Journal of General Virology (2001), **82**, 241–257. Printed in Great Britain

The sequence of the *Helicoverpa armigera* single nucleocapsid nucleopolyhedrovirus genome

Xinwen Chen,[1,2] Wilfred F. J. IJkel,[2] Renato Tarchini,[3] Xiulian Sun,[1] Hans Sandbrink,[3] Hualin Wang,[1] Sander Peters,[3] Douwe Zuidema,[2] René Klein Lankhorst,[3] Just M. Vlak[2] and Zhihong Hu[1]

[1] Joint-Laboratory of Invertebrate Virology, Institute of Virology, Chinese Academy of Sciences, Wuhan, Hubei 430071, People's Republic of China

[2] Laboratory of Virology, Wageningen University, Binnenhaven 11, 6709 PD Wageningen, The Netherlands

[3] Greenomics, Plant Research International, PO Box 16, 6700 AA Wageningen, The Netherlands

The nucleotide sequence of the *Helicoverpa armigera* single-nucleocapsid nucleopolyhedrovirus (HaSNPV) DNA genome was determined and analysed. The circular genome encompasses 131 403 bp, has a G+C content of 39·1 mol% and contains five homologous regions with a unique pattern of repeats. Computer-assisted analysis revealed 135 putative ORFs of 150 nt or larger; 100 ORFs have homologues in *Autographa californica* multicapsid NPV (AcMNPV) and a further 15 ORFs have homologues in other baculoviruses such as *Lymantria dispar* MNPV (LdMNPV), *Spodoptera exigua* MNPV (SeMNPV) and *Xestia c-nigrum* granulovirus (XcGV). Twenty ORFs are unique to HaSNPV without homologues in GenBank. Among the six previously sequenced baculoviruses, AcMNPV, *Bombyx mori* NPV (BmNPV), *Orgyia pseudotsugata* MNPV (OpMNPV), SeMNPV, LdMNPV and XcGV, 65 ORFs are conserved and hence are considered as core baculovirus genes. The mean overall amino acid identity of HaSNPV ORFs was the highest with SeMNPV and LdMNPV homologues. Other than three 'baculovirus repeat ORFs' (*bro*) and two 'inhibitor of apoptosis' (*iap*) genes, no duplicated ORFs were found. A putative ORF showing similarity to poly(ADP-ribose) glycohydrolases (*parg*) was newly identified. The HaSNPV genome lacks a homologue of the major budded virus (BV) glycoprotein gene, *gp64*, of AcMNPV, BmNPV and OpMNPV. Instead, a homologue of SeMNPV ORF8, encoding the major BV envelope protein, has been identified. GeneParityPlot analysis suggests that HaSNPV, SeMNPV and LdMNPV (group II) have structural genomic features in common and are distinct from the group I NPVs and from the granuloviruses. Cluster alignment between group I and group II baculoviruses suggests that they have a common ancestor.

Introduction

Members of the family *Baculoviridae* are rod-shaped viruses with circular, covalently closed, double-stranded DNA genomes ranging from 100 to 180 kb. The virions are occluded into large proteinaceous capsules or occlusion bodies. Two genera, nucleopolyhedrovirus (NPV) and granulovirus (GV), have been recognized. Each genus is distinguished by a particular occlusion body morphology with single (GV) and multiple (NPV) virions occluded in granules and polyhedra, respectively. The NPVs are designated as single (S) or multiple (M) depending on the potential number of nucleocapsids packaged in a virion, but this appears to have no taxonomic value (Murphy *et al.*, 1995).

Baculoviruses are frequently used as bio-insecticides of phytophagous insects, mainly belonging to the orders *Lepidoptera*, *Hymenoptera* and *Diptera* (Moscardi, 1999; Federici, 1999). The SNPV of the bollworm *Helicoverpa armigera* (HaSNPV) has been extensively used to control this insect in cotton and vegetable crops in China (Zhang, 1994). In 1999, about 100 000 hectares of cotton had been treated with a

Author for correspondence: Zhihong Hu.
Fax +86 27 87641072. e-mail huzh@pentium.whiov.ac.cn

The GenBank accession number of the sequence reported in this paper is AF271059.

Table 1. Characteristics of baculovirus genomes

The genome characteristics of the different baculoviruses are derived from the following references: AcMNPV (Ayres *et al.*, 1994); BmNPV (Gomi *et al.*, 1999); OpMNPV (Ahrens *et al.*, 1997); LdMNPV (Kuzio *et al.*, 1999); SeMNPV (IJkel *et al.*, 1999); and XcGV (Hayakawa *et al.*, 1999).

Characteristic	HaSNPV	AcMNPV	BmNPV	OpMNPV	LdMNPV	SeMNPV	XcGV
Size (kb)	131·4	133·9	128·4	132·0	161·0	135·6	178·7
G + C content (mol%)	39	41	40	55	58	44	41
Total ORFs	135	154	136	152	163	139	181
Unique ORFs	20	11	1	16	29	17	82
Hr regions	5	8	7	5	13	6	8
Early	33	65	12	61	12	34	13
Late	60	72	78	64	79	72	84
Early + late	9	29	7	26	6	14	2
Promoter not identified	49	47	35	58	78	53	84

commercial virus preparation based on HaSNPV. Recently, recombinant HaSNPVs with improved insecticidal properties have been engineered (Chen *et al.*, 2000*b*) and field tested (S. Sun, X. Chen, Z. Zhang, H. Wang, F. J. J. A. Bianchi, H. Peng, J. M. Vlak & Z. H. Hu, unpublished). However, the genetics of HaSNPV have only been partly described.

The nucleotide sequences of five MNPVs, *Autographa californica* (Ac) MNPV (Ayres *et al.*, 1994), *Bombyx mori* (Bm) NPV (Gomi *et al.*, 1999), *Orgyia pseudotsugata* (Op) MNPV (Ahrens *et al.*, 1997), *Lymantria dispar* (Ld) MNPV (Kuzio *et al.*, 1999) and *Spodoptera exigua* (Se) MNPV (IJkel *et al.*, 1999), and one granulovirus, *Xestia c-nigrum* (Xc) GV (Hayakawa *et al.*, 1999), have been determined. The size of these genomes ranges from 128 413 bp for BmNPV to 178 733 bp for XcGV. This size difference is predominantly due to the presence of gene duplications including the so-called 'baculovirus repeat ORF' or *bro* genes (Gomi *et al.*, 1999). However, no SNPV genome has been sequenced to date and it is therefore of interest to see whether the sequence of HaSNPV would reveal some unique features contributing to, among others, the SNPV phenotype and to the specificity of this virus for heliothine insects.

A physical map of HaSNPV has been previously constructed and the size was estimated to be about 130 kb (Chen *et al.*, 2000*a*). Analysis of approximately 45 kb of random sequence from the HaSNPV genome resulted in the identification of 53 ORFs with homologies to ORFs of other baculoviruses. Partial alignment of the HaSNPV genome with other baculovirus genomes using GeneParityPlot (Hu *et al.*, 1998) revealed a close relationship of HaSNPV and SeMNPV in terms of genomic organization (Chen *et al.*, 2000*a*). A few genes, notably *polyhedrin* (Chen *et al.*, 1997*b*), *ecdysteroid UDP-glucosyltransferase* (*egt*) (Chen *et al.*, 1997*a*), *DNA polymerase* (Bulach *et al.*, 1999) and '*late expression factor* 2' (*lef-2*) (Chen *et al.*, 1999), have been characterized in some detail. Phylogenetic analysis of these genes also revealed a close ancestral relationship between HaSNPV, SeMNPV and LdMNPV at the gene level.

In this paper we describe the complete nucleotide sequence and organization of the HaSNPV genome. This baculovirus is characterized by the absence of extensive gene duplications and by the presence of a limited number of homologous repeat (*hr*) regions, the structure of which is distinctly different from the *hr* sequences of other baculoviruses. Finally, a genomic comparison is made with the complete sequences of AcMNPV, SeMNPV, LdMNPV and XcGV using GeneParityPlot (Hu *et al.*, 1998).

Methods

■ **Insect and virus.** The bollworm *H. armigera* was cultured as a laboratory colony and reared on artificial diet as described by Zhang *et al.* (1981). The wild-type virus was originally isolated from diseased *H. armigera* larvae in the Hubei province of the People's Republic of China in 1981. By *in vivo* cloning, eight HaSNPV genotypes were isolated (G1–G8) (Sun *et al.*, 1998), of which the G4 strain was selected for sequencing. Polyhedra of the G4 strain were propagated in fourth instar *H. armigera* larvae.

■ **HaSNPV DNA isolation, cloning and sequence determination.** The HaSNPV G4 strain (Sun *et al.*, 1998) was sequenced to a sixfold genomic coverage using a shotgun approach. The viral DNA was caesium chloride-purified (King & Possee, 1992) and sheared by nebulization into fragments with an average size of 1200 bp. Blunt repair of the ends was performed with *Pfu* DNA polymerase (Stratagene), according to the manufacturer's directions. DNA fragments were size-fractionated by gel electrophoresis and cloned into the *Eco*RV site of pBluescriptSK (Stratagene). After transformation into *E. coli* XL2-Blue competent cells (Stratagene), 1000 recombinant colonies were picked randomly. DNA templates for sequencing were isolated using QIAprep Turbo kits (Qiagen) on a QIAGEN BioRobot 9600. Sequencing was performed using the ABI PRISM Big Dye Terminator Cycle Sequencing Ready reaction kit with FS AmpliTaq DNA polymerase (Perkin Elmer) and analysed on an ABI 3700 DNA Analyser.

