*	72915 < 73679	254	30.8	80	0	0	0	0	0	0	254	100	I
81		73681 > 74169	162	19.1	81	0	0	0	0	0	0	162	100	I
82	*odv-e25*	74215 > 74907	230	25.9	82	0	0	0	0	0	0	230	100	I
83		74939 < 75436	165	18.8	83	3	0	0	3	0	0	165	98.2	S
84	*helicase*	75455 < 79216	1253	146.0	84	8[3]	0	0	4	0	0	1253	99.7	S
85		79173 > 79694	173	19.8	85	1	0	0	1	0	0	173	99.4	S
86		79753 < 80718	321	37.9	86	0	0	0	0	0	0	321	100	I
87	*lef5*	80614 > 81561	315	37.0	87	0	0	0	0	0	0	315	100	I
88	*p6.9*	81555 < 81884	109	11.5	88	0	0	0	0	0	0	109	100	I
89		81949 < 83058	369	42.6	89	1	0	0	0	0	0	369	100	I
90	*13.1kDa*	83104 > 83472	122	13.8	90	2	0	0	0	0	0	122	100	I

(continued on next page)

Table 1 (*continued*)

ORF	Name	Position	Length (aa)	Predicted MW (kDa)	C1 compared with G4									
					G4ORF	Nucleotide				Amino acid residue				
						Substitution	Insertion	Deletion	Substitution	Insertion	Deletion	G4 length	Identity %	Note
91		83472 < 84605	377	44.0	91	6[5] (1)	0(1)	0	1	0	0	377	99.7	S
92	*vp80/p87*	84701 > 86518	605	69.7	92	13[3]	0	0	10	0	0	605	98.3	S
93		86515 > 86691	58	6.9	93	0	0	0	0	0	0	58	100	I
94	*odv-ec43*	86706 > 87791	361	41.5	94	0	0	0	0	0	0	361	100	I
95		87837 > 88121	94	11.0	95	0(1)	0	0	0	0	0	94	100	I
96	*odv-e66*	88188 < 90206	672	76.0	96	8[6]	0	0	2	0	0	672	99.7	S
97	*p13+*	90227 < 91057	276	32.5	97	0	0	0	0	0	0	276	100	I
repeat	*hr4*	91222–93474			*hr4*	4	550	0						
97a		93485 < 93652	55	6.4		0	0	0	0	0	0	55	100	I
98		93913 > 94512	199	22.4	98	0	0	0	0	0	0	199	100	I
99		94516 > 94872	118	14.4	99	0	0	0	0	0	0	118	100	I
100		94968 > 96500	510	58.1	100	0	0	0	0	0	0	510	100	I
101		96579 > 97340	253	29.0	101	0	0	0	0	0	0	253	100	I
102		97355 > 97687	110	12.8	102	0	0	0	0	0	0	110	100	I
103	*iap-3*	97745 < 98551	268	31.5	103	0	0	0	0	0	0	268	100	I
104		98548 < 98703	51	5.9	104	0	0	0	0	0	0	100	100	I
105	*bro-c*	98814 < 100319	501	58.3	105	4[2]	0	0	2	0	0	501	99.6	S
106		100487 > 100966	159	16.8	106	2	0	0	1	0	0	159	99.4	S
107	*sod*	100973 > 102346	457	51.2	107	5[2]	0	0	2	0	0	457	99.6	S
108		102399 < 102977	192	22.7	108	1	0(3)	0	1	0	0	192	99.5	S
109		103149 > 103505	118	13.6	109	2[2]	0	0	0	0	0	118	100	I
110		103516 > 103782	88	10.1	110	0	0	0	0	0	0	88	100	I
111	*pif*	103850 > 105436	528	60.3	111	2[2]	0	0	0	0	0	528	100	I
112		105433 > 105669	78	9.1	112	0	0	0	0	0	0	78	100	I
113	*fgf*	105692 < 106597	301	34.3	113	7[5] (7)	0	0	2	0	0	301	99.3	S
114	*alk-exo*	106724 < 108010	428	49.4	114	2[2]	0	0	0	0	0	428	100	I
115		108030 < 108419	129	15.2	115	3[1] (2)	0	0(1)	2	0	0	129	98.4	S
repeat	*hr5*	108500–109398			*hr5*	37(2)	43	648						
115a		109493 < 110419	308	37.1		6[5]	0	0(2)	1	0	0	308	99.7	S
116		110618 > 110833	71	8.2	116	2	0	0	2	0	0	71	97.2	S
117	*lef2*	110951 < 111679	242	27.9	117	4[4]	3[b]	0	0	1	0	241	99.6	d/i
117a		111543 < 111950	135	15.7		0	3	0	0	1	0	134	99.2	d/i
118	*p24capsid*	112041 > 112787	248	28.3	118	1	0	0	1	0	0	248	99.6	S
119	*gp16*	112849 > 113139	96	10.9	119	4[4]	6	0	0	2	0	94	97.9	d/i
120	*calyx/pep*	113191 > 114213	340	39.1	120	0	0	0	0	0	0	340	100	I
121		114292 > 114756	154	18.5	121	0(3)	0	0(1)	0	0	0	154	100	I
122	*odv-c21*	114886 > 115476	196	23.4	122	3[1] (1)	0	0	2	0	0	196	99.0	S
123	*38.7kd*	115520 < 116689	389	44.9	123	3[2]	0	1	1	0	0	385	98.7	E
124	*lef1*	116691 < 117428	245	29.0	124	2[1]	0	0	1	0	0	245	99.6	S
125		117403 < 117831	142	16.0	125	5[2] (3)	0(1)	6(1)	3	0	2	144	96.5	d/i
126	*egt*	117976 > 119523	515	58.9	126	6[2]	0	0	3	0	0	515	99.4	S
127		119723 > 120301	192	22.6	127	2[1]	0	0	1	0	0	192	99.5	S

