Virology **286,** 7–22 (2001)
doi:10.1006/viro.2001.1002, available online at http://www.idealibrary.com on IDEAL®

RAPID COMMUNICATION

The White Spot Syndrome Virus DNA Genome Sequence

Mariëlle C. W. van Hulten,* Jeroen Witteveldt,* Sander Peters,† Nico Kloosterboer,* Renato Tarchini,† Mark Fiers,† Hans Sandbrink,† René Klein Lankhorst,† and Just M. Vlak*[1]

*Laboratory of Virology, Wageningen University, Binnenhaven 11, 6709 PD Wageningen, The Netherlands; and †Greenomics, Plant Research International, P.O. Box 16, 6700 AA Wageningen, The Netherlands

Received March 19, 2001; returned to author for revision May 4, 2001; accepted May 14, 2001

White spot syndrome virus (WSSV) is at present a major scourge to worldwide shrimp cultivation. We have determined the entire sequence of the double-stranded, circular DNA genome of WSSV, which contains 292,967 nucleotides encompassing 184 major open reading frames (ORFs). Only 6% of the WSSV ORFs have putative homologues in databases, mainly representing genes encoding enzymes for nucleotide metabolism, DNA replication, and protein modification. The remaining ORFs are mostly unassigned, except for five, which encode structural virion proteins. Unique features of WSSV are the presence of a very long ORF of 18,234 nucleotides, with unknown function, a collagen-like ORF, and nine regions, dispersed along the genome, each containing a variable number of 250-bp tandem repeats. The collective information on WSSV and the phylogenetic analysis on the viral DNA polymerase suggest that WSSV differs profoundly from all presently known viruses and that it is a representative of a new virus family. © 2001 Academic Press

Key Words: White spot syndrome virus; WSSV; genome sequence; phylogeny; repeat regions.

Introduction. White spot syndrome virus (WSSV) is a pathogen of major economic importance in cultured penaeid shrimp. The virus is not only present in shrimp but also occurs in other freshwater and marine crustaceans, including crabs and crayfish (*31*). In cultured shrimp, WSSV infection can reach a cumulative mortality of up to 100% within 3–10 days (*30*) and can cause large economic losses to the shrimp-culture industry. The virus was first discovered in Taiwan, from where it quickly spread to other shrimp-farming areas in Southeast Asia (*10*). WSSV initially appeared to be limited to Asia until it was found in Texas and South Carolina in November 1995 (*43*). In early 1999, WSSV was also reported from Central and South America, and it has now also been detected in Europe and Australia (*44*). Intensive shrimp cultivation, inadequate sanitation, and worldwide trade has aggravated the disease incidence in crustaceans and enhanced disease dissemination. As such, WSSV has become an epizootic disease and is not only a major threat to shrimp culture but also to marine ecology (*18*).

WSSV virions are ovoid-to-bacilliform in shape with a tail-like appendage at one end. They circulate ubiquitously in the haemolymph of infected shrimp. The virions contain a rod-shaped nucleocapsid, typically measuring 65–70 nm in diameter and 300–350 nm in length. The nucleocapsids, which contain a DNA-protein core bounded by a distinctive capsid layer giving it a cross-hatched appearance, are wrapped singly into an envelope to shape the virion (*14, 38*). The virus contains a large double-stranded DNA of about 290 kbp, as evidenced from restriction-enzyme analysis (*64*). Based on the analysis of WSSV-specific sequences, it can be concluded that there is genetic variation among WSSV isolates (*32, 62*). This was further confirmed by analysis of WSSV structural proteins from different geographical isolates which showed differential profiles (*63*).

The shape of WSSV virions and nucleocapsids resemble baculoviruses (*60*), but the size of the viral DNA of about 300 kbp is well above the range (100–180 kbp) of baculovirus genomes (*21*). Random terminal sequencing of WSSV DNA inserts of plasmid libraries indicated surprisingly that less than 5% of the translated sequences had homologues in sequence databases (*55*). A few genes, though, were identified with homology to other genes in databases, including those encoding for the large and small subunit of ribonucleotide reductase (*56*), a thymidine–thymidylate kinase (*53*), and a protein kinase (*55*). Phylogenetic analysis of these genes indicated that WSSV and baculoviruses are not closely related. Three major structural WSSV virion protein genes have been identified and their translated proteins showed no relationship with baculovirus structural proteins (*57, 59*). Based on the limited amount of genomic

[1] To whom reprint requests should be addressed. Fax: +31-317-484820. E-mail: just.vlak@viro.dpw.wag-ur.nl.

0042-6822/01 $35.00
Copyright © 2001 by Academic Press
All rights of reproduction in any form reserved.

information available, it was postulated that WSSV may be a member of an entirely new virus family (*56*).

To further study the taxonomic position of WSSV and to allow a detailed understanding of the pathology of this virus in shrimp, we have determined the entire nucleotide sequence of the WSSV genome. Analysis of the 293-kbp circular genome revealed 184 open reading frames (ORFs) of 50 amino acids or more, an unusual long ORF (18 kbp), and 9 regions along the genome with tandem repeat sequences. Many of the predicted proteins have no homology with other viral or cellular genes and hitherto unknown properties. Although the *Paramecium bursaria* Chlorella virus of the Phycodnaviridae with a genome of 330 kbp (*29*) is the largest virus sequenced, the WSSV genome of 293 kbp is at present the largest animal virus genome that has been entirely sequenced.

Results and Discussion. Organization of the WSSV genome. The complete DNA sequence of the WSSV genome was assembled into a circular sequence of 292,967 bp in size. This is close to the 290 kbp estimated by restriction digestion (*64*), but smaller than the 305 kbp reported for a putative WSSV genome of another source (*4*). Although the WSSV sequence was not determined from a clonal WSSV isolate, the sequence heterogeneity was minimal (less than 0.01%). The adenine residue at the translation initiation codon of the major structural virion envelope protein VP28, of which the coding capacity has been confirmed by amino acid sequencing (*59*), was designated as the starting point of the physical map of the WSSV genome (Fig. 1; Table 1).

The WSSV genome has an A+T content of 58.9% uniformly distributed over the genome. The frequency of occurrence of the start codon ATG (1.9%) and stop codon TGA (1.9%) was not different from the expected random distribution (1.8% for both codons). However, a paucity of the stop codons TAA (1.8%) and TAG (1.3%), occurring less often than the expected random distribution (2.6% and 1.8%, respectively), was found. The transcription of WSSV genes has not been studied extensively, and therefore few WSSV specific promoter motifs have been identified. A transcription initiation sequence (TCAc/tTC) has been identified for the large and small subunits of ribonucleotide reductase by 5′RACE (*52*), and this sequence was present almost 50% less frequently in the WSSV genome sequence than expected based on a random distribution.

In total, 684 ORFs starting with an ATG initiation codon and 50 amino acids or larger were located on both strands of the WSSV genome. From these ORFs, 184 ORFs of 51–6,077 amino acids in size with minimal overlap were selected (Fig. 1). These 184 predicted ORFs account for 92% of the genetic information in the WSSV genome. Twenty-five of the 184 ORFs have an overlap of 1–365 bp (Fig. 1). The average distance between the 159 non-overlapping ORFs is 155 bp with a smallest distance of 1 bp and a maximum distance of 1595 bp. ORFs are present on both strands in almost equal proportions (54% forward, 46% reverse), and ORFs frequently (60%) occur in head-to-tail tandem arrays (Fig. 1). The largest cluster of consecutive genes with the same transcriptional orientation contains 12 ORFs (118–129). Based on homologies with other viral or cellular genes in GenBank, only 11 of the 184 WSSV ORFs have been assigned a putative function or have similarity with known genes (Table 1). In contrast, baculoviruses share about 50% of their genes (*21*) and this clearly separates WSSV from this group of viruses. Computer analysis using the minor ORFs overlapping the 184 WSSV ORFs showed no relevant homologies to data in GenBank.

Homologous regions. The WSSV genome was analyzed for the presence of repeats by using the repeat finder program REPuter (*28*). The complete genome was compared to itself to identify perfect direct repeats of minimally 15 bp, and subsequently a circular representation of the genome was generated where repeat regions of 30 bp or longer were connected by a line (Fig. 2a). Nine direct repeat regions with different sizes were found dispersed throughout the genome (Fig. 2a; Table 1). Analysis of these regions revealed that they all consisted of identical repeat units of 250 bp or parts thereof. In accordance with homologous regions in baculoviruses (*13*), the nine repeat regions were designated homologous region (*hr*) 1 to *hr*9. One of these repeats (*hr*4), has previously been described by Van Hulten *et al.* (*58*).

The repeat units of the identified *hrs* were found in both orientations on the WSSV genome (Fig. 2b). Four of the *hrs* consisted of repeat units all in a forward orientation (*hr*1, *hr*3, *hr*5, and *hr*9), three consisted of repeat units all in the reverse orientation (*hr*4, *hr*6 and *hr*8), and two *hrs* contained repeat units in both orientations (*hr*2 and *hr*7) (Fig. 2b). This is also shown in Fig. 2a, where repeat units in the same orientation are connected by lines.

The *hrs* all contain 3–8 repeat units of about 250 bp (Fig. 2b), with a total of 53 repeat units for the WSSV genome. The 53 repeat units were aligned, and part of this alignment is depicted in Fig. 2c with 1 representative repeat unit from each *hr*. A highly conserved domain of 115 bp is present in the center of all the repeat units and is flanked by two more variable domains ("variable domain I" and "variable domain II") of approximately 70 bp each. Based on homology in "variable region I," two types of repeats were distinguished. *Hrs* 1, 2, and 9 belong to type A, and *hrs* 3, 4, 5, 6, 7, and 8 belong to type B. The highly conserved central domain contains an imperfect palindrome of 21 bp, which mainly consists of A+T (Fig. 2c).