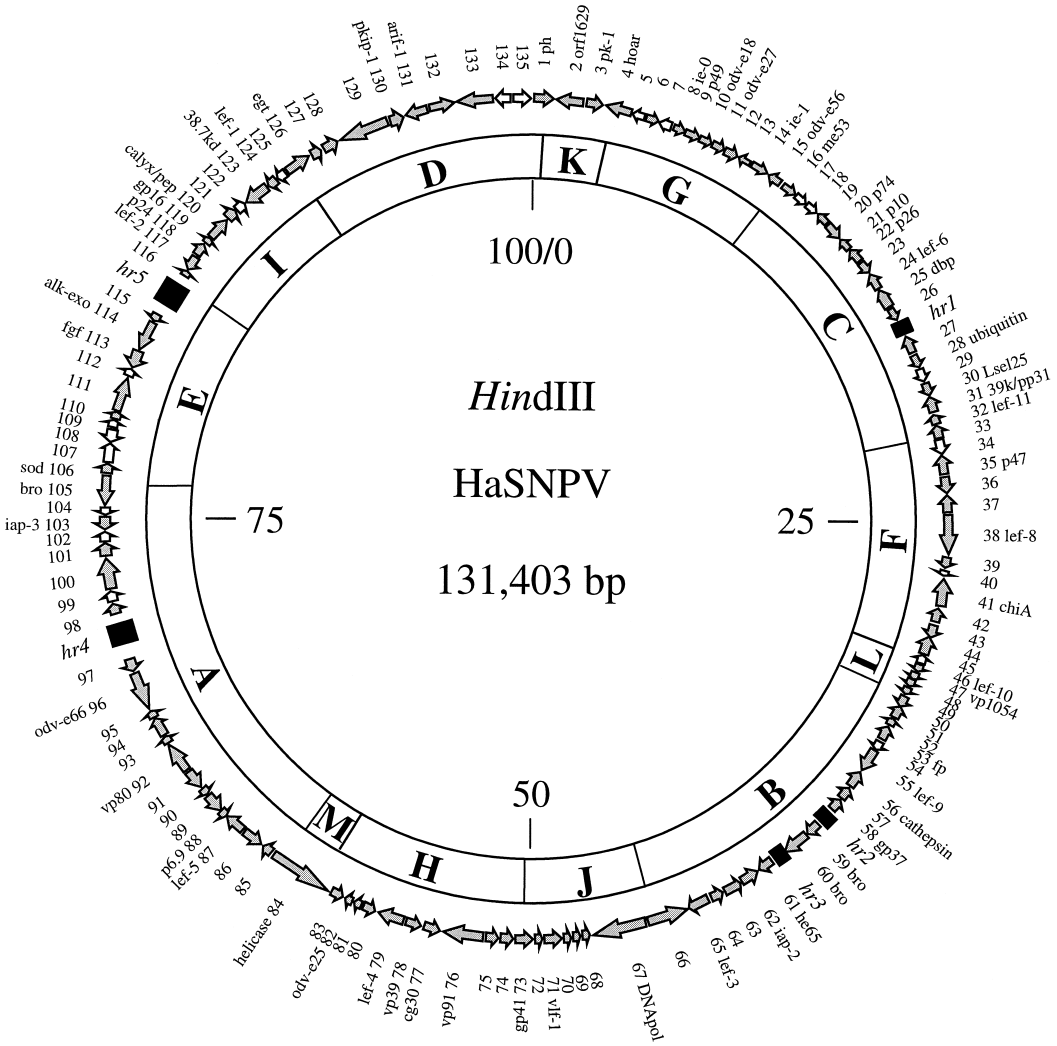

Fig. 1. Circular map and genomic organization of the HaSNPV DNA genome. The sites for restriction enzyme *Hin*dIII are presented; the fragments are indicated A to M according to size from the largest to the smallest restriction fragment (Chen *et al.*, 2000*a*). The positions of the 135 identified ORFs are indicated with arrows that also represent the direction of transcription. Shaded arrows indicate that the ORF has a homologue in other baculoviruses in the protein sequence databases. Open arrows represent ORFs unique to HaSNPV. The corresponding number along the ORF represents the HaSNPV ORF number. The positions of the *hr* sequences are indicated by black boxes. The scale on the inner circle is in map units.

Shotgun sequences were base-called by the PHRED basecaller and assembled with the PHRAP assembler (Ewing & Green, 1998; Ewing *et al.*, 1998). Using the PREGAP4 interface, PHRAP-assembled data were stored in the GAP4 assembly database (Bonfield *et al.*, 1995). The GAP4 interface and its features were then used for editing and sequence finishing. Consensus calculations with a quality cut-off value of 40 were performed from within GAP4 using a probabilistic consensus algorithm based on expected error rates output by PHRED. Sequencing the PCR products bridging the ends of existing contiguous fragments filled the remaining gaps in the sequence.

■ **DNA sequence analysis.** Genomic DNA composition, structure, repeats and restriction enzyme pattern were analysed with the University of Wisconsin Genetics Computer Group programs (Devereux *et al.*, 1984) and DNASTAR (Lasergene). ORFs encoding more than 50 amino acids (150 bp) were considered to be protein-encoding and hence designated putative genes. The maximal alignment of 115 ORFs (out of 135) was checked with known baculovirus gene homologues extracted from GenBank; ORFs with an overlap of *hr* region were excluded from the alignment analysis. The overlap between any two ORFs with known baculovirus homologues was set to a maximum of 25 amino acids; otherwise the largest ORF was selected.

DNA and protein comparisons with entries in the sequence databases were performed with FASTA and BLAST programs (Pearson, 1990; Altschul *et al.*, 1990). Multiple sequence alignments were performed with the GCG PileUp and Gap computer programs version 10.0 (Genetics Computer Group, Madison, WI, USA) with gap creation and extension penalties set to 9 and 2, respectively (Devereux *et al.*, 1984). Percentage identity indicates the percentage of identical residues between two complete sequences. The GENESCAN program was used for gene predictions (http://ccr-081.mit.edu/GENESCAN.html). The DOTTER program (http://www.cgr.ki.se/cgr/groups/sonnhammer/Dotter.html) was used to identify and classify repeat families and miniature inverted

Table 2. Listing of potentially expressed ORFs in HaSNPV

ORF	Name	Position	aa	Predicted M_r	Promoter	Homologous ORFs						Identity to homologues (%)							Cluster
						Ac	Bm	Op	Ld	Se	Xc	Ac	Bm	Op	Ld	Se	Xc	Hz*	
1	*polyhedrin*	1 → 738	245	26779	L	8	1	3	1	1	1	85	81	83	80	86	53	99	a
2	*orf1629*	735 ← 1976	413	45905		9	2	2	2	2	2	30	27	24	28	27	29	99	a
3	*pk-1*	1991 → 2794	267	31543		10	3	1	3	3	3	41	41	40	43	55	36	100	a
4	*hoar**	2917 ← 5187	756	85428	E					4						27		97	
5		5383 → 5562	59	7219															
6	*Hzorf480*	5733 → 6590	285	34459	E													100	
7		6794 ← 6961	55	6436															
8	*ie-0*	6949 → 7806	285	33186	L, L	141	117	138	21	138		31	31	32	33	31			b
9	*p49*	7823 → 9229	468	55256	L, L, L	142	118	139	20	137	13	52	52	51	56	57	35	99	b
10	*odv-e18*	9240 → 9485	81	8822	L	143	119	140	19	136	12	62	56	44	58	68	50	90	b
11	*odv-ec27*	9500 → 10354	284	33288	L	144	120	141	18	135	112	52	52	50	56	59	32	100	b
12		10399 → 10677	92	10780	L	145	121	142	17	134	11	50	48	49	58	58	23		b
13		10704 ← 11315	203	22922		146	122	144	16	133	10	32	31	32	31	34	26	92	b
14	*ie-1*	11357 ← 13324	655	75972		147	123	145	15	132	9	30	30	30	34	34	22	98	b
15	*odv-e56*	13378 ← 14442	354	38850	L, L	148	124	146	14	6	15	49	49	50	50	51	44	100	b
16	*me53*	14603 → 15457	284	33603	E	139	116	137	23	7	180	23	24	25	33	33	27		
17		15504 → 15683	59	7340															
18		15686 → 15853	55	6377															
19		15906 ← 16187	93	11110	E				26						40				
20	*p74*	16208 → 18274	688	78434		138	115	134	27	131	77	53	54	54	59	57	41		
21	*p10*	18328 ← 18591	87	9331	L	137	114	133	41	130	5	26	32	23	47	48	59		c
22	*p26*	18674 ← 19477	267	30510	L	136	113	132	40	129		38	37	34	33	23			c
23		19591 → 19794	67	8258	E	29	20	39	39	128	16	32	31	33	45	44	39		
24	*lef-6*	19870 ← 20433	187	22188		28	19	40	38	127		30	30	30	37	38			d
25	*dbp*	20447 ← 21421	324	37560		25	16	43	47	126	89	36	37	31	24	49	25		d
26	*hr1*	21638 → 22039	133	15025		26	17	42	36	125		35	32	35	39	29			
27		24316 ← 25083	255	29529	E	34	25	26	42	124		31	33	35	46	54			e
28	*ubiquitin*	24923 → 25174	83	9244	L, L	35	26	25	43	123	52	73	73	74	76	74	79		e
29		25238 → 25744	168	20406	E														
30	*Lsel25*†	25764 → 26336	190	22541	L													31†	
31	*39K/pp31*	26395 ← 27330	311	35195		36	27	24	44	120	55	36	36	33	40	36	24		e
32	*lef-11*	27296 ← 27679	127	14583		37	28	23	45	119	56	35	35	31	41	48	35		e
33		27648 ← 28364	238	28411		38	29	22	46	118	79	53	53	57	57	61	47		e
34		28595 → 29674	359	41190	E														
35	*p47*	29985 ← 30986	333	38963		40	31	45	48	115	78	53	53	47	58	56	42		
36		31059 → 31730	223	25768	E	41	32	46				26	25	26					
37		31816 → 32058	80	9543	L	43	34	48		113		25	25	25		31			
38	*lef-8*	32055 ← 34760	901	104988		50	39	54	51	112	148	64	65	60	67	70	54		
39		34813 → 35397	194	22508	L	51	40	55		111		26	24	25		26			
40		35538 → 35690	50	6299															

Table 2 (*cont.*)

