Journal of General Virology (2016), **97**, 1010–1031 **DOI** 10.1099/jgv.0.000409

Comprehensive annotation of *Glossina pallidipes* salivary gland hypertrophy virus from Ethiopian tsetse flies: a proteogenomics approach

Adly M. M. Abd-Alla,[1] Henry M. Kariithi,[1,2,3] François Cousserans,[4] Nicolas J. Parker,[5] İkbal Agah İnce,[6] Erin D. Scully,[7] Sjef Boeren,[8] Scott M. Geib,[9] Solomon Mekonnen,[10] Just M. Vlak,[3] Andrew G. Parker,[1] Marc J. B. Vreysen[1] and Max Bergoin[4]

Correspondence
Adly M. M. Abd-Alla
a.m.m.abd-alla@iaea.org

[1]Insect Pest Control Laboratories, Joint FAO/IAEA Division of Nuclear Techniques in Food and Agriculture, International Atomic Energy Agency, Vienna, Austria

[2]Biotechnology Research Institute, Kenya Agricultural and Livestock Research Organization, PO Box 57811, Loresho, Nairobi, Kenya

[3]Laboratory of Virology, Wageningen University, 6708 PB Wageningen, The Netherlands

[4]Laboratoire de Pathologie Comparée, Faculté des Sciences, Université de Montpellier, 34095 Montpellier, France

[5]10 Lockhart Close, Kenilworth, Warwickshire CV8 1RB, UK

[6]Department of Medical Microbiology, School of Medicine, Acibadem University, 34752 Ataşehir, Istanbul, Turkey

[7]Grain, Forage and Bioenergy Research Unit, USDA-ARS, University of Nebraska East Campus, Lincoln, NE 68583, USA

[8]Laboratory of Biochemistry, Wageningen University, 6703 HA Wageningen, The Netherlands

[9]Tropical Crop and Commodity Protection Research Unit, USDA-ARS Daniel K. Inouye US Pacific Basin Agricultural Research Centre, Hilo, HI 96720, USA

[10]National Institute for Control and Eradication of Tsetse and Trypanosomosis (NICETT), Addis Ababa, Ethiopia

Glossina pallidipes salivary gland hypertrophy virus (GpSGHV; family *Hytrosaviridae*) can establish asymptomatic and symptomatic infection in its tsetse fly host. Here, we present a comprehensive annotation of the genome of an Ethiopian GpSGHV isolate (GpSGHV-Eth) compared with the reference Ugandan GpSGHV isolate (GpSGHV-Uga; GenBank accession number EF568108). GpSGHV-Eth has higher salivary gland hypertrophy syndrome prevalence than GpSGHV-Uga. We show that the GpSGHV-Eth genome has 190 291 nt, a low G + C content (27.9 %) and encodes 174 putative ORFs. Using proteogenomic and transcriptome mapping, 141 and 86 ORFs were mapped by transcripts and peptides, respectively. Furthermore, of the 174 ORFs, 132 had putative transcriptional signals [TATA-like box and poly(A) signals]. Sixty ORFs had both TATA-like box promoter and poly(A) signals, and mapped by both transcripts and peptides, implying that these ORFs encode functional proteins. Of the 60 ORFs, 10 ORFs are homologues to baculovirus and nudivirus core genes, including three *per os* infectivity factors and four RNA polymerase subunits (LEF4, 5, 8 and 9). Whereas GpSGHV-Eth and GpSGHV-Uga are 98.1 % similar at the nucleotide level, 37 ORFs in the GpSGHV-Eth genome had nucleotide insertions ($n=17$) and deletions ($n=20$) compared with their homologues in GpSGHV-Uga. Furthermore, compared with the GpSGHV-Uga genome, 11 and 24 GpSGHV ORFs were deleted and novel, respectively. Further, 13 GpSGHV-Eth ORFs were non-canonical; they had either CTG or TTG start codons instead of ATG. Taken together, these data suggest that GpSGHV-Eth and GpSGHV-Uga represent two different lineages of the same virus. Genetic differences combined with host and environmental factors possibly explain the differential GpSGHV pathogenesis observed in different *G. pallidipes* colonies.

Received 21 November 2015
Accepted 20 January 2016

The GenBank/EMBL/DDBJ accession number for the GpSGHV-Eth genome sequence is KU050077.

Four supplementary tables are available with the online Supplementary Material.

INTRODUCTION

Tsetse flies (Diptera; Glossinidae) transmit African trypanosomoses, a set of neglected tropical zoonotic diseases with devastating socio-economic impacts in sub-Saharan Africa (Jordan, 1986; Steelman, 1976). Due to a lack of effective vaccines and increasing drug resistance and counterfeits (i.e. substandard drugs that mimic authentic drugs) (Barrett *et al.*, 2011), vector (tsetse fly) control is of critical importance and represents a sustainable trypanosomoses control method (Schofield & Kabayo, 2008). An effective vector control strategy is the application of the sterile insect technique (SIT) as a component of an area-wide integrated pest management approach (Vreysen *et al.*, 2013).

A prerequisite for the application of SIT programmes is mass production of healthy and productive colonies of the target tsetse species. A SIT programme was initiated to eradicate *Glossina pallidipes* from the Rift Valley of Ethiopia (Feldmann *et al.*, 2005). For this programme, a *G. pallidipes* colony was established at the UN Food and Agriculture Organization/International Atomic Energy Agency's Insect Pest Control Laboratory (IPCL), Seibersdorf, Austria, using pupae collected from the target area. However, this colony completely collapsed within 2 years of its establishment (Abd-Alla *et al.*, 2010) due to infection by *Glossina pallidipes* salivary gland hypertrophy virus (GpSGHV; family *Hytrosaviridae*) (Abd-Alla *et al.*, 2007). Up to 85 % of the flies in this colony had salivary gland hypertrophy (SGH) syndrome – a phenomenon that was first described in wild *G. pallidipes* populations (Whitnall, 1934) and later observed in other *Glossina* species (Kariithi *et al.*, 2013c).

GpSGHV can establish a chronic non-debilitating asymptomatic infection and an acute symptomatic infection in the adult insect host (Boucias *et al.*, 2013). Whereas asymptomatic virus infection has no apparent host fitness cost, symptomatic infection, which is characterized by severe and extensive SGH symptoms, causes reproductive dysfunction and colony collapse (Abd-Alla *et al.*, 2007; Jaenson, 1978a, b). Overt SGH occurs in *G. pallidipes*, although GpSGHV infection is largely asymptomatic in other *Glossina* species (Kariithi *et al.*, 2013a). Expression of overt SGH in *G. pallidipes* can be successfully suppressed by oral administration of antiviral drugs (Abd-Alla *et al.*, 2014), and by modification of fly feeding and colony handling protocols (Abd-Alla *et al.*, 2013). Asymptomatic GpSGHV infection potentially represents viral latency (Kariithi *et al.*, 2013c) and the occurrence of SGH symptoms is an exception rather than a rule.

GpSGHV pathogenesis and SGH prevalence differ from one *G. pallidipes* colony to another. For instance, despite a high prevalence of asymptomatic infection in the IPCL *G. pallidipes* colony (originating from Uganda), SGH prevalence averages at <10 % and the colony has been stable for almost three decades (Abd-Alla *et al.*, 2010). This prevalence markedly differs from the 85 % SGH prevalence in another IPCL *G. pallidipes* colony (originating from Ethiopia in 2001), despite the two colonies being maintained under the same insectary conditions (Abd-Alla *et al.*, 2010, 2011), suggesting that the different pathogenesis could be the result of two different strains of the same virus.

Here, we hypothesize that (viral and host) genetic factors contribute to the expression of overt SGH symptoms. We tested this hypothesis by complete genome sequence annotation of GpSGHV isolated from an Ethiopian *G. pallidipes* colony and compared it with the reference GpSGHV genome (GenBank accession number EF568108), which originated from Uganda (Abd-Alla *et al.*, 2008). We enhanced the GpSGHV genome annotations by mapping onto the virus genomic sequence transcripts and peptides obtained from RNA sequencing (RNA-Seq) and MS data, respectively, which were analysed from hypertrophied salivary gland (HSG) extracts.

RESULTS AND DISCUSSION

The current study was conceived from the observed differential pathogenesis of GpSGHV in different *G. pallidipes* colonies. We used a combination of proteogenomic and transcriptomic mapping approaches (Nesvizhskii, 2014) to complement the annotation of ORFs in the GpSGHV-Eth genome. In addition to using peptides identified from flies infected by GpSGHV-Eth, we also included data obtained from previous proteomics of the GpSGHV-Uga (Kariithi *et al.*, 2010, 2011, 2013b, 2016). Moreover, we compared the GpSGHV-Eth genome to the GpSGHV-Uga genome, which was annotated *in silico* (Abd-Alla *et al.*, 2008).