The *hrs* are largely located in intergenic regions (Fig. 1), although several short ORFs are present. The WSSV

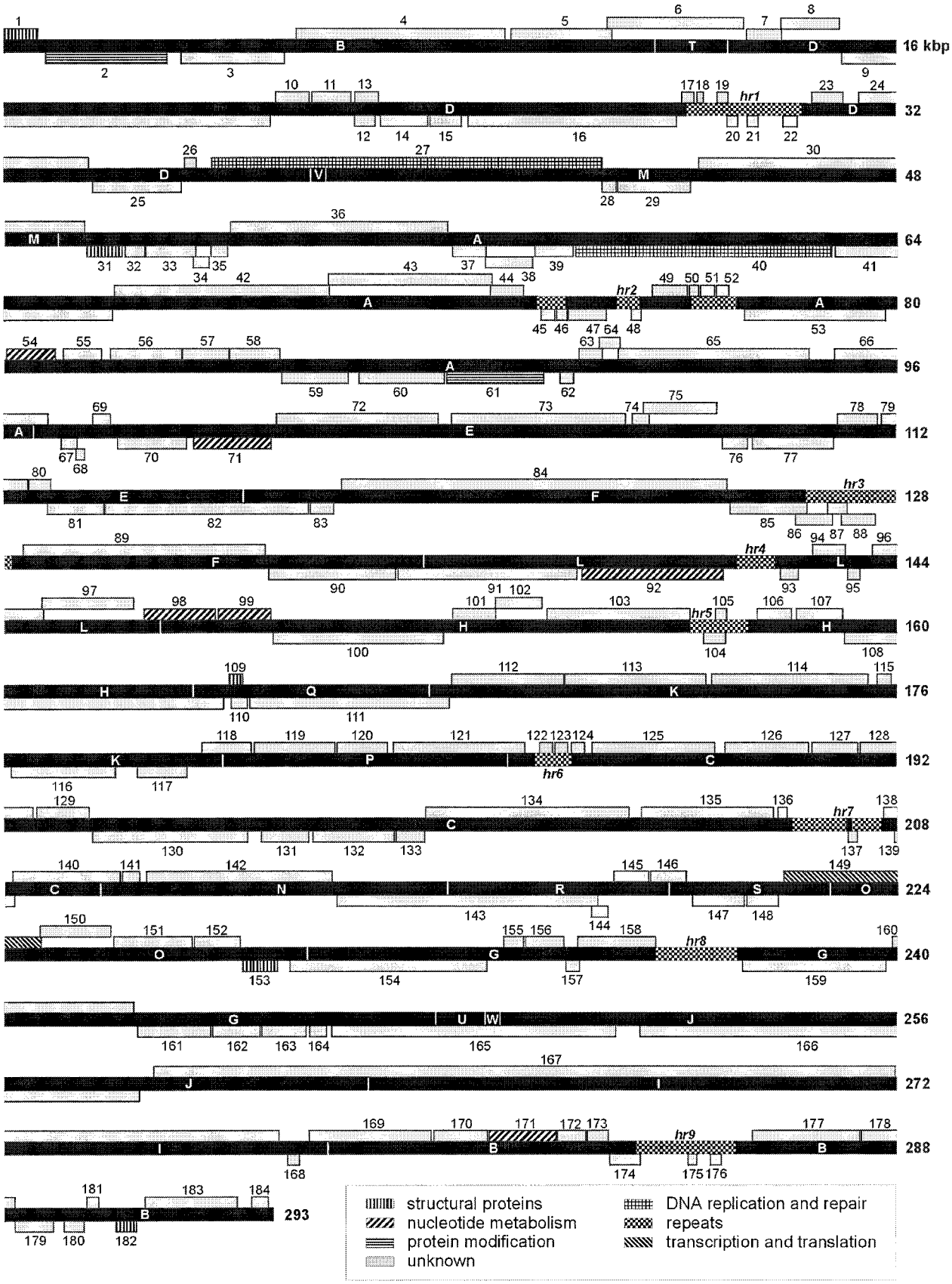

FIG. 1. Linearized map of the circular double-stranded WSSV genome showing the genomic organization. The "A" of the ATG initiation codon of VP28 (ORF1) has been arbitrarily designated position 1. Restriction *Bam*HI sites are shown in the black central bar; fragments are indicated "A" to "W" according to size from the largest (A) to the smallest (W). ORFs are numbered from left to right. ORFs transcribed forward are located above the genome; ORFs transcribed in the reverse orientation are located below. Genes with similar functions are indicated according to the figure key. *Hrs* are presented according to the figure key and numbered (1–9). Numbers on the right indicate number of nucleotides in kbp.

TABLE 1

WSSV ORFs

ORF	Position[a] Start		Stop	Size[b] aa	Mr	pI[c]	Characteristics[d]	Predicted function[e]
1	1	→	615	204	22	4.6	TM; Gene family 1	VP28; envelope protein* (*59*)
2	710	←	2902	730	82	9.3	Similar to *Homo sapiens* protein kinase (NP_055311); Gene family 2	Protein kinase (*57*)
3	3118	←	4989	623	70	7.5	EF-hand calcium-binding domain [PS00018]	
4	5185	→	8970	1261	142	9		
5	9056	→	10879	607	67	7.6		
6	10834	→	13236	800	89	6.5		
7	13311	→	13982	223	25	6.4	SP; ATP/GTP-binding site motif A [PS00017]	
8	13979	→	14890	303	35	6.3	TM	
9	14923	←	20733	1936	216	7	ATP/GTP-binding site motif A [PS00017]; Eukaryotic and viral aspartyl proteases signature and profile [PS00141]; Prenyl group binding site [PS00274]	
10	20837	→	21358	173	19	11.9	SP	
11	21364	→	22161	265	30	5	TM	
12	22201	←	22596	131	13	11.2	TM; Gene family 3	
13	22232	→	22648	138	15	10.3	SP; 2 TMs	
14	22685	←	23581	298	34	5.9		
15	23591	←	24157	188	21	6.1		
16	24265	←	27996	1243	138	6		
17	28024	→	28296	90	11	9.9		
Hr1	28250		30320					
18	28366	→	28530	54	6	9.7		
19	28760	→	28960	66	8	9.7		
20	28957	←	29142	61	7	9.1		
21	29283	←	29468	61	7	9.1		
22	29934	←	30149	71	9	9.4		
23	30426	→	31052	208	24	6.2	ATP/GTP-binding site motif A [PS00017]	
24	31320	→	33485	721	81	7.2	ATP/GTP-binding site motif A [PS00017]; Cell attachment sequence [PS00016]	
25	33532	←	35148	538	62	8.3	Gene family 4	
26	35172	→	35402	76	9	4.3	TM	
27	35571	→	42626	2351	262	7.2	DNA polymerase family B signature [PS00116]; Similar to DNA polymerase of *Saccharomyces cerevisiae* (X61920)	DNA polymerase
28	42667	←	42882	71	8	4.7	SP	
29	42935	←	44281	448	50	5.2		
30	44350	→	49404	1684	186	9.4	TM; Similar to several collagen types	Collagen
31	49448	←	50074	208	23	8.7	SP; Gene family 1	VP24 nucleocapsid protein* (*57*)
32	50129	←	50467	112	13	9.7	Microbodies C-terminal targeting signal [PS00342]	
33	50494	←	51381	295	33	4.2	TM	
34	51341	←	51628	95	11	4.6		
35	51659	←	51952	97	11	4.2		
36	52007	→	55912	1301	144	5.5	3 TMs	
37	55999	←	56601	200	23	9.4	TM	
38	56598	←	57458	286	31	4.8		
39	57509	←	58204	231	26	9.1	TM	
40	58285	←	62892	1535	172	6.2	TM; Similar to *sno* gene of *Drosophila melanogaster* (U95760)	
41	63021	←	65939	972	108	7	TM; Cell attachment sequence [PS00016]	
42	65956	→	69795	1279	143	5.2	TM	

TABLE 1—*Continued*

ORF	Position[a] Start		Stop	Size[b] aa	Mr	pI[c]	Characteristics[d]	Predicted function[e]
43	69737	→	72682	981	109	5.7	ATP/GTP-binding site motif A [PS00017]	
44	72663	→	73253	196	23	4.9		
Hr2	73550		77150					
45	73614	←	73859	81	9	9.8		
46	73915	←	74106	63	7	9.8		
47	74151	←	74831	226	26	4.5	Gene family 5	
48	75246	←	75422	58	6	12.1		
49	75584	→	76210	208	25	8.8	Gene family 6	
50	76237	→	76401	54	7	10.8	Microbodies C-terminal targeting signal [PS00342]	
51	76463	→	76714	83	10	9.7	Microbodies C-terminal targeting signal [PS00342]	
52	76776	→	77000	74	9	9		
53	77284	←	79815	843	96	6.4	Gene family 7	
54	80046	→	80915	289	33	7.1	Thymidylate synthase active site [PS00091]; Similar to *Homo sapiens* thymidylate synthase (NP_001062) and other thymidylate synthases	Thymidylate synthase
55	81077	→	81751	224	25	4.8	Gene family 5	
56	81900	→	83168	422	47	4.8	TM; Gene family 8	
57	83170	→	84000	276	32	8.6		
58	84026	→	84919	297	33	4.7	Cell attachment sequence [PS00016]	
59	85001	←	86197	398	45	9.6		
60	86334	←	87869	511	57	4.2	EF-hand calcium-binding domain [PS00018]	
61	87925	←	89667	580	66	6.9	Similar to *Homo sapiens* protein kinase (NP_009202); Gene family 2	Protein kinase
62	89955	←	90197	80	9	8.5	Cell attachment sequence [PS00016]	
63	90298	→	90744	148	17	10.6	SP	
64	90669	→	91046	125	14	8.7		
65	91003	→	94443	1146	126	4.8	TM; Cell attachment sequence [PS00016]	
66	94903	→	96777	624	69	5.1		
67	97012	←	97242	76	9	10.0		
68	97239	←	97394	51	6	4.8		
69	97587	→	97898	103	12	4.4	SP	
70	98032	←	99252	406	44	8.9		
71	99376	←	100761	461	52	5.4	Similar to fowl adenovirus dUTPase (NP_043869), and other viral and eukaryotic dUTPases	dUTPase
72	100959	→	103865	968	108	6.3	4 TMs	
73	104007	→	107141	1044	118	6.4		
74	107265	→	107570	101	12	10		
75	107467	→	108789	440	48	10.2	TM	
76	108889	←	109341	150	17	8		
77	109433	←	110887	484	53	4.8		
78	110964	→	111779	271	31	4.8		
79	111751	→	112419	222	25	6.5		
80	112426	→	112812	128	15	9	TM	
81	112771	←	113784	337	38	7.5	TM	
82	113793	←	117419	1208	138	6		
83	117465	←	117878	137	16	8	Glycosyl hydrolases family 5 signature [PS00659]	
84	118025	→	124969	2314	289	5.1		
85	125037	←	126416	459	52	7.7	Glucagon/GIP/secretin/VIP family signature [PS00260]	
86	126211	←	126876	221	26	9.8		
Hr3	126388		128112					
87	126782	←	127129	115	13	9.6		
88	127035	←	127634	199	24	9.6		