128		120252 > 121052	266	30.4	128	0	0	0	0	0	100	I
129		121133 < 123976	947	111.4	129	7[2]	0	5	0	0	99.5	S
130	*pkip-1*	124341 > 124850	169	20.3	130	5[1] (1)	0	4	0	0	97.6	S
131	*arif1*	124917 < 125714	265	30.3	131	6[3]	0(1)	3	0	0	98.9	S
132	*pif-2*	125972 > 127123	383	44.5	132	12[8]	0	3	0	0	99.2	S
133	*f-protein*	127164 < 129197	677	78.2	133	32[26] (2)	0	6	0	0	99.1	S
134		129339 < 129884	181	21.9	134	1	0	1	0	0	99.4	S
135		130066 > 130653	195	23.5	135	10[3] (1)	3	5	1	0	96.9	d/i

Note. T: truncated; S: similar; I: completely identical; d/i: deletion or insertion; E: extended. Figures in square brackets indicate silent mutations. Figures in parentheses indicate the nucleotide changes occurred in intergenic (IG) regions.

[a] Only the truncated regions were compared for ORF16, 26 and 35.

[b] Insertions or deletions in the overlapped regions of two neighboring ORFs.

Vlak, 1993; Pang et al., 2001). Thirty-nine ORFs lack any recognized consensus early or late promoter motif within 180 bp of the ATG.

Comparison of ORFs between HearSNPV-C1 and HearSNPV-G4

The genome of HearSNPV-G4 was reported 131,403 bp long and contained 135 ORFs (Chen et al., 2001). All the 135 ORFs identified in G4 were also found in C1. For comparing and to avoid confusion of ORF names in related studies, we updated the C1 genomic sequence data in GenBank (accession number AF303045). The 133 homologous ORFs in C1 were numbered similar to those of G4. ORF16/17 in G4 was found to be a single ORF (Chen et al., 2002), which was homologous to ORF16 (*me*53) in C1; thus, the ORF17 was no longer used as an ORF name for C1 in the updated sequence. Three C1 ORFs which were also present in G4 genome but not identified earlier by Chen et al. (2001) were named as ORF97a, ORF115a, and ORF117a. The C1-ORF115a homologue was designated as G4-115a (Chen et al., 2002). The ORF5b of C1 and ORF5 of G4 appeared to be different because different reading frames were used for ORF identification. Thus, G4 isolate had a total number of 137 ORFs, including the earlier reported 134 ORFs (ORF17 was not included) (Chen et al., 2001) and the three newly described ORFs, ORF97a, ORF115a, and ORF117a, which is equal to the ORF number of C1 isolate.