General features of the GpSGHV-Eth genome

The sequencing results revealed that the GpSGHV-Eth genome had a size of 190 291 nt and a low (27.9 %) G+C content (the GpSGHV-Uga genome has a size of 190 032 nt; 28.0 % G+C content). From this genomic sequence, a total of 319 ORFs (methionine-initiated ORFs; ≥ 50 aa) were predicted, of which 174 ORFs had minimal overlaps and were therefore considered to encode putative viral proteins (Table 1). The 174 putative ORFs were evenly distributed on both strands (51.15 % forward; 48.85 % reverse), mostly arranged in unidirectional clusters (Fig. 1). The BLAST analyses revealed that the majority of the putative GpSGHV-Eth ORFs were nearly identical to the homologous GpSGHV-Uga ORFs with only minor differences at the amino acid level (Table S1, available in the online Supplementary Material). Overall, the GpSGHV-Eth and GpSGHV-Uga genomes were 98.1 % identical at the nucleotide level (Fig. 2). Analysis of the nucleotide sequences upstream of the first methionine residues revealed that 89.7 % ($n=156$) of the GpSGHV-Eth ORFs contained a putative TATA-like box promoter element. Furthermore, 83.9 % ($n=146$) contained the canonical polyadenylation [poly(A)] signal, whilst 35.8 % ($n=62$) had putative

Table 1. Annotation of 174 ORFs potentially expressed in the GpSGHV-Eth genome

Homologies of the GpSGHV-Eth ORFs to other known virus and/or cellular genes and to the corresponding ORFs in GpSGHV-Uga are shown in columns 2–7. Of the 174 putative ORFs, protein products of 86 ORFs were detected by LC-MS/MS. Out of the 86 proteins, 10, 15, 20 and 12 were classified as envelope, nucleocapsid, tegument and virion proteins, respectively, based on the detection of their homologues in highly purified GpSGHV-Uga virions by LC-MS/MS (Kariithi *et al.*, 2013b). The remaining 29 of the 86 proteins without specific localization were designated infected cell-specific viral proteins (ICSVPs). The annotation of the structural features of the ORFs, the presence of putative transcriptional signals, the peptide and/or transcript mapping on the viral genome, and the expression levels of the ORFs are also indicated. ORFs under positive selection pressure are indicated in bold (see discussion in main text and details in Table S4). The gene expression levels were normalized with fragments per kilobase of exon per million mapped reads (FKPM). The mean number of reads representing a given nucleotide in the reconstructed sequence is shown as mean coverage. ORFs marked 'Yes' on transcript (T) mapping were those mapped by exonic transcripts, i.e. read alignments were completely (100%) contained within the respective exons defined by the database used in RNA-seq quantification.

		Best BLAST match (description of homologies)						Transcription signals			Transcript (T) and peptide (P) mapping		Transcript expression levels		
ORF (position and orientation)	Homology to viral and/or cellular proteins (localization)	Best match	Identity (%)	GenBank accession no.	*E* value	Score	Functional and/or structural annotation	TATA-like box (position)	Poly(A) signal (position)	G/T/ATAAG late motif (position: T of TAAG)	T	P	Mean coverage	Raw read count	FPKM
SGHV-Eth001 (1→2088)	PIF-0 (P74) *Spodoptera litura* nucleopolyhedrovirus (envelope protein)	SGHV-Uga001	99.71	YP_001686949.1	0.0	1442	TM; SP; isoleucine-rich activation motif	TATATT (190205)	AATAAT (2106)	TGAATAAA**TAAG** (190286)	Yes	Yes	1126.4	23287	16093
SGHV-Eth002 (3071←2091)	(Tegument protein)	SGHV-Uga002	100.00	YP_001686950.1	0.0	671		TATAAT (3100)	AATAAA (2046)	Not found	Yes	Yes	310.8	3019	4441
SGHV-Eth003 (3785←3237)		SGHV-Uga003	100.00	YP_001686951.1	2.00e–130	375		TATAAA (3839)	AATAAA (3238)	TATAAAAT**TAAG** (3831)	No	No	126.2	686	1803
SGHV-Eth004 (5440←4382)	ODV-E66 protein *Euproctis pseudoconspersa* nucleopolyhedrovirus	SGHV-Uga005	99.72	YP_001686953.1	0.0	724	TM helix; INM-SM; chondroitin AC/ alginate lyase	TATAAT (5453)	AATTATT (4349)	ATGTA**TAATAAG** (5448)	Yes	No	515.6	5406	7366
SGHV-Eth005 (6580←5495)	Lecithin-cholesterol acyltransferase *Pseudomonas* sp. (ICSVP)	SGHV-Uga006	98.90	YP_001686954.1	0.0	719	Lecithin : cholesterol/ phospholipid : diacylglycerol acyltransferase	Not found	AATAA (5450)	TACACAAA**TAAG** (6590)	Yes	Yes	77.4	832	1105
SGHV-Eth006 (6718→7779)	D-3-Phosphoglycerate dehydrogenase *Clostridium ultunense* (tegument protein)	SGHV-Uga007	98.87	YP_001686955.1	0.0	691		TATAAA (6646)	AATAAA (7800)	ATAAATCT**TAAG** (6655)	Yes	Yes	48.5	510	693
SGHV-Eth007 (8621←7809)	(Virion protein)	SGHV-Uga008	97.79	YP_001686956.1	0.0	529		TATATT (8712)	AATAAA (7798)	GTTGCTGA**TAAG** (8698)	Yes	Yes	35.3	284	504
SGHV-Eth008 (8634→10868)	MAL7P1.132 *Plasmodium falciparum* 3D7 (ICSVP)	SGHV-Uga009	92.73	YP_001686957.1	0.0	1347	SWI/SNF-related matrix-associated actin-dependent regulator of chromatin	TATAAT (8583)	AATAAA (10880)	Not found	Yes	Yes	719.3	15916	10276
SGHV-Eth009 (14184←10894)	**ORF MSV156 *Melanoplus sanguinipes* entomopoxvirus (tegument protein) (desmoplakin)**	**SGHV-Uga010**	**94.04**	**YP_001686958.1**	**0.0**	**1999**	**Coiled coils; proline-rich; BP-NLS; PPase-tensin**	**TATATT (14259)**	**AATAAA (10879)**	AAATTGTA**TAAG** **(14242)**	**Yes**	**Yes**	**591.2**	**19264**	**8446**
SGHV-Eth010 (14759←14184)	ORF AMV179 *Amsacta moorei* entomopoxvirus (ICSVP)	SGHV-Uga011	98.44	YP_001686959.1	2.00e–129	374		TATAAA (14813)	AATAA (14043)	Not found	Yes	Yes	160.8	917	2297
SGHV-Eth011 (15876←14818)		SGHV-Uga012	99.15	YP_001686960.1	0.0	691		TATATT (15984)	AATATAA (14815)	Not found	Yes	No	13.5	141	192
SGHV-Eth012 (16760←16554)		No hits					TM helix	Not found	Not found	Not found	No	No			
SGHV-Eth013 (16935→17501)	UDP-glucose-6 deshydrogenase *Pseudobutyrivibrio ruminis*	SGHV-Uga016	100.00	YP_001686964.1	4.00e–83	253	TM	TATATAA (16760)	AATAAA (17504)	AGATAAGA**TAAG** (16878)	Yes	No	487.7	2738	6968
SGHV-Eth014 (18701←17514)	Putative ubiquitin ligase *Feldmannia* sp. virus	SGHV-Uga017	89.39	YP_001686965.1	0.0	694	Zinc finger; RING/ U-box fold	TATAAA (18795)	AATAAA (17452)	Not found	Yes	No	1565.5	18414	22366
SGHV-Eth015 (19087←18923)		No hits					TM helix	TATAT (19214)	Not found	Not found	No	No	28.6	47	411

Table 1. cont.