TABLE 1—*Continued*

ORF	Position[a] Start		Stop	Size[b] aa	Mr	pI[c]	Characteristics[d]	Predicted function[e]
89	128334	→	132644	1436	161	5.4	Nt-dnaJ domain signature [PS00636]	
90	132697	←	134976	759	85	5.6		
91	135031	←	138249	1072	122	5.5	2 TMs	
92	138330	←	140876	848	96	7.8	Similar to ribonucleotide reductase large subunits	Ribonucleotide reductase (large subunit) (*58*)
Hr4	141139		141827					
93	141913	←	142233	106	12	8.4		
94	142498	→	143082	194	22	4.5		
95	143118	←	143342	74	9	8.5		
96	143569	→	144687	372	43	7.1		
97	144689	→	146314	541	63	9	TM	
98	146492	→	147733	413	48	4.8	Ribonucleotide reductase small subunit signature [PS00368]; Similar to viral and eukaryotic ribonucleotide reductase small subunits	Ribonucleotide reductase small subunit (*58*)
99	147798	→	148733	311	36	8.8	SP; similar to *Penaeus japonicus* deoxyribonuclease I (CAB55635); similar to eukaryotic endonucleases	Endonuclease
100	148770	←	151829	1019	117	8.1	Eukaryotic RNA Recognition Motif (RRM) RNP-1 region signature [PS00030]	
101	152015	→	152788	257	29	6.7	TM	
102	152788	→	153624	278	31	6.7		
103	153704	→	156274	856	98	8.2	Ribosomal protein L35 signature [PS00936]; Gene family 7	
Hr5	156319		157366					
104	156538	←	156927	129	14	9.6		
105	156746	→	156955	69	8	11.1		
106	157493	→	158107	204	23	9.6	Gene family 6	
107	158204	→	159031	275	32	7.9	TM	
108	159076	←	163896	1606	174	6.3	SP	
109	163996	→	164238	80	9	12.6		VP15 nucleocapsid protein* (unpublished results)
110	164030	←	164314	94	11	9.3		
111	164346	←	167930	1194	132	5.8	TM; Protein splicing signature [PS00881]; Soybean trypsin inhibitor (Kunitz) protease inhibitors family signature [PS00283]	
112	16800	→	170024	674	76	5.6	Long hematopoietin receptor, gp130 family signature [PS01353]	Class I cytokine receptor
113	170043	→	172577	844	97	6.3	ATP/GTP-binding site motif A [PS00017]; Gene family 9	
114	172701	→	175511	936	108	7.4	TM	
115	175716	→	175964	82	9	4		
116	176120	←	177967	615	71	7.1	TM; Gene family 4	
117	178367	←	179251	294	34	5.5	Gene family 4	
118	179527	→	180405	292	33	4.5	Gene family 10	
119	180442	→	181884	480	51	4.6	SP; Gene family 8	
120	181937	→	182839	300	34	5.8	Gene family 3	
121	182911	→	185286	791	90	8.9	TM	
Hr6	185500		186155					
122	185588	→	185818	76	9	10.5	Microbodies C-terminal targeting signal [PS00342]	
123	185843	→	186073	76	9	11.1	Microbodies C-terminal targeting signal [PS00342]	
124	186135	→	186374	79	9	9.8	Microbodies C-terminal targeting signal [PS00342]	
125	186534	→	188747	737	84	8.0	Vitamin K-dependent carboxylation domain (PS00011); Gene family 9	

TABLE 1—*Continued*

ORF	Position[a] Start		Stop	Size[b] aa	Mr	pI[c]	Characteristics[d]	Predicted function[e]
126	188918	→	190420	500	56	5.2	Prenyl group binding site [PS00274]; Cell attachment sequence [PS00016]; Gene family 4	
127	190500	→	191345	281	32	4.6	Cell attachment sequence [PS00016]; Gene family 10	
128	191349	→	192503	384	43	4.6	SP; Gene family 8	
129	192564	→	193493	309	35	4.6	Gene family 3	
130	193553	←	196321	922	103	4.4		
131	196571	←	197416	281	31	6.1		
132	197480	←	198949	489	56	8.8		
133	198967	←	199479	170	20	9.1		
134	199492	→	203151	1219	135	7.8	Gram-positive cocci surface proteins 'anchoring' hexapeptide [PS00343]; Cell attachment sequence [PS00016]	
135	203364	→	205739	791	87	6.3		
136	205865	→	206029	54	6	10.9		
Hr7	206140		207726					
137	207118	←	207279	53	6	10		
138	207790	→	207999	69	7	7		
139	207992	←	208159	55	6	9.2	TM	
140	208153	→	210057	634	69	5.5		
141	210064	→	210366	100	11	4.8	TM	
142	210519	→	213821	1100	123	5.1	ATP/GTP-binding site motif A [PS00017]; Cell attachment sequence [PS00016]	
143	213918	←	218612	1564	174	6.6	FGGY family of carbohydrate kinases signature 1 [PS00933]; Aminoacyl-transfer RNA synthetases class-II signature 2 [PS00339]	
144	218566	←	218859	97	11	9.5	TM	
145	218912	→	219532	206	23	5.5		
146	219631	→	220260	209	22	4		
147	220309	←	221238	309	34	8.6	TM	
148	221305	←	221874	189	21	5.1		
149	221977	→	224652	891	100	9.2	TM; similar to *Aspergillus nidulans* TATA-box binding protein (AAB57874)	TATA box binding protein
150	224639	→	225898	419	47	5.5		
151	225923	→	227323	466	52	7.2		
152	227329	→	228147	272	31	7.8		
153	228221	←	228835	204	22	9.3	TM; Gene family 1	VP26; nucleocapsid protein* (*59*)
154	229074	←	232613	1179	132	4.2		
155	232928	→	233281	117	13	9.4	TM	
156	233295	→	233978	227	26	8.8	ABC transporters family signature [PS00211]	
157	233982	←	234230	82	9	9.2		
158	234229	→	235626	465	51	8.5	TM; Gram-positive cocci surface proteins 'anchoring' hexapeptide [PS00343]	
Hr8	235672		237156					
159	237222	←	239792	856	96	9	Cell attachment sequence [PS00016]	
160	239925	→	242285	786	88	6.4	Immunoglobulins and major histocompatibility complex proteins signature [PS00290]	
161	242377	←	243678	433	48	4.6		
162	243701	←	244552	283	32	4.9		
163	244556	←	245341	261	30	6.9	Cell attachment sequence [PS00016]	
164	245444	←	245746	100	12	11.3		
165	245849	←	250966	1705	190	7.6		

TABLE 1—*Continued*

ORF	Position[a] Start		Stop	Size[b] aa	Mr	pI[c]	Characteristics[d]	Predicted function[e]
166	251400	←	258392	2330	261	5.7		
167	258666	→	276899	6077	664	6.7	Cell attachment sequence [PS00016]; Leucine zipper pattern [PS00029]	
168	277040	←	277246	68	7	8.2	2 TMs	
169	277425	→	279614	729	85	8.3		
170	279667	→	280632	321	36	5		
171	280683	→	281849	388	43	6.3	ATP/GTP-binding site motif A [PS00017]; Thymidine kinase cellular-type signature [PS00603]; Thymidylate kinase signature [PS01331]	Chimeric Thymidine kinase-Thymidylate kinase (*53*)
172	281869	→	282384	171	20	4.9	Prenyl group binding site [PS00274]	
173	282433	→	282816	127	14	9.1	SP	
174	282829	←	283380	183	22	9.1		
Hr9	283323		285125					
175	284246	←	284401	51	6	8.6		
176	284646	←	284843	65	7	9.4		
177	285406	→	287331	641	74	6.7	Gene family 9	
178	287386	→	288165	259	30	6.6		
179	288183	←	288866	227	26	6.2		
180	289149	←	289343	64	7	8.5	Leucine zipper pattern [PS00029]	
181	289474	→	289680	68	8	11.7		
182	289998	←	290363	121	13	4.2	2 TMs	VP19, envelope protein* (unpublished results)
183	290501	→	292135	544	62	7.1	MIP family signature [PS00221]	
184	292511	→	292804	97	11	8.4	TM	

[a] Position and orientation of the ORFs in the WSSV genome.
[b] Size of ORFs in amino acids (aa) and predicted molecular mass in kDa (Mr).
[c] Predicted isoelectric point (p*I*).
[d] The presence of transmembrane domains (TM) and signal peptides (SP) are indicated. Presence of motifs in the PROSITE databank is indicated and PROSITE Accession Nos. are shown in between square brackets. Similarity with proteins in GenBank, including accession number between brackets, is indicated.
[e] Predicted function; empirically demonstrated functions are indicated with an *.

repeat regions resemble baculovirus *hrs*, which also occur dispersed in all baculovirus genomes sequenced to date (*21*). However, the WSSV *hr* repeat units (250 bp) are much larger as compared to the repeat unit (about 70 bp) in the nucleopolyhedroviruses (NPVs) *hrs*. Furthermore, a key structural feature of the NPV *hrs* is a conserved 30-bp imperfect palindrome located in the center of the 70-bp repeat unit (*21*), whereas the WSSV *hrs* only have a 21-bp imperfect palindrome which is not in the center of the repeat units. The WSSV *hrs* have some resemblance to the *hrs* of granulovirus *Plutella xylostella* (PxGV), as the *hrs* of this virus are larger (105 bp) and only contain a small 15-bp palindromic region in the center of the repeat (*20*).