ORF	Name	Position	aa	Predicted M_r	Promoter	Homologous ORFs						Identity to homologues (%)							Cluster
						Ac	Bm	Op	Ld	Se	Xc	Ac	Bm	Op	Ld	Se	Xc	Hz*	
41	*chitinase*	35698 ← 37410	570	65 481		126	103	124	70	19	103	67	68	68	66	63	60	95	
42		37489 ← 38031	180	21 260	E	52	41		53	109		29	31		38	26			f
43		38148 → 38558	136	16 419		53	42	56	54	108	171	42	44	45	49	56	28		f
44		38565 ← 39701	378	42 771	L				55	107					26	30			f
45		39709 ← 39936	75	9090	L														
46	*lef-10*	39896 → 40111	71	7684		53a	42a	57	56	106	171	43	42	31	43	54	36		f
47	*vp1054*	39984 → 41039	351	41 700		54	43	58	57	105	175	44	44	40	50	55	40		f
48		41159 → 41365	68	7962		55	44	59	58	104		40	31	31	40	53			f
49		41366 → 41560	64	7406	L	56	45	60		103		26	26	28		39			f
50		41846 → 42361	171	20 671	L	57	46	61	60	102		42	42	41	44	44			f
51		42412 ← 42894	160	19 034		59		62	61	101		28		38	39	47			f
52		42906 ← 43172	88	10 219	L	60	48	63	62	100	102	43	44	31	45	57	43		f
53	*fp*	43385 ← 44038	217	25 368	L	61	49	64	63	98	140	62	62	56	52	68	37		f
54		44210 → 44395	61	7302															
55	*lef-9*	44507 → 46066	519	59 545		62	50	65	64	97	139	65	66	53	70	72	57		f
56	*cathepsin*	46150 ← 47247	365	42 021	L	127	104	125	78	16	58	47	48	48	47	46	44		
57		47288 ← 47875	195	21 292	E, L						83						33		
58	*gp37 hr2*	47946 ← 48785	279	32 099	E, L	64	52	69	68	25	107	56	56	56	58	60	45		
59	*bro-a*	49936 → 50670	244	28 269					150						53				
60	*bro-b hr3*	50794 → 52377	527	59 734					146		159‡				60		58‡		
61	*he65*	53133 → 53843	236	27 478	E	105	89				67	29	28				33		
62	*iap-2*	53920 ← 54672	250	29 254	L	71	58	74	79	88		34	35	35	41	42			g
63		54720 ← 55544	274	31 562		69	57			89		42	43			48			g
64		55513 ← 55914	133	15 561	E	68	56	73	80	90	135	42	43	35	52	56	30		g
65	*lef-3*	55934 → 57073	379	44 018		67	55	72	81	91	134	27	29	29	29	35	17		g
66		57181 ← 59538	785	88 881	L	66	54	71	82	92		28	25	24	25	27			g
67	*DNA pol*	59569 → 62631	1020	119 250		65	53	70	83	93	132	47	47	44	55	61	38		g
68		62708 ← 63166	152	17 612	L, L	74	60	77				26	26	17					
69	*hzORF384*	63232 ← 63615	127	14 880	L	75	61	78	84	94	126	24	24	25	40	38	26	100	h
70		63621 ← 63878	85	9958	L	76	62	79	85	95	125	43	42	39	74	64	37		h
71	*vlf-1*	63919 ← 65157	412	47 878	L	77	63	80	86	82	123	70	69	67	71	64	35	99	i
72		65170 ← 65502	110	12 730	L, E	78	64	81	87	81	122	44	41	41	44	50	33		I
73	*gp41*	65571 ← 66539	322	36 579	L	80	66	83	88	80	121	58	57	53	64	56	37		i
74		66469 ← 67194	241	27 681	E	81	67	84	89	79	120	54	53	50	55	58	52		i
75		67067 ← 67744	225	24 912	E	82	68	85	90	78	119	31	31	23	44	49	31		i
76	*vp91capsid*	67674 → 70124	816	93 527	L	83	69	86	91	77	118	43	43	42	42	48	33		i
77	*cg30*	70252 ← 71103	283	32 325	L	88	71	89		76		31	29	25		24			i
78	*vp39capsid*	71192 ← 72073	293	33 403		89	72	90	92	75	111	45	46	48	52	54	35		i
79	*lef-4*	72072 → 73457	461	53 977		90	73	91	93	74	110	46	46	40	46	53	37		i
80		73510 ← 74274	254	30 849		92	75	93	94	73	101	55	56	56	51	60	45		i

Table 2 (*cont.*)

ORF	Name	Position	aa	Predicted M_r	Promoter	Homologous ORFs						Identity to homologues (%)						Hz*	Cluster
						Ac	Bm	Op	Ld	Se	Xc	Ac	Bm	Op	Ld	Se	Xc		
81		74276 → 74764	162	19065	L	93	76	94	95	72	100	54	54	51	61	62	36		i
82	*odv-e25*	74810 → 75502	230	25933		94	77	95	96	71	99	44	43	40	74	69	51		i
83		75534 ← 76031	165	18793	E, L					68						26			
84	*helicase*	76050 ← 79811	1253	145955		95	78	96	97	70	98	44	44	40	49	50	30		l
85		79768 → 80289	173	19805	E	96	79	97	98	69	97	48	48	45	61	61	36		l
86		80348 ← 81313	321	37930		98	82	99	99	67	96	46	45	45	50	55	41		k
87	*lef-5*	81209 → 82156	315	37040		99	83	100	100	66	95	52	52	50	50	57	43		k
88	*p6.9*	82150 ← 82479	109	11522		100	84	101	101	65	94	44	49	53	69	59	57		k
89		82544 ← 83653	369	42553	L	101	85	102	102	64	93	43	42	39	43	53	26		k
90		83699 ← 84067	122	13830	L	102	86	103	103	63	92	26	26	32	26	33	26		k
91		84067 ← 85200	377	44040	E, L	103	87	104	104	62	91	51	51	45	52	60	41		k
92	*vp80capsid*	85295 → 87112	605	69719	E, L	104	88	105	105	61		23	24	22	26	29			k
93		87109 → 87285	58	6943		110			106	60	51	29			46	48	43		
94		87300 → 88385	361	41508		109	92	109	107	59	53	58	58	54	60	58	35		l
95		88431 → 88715	94	10974		108	91	108	108	58		41	42	34	45	51			l
96	*odv-e66*	88782 ← 90800	672	76093	L	46	37	50	131	57/114	149	43	42	43	54	45/34	60		
97	*p13*† *hr4*	90821 ← 91651	276	32453	L					56	43					59	48		
98		93957 → 94556	199	22409	L	115	95	115	143	50	32	39	39	40	47	45	40		18
99		94560 → 94916	118	14449	E														
100	*parg*	95012 → 96544	510	58136					141	52					24	24			
101		96623 → 97384	253	29046	L	106/107	90	107	140	53	50	47/32	47	47	50	56	31		
102		97399 → 97731	110	12790															
103	*iap-3*	97789 ← 98595	268	31522	E, L			35	139	110				41	29	38			
104		98592 ← 98747	51	5931															
105	*bro-c*	98858 ← 100363	501	58269	L				71		60				51		66		
106	*sod*	100531 → 101010	159	16853		31	23	29	145	48	68	72	72	71	68	68	57		
107		101017 → 102390	457	51209															
108		102443 ← 103021	192	22772															
109		103190 → 103546	118	13648															
110		103557 → 103823	88	10079		117	96	117		47		33	30	24		48			
111		103891 → 105477	528	60289		119	97	119	155	36	84	49	49	47	47	46	36		
112		105474 → 105710	78	9090	L														
113	*fgf*	105733 ← 106638	301	34358	E	32	24	27	156	38	144	29	31	29	29	27	23		
114	*alk-exo*	106765 ← 108051	428	49416		133	110	131	157	41	145	45	45	44	42	44	42		
115	*hr5*	108071 ← 108460	129	15332	L	19	11	18	159	42		28	27	27	29	32			
116		111267 → 111482	71	8204	E	111	93	112	76		160	36	34	35	32		59		
117	*lef-2*	111600 ← 112325	241	27811	E	6	135	6	137	12	35	43	43	42	42	46	29		
118	*p24capsid*	112687 → 113433	248	28373	L	129	106	127		10	80	37	40	37		51	23		
119	*gp16*	113495 → 113779	94	10669	L, L	130	107	128		9		25	25	22		32			

Table 2 (*cont.*)

ORF	Name	Position	aa	Predicted M_r	Promoter	Homologous ORFs						Identity to homologues (%)							Cluster
						Ac	Bm	Op	Ld	Se	Xc	Ac	Bm	Op	Ld	Se	Xc	Hz*	
120	*calyx / pep*	113831 → 114853	340	39058	L	131	108	129	136	46		36	38	29	38	47			
121		114932 → 115396	154	18472	E	63	51		117			29	29		23				
122		115527 → 116117	196	23477	E, L														
123	*38·7kd*	116174 ← 117331	385	44474		13	5	12	122	13		26	26	25	29	38			m
124	*lef-1*	117333 ← 118070	245	29059		14	6	13	123	14	82	39	40	44	48	51	45		m
125		118045 ← 118479	144	16114	E, L														
126	*egt*	118624 → 120171	515	58870	L	15	7	14	125	27		47	47	45	50	55		99	n
127		120371 → 120949	192	22595														24§	
128		120900 → 121700	266	30352		17	9	16	128	29		26	29	30	31	33			n
129		121781 ← 124624	947	111338	L, L				129	30					30	29			n
130	*pkip-1*	124989 → 125498	169	20282		24	15	44	110	32		23	26	27	35	38			
131	*arif-1*	125565 ← 126362	265	30355	E	21	12	19	118	34		29	24	24	25	30			
132		126619 → 127770	383	44534		22	13	22	119	35	45	60	60	57	66	66	50		
133		127811 ← 129844	677	78241	L, E	23	14	21	130	8	27	24	23	25	43	40	29		
134		129986 ← 130531	182	21930	E														
135		130713 → 131297	194	23310	E														

* Taken from HzSNPV GenBank accessions.
† LsNPV ORF name taken from Wang *et al.* (1995). Percentage amino acid identity is shown to Lsel25.
‡ Identity to C-terminal 344 amino acids.
§ *S. litoralis* ORF homologue.

repeat transposable elements (MITEs). GeneParityPlot analysis was performed on the HaSNPV genome versus the genomes of AcMNPV, SeMNPV, LdMNPV and XcGV as described previously (Hu *et al.*, 1998).

Results and Discussion

Nucleotide sequence analysis of the HaSNPV genome

The HaSNPV genome was assembled into a contiguous sequence of 131 403 bp (Table 1). This size is in good agreement with a previous estimate of 130·1 kb for HaSNPV DNA based on restriction enzyme analysis and physical mapping (Chen *et al.*, 2000*a*). AcMNPV, BmNPV, OpMNPV and SeMNPV have similar size genomes, which are much smaller than the genomes of LdMNPV and XcGV with 161 kb and 178 kb, respectively (Table 1). With a G + C content of 39·1 mol%, HaSNPV has the lowest G + C content among baculoviruses to date, which is close to that of AcMNPV (41 mol%) (Ayres *et al.*, 1994), BmNPV (Gomi *et al.*, 1999) and XcGV (Hayakawa *et al.*, 1999). The G + C contents of OpMNPV (Ahrens *et al.*, 1997) and LdMNPV (Kuzio *et al.*, 1999) are much higher with 55 and 58 mol%, respectively. According to a recently adopted convention (IJkel *et al.*, 1999; Hayakawa *et al.*, 1999), the adenine residue at the translational initiation codon of the *polyhedrin* gene was designated as the zero point of the physical map of HaSNPV DNA (Fig. 1). Taking *polyhedrin* as the first gene determines the orientation of the physical map. This map is now reversed as compared with the original map presented by Chen *et al.* (2000*a*) and positions the *p10* gene at map unit 10.