ORF (position and orientation)	Homology to viral and/or cellular proteins (localization)	Best BLAST match (description of homologies)					Functional and/or structural annotation	Transcription signals			Transcript (T) and peptide (P) mapping		Transcript expression levels		
		Best match	Identity (%)	GenBank accession no.	*E* value	Score		TATA-like box (position)	Poly(A) signal (position)	G/T/ATAAG late motif (position: T of TAAG)	T	P	Mean coverage	Raw read count	FPKM
SGHV-Eth016 (20108←19947)		No hits					NAD-specific glutamate dehydrogenase domain	TATATA (20173)	Not found	Not found	No	No	130.8	210	1870
SGHV-Eth017 (20613←20314)	Uncharacterized conserved protein (COG3868)	No hits						TATAAA (20703)	AATAAA (20306)	Not found	No	No	49.5	147	707
SGHV-Eth018 (22022←20721)	Putative ubiquitin ligase *Feldmannia* sp. virus	SGHV-Uga019	90.97	YP_001686967.1	0.0	805	Zinc finger; RING/U-box fold	TATAAA (22101)	AATAAA (20699)	Not found	Yes	No	258.7	3335	3696
SGHV-Eth019 (22943←22677)		No hits					TM helix	TATAAA (23032)	Not found	Not found	No	No			
SGHV-Eth020 (23443→23970)		SGHV-Uga020	99.43	YP_001686968.1	2.00e–123	357		TATAAA (23375)	AATAAA (24150)	Not found	Yes	No	52.3	273	746
SGHV-Eth021 (25248←23971)		SGHV-Uga021	99.06	YP_001686969.1	0.0	856		TATATAA (25289)	AATAAT (23921)	Not found	Yes	No	83.4	1055	1191
SGHV-Eth022 (25256→26284)	ORF AMV156 *Amsacta moorei* entomopoxvirus	SGHV-Uga022	98.83	YP_001686970.1	0.0	693		TATATT (25147)	AATAAA (26335)	Not found	Yes	No	9.1	93	130
SGHV-Eth023 (26341→27468)	ORF AMV253 *Amsacta moorei* entomopoxvirus (possible surface protein)	SGHV-Uga023	99.73	YP_001686971.1	0.0	742	AARP2CN (NUC121) domain SM00785	TATATA (26295)	AAATAA (27589)	Not found	Yes	No	4.4	49	63
SGHV-Eth024 (29032←27479)	ORF098 *Orgyia pseudotsugata* nucleopolyhedrovirus	SGHV-Uga024	96.14	YP_001686972.1	0.0	990	TM helix	TATAAT (29058)	AATTAA (27442?)	Not found	No	No	329.8	5075	4712
SGHV-Eth025 (29124→29873)	ORF55 *Leucania separata* nucleopolyhedrovirus	SGHV-Uga025	98.00	YP_001686973.1	2.00e–176	497		TATAAT (29055)	AATAAA (29885)	Not found	Yes	No	16.8	125	240
SGHV-Eth026 (30249→31292)	(Possible surface protein)	SGHV-Uga026	99.71	YP_001686974.1	0.0	698		TATAAA (30196)	AATAAA (31310)	Not found	Yes	No	120	1240	1714
SGHV-Eth027 (31318→32688)	ORF010 *Pieris rapae* granulosis virus type II chitinase (virion protein)	SGHV-Uga027	99.33	YP_001686975.1	0.0	900	SP; chitinase family 18	TATATA (31194)	AATAAA (32703)	CATAT**TAATAAG** (31296)	Yes	Yes	2892	39257	41317
SGHV-Eth028 (33059←32772)		No hits					Rifin_STEVOR, Rif/Stevor family	TATATA (33106)	AATAAA (32733)	Not found	No	No	88.2	251	1258
SGHV-Eth029 (33186→33776)	ORF 091R invertebrate iridovirus 25 GIY-YIG nuclease-like protein	SGHV-Uga028	99.49	YP_001686976.1	5.00e–137	393	Coiled coils; GIY-YIG superfamily family domain	TATATA (33101)	AATAAA (33876)	Not found	Yes	No	9.9	58	142
SGHV-Eth030 (33807→34487)	NADH ubiquinone oxidoreductase	SGHV-Uga029	98.68	YP_001686977.1	1.00e–156	446	Staphylococcal AgrD protein	TATAAA (33765)	AATAAA (34486)	ACTGTTTG**TAAG** (33791)	Yes	No	34.4	232	492
SGHV-Eth031 (34895←34497)	(Virion protein)	SGHV-Uga030	100.00	YP_001686978.1	2.00e–93	277	DNA breaking-rejoining enzyme	TATAT (34947)	AATAAA (34451)	Not found	Yes	Yes	5736.1	22660	81948
SGHV-Eth032 (35834←34980)	(Nucleocapsid protein)	SGHV-Uga031	100.00	YP_001686979.1	0.0	575	Leucine/isoleucine-rich; HDAC domain	TATAAA (35848)	AATAAT (34938)	Not found	Yes	Yes	211.8	1793	3026
SGHV-Eth033 (36661←35885)	LEF-9 ORF24 *Gryllus bimaculatus* nudivirus (ICSVP)	SGHV-Uga032	99.61	YP_001686980.1	0.0	526	Coiled coils	TATATA (36797)	AATAAA (35844)	Not found	Yes	Yes	87	669	1242
SGHV-Eth034 (38003←36960)	LEF-9 ORF24 *Gryllus bimaculatus* nudivirus (ICSVP)	SGHV-Uga033	99.71	YP_001686981.1	0.0	705		TATAAA (38029)	AATAAA (36870)	Not found	Yes	Yes	621.7	6427	8883
SGHV-Eth035 (39105←38056)	**ORF AMV260 *Amsacta moorei* entomopoxvirus (virion protein)**	**SGHV-Uga034**	**98.57**	**YP_001686982.1**	**0.0**	**659**		**TATATA (39172)**	**AATAAA (38046)**	TATCAAAA**TAAG** **(39124)**	**Yes**	**Yes**	**601.5**	**6253**	**8593**
SGHV-Eth036 (39378←39118)	ORF MSV238 *Melanoplus sanguinipes* entomopoxvirus (thymidylate synthase) (ICSVP)	SGHV-Uga035	100.00	YP_001686983.1	1.00e–54	175	TM helix; SP	TATAAA (39484)	AATAAA (39058)	Not found	Yes	Yes	944.8	2442	13501

Table 1. cont.