As WSSV is not closely related to baculoviruses based on genome content and is clearly phylogenetically separated from the baculoviruses on the basis of gene phylogeny (*53, 55, 58*), the presence of *hrs* could be a general feature of large circular viral DNA genomes. A possible essential function of these *hrs* might be their involvement in the replication of viral DNA (*25*), or in enhancement of transcription (*19*), as was shown for baculovirus *hrs*.

Comparison to other WSSV isolates. WSSV sequence data, available in GenBank, were compared to the complete genome sequence presented here and most sequences showed a high degree of homology. Ninety-eight to 100% homology was found with sequences from WSSV isolated from *P. chinesis* [Accession Nos.: U92007 (2424 bp) and U89843 (420bp)], with sequence data from WSSV isolated from *P. monodon* from Vietnam [Accession No.: AJ297947 (941 bp)], and with sequence data from a Taiwan isolate of WSSV (*32*) [Accession Nos.: AF272669 (1400 bp), AF272979 (1250 bp), and AF272980 (1450 bp)]. Wang *et al.* (*62*) analyzed three fragments of WSSV DNA of which two are present in GenBank (C42, Accession No. AF29524, and A6, Accession No. AF295123). The third fragment (LN4) partly overlaps with a sequence present in GenBank (Accession No. AF178573). Compared to our WSSV complete genome sequence, a 100% nt homology was found with the A6

fragment (1416 bp), but the C42 fragment (510 bp) and the LN4 overlapping fragment (2833 bp) were, surprisingly, not present in the complete WSSV genome sequence. A possible explanation is the naturally occurring genetic heterogeneity between WSSV isolates of different origin. The observation of restriction fragment length polymorphisms in different WSSV isolates supports this view (*32, 62*).

To exclude the possibility that the absence of these sequences is the consequence of a sequencing artifact, we have tested the primers used by Wang *et al.* (*62*) (A6, C42, and LN4 primer sets) to amplify the three fragments (1128 bp, 425 bp, and 750 bp, respectively). Furthermore, we tested a different set of primers (control) used for WSSV detection by PCR (*33*). The results of this PCR showed that the WSSV isolate used in this study does not contain fragments LN4 and C42, whereas the A6 and the control PCR fragments were present (data not shown). Assuming that C42 and LN4 are WSSV-specific, this result suggests the existence of WSSV variants or isolates with different genetic complexity.

Gene expression. The complete WSSV genome sequence was searched for transcriptional and translational motifs. Seventy-two percent of the ORFs selected have an ATG in a favorable Kozak context (*26*). From the 27 ORFs located in the *hrs*, only 4 have an ATG in a favorable Kozak context. Furthermore, these ORFs have a small size (average of 89 aa). A TATA box sequence was found in the promoter regions of 46% of the WSSV ORFs. Early transcribed genes, like the ribonucleotide reductase large and small subunit homologues (*52*), contain a TATA box, which is also the case for other potential early transcribed genes like the thymidine–thymidylate (ORF171; Table 1) and dUTPase (ORF73; Table 1) homologue. From the structural proteins which have been identified by N-terminal sequencing (VP28 and VP26: *57*; VP24: *59*; VP19 and VP15: unpublished results), only VP15 and VP19 contain a TATA box sequence, indicating that this sequence is not essential in WSSV for efficient transcription of these putative late genes. No putative late promoter elements have been identified in WSSV yet. The late promoter element "RTAAG," canonical in baculoviruses, was not found in putative WSSV ORF promoter regions and occurs at an average frequency in the WSSV genome sequence. Consensus poly(A) signal sequences are found located in or after the termination codon for 54% of the ORFs, indicating that the WSSV transcripts of these WSSV genes are most probably polyadenylated.

Sequence similarities to proteins in the databases. Homology searches were performed with the major ORFs of the WSSV sequence (Fig. 1; Table 1). The deduced translation products of the 184 ORFs were compared to amino acid sequences in GenBank. ORFs which produced a significant BLASTp score and ORFs with lower but interesting similarities and PROSITE motifs are listed in Table 1. For only 6% of the ORFs, a putative function could be assigned based on homology with GenBank sequences. Three percent of the ORFs were identified as major structural proteins in the WSSV virion confirmed by N-terminal amino acid sequencing (*57, 59*).

DNA replication. Genes involved in DNA replication and repair (such as DNA polymerase, DNA helicase, and DNA binding proteins) are often found in the genome of large DNA viruses. However, for WSSV, we could only identify one of these genes. The presence of a DNA polymerase family B signature (PROSITE entry: PS00116) in ORF27 led us to the identification of a putative DNA polymerase gene. In the BLAST homology search, only low homology (maximum BLASTp score 52) was found with several eukaryotic DNA polymerase genes. Alignment of the putative WSSV polymerase gene and several viral and eukaryotic DNA polymerases showed that all seven conserved DNA polymerase sequence motifs (*8*) were present on the polypeptide encoded by WSSV ORF27. Furthermore, the three conserved regions implicated in DNA polymerase 3′–5′ exonuclease activity (*7*) were also conserved. Despite the presence of these conserved domains, the overall homology of WSSV to several viral and eukaryotic DNA polymerases was only maximally 22%. A notable difference of the WSSV polymerase gene in comparison with other DNA polymerases is its size (2351 amino acids), which is about twice the size of an average DNA polymerase. The additional amino acids of the WSSV polymerase gene are located at both the N and C terminus, as well as in between the conserved DNA polymerase motifs. Except for the DNA polymerase, no other genes involved in DNA replication could be identified based on homologies with such genes in GenBank or based on the presence of conserved domain, present in the PROSITE databank.

Nucleotide metabolism. Most large DNA viruses encode a set of genes involved in nucleotide metabolism, enabling their efficient replication in non-dividing cells (*41*). WSSV encodes three key enzymes for the synthesis of deoxynucleotide precursors for DNA replication: ribonucleotide reductase, thymidine kinase, and thymidylate kinase. The large and the small subunits of ribonucleotide reductase (RR1 and RR2, respectively) were already identified previously (*58*). ORF92 (RR1) and ORF98 (RR2) are located in proximity on the WSSV genome, separated only by 5615 bp, including *hr*4. A chimeric protein consisting of a thymidine kinase and thymidylate kinase (TK-TMK) (*53*) is encoded by ORF171. This chimeric protein is a unique feature of WSSV, as these genes are normally encoded by separate ORFs in other large DNA viruses.

WSSV ORF71 contains a putative homologue of dUTP pyrophosphatase (dUTPase). This enzyme is encoded by many large DNA viruses and is responsible for regulat-