ORF115a was 308 aa in length and homologous to SpliNPV ORF70. The putative protein of ORF117a was 135 aa long and characterized by the presence of ten continuous asparaginate residues. The Blast search showed that ORF117a protein shared 38% identity to AdhoNPV ORF101 and 28% to MacoNPV ORF13. It also had 28% identity to gp160 protein from human immunodeficiency virus 1. No homologue of ORF97a was found in the database.

Single nucleotide polymorphisms

The HearSNPV-C1 genome is 644 bp smaller than that of G4. The two genomes shared 98.1% identity in nucleotide sequence, with a total number of 555 bp substitutions, 1354 bp deletions, and 710 bp insertions in HearSNPV-C1. Among 555 bp substitutions, 450 bp substitutions occurred in the coding regions of 72 ORFs and resulted in a total number of 183 aa residue changes in 51 ORFs. Two hundred and thirty-one basepairs were silent among the 450-bp substitutions in ORFs (Table 1). The G + C content of the two genomes was very close, 38.9% and 39.0%, respectively. Of 137 HearSNPV-C1 ORFs, 77 (56.2%) were completely identical to the corresponding G4 ORFs. Thirty-nine ORFs (28.5%) had the same aa number and were very similar with 98–99.8% aa identity, and 15 ORFs contained insertions and/or deletions from 1 to

177 aa residues. Among these 15 ORFs, ORF1, 2, 31, 51, 71, 117, and 135 had only a single aa change in length, ORF4 (Hoar) had 11 residue insertions and 12 residue deletions, the major change was Bro-b, with 177 aa insertion and 7 aa deletion. ORF13 contained a truncation of 2 aa residues. ORF26, 35 and 123 had extensions of 4 to 79 aa residues. ORF26 is an unknown protein, with 23% aa identity to reverse transcriptase/envelope protein of simian T-cell lymphotropic virus type 1. A "T" insertion in the HearSNPV-G4 ORF26 before the start codon led to a later translational initiation. ORF35 encodes P47, which was a putative late expression factor and transcription regulator (Lapointe et al., 2000; Lu and Miller, 1995). A single nucleotide "A" deletion of ORF35 in HearSNPV-G4 resulted in frame shifts that induced a premature stop codon, so HearSNPV-G4 ORF35 was truncated by 79 codons. When we compared HearSNPV-C1 ORF26 and 35 to the homologues of HearSNPV-G4 and other NPVs, such as AcMNPV, LdMNPV, SeMNPV, SpliMNPV, and MacoMNPV, we found that ORF26 and 35 were obviously truncated in HearSNPV-G4, which was the consequence of an insertion or deletion in the continuous AAAAA or TTTT region, causing the shifted reading frame. Furthermore, HezeSNPV was of same size like HearSNPV-C1 in these two ORFs. It is reasonable to think that these differences may be caused by sequencing errors in HearSNPV-G4, a similar case in G4-ORF16/17.

The major differences between C1 and G4 were in the *hr*1, *hr*4, *hr*5, and *bro-b* genes, both in sequence and in length (Table 1, Figs. 1 and 2).

ORF5b

HearSNPV-G4 ORF5 is 180 bp in length, which is also present in C1, with 96.6% aa identity, but it 119 bp overlaps with ORF5b in C1. So ORF5 was not listed in Table 1. The C1 ORF5b was an interesting ORF encoding 63 aa, which had the homologues of Ac152 (92 aa), Spli5 (67 aa), and Maco8 (67 aa) in AcMNPV, SpliMNPV, and MacoMNPV with aa identity of 50%, 54%, and 53%, respectively. This ORF was also present in G4 at nucleotide level, with 96.4% identity, but was apparently not recognized by Chen et al. (2001) due to frame shifts, for there are 4 bp insertion in the position of 40 bp downstream from the start site in the corresponding regions of the G4 genome, which led to a premature stop codon. It was interesting to note that the complete ORF5b was also not present in HezeSNPV (Chen et al., 2002). In order to clarify if the frame shift mutations were caused by sequencing errors, the C1 ORF5b region was further PCR amplified and sequenced. The result confirmed that our ORF5b sequence was correct. There is a need to further explore whether ORF5b is a functional gene.