ORF (position and orientation)	Homology to viral and/or cellular proteins (localization)	Best BLAST match (description of homologies)					Functional and/or structural annotation	Transcription signals			Transcript (T) and peptide (P) mapping		Transcript expression levels		
		Best match	Identity (%)	GenBank accession no.	*E* value	Score		TATA-like box (position)	Poly(A) signal (position)	G/T/ATAAG late motif (position: T of TAAG)	T	P	Mean coverage	Raw read count	FPKM
SGHV-Eth037 (40006←39662)	ORF067 white spot syndrome virus dihydrofolate reductase-thymidylate synthase (envelope protein)	SGHV-Uga036	100.00	YP_001686984.1	5.00e–77	234	SP; TM	TATAAA (40048)	AATAAT (39648)	Not found	Yes	Yes	288.4	985	4120
SGHV-Eth038 (40289←40053)		SGHV-Uga037	88.24	YP_001686985.1	1.00e–21	90.1		TATATTT (40363)	AATAAA (40038)	Not found	Yes	No	17866.9	41925	255255
SGHV-Eth039 (44292←40702)	Maltodextrin glycosyltransferase (tegument protein)	SGHV-Uga038	97.70	YP_001686986.1	0.0	2246	Coiled coils; SP; RGD motif	Not found	AAATAA (40644)	TTTAA**TAATAAG** (44304)	Yes	Yes	5406.5	192225	77240
SGHV-Eth040 (45360←44359)	**HSP90-like ATPase (envelope protein)**	**SGHV-Uga039**	**89.60**	**YP_001686987.1**	**4.00e–166**	**478**	**TM; SP; threonine-rich**	**TATAAT (45383)**	**AATAAT (44309)**	CTTTA**TAATAAG** **(45378)**	**Yes**	**Yes**	**14559.5**	**144442**	**208006**
SGHV-Eth041 (45680→48382)	LEF-8 *Spodoptera litura* nucleopolyhedrovirus (ICSVP)	SGHV-Uga040	99.78	YP_001686988.1	0.0	1849		TATATT (45572)	AATAAA (48402)	Not found	Yes	Yes	135	3614	1929
SGHV-Eth042 (49675←48437)	Casein kinase isoform I-d ORF MSV214 *Melanoplus sanguinipes* entomopoxvirus (envelope protein)	SGHV-Uga041	98.79	YP_001686989.1	0.0	831	PD-(D/E) XK nuclease fold	TATAAT (49779)	AATAAA (48413)	Not found	No	Yes	46.3	568	661
SGHV-Eth043 (50093←49722)	(ICSVP)	SGHV-Uga042	100.00	YP_001686990.1	2.00e–83	251	Coiled coils; TM helix	TATAAA (50226)	AATAAA (49586)	AATGT**TAATAAG** (50102)	Yes	Yes	2189.8	8066	31287
SGHV-Eth044 (50562←50131)	(Tegument protein)	SGHV-Uga043	99.31	YP_001686991.1	4.00e–97	288	SP; glutamine-rich region	TATATA (50672)	AATATT (50088)	Not found	Yes	Yes	1213.6	5191	17339
SGHV-Eth045 (51813←50731)	(Nucleocapsid protein)	SGHV-Uga044	98.34	YP_001686992.1	0.0	672	Coiled coils; TM; t-SNAREs; SF3 helicase	TATAAT (51848)	AATAAT (50690)	Not found	Yes	Yes	267.3	2866	3819
SGHV-Eth046 (56956←51773)	LEF-3 ORF51 *Maruca vitrata* multiple nucleopolyhedrovirus (envelope protein)	SGHV-Uga045	99.31	YP_001686993.1	0.0	3511	PPase-tensin	TATAAA (57002)	AATATA (51736)	TTAGCCCA**TAAG** (56975)	Yes	Yes	72.6	3726	1037
SGHV-Eth047 (57195←57001)		No hits					GTPase-activator protein for Ras-like GTPase; TM helix	TATATA (57313)	AATAAA (56936)	CTTATTTA**TAAG** (57210)	No	No			
SGHV-Eth048 (57226→58824)	Glutathione *S*-transferase (tegument protein)	SGHV-Uga046	99.81	YP_001686994.1	0.0	1098	PPase (inorganic pyrophosphatase)	TATAAA (57164)	AATAAT (58999)	CTTATAAA**TAAG** (57215)	Yes	Yes	395.2	6257	5646
SGHV-Eth049 (58843→60060)	**Cellular protein CBG22662 *Coenorhabditis briggsae* (tegument protein)**	**SGHV-Uga047**	**98.52**	**YP_001686995.1**	**0.0**	**758**	**NUDIX hydrolase domain-like; coiled coils; pre-mRNA splicing factor 9-like protein**	**TATATA (58756)**	**AATAAA (60131)**	AATAA**TAATAAG** **(58837)**	**Yes**	**Yes**	**180.8**	**2181**	**2584**
SGHV-Eth050 (60099→60881)	Budded virus envelope protein 31 *Helicoverpa armigera* nucleopolyhedrovirus NNg1	SGHV-Uga048	97.72	YP_001686996.1	2.00e–180	508	NUDIX superfamily hydrolases motif	TATAAT (60052)	AAATAT (60883)	ATTTT**TAATAAG** (60076)	No	No	63.7	494	910
SGHV-Eth051 (60833→62080)	**(Tegument protein)**	**SGHV-Uga049**	**90**	**YP_001686997.1**	**6.00e–49**	**168**		**TATATT (60755)**	**AAATAT (62114)**	TATAT**TAATAAG** **(60763)**	**Yes**	**Yes**	**5246**	**64822**	**74947**
		SGHV-Uga050	98.97	YP_001686998.1	0.0	587									
SGHV-Eth052 (64069←62096)	**LEF-4 ORF96 *Gryllus bimaculatus* nudivirus (ICSVP)**	**SGHV-Uga051**	**98.94**	**YP_001686999.1**	**0.0**	**1237**	**TM helix**	**TATAAA (64108)**	**AATAAA (62097)**	**Not found**	**Yes**	**Yes**	**10.7**	**208**	**152**
SGHV-Eth053 (65113←64196)	ORF MSV152 putative core protein (nucleocapsid protein)	SGHV-Uga052	100.00	YP_001687000.1	0.0	603	TM helix	TATAA (65130)	AATAAA (64126)	TGAAGATA**TAAG** (65128)	Yes	Yes	46.3	421	662
SGHV-Eth054 (66208←65129)	PIF-2 *Apis mellifera* filamentous virus (envelope protein)	SGHV-Uga053	98.89	YP_001687001.1	0.0	731	SP	TATAT (66295)	AATAAA (65003)	Not found	Yes	Yes	1920.3	20534	27435
SGHV-Eth055 (67392←66361)	ORF AMV054 *Amsacta moorei* entomopoxvirus putative RNA polymerase TF (ICSVP)	SGHV-Uga054	99.13	YP_001687002.1	0.0	665		TATAAA (67507)	AATAAA (66362)	Not found	Yes	Yes	12.6	129	180

Table 1. cont.

ORF (position and orientation)	Homology to viral and/or cellular proteins (localization)	Best BLAST match (description of homologies)					Functional and/or structural annotation	Transcription signals			Transcript (T) and peptide (P) mapping		Transcript expression levels		
		Best match	Identity (%)	GenBank accession no.	*E* value	Score		TATA-like box (position)	Poly(A) signal (position)	G/T/ATAAG late motif (position: T of TAAG)	T	P	Mean coverage	Raw read count	FPKM
SGHV-Eth056 (69283←67388)	ORF AMV253 *Amsacta moorei* entomopoxvirus (possible surface protein)	SGHV-Uga055	99.37	YP_001687003.1	0.0	1257		TATAAA (69337)	AAATAA (67369)	Not found	Yes	No	2	37	0
SGHV-Eth057 (69349→69993)		SGHV-Uga056	99.07	YP_001687004.1	5.00e–150	427		TATAAT (69292)	AAATAA (70097)	Not found	Yes	No	0.8	5	0
SGHV-Eth058 (70035→70988)	Rhoptry protein *Plasmodium yolei* strain 17XNL	SGHV-Uga057	99.69	YP_001687005.1	0.0	617		TATATA (69943)	AATTAAA (71059)	ATTTAAA**TAAG** (70031)	Yes	No	2.1	19	0
SGHV-Eth059 (70997→71203)		No hits					NifE/NifN	TATAAT (70977)	AATAA (71224)	TAATT**TAATAAG** (70992)	Yes	No	4721.1	9676	67449
SGHV-Eth060 (71212→71409)	(ICSVP)	No hits					Amino acid transporter	TATAAT (71189)	AATAAA (71408)	Not found	Yes	Yes	1220.8	2393	17439
SGHV-Eth061 (71482→71676)	(ICSVP)	No hits					Domain of unknown function (DUF3557)	TATAAT (71368)	AATAAA (71704)	Not found	Yes	Yes	5.3	10	74
SGHV-Eth062 (71936←71775)		No hits						TATAAA (72040)	Not found	Not found	No	No			
SGHV-Eth063 (72216→72488)		No hits						Not found	Not found	Not found	No	No	16.3	44	233
SGHV-Eth064 (73411→73713)	**Signalling mucin HKR1-like *Xenopus (Silurana) tropicalis***	**No hits**	**37.00**	**XP_006821602.1**	**9.00e–07**	**53.9**		**TATAAT (73344)**	**AATAAA (73558)**	**Not found**	**No**	**No**	**6.6**	**20**	**95**
SGHV-Eth065 (74476←74321)		No hits					SplB; DNA repair; TM helix	TATAAA (74514)	Not found	Not found	No	No			
SGHV-Eth066 (75576←74602)	RpoD-like protein *P. falciparum*	SGHV-Uga059	98.15	YP_001687007.1	0.0	619		TATAAT (75627)	AATAAA (74591)	Not found	Yes	No	3.8	37	55
SGHV-Eth067 (76209←75586)	(ICSVP)	SGHV-Uga060	100.00	YP_001687008.1	5.00e–148	422		TATATA (76248)	AATAAT (75447)	Not found	Yes	Yes	367.9	2273	5256
SGHV-Eth068 (77752←76268)	***Acanthamoeba polyphaga* mimivirus (nucleocapsid protein)**	**SGHV-Uga061**	**99.60**	**YP_001687009.1**	**0.0**	**1007**	**TM helix**	**Not found**	**AATAAA (76136)**	**Not found**	**Yes**	**Yes**	**106.8**	**1571**	**1527**
SGHV-Eth069 (77713→90453)	**ORF147 *Trichoplusia ni* ascovirus-2c (nucleocapsid protein)**	**SGHV-Uga062**	**91.59**	**YP_001687010.1**	**0.0**	**7128**	**SP; BP-NLS; GBD-FH3; ezrA; NUMOD3 motifs; spectrin repeats; t-SNAREs**	**TATAAA (77660)**	**AATATAA (90504)**	TATAA**TAATAAG** **(77704)**	**Yes**	**Yes**	**69**	**8707**	**986**
SGHV-Eth070 (91852←90461)		SGHV-Uga063	100.00	YP_001687011.1	0.0	924	TM helix	TATAAT (91903)	Not found	Not found	Yes	No	54.6	735	781
SGHV-Eth071 (93816←92032)	ORF AMV130 *Amsacta moorei* entomopoxvirus ABC transporter protein (tegument protein)	SGHV-Uga064	99.83	YP_001687012.1	0.0	1192	Coiled coils; PUM repeats; zinc finger domain	TATAAT (93944)	AATAAA (91991)	TCTCTAA**TAAG** (93826)	Yes	Yes	123.3	2178	1761
SGHV-Eth072 (98093←93831)	ORF AMV039 *Amsacta moorei* entomopoxvirus ATPase/DNA helicase protein	SGHV-Uga065	99.01	YP_001687013.1	0.0	2825		TATATA (98104)	Not found	AATATATA**TAAG** (98098)	No	No	12.4	524	177
SGHV-Eth073 (98139→98453)		SGHV-Uga066	99.06	YP_001687014.1	2.00e–67	209	α-Helix–β-strand–α-helix	TATATA (98097)	AATAAA (98459)	Not found	Yes	No	831.6	2594	11883
SGHV-Eth074 (99279←98503)	Helicase (tegument protein)	SGHV-Uga067	98.46	YP_001687015.1	2.00e–180	508	TM helix	TATATA (99305)	AATAAA (98504)	ATAA**TAAGTAAG** (99290)	Yes	Yes	281.5	2165	4021
SGHV-Eth075 (99625←99302)	Riboflavin uptake protein, chain A (ECF transporter) (envelope protein)	SGHV-Uga068	100.00	YP_001687016.1	2.00e–70	217	TM helix; SP	TATAAT (99746)	AATAAT (99299)	TTAATAA**TAAG** (99672)	Yes	Yes	222.8	715	3184
SGHV-Eth076 (100496←99687)	Ca^{2+}- and Zn^{2+}-binding protein (tegument protein)	SGHV-Uga069	100.00	YP_001687017.1	0.0	542		TATATT (100581)	AATAAA (99644)	TATATAAA**TAAG** (100500)	Yes	Yes	1340.2	10748	19147