Using computer-assisted analysis, 326 ORFs defined as methionine-initiated ORFs larger than 50 amino acids were found. From these, 135 ORFs with fewer than 25 amino acids or no overlap with other ORFs have been identified on the HaSNPV genome (Fig. 1; Tables 1 and 2) and were further analysed. This number of 135 ORFs is roughly proportional to the size of the HaSNPV genome as compared with the other six completely sequenced baculovirus genomes AcMNPV, BmNPV, OpMNPV, SeMNPV, LdMNPV and XcGV. The HaSNPV ORFs are in general tightly packed with minimal intergenic distances; their orientation is almost evenly distributed along the genome (52% clockwise, 48% anticlockwise; Fig. 1). The locations, orientations and sizes of the predicted ORFs are shown in detail in Table 2. The 135 predicted ORFs account for 87% of the genome versus 8% for intergenic sequences and 6% for the *hr* region. The HaSNPV ORFs have an average length of 844 nt with Ha84 (*helicase*) being the largest (3758 nt) and Ha40, without a homologue in other baculoviruses, being the smallest (150 nt). Of the 135 HaSNPV ORFs, 115 (86%) have an assigned function or have homologues with other putative baculovirus genes (Table 2). So far it appears that 20 ORFs are unique to HaSNPV. These ORFs accounted for 6% (7·3 kb) of the genome in total.

The HaSNPV nucleotide sequence was determined from an isolate cloned *in vivo* (Sun *et al.*, 1998). Based on restriction enzyme analysis and Southern hybridization, no fragments in a less than molar ratio were observed in this isolate. However, sequence analysis showed that at approximately 100 nucleotide locations (0·07% of the genome) along the genome a polymorphism was observed in the nucleotide usage. None of these affected the ORFs. This polymorphism may be partly the result of the sequencing (error 10^{-5}), but also the consequence of the intrinsic genetic variation that exists either in natural HaSNPV isolates (Gettig & McCarthy, 1982; Figueiredo *et al.*, 1999) or in *in vivo* cloned isolates of HaSNPV (Sun *et al.*, 1998), GV (Smith & Crook, 1988) and MNPVs (Muñoz *et al.*, 1998). Despite the *in vivo* cloning and the apparent lack of genetic heterogeneity as evidenced from restriction enzyme analysis (Sun *et al.*, 1998), microvariation may thus exist. This suggests that the quasispecies concept for RNA viruses, i.e. a virus species is defined not as a single nucleotide sequence but as a mixture of genotypes (Domingo *et al.*, 1995), may also apply to DNA viruses including baculoviruses.

Homologous repeat (*hr*) regions

Regions with homologous repeats were first found in AcMNPV (Cochran & Faulkner, 1983) and appear to be present in all baculoviruses. They occur at multiple locations along the genome and may serve as origins of DNA replication (Kool *et al.*, 1995) and as enhancers of transcription (Guarino & Summers, 1986; Guarino *et al.*, 1986). *Hr* regions are characterized by the presence of multiple, often imperfect, tandemly repeated palindromic sequences (AcMNPV). Five *hr* regions were previously identified on the genome of HaSNPV by direct sequencing and Southern blot hybridization (Chen *et al.*, 2000*a*). No further *hr* regions were detected in the complete sequence (Fig. 1; Table 1). These five *hr* regions were found dispersed along the HaSNPV genome around map positions 17.5 (*hr*1), 37.7 (*hr*2), 40.2 (*hr*3), 70.8 (*hr*4) and 83.6 (*hr*5) and are located in AT-rich intergenic regions. Their sizes vary from 750 (*hr*3) to 2800 nt (*hr*5). It is interesting to note that *hr*2 and *hr*3 are separated by two *bro*-related genes (Fig. 1). This configuration might have been the result of an insertion of two *bro* genes into what originally may have been a single *hr*. Assuming that *hr*2 and *hr*3 have been a single *hr*, the *hr* regions of HaSNPV are remarkably similar is size (2100–2800 nt).

Using a dot matrix analysis, the HaSNPV sequence was compared to itself and its complementary strand. Two types of repeats were identified, type A and type B, with imperfect 40 and 107 bp long repeats, respectively, or truncated versions thereof (Fig. 2). The type A and type B repeats are found in each of the *hr* regions. There is no sequence homology with other known baculovirus *hr* regions. The type B repeats contain short internal stretches of palindromic and direct repeats. Not only is the sequence of the HaSNPV *hr* regions different from those of other baculovirus *hr* regions, but their structure is also rather unique. The function of the type A and type B repeats remains to be determined.

Table 3. Number of ORFs with homologues in baculoviruses and percentage amino acid identity

The numbers of ORFs with homologues in baculoviruses are shown above the diagonal and the percentage amino acid identity is shown below the diagonal.

	AcMNPV	BmNPV	OpMNPV	LdMNPV	SeMNPV	XcGV	HaSNPV
HaSNPV	100	98	94	94	103	69	–
XcGV	84	80	76	93	72	–	40
SeMNPV	103	99	102	104	–	ND	47
LdMNPV	94	91	95	–	45	ND	46
OpMNPV	126	121	–	ND	40	34	41
BmNPV	115	–	55	ND	41	ND	41
AcMNPV	–	93	56	41	41	33	41

ND, Not determined.

Comparison of the gene content of HaSNPV and other baculoviruses

The sequence of the HaSNPV genome was compared with those of AcMNPV (Ayres *et al.*, 1994), BmNPV (Gomi *et al.*, 1999), OpMNPV (Ahrens *et al.*, 1997), SeMNPV (IJkel *et al.*, 1999), LdMNPV (Kuzio *et al.*, 1999) and XcGV (Hayakawa *et al.*, 1999) for the presence or absence of putative ORFs (Table 2). These seven baculovirus genomes have a cumulative total of 354 different ORFs, of which 183 are unique to individual baculovirus genomes. Among the seven baculoviruses, including HaSNPV 65 ORFs are conserved. Among all NPVs, 84 ORFs are conserved (data not shown). This suggests that about 70 ORFs are the minimal requirement for basic baculovirus features, such as virus structure, transcription, DNA replication, auxiliary functions on the cellular or organism level and occlusion body morphogenesis (Table 2). Putative functions have been assigned to approximately 61% of these common baculovirus genes. Twenty ORFs larger than 50 amino acids were unique to HaSNPV. Nine of these, 50 to 100 amino acids long, have no consensus baculovirus promoter (Table 2). Most likely, these small ORFs in HaSNPV are not functional, but this has to be tested experimentally.

Of the 135 HaSNPV ORFs identified, 100 have homologues in AcMNPV and a further 15 have homologues in other baculoviruses (Tables 1 and 2). HaSNPV shares the largest number of homologues (103) with SeMNPV, underscoring the close relationship between these two viruses as evidenced from gene phylogeny analyses involving *polyhedrin*, *egt*, *lef-2* (Chen *et al.*, 1997*a*, *b*, 1999) and *DNA polymerase* (Bulach *et al.*, 1999). *Polyhedrin* is the most conserved ORF of the six NPVs, with a mean amino acid identity of 83% to other NPV *polyhedrin* genes; the identity to GV *granulin* is much less (51% for XcGV). *Ubiquitin* (*ubi*), which is involved in the targetting of proteins for degradation, is the next most conserved gene among the seven sequenced baculoviruses, with 75% amino acid identity, followed by *superoxide dismutase* (*sod*) with 70% amino acid identity. The mean ORF amino acid identity between HaSNPV and the group II baculoviruses SeMNPV and LdMNPV is similar (46%) and higher than to group I baculoviruses (41%). This is in support of the distinct phylogenetic relationship between group I and group II NPVs (Zanotto *et al.*, 1993; Bulach *et al.*, 1999).

Structural virion genes

The HaSNPV genome contains the known genes encoding the common virion structural proteins of NPVs (Table 2). In contrast to SeMNPV, where *odv-e66* is duplicated (IJkel *et al.*, 1999), HaSNPV does not contain duplicate genes for virion structural proteins. However, HaSNPV apparently lacks a homologue of the BV envelope surface glycoprotein gene *gp64* (Ac128). The product of this gene, GP64, is acquired by virions during budding through the plasma membrane and is involved in the association with cell receptors upon invasion and fusion in endosomes (Oomens & Blissard, 1999). However, an ORF has been identified in HaSNPV (Ha133) with an average amino acid identity with Ld130 (43%) and Se8 (40%). The latter viruses also lack a *gp64* homologue and it has been suggested that Ld130 and Se8 are the functional homologues of AcMNPV *gp64* (Kuzio *et al.*, 1999; IJkel *et al.*, 1999). Recently, direct evidence was obtained that the products of Ld130 and Se8 are the major constituents of the BV envelope and are responsible for the fusogenic activity of SeMNPV (Pearson *et al.*, 2000; IJkel *et al.*, 2000).