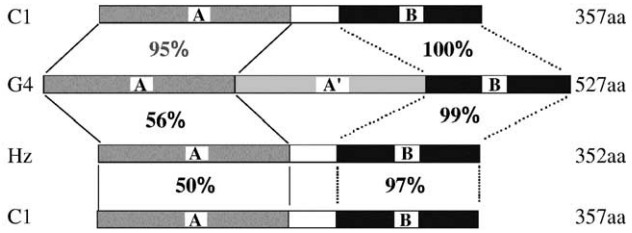

Fig. 2. Comparison of the *bro-b* gene from HearSNPV-C1, HearSNPV-G4, and HezeSNPV. Identities shown in percent are based on the aa sequences. A and A′ represent homologous repeat regions. The length of each ORF is indicated on the right-hand side.

Homologous repeat (hr) regions

A common feature of all the NPV genomes is the presence of homologous repeat (*hr*) regions that are located along the genome. They have been shown to serve as origins of DNA replication in transient assays (Possee and Rohrmann, 1997), enhancers of RNA polymerase II-mediated transcription of baculovirus early promoters (Guarino and Summers, 1986; Theilmann and Stewart, 1992) and sites of frequent recombinant and rearrangement

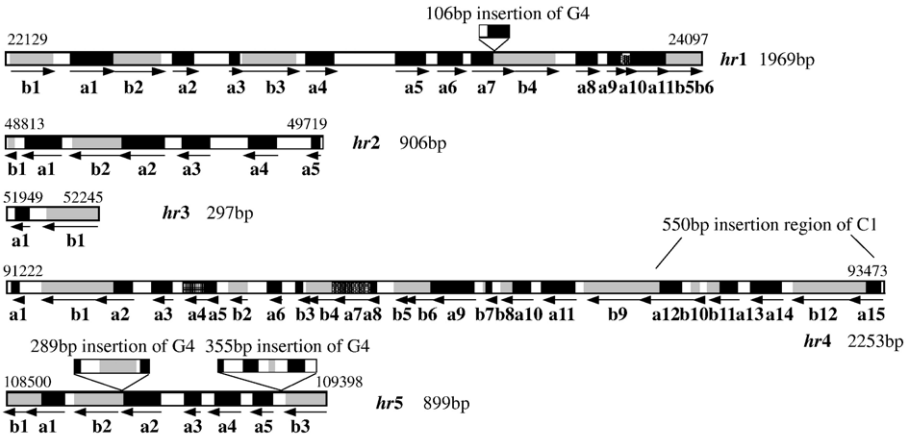

Fig. 1. Comparison of the *hr* regions between HearSNPV-C1 and HearSNPV-G4. Arrows that represent the direction indicate the positions of the repeat-A and repeat-B region. Shaded box represent type B repeat and black box represent type A repeat.

in baculovirus genomes (Chen et al., 2002; Harrison and Bonning, 2003; Hayakawa et al., 2000; Hyink et al., 2002). The HearSNPV-C1 genome contained five *hr* regions that were dispersed throughout the C1 genome with sizes of 1969, 906, 297, 2252, and 899 bp, respectively. The five *hr*s occupy the positions 22.1 kb (*hr*1), 48.8 kb (*hr*2), 51.9 kb (*hr*3), 91.2 kb (*hr*4), and 108.5 kb (*hr*5) on the genome. The *hr*s included two types of repeats, type A and type B, consisting of 64 and 147 bp, respectively, or truncated versions thereof (Fig. 1). The functions of these repeats remain to be determined. Both type A and B repeats were found in each of the *hr*s in HearSNPV-C1. The five *hr*s of C1 isolate were located at the similar positions in the genome as those of G4 isolate. Two of the five *hr*s, *hr*2 and *hr*3, were the most stable *hr*s with 99.6% and 99.1% sequence identity to that of HearSNPV-G4. Three other *hr*s (*hr*1, *hr*4, and *hr*5) exhibited high variability with different numbers of the repeat units in each *hr* region (Fig. 1). Sequence alignment between C1 and G4 *hr*s indicated that these three homologous regions had a distinctly lower nucleotide identity (82.6–90.2%), with a few insertions/deletions of different sizes. *hr*1 contained a 106-bp deletion (only a type A repeat) in C1 compared with isolate G4. *hr*4 contained the biggest insertion in C1 genome, an insertion of 550 bp that contained both type A and type B repeats. *hr*5 contained two bigger deletions (289 and 355 bp) that also contained type A and type B repeats and a small insertion of 40 bp (Fig. 1), as well as three single base pair insertions and four single base pair deletions, remarkably different from other *hr*s. The sequence data revealed that *hr*1, *hr*4, and *hr*5 are three highly variable genomic regions located in AT-rich intergenic regions. The comparison of C1 and G4 *hr*s suggested that *hr*s were possibly associated with the mechanisms of recombination.