Table 1. cont.

ORF (position and orientation)	Homology to viral and/or cellular proteins (localization)	Best BLAST match (description of homologies)					Functional and/or structural annotation	Transcription signals			Transcript (T) and peptide (P) mapping		Transcript expression levels		
		Best match	Identity (%)	GenBank accession no.	*E* value	Score		TATA-like box (position)	Poly(A) signal (position)	G/T/ATAAG late motif (position: T of TAAG)	T	P	Mean coverage	Raw read count	FPKM
SGHV-Eth077 (101994←100687)	ORF AMV156 *Amsacta moorei* entomopoxvirus (nucleocapsid protein)	SGHV-Uga070	98.39	YP_001687018.1	0.0	858		TATATA (102020)	Not found	Not found	Yes	Yes	10.5	136	150
SGHV-Eth078 (102058→103881)	Probable capsid protein 3 *Acanthamoeba polyphaga* mimivirus (tegument protein)	SGHV-Uga071	99.34	YP_001687019.1	0.0	1225	NUDIX hydrolase domain-like	TATATA (102001)	AATAAA (103927)	ACTTATTA**TAAG** (102040)	Yes	Yes	29.1	525	415
SGHV-Eth079 (104676←103870)	FAD-dependent thiol oxidase African swine fever virus Malawi LIL 20/1 (envelope protein)	SGHV-Uga072	99.63	YP_001687020.1	0.0	540	TM; ERV/ALR thiol oxidase domain (mitochondrial)	TATAAA (104760)	AATAAA (103871)	Not found	Yes	Yes	879.1	7024	12559
SGHV-Eth080 (104928←104692)		SGHV-Uga073	98.73	YP_001687021.1	2.00e–46	154		TATAAA (104965)	Not found	Not found	Yes	No	1023.6	2402	14624
SGHV-Eth081 (104975→107110)	Helicase 2-like protein *Spodoptera frugiperda* granulosis virus	SGHV-Uga074	99.16	YP_001687022.1	0.0	1439	P-loop NTPase fold (Walker A/B motifs)	TATAAA (104771)	AATAAA (107120)	Not found	Yes	No	32.5	687	464
SGHV-Eth082 (107969←107157)		SGHV-Uga075	99.26	YP_001687023.1	0.0	526		TATAAA (108021)	Not found	Not found	Yes	No	65.6	538	937
SGHV-Eth083 (107993→108625)	PIF-3 *Euproctis pseudoconspersa* single nucleopolyhedrovirus (virion protein)	SGHV-Uga076	100.00	YP_001687024.1	3.00e–148	423	TM helix	Not found	AATAAA (108645)	Not found	Yes	Yes	478.9	3002	6843
SGHV-Eth084 (108695→111871)	ORF AMV130 *Amsacta moorei* entomopoxvirus ABC transporter protein (ICSVP)	SGHV-Uga077	99.72	YP_001687025.1	0.0	2117	DNA polymerase (B); exonuclease/H-like domains	TATATA (108633)	AATAAA (111979)	CAAACAAA**TAAG** (108691)	Yes	Yes	6.6	208	94
SGHV-Eth085 (111886→112041)		No hits					Ycf48-like protein (provisional); TM helix	Not found	AATAAA (112241)	Not found	No	No			
SGHV-Eth086 (112845←112138)	ORF004 *Oryctes rhinoceros* nudivirus Ac81-like protein (ICSVP)	SGHV-Uga078	98.31	YP_001687026.1	2.00e–166	471	TM	TATATT (113012)	AATAAA (112133)	ATAATAAA**TAAG** (112856)	Yes	Yes	325.7	2283	4653
SGHV-Eth087 (112896→115754)	ORF9 human herpesvirus 8 type M DNA polymerase	SGHV-Uga079	99.58	YP_001687027.1	0.0	1955	DNA/RNA polymerases (α/β multidomains)	TATAAA (112820)	AATAAA (115763)	Not found	Yes	No	188.2	5327	2689
SGHV-Eth088 (116378←115776)	(ICSVP)	SGHV-Uga080	99.50	YP_001687028.1	1.00e–141	405		TATAAA (116401)	AATAAA (115777)	Not found	Yes	Yes	140.7	840	2010
SGHV-Eth089 (116899←116387)		SGHV-Uga081	99.42	YP_001687029.1	5.00e–117	340	TM helix	TATAAA (117011)	AATAAA (116385)	ATTAACTA**TAAG** (116909)	Yes	No	28.4	144	405
SGHV-Eth090 (117392←116916)	(Virion protein)	SGHV-Uga082	100.00	YP_001687030.1	3.00e–110	322	TM	TATAAT (117445)	AATAAA (116776)	ATAGCAAA**TAAG** (117397)	Yes	Yes	112.9	533	1612
SGHV-Eth091 (119479←117398)	ORF AMV214 *Amsacta moorei* entomopoxvirus (nucleocapsid protein)	SGHV-Uga083	99.42	YP_001687031.1	0.0	1402	BP-NLS	TATAAT (119523)	AATAAA (117389)	Not found	Yes	Yes	116.5	2401	1664
SGHV-Eth092 (119571→120227)	(Virion protein)	SGHV-Uga084	99.09	YP_001687032.1	3.00e–149	426		TATATT (119537)	AATAAA (120246)	TATAAATG**TAAG** (119556)	Yes	Yes	13.9	90	198
SGHV-Eth093 (120447→121211)	Signalling protein (tegument protein)	SGHV-Uga085	96.47	YP_001687033.1	0.0	509	α-Helix–β-strand–α-helix	Not found	AATAAA (121218)	Not found	Yes	Yes	644	4877	9199
SGHV-Eth094 (121304→123079)	ORF398 megavirus chiliensis ankyrin repeat protein (tegument protein)	SGHV-Uga086	100.00	YP_001687034.1	0.0	1201	α-Helix–β-strand–α-helix	TATAAT (121236)	AATAAA (123078)	TAGATATA**TAAG** (121292)	Yes	Yes	1031.3	18134	14733

Table 1. cont.