DNA replication and late gene expression

There are 19 *lef* genes in AcMNPV that have been implied in DNA replication and late gene expression (Kool *et al.*, 1995; Lu & Miller, 1995). They were all required for late and very late gene expression. Of these, six (*lef-1*, *lef-2*, *lef-3*, *dnapol*, *helicase*, *ie-1*) are essential for DNA replication, whereas others are involved in transcription (*ie-2*, *lef-4*, *lef-5*, *lef-8*, *lef-9*) (Guarino *et*

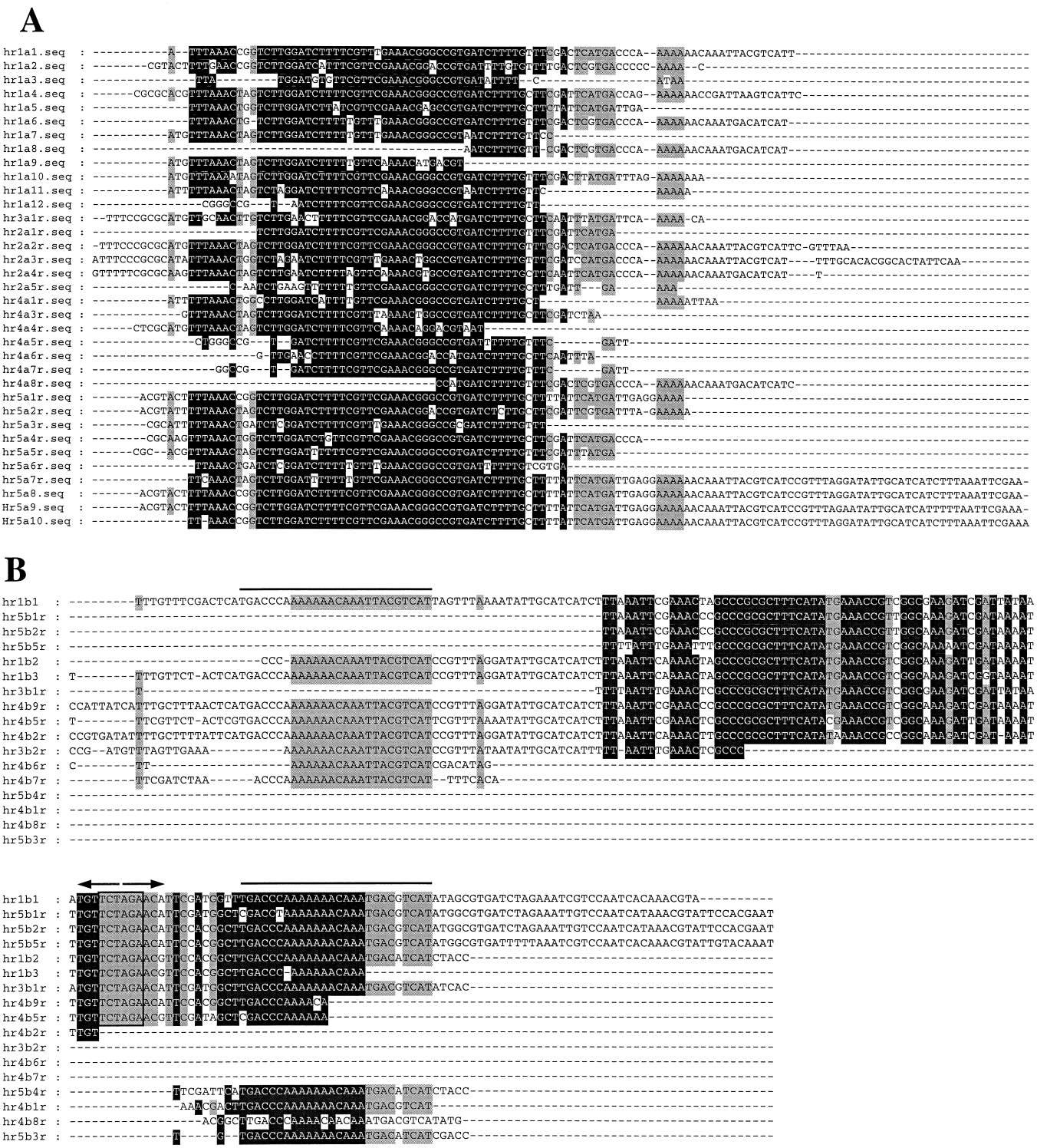

Fig. 2. Alignment of HaSNPV repeated sequences. The nucleotide sequences of the type A repeats (A) and type B repeats (B) are aligned to obtain maximum similarity. The repeats are denoted according to their presence in a homologous region (*hr*1–*hr*5), their type (a or b), their order number in the *hr* and whether they occur in the reverse orientation (r) or not. Shading is used to indicate the relevant occurrence of similar nucleotides in the repeats: black indicates > 59%, grey 53% and white < 47% representation. Short palindromes (by arrows), direct repeats (lines above) and *Xba*I sites (boxed) are indicated.

Table 4. Baculovirus ORFs without homologues in HaSNPV

The AcMNPV ORFs that have no homologue in HaSNPV are shown. ORFs from BmNPV, OpMNPV, LdMNPV and SeMNPV that have no homologue in either AcMNPV or HaSNPV are also shown. Superscripts show that the ORF is present in the indicated virus; AcMNPC (*Ac*), BmNPV (*Bm*), OpMNPV (*Op*), LdMNPV (*Ld*) or SeMNPV (*Se*).

AcMNPV	1 *ptp-1*^{Bm,Op}	20^{Bm}	48 *etm*^{Op}	86 *pnk/pnl*	121	140
	3 *ctl*^{Op,Ld}	27 *iap-1*^{Bm,Op}	49 *pcna*^{Op}	87^{Bm,Op}	122^{Bm,Op}	149^{Bm}
	4^{Bm,Op,Ld}	30^{Bm,Op}	58^{Bm}	91^{Bm,Op}	123 *pk-2*^{Bm}	151 *ie-2*^{Bm,Op}
	5^{Bm,Op}	33^{Se}	70 *hcf-1*	97	124^{Bm,Op}	152
	7 *orf603*	39 *p43*^{Bm}	72^{Bm,Op}	112^{Bm,Ld}	125 *lef-7*^{Bm,Op}	153 *pe38*^{Bm,Op}
	11^{Bm,Op,Ld}	42 *gta*^{Bm,Op}	73^{Bm,Op}	113	128 *gp64*^{Bm,Op}	154^{Bm}
	12^{Bm,Ld}	44^{Bm,Op,Se}	79^{Bm,Op}	116	132^{Bm,Op}	
	16^{Bm,Op}	45^{Bm}	84	118	134^{Se}	
	18^{Bm,Op,Ld,Se}	47^{Bm,Op}	85^{Op}	120^{Bm,Op,Ld,Se}	135 *p35*^{Bm,Op}	
BmNPV	111					
OpMNPV	4	28	37	106 *iap-4*^{Ld}	118	147 Opep32
	5	33	68	110^{Ld}	135	148 Opep25
	9^{Se}	36	98	113	143 *hrf-1*^{Ld}	149 *p8.9*
LdMNPV	4	11	31	69	127^{Se}	141^{Se}
	5	12	34	77	132	142^{Se}
	6	13	49	111^{Se}	133	144^{Se}
	7 *g22*	22 *ligase*	50 *helic-2*	120	134	152
	8	24	52	121	135	160 *vef-2*
	9	25	59	124^{Se}	137a^{Se}	163
	10	28	65 *vef-1*	126	138	
SeMNPV	5	21	31	83	117	
	17	22	39	85	121	
	18	23	40	86	122	
	20	24	44	116		

al., 1998) or in inhibition of apoptosis (such as *p35* and *iap* genes) (Clem & Miller, 1994). *In silico* analysis indicated that the genome of HaSNPV contains homologues of 16 of the above AcMNPV *lef* genes and lacks *ie-2*, *p35* and *lef-12* (Table 4). The latter genes are also absent in LdMNPV, SeMNPV and XcGV, suggesting that they occur only in the group I NPVs. HaSNPV also has a homologue (Ha8) to the first exon of a spliced transcript from Ac141 (*ie-0*). This transcript also includes Ac147 located 4 kb downstream of *ie-0* (Chisholm & Henner, 1988). In contrast, this exon encoded by Se138 is not functional in SeMNPV (Van Strien *et al.*, 2000).

The percentage amino acid identity of HaSNPV *lef-8* (Ha38) and *lef-9* (Ha55) with AcMNPV *lef-8* and *lef-9*, encoding subunits of the RNA polymerase complex (Guarino *et al.*, 1998), was the highest among the *lef*s at about 65%, whereas HaSNPV *lef-3* (Ha65) and AcMNPV *lef-3* shared only 27% of their amino acids. HaSNPV LEF3 has a low degree of homology with other NPVs as well (Table 2) and a *lef-3* gene is not assigned in XcGV (Hayakawa *et al.*, 1999). It has been suggested that LEF3 is chaperoning other replication factors, such as helicase and LEF2, across the nuclear membrane in infected cells (Wu & Carstens, 1998). Since this membrane is almost eliminated upon infection of cells with GV (Federici, 1999), LEF3 may not be required for GVs to replicate. However, there is a very low degree of homology of *lef-3* to Xc134 and this ORF is also of roughly the same size and has a conserved location in the genome compared with the other baculovirus *lef* genes. Further experimentation is required to clarify this assumption. Ha25 shows approximately 36% amino acid identity to Ac25 and Bm16, which encode a putative DNA-binding protein (DBP) (Okano *et al.*, 1999; Mikhailov *et al.*, 1998). An AcMNPV gene involved in the modulation of very late gene expression (*vlf-1*) (Todd *et al.*, 1996) has also been found in HaSNPV (Ha71).

Similar to SeMNPV, LdMNPV and XcGV, HaSNPV also lacks a *p35* homologue (Table 4). Instead, two members of the *iap* (Crook *et al.*, 1993) gene family were observed in HaSNPV, *iap-2* (Ha62) and *iap-3* (Ha103). Homologues of *iap* subclasses (1–4) have been found in AcMNPV (Ac27, *iap-1* and Ac71, *iap-2*), OpMNPV (Op41, *iap-1*; Op74, *iap-2*; Op35, *iap-3* and

ORF106, *iap-4*), SeMNPV (Se88, *iap-2* and Se110, *iap-3*), LdMNPV (Ld79, *iap-2* and Ld139, *iap-3*) and XcGV (Xc137, *iap-3*). The HaSNPV *iap-3* gene has high homology to the CpGV *iap* gene, which could functionally complement an AcMNPV *p35* deletion mutant (Crook *et al*., 1993). OpMNPV *iap-3* can also complement AcMNPV *p35* null mutants (Birnbaum *et al*., 1994). The function of the *iap-1*, *iap-2* and *iap-4* genes is unknown. Through partial DNA sequence analysis, three *iap* gene homologues (*iap-1*, *iap-2* and *iap-3*) were found in *Buzura suppressaria* SNPV (Hu *et al*., 1998).