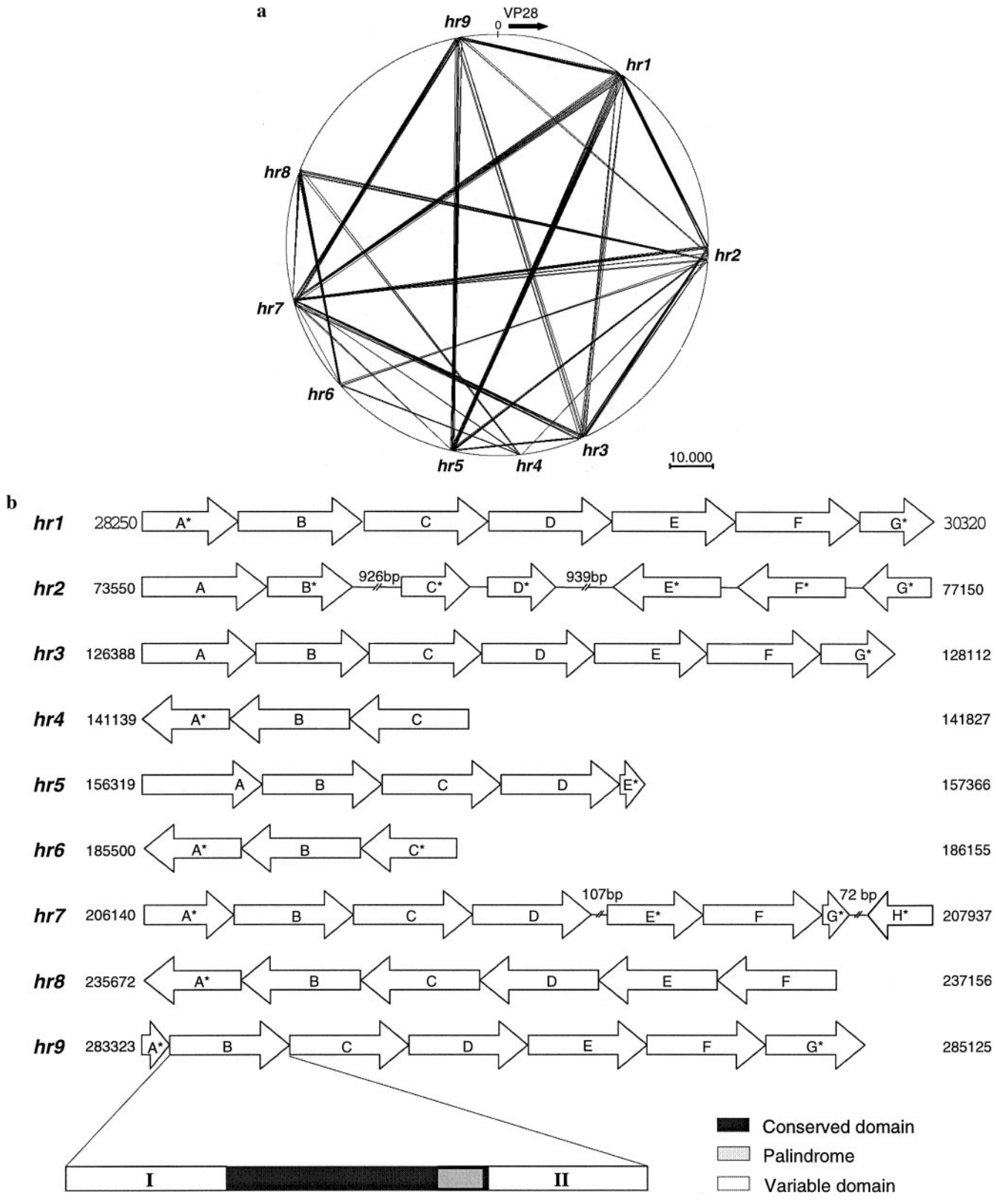

FIG. 2. (a) Circular display of the WSSV genome showing direct repeat regions predicted by REPuter (*28*). Lines connect regions with minimal 30-bp homology. The location and orientation of VP28 is shown as an arrow above the circular genome presentation. A bar at the bottom indicates the scale. The homologous regions (*hrs*) identified on the genome are numbered 1–9. (b) Schematic representation of the repeat structure of the WSSV *hrs*. The repeat units are depicted as arrows, indicating their respective orientation on the genome. Partial repeats are shown by a shorter arrow and an asterisk (*) following its letter. At the bottom of the figure, a schematic representation of the repeat domains is shown as a linear bar with the conserved domain including the perfect palindrome and variable domains indicated as shown in the legend and detailed in (c). (c) Nucleotide sequence alignment of one representative repeat unit from each of the nine *hrs* of the WSSV genome. Shading is used to indicate the occurrence (black, 90%; dark gray, 70%; light gray, 30%) of identical nucleotides. The conserved domain (gray bar), palindrome (black bar), and variable domains (white bar) are indicated underneath the alignment.

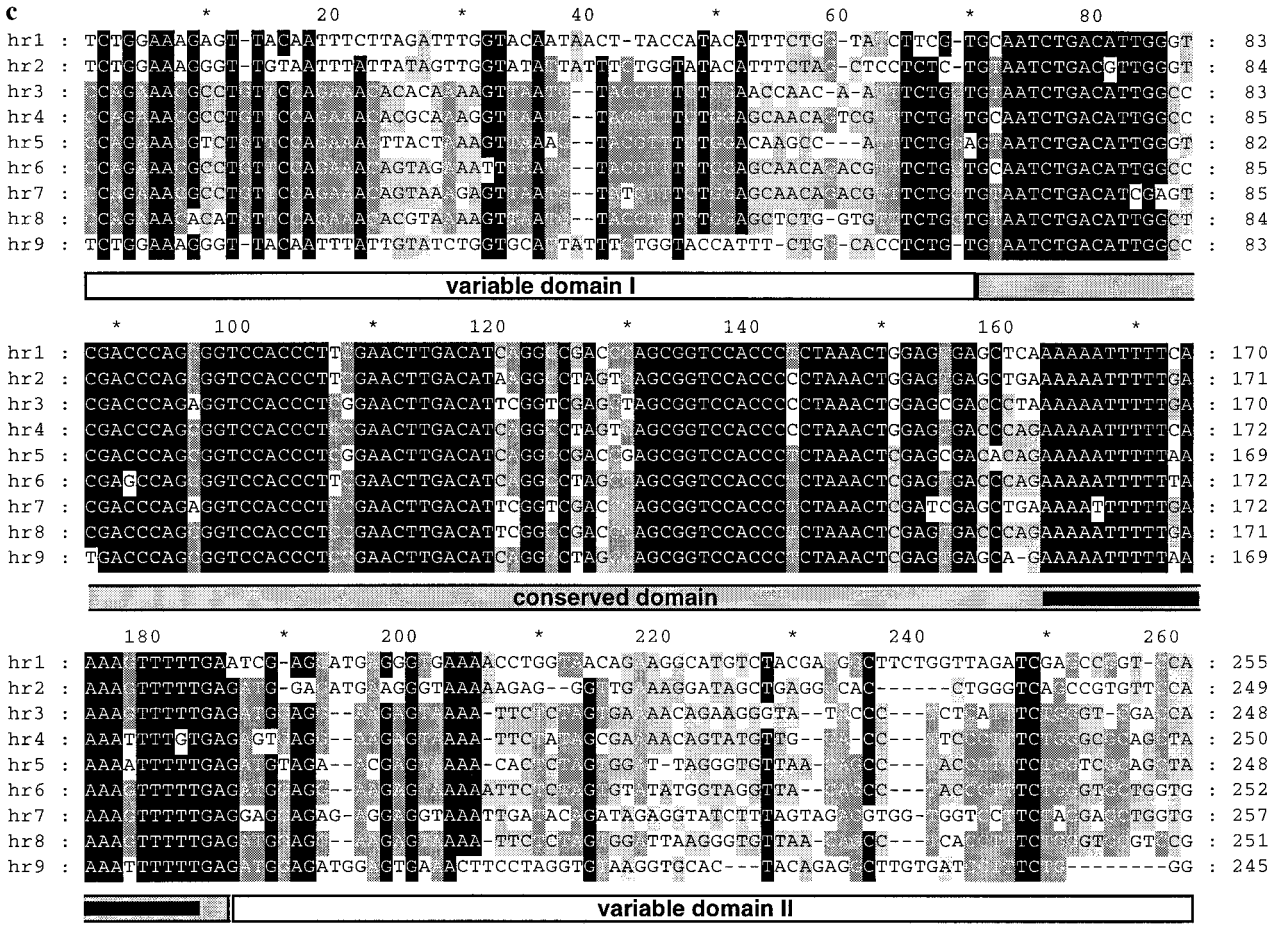

FIG. 2—*Continued*

ing cellular levels of dUTP (*5*). High homology was found with other viral and eukaryotic dUTPase genes. The highest BLASTp score (84) was found with fowl adenovirus dUTPase (Accession No. NP 043869), which showed a 46% amino acid similarity over a stretch of 200 amino acids.

The WSSV genome also contains a highly conserved gene (ORF54) for thymidylate synthase (TSY). Such a gene is, until now, only observed in *Melanoplus sanguinipes* entomopoxvirus (MsEPV), several herpesviruses, and bacteriophages (*1*). Homodimeric TSY catalyzes the methylation of dUMP to the nucleotide precursor dTMP, thus representing an important part of the *de novo* pathway of pyrimidine biosynthesis (*11*). The polypeptide encoded by ORF54 contains the PROSITE motif for the TSY active site (PS00091) and is very similar to TSY from eukaryotes and large DNA viruses. The highest BLASTp score (392) was found with the human TSY, which had a 74% overall amino acid similarity (61% identity) to the WSSV TSY.

ORF99 encodes a putative non-specific endonuclease and can be translated as a 311-amino-acid polypeptide, which includes a putative hydrophobic signal peptide. DNases are encoded by several other large DNA viruses (*2, 29*). The function of an endonuclease in the viral replication cycle is unknown, but it may serve in DNA catabolism during apoptosis (*27*). The WSSV endonuclease homologue has the highest homology (34% overall amino acid similarity and 17% identity) with the DNase I gene of *Penaeus japonicus* (GenBank Accession No. CAB55635), suggesting that this WSSV gene may have been obtained from a crustacean host.

Transcription and mRNA biosynthesis. The N-terminal part of ORF149 encodes a polypeptide with a high similarity to a transcription initiation factor [TATA-box binding protein (TBP)] of eukaryotes (*6*). The highest BLASTp score (41) was found with the *Aspergillus nidulans* TBP (Accession No. AAB57874), where a 150-amino-acid part of the N-terminal region of the WSSV ORF149 has a 40% similarity (22% identity) with this protein. Despite this homology, the internal repeat that is conserved in TBPs (*23*) is not present in WSSV ORF149. Therefore, it is not clear whether ORF149 could have a similar function as eukaryotic TBPs, which play a major role in the activation of eukaryotic genes transcribed by RNA polymerase II and bind to the TATA box promoter element (*6*).

Many viruses encode RNA polymerase subunits, which are involved in mRNA transcription, initiation, elon-

gation, and termination. However, no homologues of these enzymes have yet been identified in the WSSV genome. Furthermore, no RNA helicase, poly(A) polymerase, or other genes involved in transcription and mRNA biogenesis were found and may therefore be absent or too diverged from known homologues to be found based on amino acid homology.

Protein modification. A gene (ORF2) coding for a serine/threonine protein kinase (PK) has recently been identified (*55*). Such enzymes are responsible for the phosphorylation of proteins. Phylogenetic analysis using this gene underscored the unique taxonomic position of WSSV relative to baculoviruses and other large DNA viruses (*55*). All 12 conserved domains of a PK were present in the polypeptide encoding this ORF. A second PK gene homologue has been identified as ORF61. For both proteins, the highest homology in the BLASTp search was found with a *Homo sapiens* PK gene. ORF2 had 30% amino acid similarity (*H. sapiens* PK: NP 055311) and ORF61 29% (*H. sapiens* PK: NP 009202). These two WSSV PK genes have a pairwise amino acid sequence similarity of 45% (27% identity). When included in the unrooted parsonimous phylogenetic tree of PK described by Van Hulten and Vlak (*55*), the two genes have a most recent common ancestor and could therefore be the result of gene duplication.

Immune evasion functions. ORF43 has 43% similarity (22% identity) in a 220-amino-acid-long overlap with a *sno* gene of *Drosophila melanogaster*. The *sno* product is part of a complex which negatively regulates transforming growth factor-β (TGF-β) signaling. This process is important in mediating inflammatory and cytotoxic reactions (*46*). As not much is known about the shrimp immune system, the presence of a putative *sno* gene in the WSSV genome cannot be fully explained, but might be involved in abrogating the host defense response.