ORF (position and orientation)	Homology to viral and/or cellular proteins (localization)	Best BLAST match (description of homologies)					Functional and/or structural annotation	Transcription signals			Transcript (T) and peptide (P) mapping		Transcript expression levels		
		Best match	Identity (%)	GenBank accession no.	*E* value	Score		TATA-like box (position)	Poly(A) signal (position)	G/T/ATAAG late motif (position: T of TAAG)	T	P	Mean coverage	Raw read count	FPKM
SGHV-Eth095 (123392←123105)	LEF-5 ORF99 *Culex nigripalpus* nucleopolyhedrovirus	SGHV-Uga087	96.88	YP_001687035.1	7.00e–62	194	TFIIS-type zinc finger; RNA polymerase subunit RPO30	TATAAT (123456)	AATAAA (123074)	Not found	Yes	No	93.1	265	1328
SGHV-Eth096 (123514→125466)	ORF AMV258 *Amsacta moorei* entomopoxvirus vaccinia G1L metalloprotease (envelope protein)	SGHV-Uga088	98.16	YP_001687036.1	0.0	1226	TM helix	TATAAT (123434)	Not found	TGTTGTT**ATAAG** **(123503)**	Yes	Yes	180.6	3492	2580
SGHV-Eth097 (125484→126023)	ORF AMV134 *Amsacta moorei* entomopoxvirus leucine-rich repeat gene family protein	SGHV-Uga089	97.78	YP_001687037.1	2.00e–117	342	Coiled coils	TATATA (125431)	AATAAA (126050)	Not found	Yes	No	14.4	77	206
SGHV-Eth098 (126211←126026)		No hits						TATAAT (126291)	AATAAA (125962)	Not found	No	No			
SGHV-Eth099 (126398→126631)		SGHV-Uga090	100.00	YP_001687038.1	2.00e–43	146	TM helix	TATAAA (126646)	AATAAA (127611)	Not found	Yes	No			
SGHV-Eth100 (126731→127534)	Putative capsid protein 3 *Cafeteria roenbergensis* virus BV_PW1 (virion protein)	SGHV-Uga091	99.25	YP_001687039.1	0.0	541	TM; PS; RNA-dependent RNA polymerase	TATAAA (127546)	AATAAA (127828)	GCAATTAT**TAAG** (126556)	Yes	Yes	33.1	264	474
SGHV-Eth101 (127557→127805)	(ICSVP)	SGHV-Uga092	100.00	YP_001687040.1	1.00e–50	165		TATATA (127252)	AATAAT (126196)	Not found	Yes	Yes	5496.9	13552	78533
SGHV-Eth102 (127816→128802)	Hydrolase (IIe RE) (tegument protein)	SGHV-Uga093	100.00	YP_001687041.1	0.0	665	TM	Not found	AATAAA (128797)	Not found	Yes	Yes	1188.1	11611	16975
SGHV-Eth103 (128827→129645)	(Tegument protein)	SGHV-Uga094	98.90	YP_001687042.1	0.0	541	TM; α-helix–β-strand–α-helix	Not found	AATAAT (129731)	Not found	Yes	Yes	278.7	2260	3982
SGHV-Eth104 (129659→130129)	(Virion protein)	SGHV-Uga095	98.09	YP_001687043.1	2.00e–103	305		TATATA (129559)	AATAAA (130132)	TAAAGAAG**TAAG** (129590)	Yes	Yes	33.1	154	472
SGHV-Eth105 (130169→131311)	Metal-binding protein (tegument protein)	SGHV-Uga096	98.16	YP_001687044.1	0.0	747	TM helix	TATAAA (130096)	AATAAA (131321)	TCATTAGA**TAAG** (130156)	Yes	Yes	796.3	9012	11377
SGHV-Eth106 (131301→132482)	Vesicle-associated membrane protein (tegument protein)	SGHV-Uga097	98.98	YP_001687045.1	0.0	800	TM helix; RNA polymerase	TATATT (131262)	AATAAA (132580)	Not found	Yes	Yes	1843.2	21571	26333
SGHV-Eth107 (132505→132849)	(Nucleocapsid protein)	SGHV-Uga098	96.52	YP_001687046.1	9.00e–73	223	TM	Not found	AATAAA (132874)	Not found	Yes	Yes	1015.7	3469	14509
SGHV-Eth108 (133544←133065)		**SGHV-Uga099**	**98.75**	**YP_001687047.1**	**3.00e–109**	**320**	**PD-(D/E) XK nuclease fold**	**TATATA (133638)**	**AATAAT (133030)**	TATAACTA**TAAG** **(133555)**	**Yes**	**No**	**399.8**	**1900**	**5712**
SGHV-Eth109 (133566→133856)	Hypothetical conserved protein PRK06126 (ICSVP)	No hits						Not found	Not found	Not found	Yes	Yes	6160.8	17750	88015
SGHV-Eth110 (133970→134227)	(ICSVP)	No hits					Serpentine type 7TM GPCR chemoreceptor Srx; TM helix	Not found	AATAAA (134356)	Not found	Yes	Yes	2063.8	5272	29485
SGHV-Eth111 (134308→134625)	(Nucleocapsid protein)	SGHV-Uga101	100.00	YP_001687049.1	3.00e–67	209	TM helix	TATAAT (134265)	AATAAA (134673)	Not found	Yes	Yes	473.3	1493	6775
SGHV-Eth112 (134681→136636)	PIF-1 *Neodiprion abietis* nucleopolyhedrovirus (envelope protein)	SGHV-Uga102	99.54	YP_001687050.1	0.0	1332	TM helix; SP; EGF-like domain	TATATA (134644)	AATAAA (136700)	TTATA**TAATAAG** (134651)	Yes	Yes	608.1	11776	8687
SGHV-Eth113 (137858←136731)	Viral capsid-associated protein 1054 *Neodipron abietis* nucleopolyhedrovirus	SGHV-Uga103	97.87	YP_001687051.1	0.0	743		TATAAA (137935)	AATAAA (136732)	Not found	Yes	No	29.6	331	423
SGHV-Eth114 (137907→139886)	(Nucleocapsid protein)	SGHV-Uga104	99.39	YP_001687052.1	0.0	1335	TM; coiled coil region	TATAAA (137859)	AATAAA (139896)	Not found	Yes	Yes	31.9	625	455
SGHV-Eth115 (140947←140072)	(Virion protein)	SGHV-Uga105	98.97	YP_001687053.1	0.0	562	Coiled coils	TATATT (141031)	AATAAA (139965)	TAATATAT**TAAG** (140963)	Yes	Yes	110.1	955	1573

Table 1. cont.