HaSNPV lack genes for enzymatic functions in nucleotide metabolism, such as ribonucleotide reductase (*rr*) and deoxyuridyltriphosphatase (*dUTPase*). The products of *rr* and *dUTPase* allow the virus to convert rNTPs into dNTPs to the benefit of virus DNA replication. RR reduces NDPs into dNDPs and dUTPase converts dUTP into dUMP, thereby excluding dUTP from incorporation into DNA and providing dUMP as a precursor for dTTP. *dUTPase* and *rr* are present in SeMNPV (IJkel *et al*., 1999), OpMNPV (Ahrens *et al*., 1997) and LdMNPV (Kuzio *et al*., 1999) but are absent from AcMNPV and BmNPV and also from XcGV. The latter virus contained a DNA ligase (Xc141), which appeared to be absent from NPVs except LdMNPV.

Genes with auxiliary functions

Baculovirus auxiliary genes are not essential for virus replication per se but are important, for example, for interaction with the insect host (O'Reilly, 1997). HaSNPV has a very similar set of auxiliary genes as SeMNPV, encoding for example *chitinase* (*chitA*, Ha41), *cathepsin* (*v-cath*, Ha56) and *egt* (Ha126) (IJkel *et al*., 1999). These genes are quite well conserved, with 66, 47 and 49% amino acid identity, respectively, whereas the fibroblast growth factor (*fgf*, Ha113) is poorly conserved among baculoviruses with 28% amino acid identity.

HaSNPV lacks a gene for protein tyrosine/serine phosphatase (*ptp*) with dual-specificity (dsPTP) (Tilakaratne *et al*., 1991; Kim & Weaver, 1993). This protein specifically removes phosphates from both tyrosine and serine/threonine residues (Wishart *et al*., 1995). The absence of a *ptp* gene homologue in HaSNPV is striking, since such a gene is present in all NPV genomes sequenced to date and is thought to be involved in the regulation of the phosphorylation status of viral and host proteins during infection.

Duplicated *bro* genes

A common characteristic of baculovirus genomes is the presence of a group of related genes, the so-named *bro* genes. Five homologues of AcMNPV ORF2 (Ac2) are present in BmNPV (Gomi *et al*., 1999). In LdMNPV, SeMNPV and XcGV sixteen, one and five *bro*-related genes are found, respectively (Kuzio *et al*., 1999; IJkel *et al*., 1999; Hayakawa *et al*., 1999). In OpMNPV, a truncated version and two smaller *bro*-related ORFs are present (Ahrens *et al*., 1997). Three *bro*-related genes were identified in HaSNPV, named *bro-a* (Ha59), *bro-b* (Ha60) and *bro-c* (Ha105). Ha59 is most closely related to Ld150 (*bro-m*), belonging to the group II *bro* family (Kuzio *et al*., 1999), with 50% amino acid identity. Ha60 also belongs to the group II *bro* genes and shares the largest homology to Ld140 (*bro-l*) and Xc159 (*bro-g*), but has an N-terminal duplication of 183 amino acids. It thus seems unlikely that the Ha59 and Ha60 *bro* genes have a common recent ancestor and therefore might have been spliced in tandem into an *hr* sequence (*hr*3 and *hr*4). Ha105 and Xc60 are 66% identical and related to the group III *bro* genes (Kuzio *et al*., 1999).

HaSNPV ORFs with homologues in a few other baculoviruses

HaSNPV possesses 22 ORFs that have no homologues in AcMNPV, BmNPV, OpMNPV, SeMNPV, LdMNPV or XcGV (Table 2). Of these, Ha6 is identical to Hz480 from *Helicoverpa zea* SNPV (HzSNPV) (Le *et al*., 1997). In HaSNPV, a homologue of the *Leucania separata* NPV (LsNPV) *p13* gene (Ha97) is found. This homologue is, in contrast to the SeMNPV homologue, not C-terminally extended (Wang *et al*., 1995; IJkel *et al*., 1999). The two leucine-zipper-like structures present in LsNPV P13 (Wang *et al*., 1995) are also conserved in Ha97. The function of this ORF in LsNPV as well as in SeMNPV (Se59) and XcGV (Xc48) is unknown.

Three HaSNPV ORFs have a homologue in only one other baculovirus. None of these genes has yet been assigned functions. Ha19 has a gene homologue in LdMNPV (Ld26) with an amino acid identity of 40%. This ORF, however, is rather small, encoding an 11 kDa protein. Ha57, encoding a putative 21 kDa product, has a homologue in XcGV (Xc83) with an amino acid identity of 33%. An Se68 homologue is identified in HaSNPV as Ha83 encoding a putative protein of 18·8 kDa but with a low amino acid identity of 26%. All ORFs, however, have baculovirus early and/or late transcription motifs and may therefore be functional.

HaSNPV ORF100 (Ha100) was found to encode a putative poly(ADP-ribose) glycohydrolase (*parg*). The homology with *Drosophila melanogaster* (24% identity) and *Homo sapiens* (23% identity) genes was found in the C-terminal portion of the putative protein. Homologues of Ha100 were also found in LdMNPV (Ld141) and SeMNPV (Se52), so that their presence appears to be limited to group II NPVs. In eukaryotes this enzyme is involved in the breakdown of polyribose and recruitment of this compound for nuclear functions such as DNA replication and repair (D'Amours *et al*., 1999). The function of this enzyme in baculovirus group II morphogenesis or pathology is not known, but it is possible that it is involved in similar capacity during the NPV infection process. The baculovirus *parg* gene is much longer than the eukaryotic counterpart and thus may have additional activities.

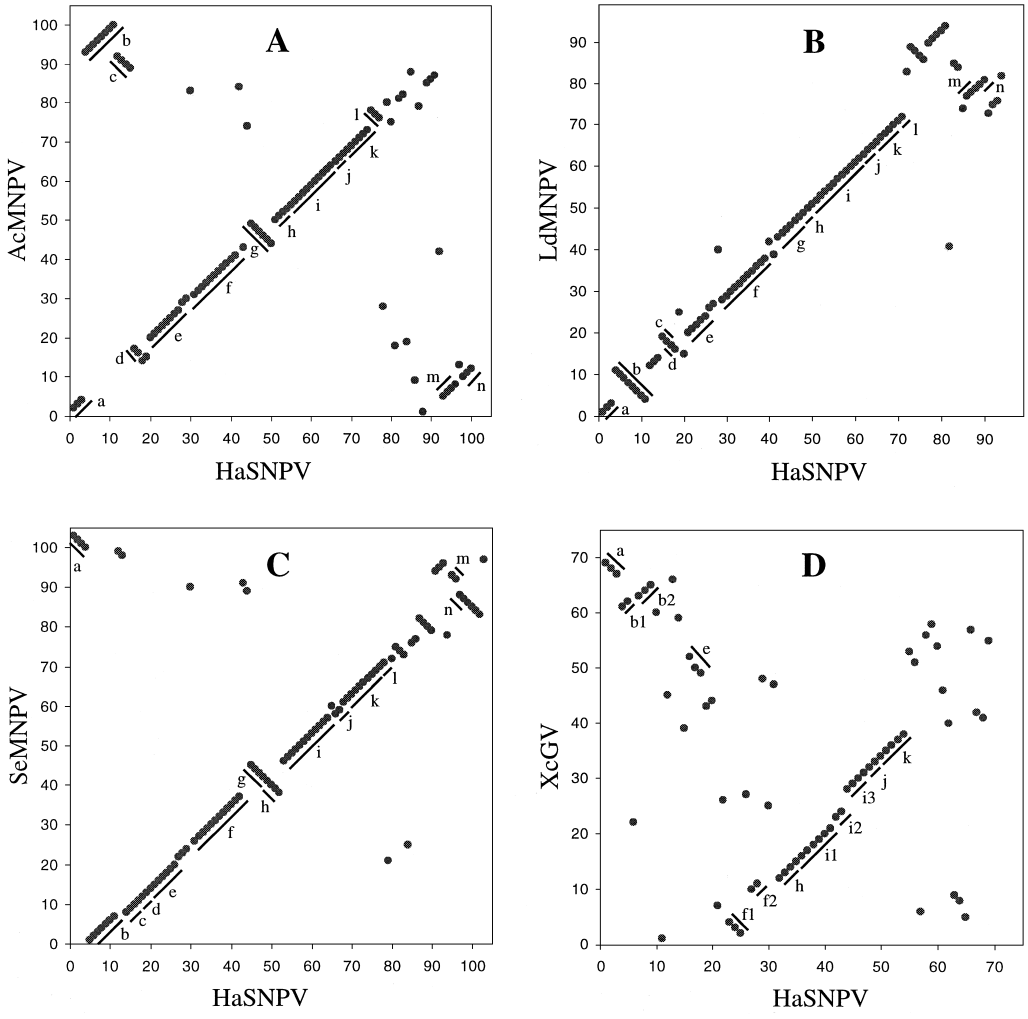

Fig. 3. GeneParityPlots of HaSNPV versus three other baculoviruses. Graphic representation of the collinearity of baculovirus genomes of AcMNPV (A), LdMNPV (B), SeMNPV (C) and XcGV (D) obtained by GeneParityPlot analysis (see Methods). Fourteen putative gene clusters of the HaSNPV genome, which are similar to those of AcMNPV (A), LdMNPV (B) and SeMNPV (C), were ordered alphabetically and underlined. The putative gene clusters are indicated in Table 2.

A few HaSNPV ORFs (Ha1–4, Ha6, Ha9–11, Ha13–15, Ha41, Ha69, Ha71 and Ha126; Table 2) have a high degree of amino acid identity (> 90%) to sequences available from HzSNPV (Ma *et al.*, 1993; Cowan *et al.*, 1994; Le *et al.*, 1997). This suggests that the overall homology between HaSNPV and HzSNPV is very high and that they are most likely variants of the same virus species. Sequencing of the HzSNPV genome would reveal whether this assumption is correct.

Unique HaSNPV ORFs

To date, 20 ORFs in the HaSNPV genome are unique to this virus and also do not exhibit significant homology to any other sequences in the GenBank. Most of these ORFs are either very small, encoding putative proteins of up to 100 amino acids (Ha5, Ha7, Ha17, Ha18, Ha40, Ha45, Ha54, Ha104 and Ha112), or contain no common baculovirus transcription initiation sites for early or late gene expression (Ha102, Ha108 and Ha109). Eight ORFs (Ha29, Ha34, Ha99, Ha107, Ha122, Ha125, Ha134 and Ha135) are larger than 100 amino acids and have early and late baculovirus promoter motifs. Ha34 and Ha107 are of interest as they encode putative proteins of 41·1 and 51·2 kDa, respectively. The possible functions of these ORFs are being investigated. For convenience, the ORFs present in the other baculovirus sequences, AcMNPV, BmNPV, OpMNPV, LdMNPV and SeMNPV, are listed in Table 4.