The major difference occurred not only in the organization of homologous regions of different isolates, but also in the *bro-b* gene, which separated *hr*2 and *hr*3 with *bro-a* gene in both C1 and G4 genome.

bro-b gene

The occurrence of baculovirus repeat ORF (*bro*) gene family was a striking feature in many baculovirus genomes (Chen et al., 2001; Gomi et al., 1999; Hayakawa et al., 1999; Ijkel et al., 1999; Kuzio et al., 1999). There were two specific regions in HearSNPV-C1 *bro-b* ORF, named A region (182aa) and B region (134aa) (Fig. 2). The HearSNPV-G4 *bro-b* gene had an A region with an aa identity of 95% to C1 *bro-b* A and a B region with 100% identity to C1 *bro-b* B. Furthermore, it contained an additional homologous repeat region of 177 aa residues named A′, which had low aa identity (48%) to C1 A and 51% aa identity to HezeSNPV A region. The A region of HezeSNPV genome shared sequence identity of 50% to C1 A, 56% to G4 A, and 51% to G4 A′ region. We could see the identity of A region between C1 and G4 genome was much higher than that between HezeSNPV and HearSNPV-C1 or G4. The B region of HearSNPV-C1 shared 97% sequence identity to that of HezeSNPV. HezeSNPV B region had sequence identity of 99% to that of HearSNPV-G4 B. This suggested that *bro-b* B regions were the highly conserved portions in these three genomes. There is a low degree of identity in A region between HezeSNPV and HearSNPV-C1, HezeSNPV, and HearSNPV-G4 as seen by comparing the three *bro-b* genes. The occurrence of additional homologous region A′ in HearSNPV-G4 differs significantly from HearSNPV-C1 A, which suggested that they may have been acquired independently in the ancestral past.

Our sequence data suggested that HearSNPV-C1 was closely related to HearSNPV-G4 but also with differentiation in some degree, even though they both infected the same host species, *H. armigera*, in the same geographic location. The results also support the opinion that baculovirus strain heterogeneity is often caused by SNPs and changes in the *hr*s and *bro* genes. Comparison of the HearSNPV-C1 and HearSNPV-G4 genome showed that there were four highly variable regions *hr*1, *hr*4, *hr*5, and *bro-b*, all in repeat regions. Comparison of BV production in vitro between C1 and G4 showed that C1 replicated more quickly than G4 at the first 2 days after infection, though they both reached to a final titer of about 8.0×10^8 TCID50/ml. When HzAM1 cells were infected with viruses at an MOI of 5 $TCID_{50}$ units per cell, the virus titers in the culture media increased from 5.15×10^6 at beginning of infection to $9.55 \pm 3.4 \times 10^7$ and $4.47 \pm 0.93 \times 10^8$ TCID50/ml at 24 and 48 h.p.i for C1, and to $4.78 \pm 0.68 \times 10^7$ and $1.62 \pm 0.5 \times 10^8$ TCID50/ml at the same time points for G4, respectively. Since *hr*s are involved in regulatory processes (enhancing transcription, DNA replication), recombination or rearrangement in these regions may be responsible for the strain difference in replication, which may further be responsible for variation in virulence in the strains. However, further studies are needed to determine which of *hr*s and *bro-b* or single mutations is responsible for strain virulence.

Materials and methods

Virus

The HearSNPV-C1 was originally isolated from Hubei Province of China and was further plaque-purified and maintained in the *H. zea* cell line HzAM1. The occlusions were purified from the infected *H. armigera* larvae by sucrose-gradient centrifugation (O'Reilly et al., 1992).