ORF (position and orientation)	Homology to viral and/or cellular proteins (localization)	Best BLAST match (description of homologies)					Functional and/or structural annotation	Transcription signals			Transcript (T) and peptide (P) mapping		Transcript expression levels		
		Best match	Identity (%)	GenBank accession no.	*E* value	Score		TATA-like box (position)	Poly(A) signal (position)	G/T/ATAAG late motif (position: T of TAAG)	T	P	Mean coverage	Raw read count	FPKM
SGHV-Eth116 (142410←140953)	**ORF MSV156 *Melanoplus sanguinipes* entomopoxvirus (nucleocapsid protein)**	**SGHV-Uga106**	**98.29**	**YP_001687054.1**	**0.0**	**919**	**Coiled coils; serine/threonine/glutamine-rich stretches**	**TATAAT (142466)**	**AATAAA (140839)**	**Not found**	**Yes**	**Yes**	**522.8**	**75.47**	**74.69**
SGHV-Eth117 (143935←142373)	Cell division protein 48 lymphocystis disease virus (China isolate) (nucleocapsid protein)	SGHV-Uga107	97.89	YP_001687055.1	0.0	1046	SP; AAA^+-ATPase; PAN	TATAAA (143960)	AAATAA (142334)	Not found	Yes	Yes	806.7	12484	11525
SGHV-Eth118 (145550←143916)	Cell division protein 48 lymphocystis disease virus (China isolate) (nucleocapsid protein)	SGHV-Uga108	99.08	YP_001687056.1	0.0	1096	SP; P-loop/AAA^+-ATPase; PAN	TATACA (145574)	Not found	Not found	Yes	Yes	30	486	429
SGHV-Eth119 (145628→145804)		No hits					Phage shock protein B; TM helix	TATAAA (145571)	AATAAA (145900)	Not found	No	No			
SGHV-Eth120 (146791←145937)	(Virion protein)	SGHV-Uga109	99.30	YP_001687057.1	0.0	567	TM helix; CBS domain	Not found	AATAAA (145853)	Not found	Yes	Yes	601.6	5093	8595
SGHV-Eth121 (147433←146831)	*Spodoptera litura* granulosis virus matrix metalloproteinase-38 (MP-NASE-like protein) (tegument protein)	SGHV-Uga110	99.50	YP_001687058.1	9.00e–145	413	TM helix; SP; helix–turn–helix domain	TATAAT (147490)	AATAAA (146701)	TTGTATTT**TAAG** (147463)	Yes	Yes	2331.8	13921	33312
SGHV-Eth122 (148189←147533)	MP-NASE-like protein (ICSVP)	SGHV-Uga111	98.63	YP_001687059.1	7.00e–153	435	TM; SP; zinc protease	TATATT (148280)	Not found	CTGTT**TAATAAG** (148194)	Yes	Yes	6661.1	43330	95164
SGHV-Eth123 (148826←148302)	**Regulatory protein (tegument protein)**	**SGHV-Uga112**	**94.86**	**YP_001687060.1**	**2.00e–110**	**323**	**TM helix; SP**	**Not found**	**Not found**	**Not found**	**No**	**Yes**	**2444.4**	**12706**	**34922**
SGHV-Eth124 (149662←148820)	Cellular protein PY00593 *Plasmodium yoelli* strain 17XNL (nucleocapsid protein)	SGHV-Uga113	98.93	YP_001687061.1	0.0	553		TATAT (149806)	Not found	TTTAATAG**TAAG** (149788)	Yes	Yes	1162	9699	16602
SGHV-Eth125 (149925←149668)		No hits					Vps52/Sac2 family	TAATAT (150036?)	AATAAT (149666?)	Not found	No	No			
SGHV-Eth126 (149956→151275)	ORF MSV016 *Melanoplus sanguinipes* entomopoxvirus leucine-rich repeat gene family protein	SGHV-Uga114	99.77	YP_001687062.1	0.0	868	Leucine-rich repeat protein-like protein	TATATT (149803)	AATAAA (151306)	TGTTCTTA**TAAG** (149937)	Yes	No	8.9	116	127
SGHV-Eth127 (151332→152561)		SGHV-Uga115	99.27	YP_001687063.1	0.0	830	F-box/RNI superfamily	TATATA (151281)	AATAAA (152653)	Not found	No	No	82.5	1004	1178
SGHV-Eth128 (152582→153655)	ORF AMV134 *Amsacta moorei* leucine-rich repeat gene family protein	SGHV-Uga116	98.88	YP_001687064.1	0.0	696	RNase inhibitor	TATATA (152445)	AATAAA (153722)	Not found	Yes	No	21.4	227	305
SGHV-Eth129 (154378←153662)	ISKNV ORF 099L RING finger protein *Phaeocystis globosa* virus	SGHV-Uga117	97.91	YP_001687065.1	1.00e–158	451	E3-ubiquitin-protein ligase TRAIP	TATAAA (154406)	Not found	Not found	Yes	No	578.3	4105	8261
SGHV-Eth130 (154650←154405)	ORF103L scale drop disease virus	SGHV-Uga118	98.78	YP_001687066.1	5.00e–52	168	Coiled coils; E3-ubiquitin-protein ligase TRAIP	TATAAA (154742)	AATAAA (154264)	Not found	Yes	No	906.3	2207	12945
SGHV-Eth131 (155345←154695)		SGHV-Uga119	100.00	YP_001687067.1	6.00e–152	432	RNase inhibitor	Not found	Not found	Not found	Yes	No	159.3	1026	2274
SGHV-Eth132 (155374→156426)		SGHV-Uga120	97.72	YP_001687068.1	0.0	696		TATAT (155238)	AAATAA (156489)	TAATGTTT**TAAG** (155291)	Yes	No	114.2	1190	1631
SGHV-Eth133 (156715→157473)		SGHV-Uga121	100.00	YP_001687069.1	0.0	523	TM	TATAAT (156598)	AAATAA (157541)	ATTATTTT**TAAG** (156610)	Yes	No	52.1	391	743

Table 1. cont.

ORF (position and orientation)	Homology to viral and/or cellular proteins (localization)	Best BLAST match (description of homologies)					Functional and/or structural annotation	Transcription signals			Transcript (T) and peptide (P) mapping		Transcript expression levels		
		Best match	Identity (%)	GenBank accession no.	*E* value	Score		TATA-like box (position)	Poly(A) signal (position)	G/T/ATAAG late motif (position: T of TAAG)	T	P	Mean coverage	Raw read count	FPKM
SGHV-Eth134 (57669→158016)		SGHV-Uga122	100.00	YP_001687070.1	8.00e–79	239	TM helix	TATAAA (157605)	AATAAA (158193)	Not found	Yes	No	170.1	586	2430
SGHV-Eth135 (158076→158225)		No hits					CRISPR/Cas system-associated RAMP superfamily protein Csb3	TATAAT (158058)	AATAAA (158236)	Not found	Yes	No	11.5	17	164
SGHV-Eth136 (158840→159307)	**chmu148 *Choristoneura murinana* alphabaculovirus**	**SGHV-Uga123**	**94.87**	**YP_001687071.1**	**1.00e–102**	**303**	**Coiled coils; zinc finger (C2H2) domain**	**TATATA (158794)**	**Not found**	**Not found**	**Yes**	**No**	**34.5**	**160**	**493**
SGHV-Eth137 (160866→162386)	chmu148 *Choristoneura murinana* alphabaculovirus	SGHV-Uga124	34.95	YP_001687072.1	4.00e–86	289	Coiled coils	TATAAT (160695)	AATAAA (162437)	Not found	Yes	No	0.6	9	0
SGHV-Eth138 (162549←162373)		No hits					Protein trafficking PGA2	TATAAA (162638)	AATAAA (162259)	Not found	No	No			
SGHV-Eth139 (162679→164472)	**ORF067 *Ecotropis obliqua* nucleopolyhedrovirus Cg30 protein**	**SGHV-Uga125**	**83.82**	**YP_001687073.1**	**6.00e–85**	**273**	**Coiled coils**	**TATAAC (162635)**	**AATAAA (164478)**	GTTAATCT**TAAG (162668)**	**Yes**	**No**	**22.3**	**396**	**319**
SGHV-Eth140 (164571→164813)	**ORF149 *Anticarsia gemmatalis* nucleopolyhedrovirus Pe38-like protein**	**SGHV-Uga126**	**98.72**	**YP_001687074.1**	**2.00e–43**	**147**	**Coiled coils**	**TATATG (164516)**	**AATAAA (164819)**	AATAATTA**TAAG (164563)**	**Yes**	**No**	**67.2**	**162**	**962**
SGHV-Eth141 (165783→166046)		**SGHV-Uga127**	**88.89**	**YP_001687075.1**	**8.00e–46**	**152**		**TATAAA (165725)**	**AATAAA (166059)**	CTTATAGT**TAAG (165757)**	**Yes**	**No**	**265.7**	**694**	**3793**
SGHV-Eth142 (166616←166461)		No hits					Zinc finger RING of BARD1-type protein domain	TATAAT (166659)	AATAAA (166462)	Not found	Yes	No	147.9	228	2109
SGHV-Eth143 (166602→166784)		SGHV-Uga128	87.30	YP_001687076.1	2.00e–22	91.7	Zinc finger (C2H2) domain	TATATT (166532)	AATAAA (166913)	Not found	No	No	1894.6	3433	27069
SGHV-Eth144 (167032→167292)	(ICSVP)	SGHV-Uga129	100.00	YP_001687077.1	7.00e–43	145	Coiled coils	TATAAA (167007)	Not found	GTTAACAA**TAAG (167026)**	Yes	Yes	4.1	11	61
SGHV-Eth145 (167445→167969)		SGHV-Uga130	93.71	YP_001687078.1	2.00e–111	326		TATATA (167300)	AATAAA (168020)	CCGAGTTG**TAAG (167348)**	Yes	No	730.3	3796	10433
SGHV-Eth146 (168084→170510)	ORF MSV156 *Melanoplus sanguinipes* entomopoxvirus	SGHV-Uga131	86.66	YP_001687079.1	0.0	1415	Coiled coils	TATAAA (168036)	Not found	Not found	Yes	No	85.5	2055	1222
SGHV-Eth147 (170635→170832)		SGHV-Uga132	100.00	YP_001687080.1	1.00e–39	135		TATATA (170587)	AATAAA (170859)	Not found	No	No	2125.9	4168	30375
SGHV-Eth148 (171023→172279)		SGHV-Uga133	99.28	YP_001687081.1	0.0	865	EGF-like domain signature 1	TATAAT (170966)	Not found	Not found	Yes	No	552.5	6876	7893
SGHV-Eth149 (172423→172719)	Tail length tape-measure protein *Oenococcus* phage phi9805 (ICSVP)	SGHV-Uga134	97.98	YP_001687082.1	2.00e–61	194	Coiled coils	TATAAA (172395)	Not found	Not found	Yes	Yes	418.9	1232	5986
SGHV-Eth150 (172742→172969)	(ICSVP)	SGHV-Uga135	100.00	YP_001687083.1	2.00e–45	151	Coiled coils; helix hairpin bin	TATAAG (172718)	AATAAA (172990)	TAGTTTTA**TAAG (172720)**	No	Yes	337.6	762	4822
SGHV-Eth151 (173210←173010)		No hits						TATATT (173349)	AATAAA (173004)	TTTTCGTT**TAAG (173320)**	Yes	No	1359.2	2705	19419
SGHV-Eth152 (173287→173448)	(ICSVP)	SGHV-Uga136	100.00	YP_001687084.1	1.00e–29	109	Coiled coils	TATAAA (173247)	AATAAA (173447)	GATATAAA**TAAG (173253)**	No	Yes	286.2	459	4088
SGHV-Eth153 (173515→173784)	**(ICSVP)**	**SGHV-Uga137**	**93.33**	**YP_001687085.1**	**3.00e–53**	**172**	**Coiled coils**	**TATAAA (173457)**	**AATAAA (173783)**	CATATAAA**TAAG (173463)**	**No**	**Yes**	**33.8**	**90**	**481**
SGHV-Eth154 (173800→174159)	(ICSVP)	SGHV-Uga138	100.00	YP_001687086.1	6.00e–80	242	Coiled coils	TATATA (173747)	AATAAA (174209)	Not found	Yes	Yes	21	75	301
SGHV-Eth155 (174442→174636)	(ICSVP)	SGHV-Uga139	100.00	YP_001687087.1	1.00e–37	130	Coiled coils	TATAAA (174313)	AATAAA (174635)	AGTAGTGGT**AAG (174381)**	No	Yes	482.5	932	6897