The HaSNPV genome organization

The genomic organization, i.e. the order of genes, of HaSNPV has been studied in a comparative manner using GeneParityPlot analysis (Hu *et al.*, 1998). As the gene order between AcMNPV, BmNPV and OpMNPV is basically

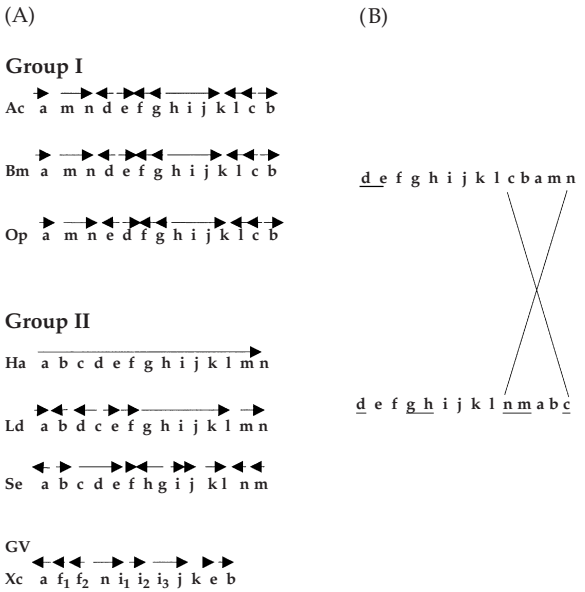

Fig. 4. Alignment of conserved genome clusters of AcMNPV, BmNPV, OpMNPV, SeMNPV, LdMNPV, HaSNPV and XcGV (A) and comparison between group I and group II baculoviruses (B). The arrows indicate the orientation of the cluster and the cluster inversions are underlined.

identical, except for a small number of rearrangements (Ahrens *et al*., 1997; Hu *et al*., 1998; Gomi *et al*., 1999), AcMNPV was taken as a representative example of this group in the analysis (Fig. 3A). A comparison was made between the recently sequenced MNPVs, SeMNPV (IJkel *et al*., 1999) and LdMNPV (Kuzio *et al*., 1999), and XcGV (Hayakawa *et al*., 1999) (Fig. 3B–D). To obtain maximum alignment in the GeneParityPlot analysis, the order of genes had to be reversed for the calculation. By convention, the orientation of a circular baculovirus genome is determined by the relative position of two genes, *polyhedrin* at map unit 0 and *p10* approximately at map unit 90 (Vlak & Smith, 1982). In the initial GeneParityPlot analysis, the orientation of the HaSNPV genome appeared to be reversed for more than 50% of the ORFs compared with AcMNPV and LdMNPV in order to obtain maximum alignment compared with the physical map constructed previously (Chen *et al*., 2000*a*). A similar situation exists for SeMNPV (IJkel *et al*., 1999). The gene organization of HaSNPV is most conserved in the 'central region' of the linearized baculovirus genomes and confirms the supposition of Heldens *et al*. (1998). The left region of the linearized HaSNPV genome displays a considerable number of gene inversions and translocations as deduced from the GeneParityPlot analyses. The right region showed a high degree of gene scrambling (Fig. 3A–D). From these analyses it is concluded that the organization of HaSNPV is highly characteristic and distinct from those of AcMNPV, SeMNPV, LdMNPV and XcGV.

Comparison of the relative gene order between HaSNPV and AcMNPV, SeMNPV, LdMNPV and XcGV revealed the presence of certain gene clusters that are conserved in all baculovirus genomes (Fig. 3, Table 2). The juxtaposition of ORFs can be used as a phylogenetic marker to study the ancestral relationship of baculoviruses, independent of the evolution of individual genes. These clusters are numbered according to their sequential appearance in the GeneParityPlots. Fourteen clusters conserved in all five baculoviruses have been identified (Fig. 3, Table 2). In comparison with a previous analysis (IJkel *et al*., 1999), a small additional cluster, named 12a (Ac28/Ac29 and their homologues), has been identified. Cluster 12, which was conserved in AcMNPV, LdMNPV and SeMNPV, was interrupted in HaSNPV by a Lesel25 homologue. Furthermore, *chitA* (Ha41) has been inserted into cluster 11, whereas Ha40, Ha54 and Ha83 also intervened in this cluster. However, the latter three ORFs are very putative and relatively small genes and, in the cases of Ha40 and Ha54, without apparent transcription control motifs. One additional cluster has been identified in the GeneParityPlot of HaSNPV versus SeMNPV and LdMNPV encompassing Ha126, Ha128 and Ha129 (cluster n, Fig. 3).

Comparison of the cluster organization of HaSNPV with that of other baculoviruses (Fig. 4) suggests that the genomic organization of HaSNPV is more closely related to that of SeMNPV and LdMNPV than to that of group I NPVs (AcMNPV, BmNPV and OpMNPV) or XcGV. This is in agreement with the phylogenetic analysis of single genes such as *egt*, *lef-2*, *dnapol* and *rr* (Chen *et al*., 1997*a*, *b*, 1999; Bulach *et al*., 1999). When the order of gene clusters is taken to represent the baculovirus genome organization, the common structure of group II baculoviruses becomes apparent (Fig. 4A). Within each group, the structural difference is relatively small and predominantly determined by inversions of gene clusters as well as inversions of individual genes (e.g. *polyhedrin*).

Comparison of the two groups showed extensive genomic translocations in addition to cluster inversions. When the inverted genes remained functional, they could be translocated to other genomic regions. These 'jumping' genes can be used as phylogenetic markers to follow baculovirus evolution in retrospect. A common genome structure for group I and group II viruses can be derived, showing a major inversion of a genomic segment containing the cluster c-b-a-m-n (Fig. 4B).

In conclusion, the sequence of the genome of HaSNPV is distinct from other baculoviruses both in gene content and in gene arrangement. Except for three *bro*-related genes and two *iap*-related genes, the HaSNPV genome contains 130 unique ORFs, many of which are shared with other NPVs. Based on the percentage identity of gene homologues, on the phylogeny of some particular genes and on the gene arrangement along the HaSNPV genome, we conclude that HaSNPV, SeMNPV and LdMNPV must have had a common ancestor. The HaSNPV sequence further confirmed the observation that the part of baculovirus genomes flanking DNA *helicase* is highly conserved, possibly as a result of transcriptional or regulatory constraints. By comparing gene clusters, a common structural genomic feature is revealed in group II baculoviruses. A study of the 11 unique putative ORFs (> 100 amino acids) may provide insight in the determinants specifying the SNPV morphotype. From sequence analysis it is also clear that the SNPV and MNPV morphotype is the only taxonomic determinant and it is likely that SNPVs and MNPVs do not represent separate phylogenetic clades.

This research was supported in part by the Royal Academy of Sciences of the Netherlands and the Chinese Academy of Sciences (98CDP008), the Dutch–Israeli Agricultural Research Program (DIARP) (93/20 and 97/29) and the Natural Science Foundation of China (NSFC). Marjo van Staveren, Marleen Abma-Henkens and Paul Mooijman are thanked for their skilful technical assistance. W.F.J.IJ. is supported by a fellowship from the Netherlands Foundation for Chemical Sciences (CW) with financial aid from the Netherlands Organization for Scientific Research (NWO). X.C. received a grant from the Royal Academy of Sciences of the Netherlands (KNAW) and a PhD sandwich fellowship from Wageningen University. Z.H. is a recipient of Hundreds of Talents Program award (STZ-3-01).

References

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

Altschul, S. F., Gish, W., Miller, W., Meyers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–410.

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.

Birnbaum, M. J., Clem, R. J. & Miller, L. K. (1994). An apoptosis-inhibiting gene from a nuclear polyhedrosis virus encoding a polypeptide with Cys/His sequence motifs. *Journal of Virology* **68**, 2521–2528.

Bonfield, J. K., Smith, K. F. & Staden, R. (1995). A new DNA sequence assembly program. *Nucleic Acids Research* **24**, 4992–4999.

Bulach, D. M., Kumar, C. A., Zaia, A., Liang, B. F. & Tribe, D. E. (1999). Group II nucleopolyhedrovirus subgroups revealed by phylogenetic analysis of polyhedrin and DNA polymerase gene sequences. *Journal of Invertebrate Pathology* **73**, 59–73.

Chen, X., Hu, Z., Jehle, J. A., Zhang, Y. & Vlak, J. M. (1997*a*). Analysis of the ecdysteroid UDP-glucosyltransferase gene of *Heliothis armigera* single-nucleocapsid baculovirus. *Virus Genes* **15**, 219–225.

Chen, X., Hu, Z. H. & Vlak, J. M. (1997*b*). Nucleotide sequence analysis of the polyhedrin gene of *Heliothis armigera* single nucleocapsid nuclear polyhedrosis virus. *Virologica Sinica* **12**, 346–353.

Chen, X., IJkel, W. F. J., Dominy, C., Zanotto, P. de A., Hashimoto, Y., Faktor, O., Hayakawa, T., Wang, C. H., Krell, P. J., Hu, Z. & Vlak, J. M. (1999). Identification, sequence and phylogeny of the *lef-2* gene of *Helicoverpa armigera* single-nucleocapsid baculovirus. *Virus Research* **65**, 21–32.

Chen, X., Li, M., Sun, X., Arif, B. M., Hu, Z. H. & Vlak, J. M. (2000*a*). Genomic organization of *Helicoverpa armigera* single-nucleocapsid nucleopolyhedrovirus. *Archives of Virology* **145**, (in press).

Chen, X., Sun, X., Hu, Z. H., Li, M., O'Reilly, D. R., Zuidema, D. & Vlak, J. M. (2000*b*). Genetic engineering of *Helicoverpa armigera* single-nucleocapsid nucleopolyhedrovirus as an improved bioinsecticide. *Journal of Invertebrate Pathology* **76**, 140–146.

Chisholm, G. E. & Henner, D. J. (1988). Multiple early transcripts and splicing of the *Autographa californica* nuclear polyhedrosis virus IE-1 gene. *Journal of Virology* **62**, 3193–3200.

Clem, R. J. & Miller, L. K. (1994). Control of programmed cell death by the baculovirus genes *p35* and *iap*. *Molecular Cell Biology* **14**, 5212–5222.

Cochran, M. A. & Faulkner, P. (1983). Location of homologous DNA sequences interspersed at five regions in the baculovirus AcMNPV genome. *Journal of Virology* **45**, 961–970.