The polypeptide encoded by ORF112 contains a "long hematopoietin receptor, gp130 family signature" (PROSITE: PS01353). Genes containing this motif all belong to the class-I cytokine family of receptors in higher eukaryotes (*22*). ORF112 contains sequences similar to a signal, an immunoglobulin-like C2-type domain and a number of fibronectin type III-like modules. Compared to other cytokine receptor genes, ORF112 is somewhat shorter and lacks a transmembrane region. Absence of the transmembrane region may suggest that the protein is produced in a soluble form. Such forms normally arise via both proteolytic processing and alternative splicing (*49*). Because of the homology with the class-I cytokine genes and the presence of the motifs typical for cytokine receptors, it is possible that the ORF112 product is involved in signal transductions related to the defense response system in shrimp.

Structural WSSV virion proteins. Five structural WSSV virion proteins have been identified so far by amino acid sequencing of the individual proteins and reverse genetics. The major envelope protein, VP28 (ORF1), and two major nucleocapsid proteins, VP26 (ORF153) and VP24 (ORF31), have been described before (*57, 59*). Nucleotide and amino acid comparison revealed that these three proteins are homologues and that they may be the result of gene duplication and divergence into proteins with different functions in the nucleocapsid (VP24, VP26) and the envelope (VP28) (*57*). All three ORFs have an initiation codon in a favorable Kozak (*26*) context. Their promoter regions contain stretches of A/T-rich sequences but no consensus TATA box sequence. As a polyadenylation consensus [poly(A)] signal is present for all ORFs, these transcripts are most probably poly-adenylated. Nucleotide sequencing of WSSV confirmed that these three virion structural protein genes are present in single copies.

Internal amino acid sequencing was performed on the envelope protein of 19 kDa (VP19) and N-terminal sequencing on the major nucleocapsid protein of 15 kDa (VP15) (M. C. W. van Hulten, unpublished results). The ORF encoding VP19 is ORF182. The initiation codon of this major envelope protein ORF is in a favorable Kozak context (AAA**ATG**G), a TATA box was identified 254 nucleotides upstream of the ATG and a poly(A) signal sequence is located 59 nt downstream of the termination codon. Two putative transmembrane domains were identified in the amino acid sequence, which could anchor VP19 in the virion envelope.

The amino acid sequence obtained for VP15 showed that this nucleocapsid protein is encoded by ORF109. The first initiation codon in this ORF is not in a favorable Kozak context (position 163996, TTC**ATG**A), whereas the second ATG (position 164052), 57 nt downstream of this ATG, is in a favorable Kozak context AAA**ATG**A for efficient translation. The N-terminal amino acid sequence data suggest that the second ATG is used for translation of this ORF. The TATA box is present 87 nt upstream of the second ATG, and has a preferred location for this ATG. A poly(A) signal is present 62 nt downstream of the translation stop codon. The very basic nature of the ORF109 product (p*I* = 12.6) and its association with the nucleocapsid of the virion may suggest that it is a basic DNA binding protein.

Putative membrane-associated and secreted proteins. The ORFs were analyzed for the presence of putative transmembrane domains (TMs) and signal peptide (SP) sequences. One or more putative TMs were found in 45 ORFs and putative SPs were located in 14 ORFs (Table 1). The proteins containing a putative TM may be associated with membrane structures. Of the structural proteins, the envelope proteins both contain one (VP28) or two (VP19) TMs, which is expected as these proteins are present in the WSSV virion envelope. Also, for VP26 and VP24, a TM was identified, although these proteins are

not located in a membrane but in the nucleocapsid of the WSSV virion. The presence of these hydrophobic domains may well be involved in protein–protein interactions which are necessary for the formation of the nucleocapsid which consists of globular subunits (*14, 38*).

Other genes with interesting properties. ORF3 and ORF60 contain an EF-hand calcium-binding domain (PROSITE Accession No. PS00018), suggesting that their products may belong to the class of the calcium-binding proteins. A further function cannot be assigned to these proteins as no homologues were found in GenBank.

ORF30 encodes a large putative protein (168 kDa) from the collagen family, as the collagen family GXY repeat motif G-x(2)-G-x(2)-G-x(2)-G-x(2)-G-x(2)-G (*54*) is present from amino acid position 161 to 1326 in the 1684-amino-acid-long polypeptide. At the N-terminal side, a predicted transmembrane region is found at position 54–70. No N-terminal signal peptide was found. The function of this collagen homologue in the WSSV genome is not clear, but it is interesting to note that, in viruses, only lymphocystis virus (iridovirus) has a homologue of this protein (*51*).

An extremely large ORF (ORF167) of 18,234 bp coding for a polypeptide of 6077 aa with a theoretical mass of 664 kDa was found on the WSSV genome. The ATG of this ORF (ATC*ATG*G) is in a favorable Kozak context and a poly(A) signal is present 87 nt downstream of this ORF. No consensus sequence for a TATA box and only one poly(A) signal is located in the coding region of this ORF. Several methionines encoded in this ORF are in a favorable Kozak context. Additional analysis will have to prove whether this giant protein encoded by this ORF is indeed expressed. This is the largest ORF to date in viruses. ORFs of about half the size have been identified in herpesviruses, and the proteins encoded by these ORFs are located in the tegument. ORFs of similar sizes as WSSV ORF167 are found in eukaryotes and are members of the family of giant actin-binding/cytoskeletal cross-linking proteins (*47*). No homology with sequences in the GenBank was found and therefore the function of this exceptionally large ORF remains unresolved.

Gene families. The FASTA3 program package (*40*) was used to identify gene families in the WSSV genome. All ORFs, except for those in the *hr* regions, were compared to find genes with homology to each other. Alignments were made of 10 groups of ORFs belonging to the same putative gene family, and which are possibly duplicated in the WSSV genome as they showed pairwise similarities of 40% or higher (Table 1). Gene family 1 consists of three duplicated ORFs (ORF1, 31, and 153) with an amino acid similarity of about 42%, which are the major envelope protein VP28 and two nucleocapsid proteins VP26 and VP24 (*57*). Family 2 contains both putative protein kinase genes (ORFs 2 and 61) which have a 46% similarity. All other families contained ORFs with unknown functions. The highest homology (55% identity, 73% similarity) was found for gene family 7.

Comparison of the WSSV genome with other virus families. WSSV resembles baculoviruses in overall genome structure based on the large circular DNA genome and presence of *hrs*. Baculoviruses share about 50% of their genes, which separates WSSV from baculoviruses, as for WSSV only 6% of its genes have a viral or cellular homologue in GenBank. Furthermore, no specific similarity with other viruses was observed based on gene content. The genes for RR1, RR2, TK, TMK, and PK were used in phylogenetic analysis to compare the position of WSSV relative to other viruses (*53, 55, 58*). Here, we report the analysis of WSSV DNA polymerase, an enzyme that is the prototype for phylogenetic analysis.

Phylogeny of DNA polymerase. The putative DNA polymerase gene (ORF27) was used in an alignment with 14 other viral and 2 eukaryotic polymerases. All seven conserved DNA polymerase sequence motifs and the three conserved regions implicated in DNA polymerase 3′–5′ exonuclease activity (*7, 8*) were identified. Phylogenetic analysis was performed by using the region containing the conserved DNA polymerase motifs. Maximum parsimony phylogenetic trees were obtained by using PAUP, followed by 100 bootstrap replicates to determine the 50% majority-rule consensus tree. Typically for maximum parsimony, bootstrap values of $\geq$70% correspond to a probability of $\geq$95% that the respective clade is a historical lineage.

In the DNA polymerase tree (Fig. 3) the different virus families are all present in clades which are high bootstrap-supported. The herpesviruses included in the tree are present in a branch, which is 100% bootstrap-supported. The same strong support exists for the poxviruses (99%), which are further separated into a chordopoxvirus and an entomopoxvirus branch. The baculoviruses are present in a well bootstrap-supported branch (84%) of the tree and further divided in NPVs and GVs. The *Culex nigripalpus* baculovirus (CuniNPV) was also included, but is not located in the NPV branch, as has been shown before (*37*). The remaining viral DNA polymerase genes, including those from WSSV and other viruses from different virus families, all have a unique position in the tree and do not share a most recent common ancestor. This DNA polymerase tree strengthens the proposition that WSSV is a member of a new virus family.

Conclusions. With a size of 293 kbp, WSSV is the largest animal DNA virus sequenced to date and second in size overall after *Chlorella* virus PBCV-1 (331 kbp) (*29*). The largest animal DNA virus so far sequenced has been the fowlpox virus (FPV), infecting chickens and turkeys, with a size of 288 kbp (*2*). Known large DNA viruses, such as herpesviruses (108–229 kbp) (*34, 36*), iridoviruses (102

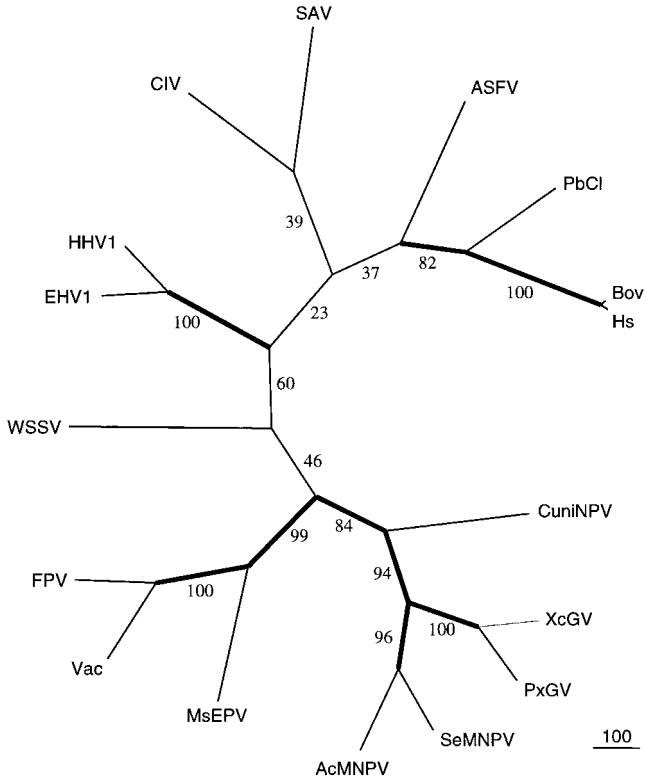

FIG. 3. Bootstrap analysis (100 replicates) of an unrooted phylogenetic tree of DNA polymerase proteins constructed with the PAUP heuristic search algorithm. Numbers at the branches indicate frequency of clusters and frequencies over 70% are indicated by thick lines. The bar at the bottom equals a branch length of 100. DNA polymerase genes used and their accession numbers in brackets: SAV: *Spodoptera frugiperda* ascovirus 1 (CAC19170), EHV1: Equine herpesvirus 1 (NP_041039), HHV1: Human herpesvirus 1 (NP_044632), PbCl: *Paramecium bursaria Chlorella* virus 1 (NP_048532), Bov: *Bos taurus* (P28339), Hs: *Homo sapiens* (S35455), CuniNPV: *Culex nigripalpus* baculovirus (AF274291), XcGV: *Xestia c-nigrum* GV (NP_059280), PxGV: (NP_068312), SeMNPV: *Spodoptera exigua* MNPV (NP_037853), AcMNPV: *Autographa californica* MNPV, (NP_054095), MsEPV: (NP_048107), Vac: Vaccinia virus (NP_063712), FPV: Fowlpox virus (NP_039057), ASFV: African swine fever virus (NP_042783), CIV: Chilo iridescent virus (AAD48150).