DNA extraction, cloning, and sequence determination

The viral DNA was isolated from purified occlusions by using alkaline treatment. The viral DNA was prepared as described (O'Reilly et al., 1992). The purified genomic DNA was sheared by ultrasonication into fragments with

sizes of 1–1.5 kb. The ends of the random fragments were repaired with the large fragment of T4 DNA polymerase (Klenow) according to the manufacturer's protocol and cloned into pUC19. The ligation products were transformed into *Escherichia coli* JM109. DNA templates for sequencing were prepared from over 2000 clones. Sequencing was performed using the ABI PRISM TM 3700 DNA Analyser and Big Dye Terminator Cycle Sequencing Ready Reaction Kit (Perkin Elmer). The combined sequence generated from these clones represented an eightfold genomic coverage.

Sequence analysis

Genomic DNA composition and ORFs were analyzed with Wisconsin Genetic Computer Group program and Genetyx-win (Software Development Co. Ltd, Japan). ORFs encoding more than 50 amino acids (aa) were designated putative genes. Relevant ORFs were checked for maximum alignment with known baculovirus gene homologues from GenBank. DNA and protein comparisons with entries in the GenBank were performed with BLAST and Genetyx program. Homologous repeat regions were detected using the Search Direct, Inverted and complementary Repeat programs of Genetyx-win (parameters: minilength 20, maxilength 160, matching percentage of sites 75%), and further analyzed by directly searching the genomic sequence with the following repeat core sequences as described in G4 by Chen et al. (2001). Repeat-A sequence: tttaaaccggtcttggatcttttcgttcgaaacgggccgtgatcttttgtttcgactcgtgacc; repeat-B sequence aaaaaacaaattacgtcatcgacatagaatattgcatcatttttaaattcgaaactagcccgctttcatatgaaaccct-cggcgaagatcgattatatttgttctagaacattcgacggcttgacccaaaaaaacaaatgacgtcat.

Nucleotide sequences for each ORF and intergenic region, and amino acid sequences for each putative protein from C1 were compared with corresponding regions of G4, using Genetyx-win software.

Acknowledgments

We are grateful to Professor Yu ZH at Normal University of Central China for providing the virus C1 occlusion bodies and Dr. Shyam Kumar at Karnatak University, India, for helpful suggestions. This work was supported by the grants from the Foundation for the Author of National Excellent Doctoral Dissertation of P. R. China (Project no. 200055) and the National Natural Science Foundation of China (Project No. 30070032 and 39770031).

References

Adams, J.R., McClintock, J.T., 1991. Nuclear polyhedrosis viruses of insects. In: Adams, J.R., Bonami, J.R. (Eds.), Atlas of Invertebrate Viruses. CRC Press, Boca Raton, pp. 87–204.

Afonso, C.L., Tulman, E.R., Lu, Z., Balinsky, C.A., Moser, B.A., Becnel, J.J., Rock, D.L., Kutish, G.F., 2001. Genome sequence of a baculovirus pathogenic for *Culex nigripalpus*. J. Virol. 75, 11157–11165.

Ahrens, C.H., Russell, R.L.Q., Funk, C.J., Evans, J.T., Harwood, S.H., Rohrmann, G.F., 1997. The sequence of the *Orgyia pseudotsugata* multicapsid nuclear polyhedrosis virus genome. Virology 229, 381–399.

Ayres, M.D., Howard, S.C., Kuzio, J., Lopez-Ferber, M., Possee, R.D., 1994. The complete DNA sequence of *Autographa californica* nuclear polyhedrosis virus. Virology 202, 586–605.

Blissard, G., Black, B., Crook, N., Keddie, B.A., Possee, R., Rohrmann, G., Theilmann, D., Volkman, L., 2000. Family Baculoviridae. In: van Regenmortel M.H.V., et al. (Eds.), Virus Taxonomy. Seventh Report of the International Committee on Taxonomy of Viruses. Academic Press, San Diego, pp. 195–202.

Chen, X., Ljkel, W.F.J., Tarchini, R., et al., 2001. The sequence of the *Helicoverpa armigera* single nucleocapsid nucleopolyhedrovirus genome. J. Gen. Virol. 82, 241–257.