Cowan, P., Bulach, D., Goodge, K., Robertson, A. & Tribe, D. E. (1994). Nucleotide sequence of the polyhedrin gene region of *Helicoverpa zea* single nucleocapsid nuclear polyhedrosis virus: placement of the virus in lepidopteran nuclear polyhedrosis group II. *Journal of General Virology* **75**, 3211–3218.

Crook, N. E., Clem, R. J. & Miller, L. K. (1993). An apoptosis-inhibiting baculovirus gene with a zinc finger-like motif. *Journal of Virology* **67**, 2168–2174.

D'Amours, D., Desnoyers, S., d'Silva, I. & Poirier, G. G. (1999). Poly (ADP-ribosyl)ation reactions in the regulation of nuclear functions. *Biochemical Journal* **342**, 249–268.

Devereux, J., Haeberli, P. & Smithies, O. (1984). A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Research* **12**, 387–395.

Domingo, E., Holland, J. J., Biebricher, C. & Eigen, M. (1995). Quasispecies: the concept and the world. In *Molecular Basis of Evolution*, pp. 171–180. Edited by A. Gibbs, C. Calisher & F. Garcia-Arenal. Cambridge: Cambridge University Press.

Ewing, B. & Green, P. (1998). Basecalling of automated sequencer traces using PHRED. II. Error probabilities. *Genome Research* **8**, 186–194.

Ewing, B., Hillier, L., Wendl, M. C. & Green, P. (1998). Basecalling of automated sequencer traces using PHRED. I. Accuracy assessment. *Genome Research* **8**, 175–185.

Federici, B. A. (1999). Naturally occurring baculoviruses for insect pest control. *Methods in Biotechnology* **5**, 301–320.

Figueiredo, E., Muñoz, D., Escribano, A., Mexia, A., Vlak, J. M. & Caballero, P. (1999). Biochemical identification and comparative

insecticidal activity of nucleopolyhedrovirus isolates pathogenic for *Heliothis armigera* (Lep., Noctuidae) larvae. *Journal of Applied Entomology* **123**, 165–169.

Gettig, R. R. & McCarthy, W. J. (1982). Genotypic variation among wild isolates of *Heliothis spp* nuclear polyhedrosis viruses from different geographical regions. *Virology* **117**, 245–252.

Gomi, S., Majima, K. & Maeda, S. (1999). Sequence analysis of the genome of *Bombyx mori* nucleopolyhedrovirus. *Journal of General Virology* **80**, 1323–1337.

Guarino, L. A. & Summers, M. D. (1986). Interspersed homologous DNA of *Autographa californica* nuclear polyhedrosis virus enhances delayed-early gene expression. *Journal of Virology* **60**, 215–223.

Guarino, L. A., Gonzales, M. A. & Summers, M. D. (1986). Complete sequence and enhancer function of the homologous DNA regions of *Autographa californica* nuclear polyhedrosis virus. *Journal of Virology* **60**, 224–229.

Guarino, L. A., Xu, B., Jin, J. P. & Dong, W. (1998). A virus-encoded RNA polymerase purified from baculovirus-infected cells. *Journal of Virology* **72**, 7985–7991.

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

Heldens, J. G. M., Yi, L., Zuidema, D., Goldbach, R. W. & Vlak, J. M. (1998). A highly conserved genomic region in baculoviruses: sequence and transcriptional analysis of an 11·3 kbp DNA fragment (46·5–55·1 mu) from the *Spodoptera exigua* multicapsid nucleopolyhedrovirus. *Virus Research* **55**, 187–198.

Hu, Z. H., Arif, B. M., Jin, F., Martens, J. W. M., Chen, X. W., Sun, J. S., Zuidema, D., Goldbach, R. W. & Vlak, J. M. (1998). Distinct gene arrangement in the *Buzura suppressaria* single-nucleocapsid nucleopolyhedrovirus genome. *Journal of General Virology* **79**, 2841–2851.

IJkel, W. F. J., Van Strien, E. A., Heldens, J. G. M., Broer, R., Zuidema, D., Goldbach, R. W. & Vlak, J. M. (1999). Sequence and organization of the *Spodoptera exigua* multicapsid nucleopolyhedrovirus genome. *Journal of General Virology* **80**, 3289–3304.

IJkel, W. F. J., Westenberg, M., Goldbach, R. W., Blissard, G. W., Vlak, J. M. & Zuidema, D. (2000). A novel baculovirus envelope fusion protein with a proprotein convertase cleavage site. *Virology* **275**, 30–41.

Kim, D. & Weaver, R. F. (1993). Transcription mapping and functional analysis of the protein tyrosine/serine phosphatase (PTPase) gene of the *Autographa californica* nuclear polyhedrosis virus. *Virology* **195**, 587–595.

King, L. A. & Possee, R. D. (1992). *The Baculovirus Expression System: A Laboratory Guide*, pp. 180–194. London: Chapman and Hall.

Kool, M., Ahrens, C. H., Vlak, J. M. & Rohrmann, G. F. (1995). Replication of baculovirus DNA. *Journal of General Virology* **76**, 2103–2118.

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.

Le, T. H., Wu, T., Robertson, A., Bulach, D., Cowan, P., Goodge, K. & Tribe, D. (1997). Genetically variable triplet repeats in a RING-finger ORF of *Helicoverpa* species baculoviruses. *Virus Research* **49**, 67–77.

Lu, A. & Miller, L. K. (1995). The roles of eighteen baculovirus late expression factor genes in transcription and DNA replication. *Journal of Virology* **69**, 975–982.

Ma, S.-W., Corsaro, B. G., Klebba, P. E. & Fraser, M. J. (1993). Cloning and sequence analysis of a p40 structural protein of *Helicoverpa zea* nuclear polyhedrosis virus. *Virology* **192**, 224–233.

Mikhailov, V. S., Mikhailova, A. L., Iwanaga, M., Gomi, S. & Maeda, S. (1998). *Bombyx mori* nucleopolyhedrovirus encodes a DNA-binding protein capable of destabilizing duplex DNA. *Journal of Virology* **72**, 3107–3116.

Moscardi, F. (1999). Assessment of the application of baculoviruses for control of Lepidoptera. *Annual Reviews of Entomology* **44**, 257–289.

Muñoz, D., Castillejo, J. I. & Caballero, P. (1998). Naturally occurring deletion mutants are parasitic genotypes in a wild-type nucleopolyhedrovirus population of *Spodoptera exigua*. *Applied and Environmental Microbiology* **64**, 4372–4377.

Murphy, F. A., Fauquet, C. M., Bishop, D. H. L., Ghabrial, S. A., Jarvis, A. W., Martelli, G. P., Mayo, M. A. & Summers, M. D. (editors) (1995). *Virus Taxonomy. Sixth Report of the International Committee on Taxonomy of Viruses*. New York: Springer–Verlag.

Okano, K., Mikhailov, V. S. & Maeda, S. (1999). Colocalization of baculovirus IE-1 and two DNA-binding proteins, DBP and LEF-3, to viral replication factories. *Journal of Virology* **73**, 110–119.

Oomens, A. G. P. & Blissard, G. W. (1999). Requirement for gp64 to drive efficient budding of *Autographa californica* multicapsid nucleopolyhedrovirus. *Virology* **254**, 297–314.

O'Reilly, D. R. (1997). Auxiliary genes of baculoviruses. In *The Baculoviruses*, pp. 267–295. Edited by L. K. Miller. New York: Plenum.

Pearson, W. R. (1990). Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods in Enzymology* **183**, 63–98.

Pearson, M. N., Groten, C. & Rohrmann, G. F. (2000). Identification of the *Lymantria dispar* nucleopolyhedrovirus envelope fusion protein provides evidence for a phylogenetic division of the Baculoviridae. *Journal of Virology* **74**, 6126–6131.

Smith, I. R. & Crook, N. E. (1988). *In vivo* isolation of baculovirus genotypes. *Virology* **166**, 240–244.

Sun, X. L., Zhang, G. Y., Zhang, Z. X., Hu, Z. H., Vlak, J. M. & Arif, B. M. (1998). *In vivo* cloning of *Helicoverpa armigera* single nucleocapsid nuclear polyhedrosis virus genotypes. *Virologica Sinica* **13**, 83–88.

Tilakaratne, N., Hardin, S. E. & Weaver, R. F. (1991). Nucleotide sequence and transcript mapping of the *Hin*dIII F region of the *Autographa californica* nuclear polyhedrosis virus. *Journal of General Virology* **72**, 285–291.

Todd, J. W., Passarelli, A. L. & Miller, L. K. (1996). Factors regulating baculovirus late and very late gene expression in transient-expression assays. *Journal of Virology* **70**, 2307–2317.

Van Strien, E. A., IJkel, W. F. J., Gerrits, H., Vlak, J. M. & Zuidema, D. (2000). Characteristics of the transactivator gene *ie-1* of *Spodoptera exigua* multiple-nucleocapsid nucleopolyhedrovirus. *Archives of Virology* **145**, 2115–2133.

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

Wang, J. W., Qi, Y. P., Huang, Y. X. & Li, S. D. (1995). Nucleotide sequence of a 1446 base pair *Sal*I fragment and structure of a novel early gene of *Leucania separata* nuclear polyhedrosis virus. *Archives of Virology* **140**, 2283–2291.

Wishart, M. J., Denu, J. M., Williams, J. A. & Dixon, J. E. (1995). A single mutation converts a novel phosphotyrosine binding domain into a dual-specificity phosphatase. *Journal of Biological Chemistry* **270**, 26782–26785.

Wu, Y. & Carstens, E. H. (1998). A baculovirus single-stranded DNA binding protein, LEF-3, mediates the nuclear localization of the putative helicase P143. *Virology* **247**, 32–40.

Zanotto, P. M. de A., Kessing, B. D. & Maruniak, J. E. (1993). Phylogenetic interrelationships among baculoviruses: evolutionary rates and host associations. *Journal of Invertebrate Pathology* **62**, 147–164.

Zhang, G. (1994). Research, development and application of *Heliothis* viral pesticide in China. *Resources and Environment in the Yangtze Valley* **3**, 1–6.

Zhang, G., Zhang, Y., Ge, L. & Shan, Z. (1981). The production and application of the nuclear polyhedrosis virus of *Heliothis armigera* (Hübner) in biological control. *Acta Phytophylacica Sinica* **8**, 235–240.

Received 9 June 2000; Accepted 15 September 2000