kbp) (*51*), baculoviruses (100–180 kbp) (*21*), and poxviruses (145–288 kbp) (Accession No. NC 002642) (*2*) have genomes of considerable size and genetic complexity. The most remarkable property of WSSV is the lack of significant gene sequence homology to any member of these recognized virus families. The presence of an extremely large gene (ORF167), encoding a putative 664-kDa protein adds to the unique character of this virus. The available data including the sequence and phylogeny on DNA polymerase strongly suggest that WSSV is a member of a new virus family. The presence of hrs dispersed along the WSSV genome, a property shared with baculoviruses, may "supergroup" the large circular DNA viruses of arthropods. The analysis of the WSSV genome provides the first complete information of such a large DNA virus of crustaceans, and shows that this virus is distinct from previously identified DNA viruses. An improved understanding of the structure of this virus and its replication, its pathology and gene functions may permit the development of novel intervention strategies.

Materials and Methods. WSSV isolation. The virus isolate used in this study originates from WSSV-infected *Penaeus monodon* shrimps imported from Thailand in 1996 and was obtained as described before (*58*). Crayfish *Procambarus clarkii* were injected intramuscularly with a lethal dose of WSSV. After 1 week, the haemolymph was withdrawn from moribund crayfish and mixed with modified Alsever solution (*42*) as anticoagulant. The virus was purified by centrifugation at 80,000 *g* for 1.5 h at 4°C on a 20–45% continuous sucrose gradient in TN (20 mM Tris, 400 mM NaCl, pH 7.4). The visible virus bands were removed and the virus particles were subsequently sedimented by centrifugation at 45,000 *g* at 4°C for 1 h after dilution with TN. The virus pellet was resuspended in TE (pH 7.5).

WSSV DNA isolation, cloning, and sequence determination. The WSSV DNA was sequenced to a sixfold genomic coverage by using a shotgun approach essentially as described by Chen *et al.* (*12*) for baculovirus *Helicoverpa armigera* NPV. The viral DNA was purified as described in Van Hulten *et al.* (*59*) 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 dephosphorylated *Eco*RV site of pBluescriptSK (Stratagene). After transformation into XL2 blue competent cells (Stratagene), 1510 recombinant colonies were picked randomly. DNA templates for sequencing were isolated by using QIAprep Turbo kits (Qiagen) on a QIAGEN BioRobot 9600. Sequencing was performed by using the ABI PRISM Big Dye Terminator Cycle Sequencing Ready reaction kit with FS AmpliTaq DNA polymerase (Perkin–Elmer) and analyzed on an ABI 3700 DNA Analyzer.

Sequences were base-called by the PRED basecaller and assembled with the PHRAP assembler (*15, 16*). Using the PREGAP4 interface, PHRAP-assembled data were stored in the GAP4 assembly database (*9*). The GAP4 interface and its features were then used for editing and sequence finishing. Consensus calculations with a quality cutoff value of 40 were performed from within GAP4 by using a probabilistic consensus algorithm based on expected error rates output by PHRED. Sequencing PCR products bridging the ends of existing contigs filled remaining gaps in the sequence.

DNA sequence analysis. Genomic DNA composition, structure, and restriction enzyme pattern were analyzed with DNASTAR (Lasergene). Open reading frames (ORFs)

encoding more than 50 amino acids were considered to be protein encoding and hence designated putative genes. DNA and protein comparisons with entries in the sequence databases were performed with FASTA and BLAST programs (*3, 39*). Multiple sequence alignments were performed with the ClustalX computer program (*50*). Phylogenetic analysis was performed with PAUP3.1 program (*48*), using ClustalX to produce input files of aligned protein sequences. A heuristic search was performed, where starting trees were obtained by stepwise addition (starting seed 1), and tree-bisection-reconnection branch-swapping was performed with the MULPARS function. Bootstrap analysis according to Felsenstein (*17*), included in the PAUP package, was used to assess the integrity of the produced phylogeny.

Prediction of signal sequences and transmembrane domains was accomplished by using the PSORT II prediction program which uses the McGeoch's method (*35*) and Von Heijne's method (*61*) for the signal sequence recognition and the Klein *et al.*'s method (*24*) to detect potential transmembrane domains. The program RE-Puter (*28*) was used to identify direct, reversed, and palindromic repeat families.

PCR primers. PCR was performed by using three primer pairs, which amplify the C42, A6, and LN4 fragments described by Wang *et al.* (*62*). The 146F and 146R primer pair described by Lo *et al.* (*33*) was used as a control in the PCR. Purified WSSV DNA, used for sequencing, was used as template in the PCRs. The PCR products were separated in 0.8% agarose gels (*45*).

Nucleotide sequence accession number. The WSSV genome sequence has been deposited in GenBank under Accession No. AF369029.

ACKNOWLEDGMENTS

We thank Marleen Abma-Henkens, Paul Mooyman, and Joost de Groot for their skillful technical assistance. We are grateful to Dr. Douwe Zuidema for reading the manuscript. This research was supported by Intervet International BV, Boxmeer, The Netherlands.

REFERENCES

1. Afonso, C. L., Tulman, E. R., Lu, Z., Oma, E., Kutish, G. F., and Rock, D. L. (1999). The genome of *Melanoplus sanguinipes* entomopoxvirus. *J. Virol.* **73,** 533–552.
2. Afonso, C. L., Tulman, E. R., Lu, Z., Zsak, L., Kutish, G. F., and Rock, D. L. (2000). The genome of Fowlpox virus. *J. Virol.* **74,** 3815–3831.
3. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402.
4. Anonymous. (1999). Genome project aims to combat prawn scourge. *Nature* **397,** 465.
5. Baldo, A. M., and McClure, M. A. (1999). Evolution and horizontal transfer of dUTPase-encoding genes in viruses and their hosts. *J. Virol.* **73,** 7710–7721.
6. Berk, A. J. (2000). TBP-like factors come into focus. *Cell* **103,** 5–8.
7. Bernad, A., Blanco, L., Lazaro, J. M., Marin, G., and Salas, M. (1989). A conserved 3′–5′ exonuclease active site in prokaryotic and eukaryotic DNA polymerases. *Cell* **59,** 219–228.
8. Bernad, A., Zaballos, Z., Salas, M., and Blanco, L. (1987). Structural and functional relationships between prokaryotic and eukaryotic DNA polymerases. *EMBO J.* **6,** 4219–4225.
9. Bonfield, J. K., Smith, K. F., and Staden, R. (1995). A new DNA sequence assembly program. *Nucleic Acids Res.* **24,** 4992–4999.
10. Cai, S. L., Huang, J., Wang, C. M., Song, X. L., Sun, X., Yu, J., Zhang, Y., and Yang, C. H. (1995). Epidemiological studies on the explosive epidemic disease of prawn in 1993–1994. *J. China Fish* **19,** 112–117.
11. Carreras, C. W., and Santi, D. V. (1995). The catalytic mechanism and structure of thymidylate synthase. *Annu. Rev. Biochem.* **64,** 721–762.
12. Chen, X. W., Ijkel, W. F. J., Tarchini, R., Sun, X. L., Sandbrink, H., Wang, H. L., Peters, S., Zuidema, D., Lankhorst, R. K., Vlak, J. M., and Hu, Z. H. (2001). The sequence of the *Helicoverpa armigera* single nucleocapsid nucleopolyhedrovirus genome. *J. Gen. Virol.* **82,** 241–257.
13. Cochran, M. A., and Faulkner, P. (1983). Location of homologous DNA sequences interspersed at five regions in the baculovirus AcMNPV genome. *J. Virol.* **45,** 961–970.
14. Durand, S., Lightner, D. V., Redman, R. M., and Bonami, J. R. (1997). Ultrastructure and morphogenesis of white spot syndrome baculovirus (WSSV). *Dis. Aquat. Org.* **29,** 205–211.
15. Ewing, B., and Green, P. (1998). Basecalling of automated sequencer traces using PHRED. II. Error probabilities. *Genome Res.* **8,** 186–194.
16. Ewing, B., Hillier, L., Wendl, M. C., and Green, P. (1998). Basecalling of automated sequencer traces using PHRED. I. Accuracy assessment. *Genome Res.* **8,** 175–185.
17. Felsenstein, J. (1993). PHYLIP (Phylogeny Interference Package), Version 3.5. Department of Genetics, University of Washington, Seattle, WA.
18. Flegel, T. W. (1997). Major viral diseases of the black tiger prawn (*Penaeus monodon*) in Thailand. *World J. Microbiol. Biotechnol.* **13,** 433–442.
19. Guarino, L. A., and Summers, M. D. (1986). Interspersed homologous DNA of *Autographa californica* nuclear polyhedrosis virus enhances delayed early gene expression. *J. Virol.* **60,** 215–223.
20. Hashimoto, Y., Hayakawa, T., Ueno, Y., Fujita, T., Sano, Y., and Matsumoto, T. (2000). Sequence analysis of the *Plutella xylostella* granulovirus genome. *Virology* **275,** 358–372.
21. Hayakawa, T., Rohrmann, G. F., and Hashimoto, Y. (2000). Patterns of genome organization and content in lepidopteran baculoviruses. *Virology* **278,** 1–12.
22. Hibi, M., Nakajima, K., and Hirano, T. (1996). IL-6 cytokine family and signal transduction: A model of the cytokine system. *J. Mol. Med.* **74,** 1–12.
23. Kim T. K., Nikolov D. B., and Burley S. K. (1993). Co-crystal structure of TBP recognizing the minor groove of a TATA element. *Nature* **365,** 520–527.
24. Klein, P., Kanehisa, M., and DeLisi, C. (1985). The detection and classification of membrane-spanning proteins. *Biochim. Biophys. Acta* **815,** 468–476.
25. Kool, M., Ahrens, C. H., Vlak, J. M., and Rohrmann, G. F. (1995). Replication of baculovirus DNA. *J. Gen. Virol.* **76,** 2103–2118.
26. Kozak, M. (1989). The scanning model for translation: An update. *J. Cell Biol.* **108,** 229–241.
27. Krieser, R. J., and Eastman, A. (1998). The cloning and expression of human deoxyribonuclease. II. Role in apoptosis. *J. Biol. Chem.* **272,** 30909–30914.
28. Kurtz, S., and Schleiermacher, C. (1999). REPuter: Fast computation of maximal repeats in complete genomes. *Bioinformatics*, **15,** 426–427.
29. Li, Y., Lu, Z., Sun, L., Ropp, S., Kutish, G. F., Rock, D. L., and Van