Chen, X., Zhang, W.J., Wong, J., et al., 2002. Comparative analysis of the complete genome sequences of *Helicoverpa zea* and *Helicoverpa armigera* single nucleocapsid nucleopolyhedroviruses. J. Gen. Virol. 83, 673–684.

Garcia-Maruniak, A., Maruniak, J.E., Zanotto, P.M.A., Doumbouya, A.E., Liu, J.C., Merritt, T.M., Lanoie, J.S., 2004. Sequence analysis of the genome of the *Neodiprion sertifer* nucleopolyhedrovirus. J. Virol. 78, 7036–7051.

Gomi, S., Majima, K., Maeda, S., 1999. Sequence analysis of the genome of *Bombyx mori* nucleopolyhedrovirus. J. Gen. Virol. 80, 1323–1337.

Guarino, L.A., Summers, M.D., 1986. Functional mapping of a trans-activating gene required for expression of a baculovirus delayed-early gene. J. Virol. 57, 563–571.

Harrison, R.L., Bonning, B.C., 2003. Comparative analysis of the genomes of *Rachiplusia ou* and *Autographa californica* multiple nucleopolyhedroviruses. J. Gen. Virol. 84, 1827–1842.

Hashimoto, Y., Hayakawa, T., Ueno, Y., Fujita, T., Sano, Y., Matsumoto, T., 2000. Sequence analysis of the Plutella xylostella granulovirus genome. Virology 275, 358–372.

Hayakawa, T., Ko, R., Okano, K., Seong, S.I., Goto, C., Maeda, S., 1999. Sequence analysis of the Xestia c-nigrum granulovirus genome. Virology 262, 277–297.

Hayakawa, T., Rohrmann, G.F., Hashimoto, Y., 2000. Patterns of genome organization and content in lepidopteran baculoviruses. Virology 278, 1–12.

Huser, A., Hofmann, C., 2003. Baculovirus vector: novel mammalian cell gene delivery vehicle and their applications. Am. J. Pharmacogenomics 3, 53–63.

Hyink, O., Dellow, R.A., Olsen, M.J., Caradoc-Davies, K.M., Drake, K., Herniou, E.A., Cory, J., O'Reilly, D.R., Ward, V.K., 2002. Whole genome analysis of the *Epiphyas postvittana* nucleopolyhedrovirus. J. Gen. Virol. 83, 957–971.

Ijkel, W.F.J., van Strien, E.A., Heldens, J.G.M., Broer, R., Zuidema, D., Golbach, R.W., Vlak, J.M., 1999. Sequence and organization of the *Spodoptera exigua* multinucleocapsid nucleopolyhedrovirus genome. J. Gen. Virol. 80, 3289–3304.

Jia, F., Xu, J., Nie, Y.C., Xiao, Y., Zhou, R., Li, X., Yu, Z.H., 2003. Comparison of *Heliothis armigera* NPV VHA273 wild isolate with its clone. Virol. Sin. 18, 241–245.

Kool, M., Vlak, J.M., 1993. The structural and functional organization of the *Autographa californica* nuclear polyhedrosis virus genome. Arch. Virol. 130, 1–16.

Kuzio, J., Pearson, M.N., Harwood, S.H., Funk, C.J., Evans, J.T., Slavicek, J.M., Rohrmann, G.F., 1999. Sequence and analysis of the genome of a baculovirus pathogenic for *Lymantria dispar*. Virology 253, 17–34.

Lange, M., Jehle, J.A., 2003. The genome of the *Cryptophlebia leucotreta* granulovirus. Virology 317, 220–236.

Lapointe, R., Back, D.W., Ding, Q., Carstens, E.B., 2000. Identification and molecular characterization of the *Choristoneura fumiferana* multicapsid nucleopolyhedrovirus genomic region encoding the regulatory genes pkip, p47, lef-12, and gta. Virology 271, 109–121.

Lauzon, H.A.M., Lucarotti, C.J., Krell, P.J., Feng, Q., Retnakaran, A., Arif, B.M., 2004. Sequence and organization of the *Neodiprion lecontei* nucleopolyhedrovirus genome. J. Virol. 78, 7023–7035.