Etten, J. L. (1997). Analysis of 74 kb of DNA located at the right end of the 330-kb chlorella virus PBCV-1 genome. *Virology* **237,** 360–377.

30. Lightner, D. V. (1996). "A Handbook of Shrimp Pathology and Diagnostic Procedures for Diseases of Cultured Penaeid Shrimp." World Aquatic Society, Baton Rouge, LA.

31. Lo, C. F., Ho, C. H., Peng, S. E., Chen, C. H., Hsu, H. C., Chiu, Y. L., Chang, C. F., Liu, K. F., Su, M. S., Wang, C. H., and Kou, G. H. (1996a). White spot syndrome baculovirus (WSBV) detected in cultured and captured shrimp, crabs and other arthropods. *Dis. Aquat. Org.* **27,** 215–225.

32. Lo, C. F., Hsu, H. C., Tsai, M. F., Ho, C. H., Peng, S. E., Kou, G. H., and Lightner, D. V. (1999). Specific genomic fragment analysis of different geographical clinical samples of shrimp white spot syndrome virus. *Dis. Aquat. Org.* **35,** 175–185.

33. Lo, C. F., Lei, J. H., Ho, C. H., Chen, C. H., Peng, S. E., Chen, Y. T., Chou, C. M., Yeh, P. Y., Huang, C. J., Chou, H. Y., Wang, C. H., and Kou, G. H. (1996b). Detection of baculovirus associated with white spot syndrome (WSBV) in penaeid shrimps using polymerase chain reaction. *Dis. Aquat. Org.* **25,** 133–141.

34. Mar Albà, M., Rhiju Das, C., Orengo, A., and Kellam, P. (2001). Genomewide function conservation and phylogeny in the Herpesviridae. *Genome Res.* **11,** 43–54.

35. McGeoch, D. J. (1985). On the predictive recognition of signal peptide sequences. *Virus Res.* **3,** 271–286.

36. Montague, M. G., and Hutchison III, C. A. (2000). Gene content phylogeny of herpesviruses. *Proc. Natl. Acad. Sci. USA*, **97,** 5334–5339.

37. Moser, B. A., Becnel, J. J., White, S. E., Afonso, C., Kutish, G., Shanker, S., and Almira, E. (2001). Morphological and molecular evidence that *Culex nigripalpus* baculovirus is an usnusual member of the *Baculoviridae. J. Gen. Virol.* **82,** 283–297.

38. Nadala, E. C. B., Tapay, L. M., and Loh, P. C. (1998). Characterization of a non-occluded baculovirus-like agent pathogenic to penaeid shrimp. *Dis. Aquat. Org.* **33,** 221–229.

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

40. Pearson, W. R. (1999). Flexible similarity searching with the FASTA3 program package. *In* "Bioinformatics Methods and Protocols" (S. Misener and S. A. Krawetz, Eds.) pp. 185–219. Humana Press, Totowa, NJ.

41. Reichard, P. (1988). Interactions between the deoxyribonucleotide and DNA synthesis. *Annu. Rev. Biochem.* **57,** 349–374.

42. Rodriguez, J., Boulo, V., Mialhe, E., and Bachère, E. (1995). Characterisation of shrimp haemocytes and plasma components by monoclonal antibodies. *J. Cell Sci.* **108,** 1043–1050.

43. Rosenberry, B. (1996). "World Shrimp Farming 1996." Shrimp News International, San Diego, CA, pp. 29–30.

44. Rosenberry, B. (2000). "World Shrimp Farming 2000." Shrimp News International, San Diego, CA.

45. Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989). "Molecular Cloning: A Laboratory Manual," 2nd edition, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

46. Shinagawa, T., Dong, H., Ming, M., Maekawa, T., and Ishii, S. (2000). The sno gene, which encodes a component of the histone deacetylase complex, acts as a tumor suppressor in mice. *EMBO J.* **19,** 2280–2291

47. Sun, Y., Zhang, J., Kraeft, S. K., Auclair, D., Chang, M. S., Liu, Y., Sutherland, R., Salgia, R., Griffin, J. D., Ferland, L. H., and Chen, L. B. (1999). Molecular cloning and characterization of human trabeculin-α, a giant protein defining a new family of actin-binding proteins. *J. Biol. Chem.* **274,** 33522–33530.

48. Swofford, D. L. (1993). PAUP: Phylogenetic Analysis Using Parsimony, Version 3.1. Illinois Natural History Survey, Champaign, IL.

49. Taga T., and Kishimoto T. (1997). Gp130 and the interleukin-6 family of cytokines. *Annu. Rev. Immunol.* **15,** 797–819.

50. Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. (1997). The ClustalX windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **24,** 4876–4882.

51. Tidona, C. A., and Darai, G. (1997). The complete DNA sequence of lymphocystis disease virus. *Virology* **230,** 207–216.

52. Tsai, M. F., Lo, C. F., Van Hulten, M. C. W., Tzeng, H. F., Chou, C. M., Huang, C. J., Wang, C. H., Lin, J. Y., Vlak, J. M., and Kou, G. H. (2000a). Transcriptional analysis of the ribonucleotide reductase genes of shrimp white spot syndrome virus. *Virology* **277,** 92–99.

53. Tsai, M. F., Yu, H. T., Tzeng, H. F., Leu, J. H., Chou, C. M., Huang, C. J., Wang, C. H., Lin, J. Y., Kou, G. H., and Lo, C. F. (2000b). Identification and characterization of a shrimp white spot syndrome virus (WSSV) gene that encodes a novel chimeric polypeptide of cellular-type thymidine kinase and thymidylate kinase. *Virology* **277,** 100–110.

54. Van der Rest, M., and Garrone, R. (1991). Collagen family of proteins. *FASEB J.* **5,** 2814–2823.

55. Van Hulten, M. C. W., and Vlak, J. M. (2001a). Identification and phylogeny of a protein kinase gene of White Spot Syndrome Virus. *Virus Genes* **22,** 201–207.

56. Van Hulten, M. C. W., and Vlak, J. M. (2001b). Genetic evidence for a unique taxonomic position of white spot syndrome virus of shrimp: Genus whispovirus. *In* "Proceedings of the Fourth Asean Conference on Diseases in Aquaculture" (Flegel *et al*., Eds.), in press.

57. Van Hulten, M. C. W., Goldbach, R. W., and Vlak, J. M. (2000a). Three functionally diverged major structural proteins of white spot syndrome virus evolved by gene duplication. *J. Gen. Virol.* **81,** 2525–2529.

58. Van Hulten, M. C. W., Tsai, M. F., Schipper, C. A., Lo, C. F., Kou, G. H., and Vlak, J. M. (2000b). Analysis of a genomic segment of white spot syndrome virus of shrimp containing ribonucleotide reductase genes and repeat regions. *J. Gen. Virol.* **81,** 307–316.

59. Van Hulten, M. C. W., Westenberg, M., Goodall, S. D., and Vlak, J. M. (2000c). Identification of two major virion protein genes of White Spot Syndrome virus of shrimp. *Virology* **266,** 227–236.

60. Van Regenmortel M. H. V., Fauquet, C. M., Bishop, D. H. L., Carstens, E. B., Estes, M. K., Lemon, S. M., Maniloff, J., Mayo, M. A., McGeoch, D. J., Pringle, C. R., and Wickner, R. B. (2000). Virus taxonomy, classification and nomenclature of viruses. *In* "Seventh Report of the International Committee on Taxonomy of Viruses." Academic Press, San Diego, CA.

61. Von Heijne, G. (1986). A new method for predicting signal sequence cleavage sites. *Nucleic Acids Res.* **14,** 4683–4690.

62. Wang, Q., Nunan, L. M., and Lightner, D. V. (2000). Identification of genomic variations among geographic isolates of white spot syndrome virus using restriction analysis and Southern blot hybridization. *Dis. Aquat. Org.* **43,** 175–181.

63. Wang, Q, Poulos, B. T., and Lightner, D. V. (1999). Protein analysis of geographic isolates of shrimp white spot syndrome virus. *Arch. Virol.* **145,** 263–274.

64. Yang, F., Wang, W., Chen, R. Z., and Xu, X. (1997). A simple and efficient method for purification of prawn baculovirus DNA. *J. Virol. Methods* **67,** 1–4.