Li, L., Donly, C., Li, Q., Willis, L.G., Keddie, B.A., Erlandson, M.A., Theilmann, D.A., 2002. Identification and genomic analysis of a second species of nucleopolyhedrovirus isolated from *Mamestra configurata*. Virology 297, 226–244.

Li, Q., Donly, C., Li, L., Willis, L.G., Theilmann, D.A., Erlandson, M., 2002. Sequence and organization of the *Mamestra configurata* nucleopolyhedrovirus genome. Virology 294, 106–121.

Lopez Ferber, M., Argaud, O., Croizier, L., Croizier, G., 2001. Diversity, distribution, and mobility of bro gene sequences in *Bombyx mori* nucleopolyhedrovirus. Virus Genes 22, 247–254.

Lu, A., Miller, L.K., 1995. The roles of eighteen baculovirus late expression factor genes in transcription and DNA replication. J. Virol. 69, 975–982.

Luque, T., Finch, R., Crook, N., O'Reilly, D.R., Winstanley, D., 2001. The complete sequence of the *Cydia pomonella* granulovirus genome. J. Gen. Virol. 82, 2531–2547.

Murphy, F.A., Fauquet, C.M., Biship, D.H.L., Ghabrial, S.A., Jarvis, A.W., Martelli, G.P., Mayo, M.A., Summers, M.D., 1995. Virus taxonomy: the classification and nomenclature of viruses. Sixth Report of the International Committee on Taxonomy of Viruses. Springer-Verlag, Vienna, pp. 104–113.

Nakai, M., Goto, C., Kang, W., Shikata, M., Luque, T., Kunimi, Y., 2003. Genome sequence and organization of a nucleopolyhedrovirus isolated from the smaller tea tortrix, *Adoxophyes honmai*. Virology 316, 171–183.

O'Reilly, D.R., Miller, L.K., Lucknow, V.A., 1992. Baculovirus Expression Vectors: A Laboratory Manual. Freeman WH and Company, New York.

Pang, Y., Yu, J.X., Wang, L.H., et al., 2001. Sequence analysis of the *Spodoptera litura* multicapsid nucleopolyhedrovirus genome. Virology 287, 391–404.

Possee, R.D., Rohrmann, G.F., 1997. Baculovirus genome organization and evolution. In: Miller, L.K. (Ed.), The Baculoviruses. Plenum Press, New York, N.Y., pp. 109–140.

Smith, G.E., Summers, M.D., Fraser, M., 1983. Production of human belta interferon in insect cells with a baculovirus expression vector. J. Mol. Cell Biol. 3, 2156–2165.

Stewart, L.M.D., Hirst, M., Ferber, M.L., Merryweather, A.T., Cayley, P.J., Possee, R.D., 1991. Construction of an improved baculovirus insecticide containing an insect-specific toxin gene. Nature 352, 85–88.

Sun, X.L., Zhang, G.Y., 1994. Comparison of four *Haliothis armigera* NPV isolates. Virol. Sin. 12, 309–318.

Tani, H., Limn, C.K., Yap, C.C., et al., 2003. In vitro and in vivo gene delivery by recombinant baculoviruses. J. Virol. 77, 9799–9808.

Theilmann, D.A., Stewart, S., 1992. Tandemly repeated sequence at the 39 end of the ie-2 gene of the baculovirus *Orgyia pseudotsugata* multicapsid nuclear polyhedrosis virus is an enhancer element. Virology 187, 97–106.

Vlak, J.M., Smith, G.E., 1982. Orientation of the genome of *Autographa californica* nuclear polyhedrosis virus: a proposal. J. Virol. 41, 1118–1121.

Wormleaton, S., Kuzio, J., Winstanley, D., 2003. The complete sequence of the *Adoxophyes orana* granulovirus genome. Virology 311, 350–365.

Zhang, G., 1994. Research, development and application of Heliothis viral pesticide in China. Resour. Environ. Yangtze Valley 3, 1–6.

Zhang, C.X., Wu, J.C., 2001. Genome structure and the p10 gene of the *Helicoverpa armigera* nucleopolyhedrovirus. Acta Biochim. Biophys. Sin. 33, 179–184.