Table 1. cont.

ORF (position and orientation)	Homology to viral and/or cellular proteins (localization)	Best BLAST match (description of homologies)					Functional and/or structural annotation	Transcription signals			Transcript (T) and peptide (P) mapping		Transcript expression levels		
		Best match	Identity (%)	GenBank accession no.	*E* value	Score		TATA-like box (position)	Poly(A) signal (position)	G/T/ATAAG late motif (position: T of TAAG)	T	P	Mean coverage	Raw read count	FPKM
SGHV-Eth156 (174647→175870)	ORF148 *Choristoneura murinana* alphabavculovirus (virion protein)	SGHV-Uga140	99.51	YP_001687088.1	0.0	840	Coiled coils	TATAAA (174520)	AATAAA (175908)	GAAATAAA**TAAG** (174641)	Yes	Yes	341.4	4137	4877
SGHV-Eth157 (176052←175789)	LEF-8 *Penaeus monodon* nudivirus	SGHV-Uga141	96.59	YP_001687089.1	3.00e–51	167	TM	TATATT (176190)	AATAAA (175620)	Not found	Yes	No	124.1	324	1771
SGHV-Eth158 (176252→177322)		SGHV-Uga142	99.72	YP_001687090.1	0.0	730		TATAAA (176200)	AATAAA (177377)	TTATAATT**TAAG** (176230)	Yes	No	53.1	563	759
SGHV-Eth159 (177366→178655)		SGHV-Uga143	98.60	YP_001687091.1	0.0	866	Coiled coils	TATAAA (177321)	Not found	Not found	Yes	No	10.2	130	145
SGHV-Eth160 (179750←179379)	ORF AMV193 protein phosphatase 1 *Amsacta moorei* entomopoxvirus	SGHV-Uga144	100.00	YP_001687092.1	5.00e–80	243	PP1 (regulatory subunits 15A/B)	TATAAA (179853)	AATAAA (179387)	Not found	Yes	No	371.7	1369	5310
SGHV-Eth161 (180040→180726)	ORF179 shrimp white spot syndrome virus	SGHV-Uga145	70.74	YP_001687093.1	6.00e–96	291		TATAAA (180029)	AATAAA (180756)	Not found	Yes	No	5.9	40	84
SGHV-Eth162 (180745→181377)	(Virion protein)	SGHV-Uga146	98.58	YP_001687094.1	2.00e–139	400	RNA-dependent RNA polymerase	TATAAT (180690)	AATAAA (181376)	TATAA**TAATAAG** (180733)	No	Yes	6.2	39	89
SGHV-Eth163 (181828→182496)	ORF179 shrimp white spot syndrome virus	SGHV-Uga148	70.23	YP_001687096.1	1.00e–90	279		TATAGA (181701)	AAATAA (182525)	Not found	No	No	5.9	39	84
SGHV-Eth164 (182922→183491)	**ORF179 shrimp white spot syndrome virus**	**SGHV-Uga148**	**64.58**	**YP_001687096.1**	**7.00e–69**	**222**		**Not found**	**Not found**	**Not found**	**No**	**No**	**5.9**	**36**	**83**
SGHV-Eth165 (185006←184023)	**(ICSVP)**	**SGHV-Uga150**	**97.57**	**YP_001687098.1**	**0.0**	**653**		**TATATT (185048)**	**AATAAA (184021)**	**Not found**	**Yes**	**Yes**	**362.8**	**3545**	**5183**
SGHV-Eth166 (185128→185373)		SGHV-Uga151	98.78	YP_001687099.1	3.00e–49	161		TATAAA (185081)	AATAAA (185414)	Not found	No	No	324.9	791	4640
SGHV-Eth167 (185476→185844)	Bcr–Abl-like protein	SGHV-Uga152	100.00	YP_001687100.1	1.00e–82	249	Coiled coils; Bcr–Abl (α-1/2)	TATAAA (185461)	AATAAA (185942)	Not found	Yes	No	109.8	401	1568
SGHV-Eth168 (185847→186077)		SGHV-Uga153	100.00	YP_001687101.1	2.00e–49	161		TATATA (185823)	AATAAA (186080)	Not found	Yes	No	132.5	303	1893
SGHV-Eth169 (186134→187147)	ORF026 *Wiseana* iridescent virus (nucleocapsid protein)	SGHV-Uga154	99.11	YP_001687102.1	0.0	668	Glutamine-rich region	TATAAA (186098)	Not found	Not found	Yes	Yes	37.2	374	532
SGHV-Eth170 (187212→187409)		SGHV-Uga155	98.48	YP_001687103.1	1.00e–35	125		TATAAA (187151)	AATAAA (187408)	Not found	Yes	No	52.2	108	787
SGHV-Eth171 (187737←187570)		**SGHV-Uga156**	**96.49**	**YP_001687104.1**	**5.00e–26**	**100**		**TATAAA (187796)**	**AATAAA (187433)**	**Not found**	**Yes**	**No**	**282.3**	**470**	**4037**
SGHV-Eth172 (188119←187892)		SGHV-Uga157	100.00	YP_001687105.1	9.00e–46	152		Not found	AATAAA (187792)	Not found	Yes	No	9.8	22	139
SGHV-Eth173 (188916←188248)		SGHV-Uga158	100.00	YP_001687106.1	6.00e–154	439		TATATA (188947)	AATAAA (188066)	Not found	Yes	No	17.7	117	252
SGHV-Eth174 (189097→190281)	ORF AMV253 *Amsacta moorei* entomopoxvirus (possible surface protein)	SGHV-Uga160	99.49	YP_001687108.1	0.0	789	F-box protein 7-like domain	TATATA (189034)	AATAAA (190280)	Not found	Yes	No	33.9	397	483

TM, Transmembrane domain; SP, signal peptide; INM-SM; inner nuclear membrane sorting motif; SWI/SNF, gene family that affects mating-type switching (SWI) and sucrose fermentation [sucrose non-fermenting (SNF)] pathways; BP-NLS, bipartite nuclear localization signal/sequence; AARP2, asparagine and aspartate-rich protein 2 domain; AgrD, pro-peptide precursor of the auto-inducing peptide (AIP) of the accessory gene regulatory protein (Agr); HDAC, histone deacetylase; RGD, arginyl-glycyl-aspartic acid; PD-(D/E) XK, conserved domain of nuclease superfamily involved in various aspects of nucleic acid metabolism; ABC, ATP-binding cassette; GPCR, G-protein-coupled receptor; EGF, epidermal growth factor; EzrA, septation FtsZ ring formation regulator; NUMOD3, nuclease-associated modular domain 3; PUM, pumilio homology domain; ERV/ALR domain, Erv1 (essential for respiratory and vegetative growth 1) and ALR (augmenter of liver regeneration) domain; TFIIS-type zinc finger; zinc finger motif found in transcription factor IIs; Bcr–Abl, an oncogene arising from the fusion of the breakpoint cluster (*bcr*) gene with chromosomal Abelson murine leukaemia (c-*abl*) proto-oncogene; GADD, growth arrest and DNA damage protein; Rifin, repetitive interspersed family; Stevor, subtelomic variant ORF; NifE/NifN, bifunctional nitrogenase fixation proteins E/N.

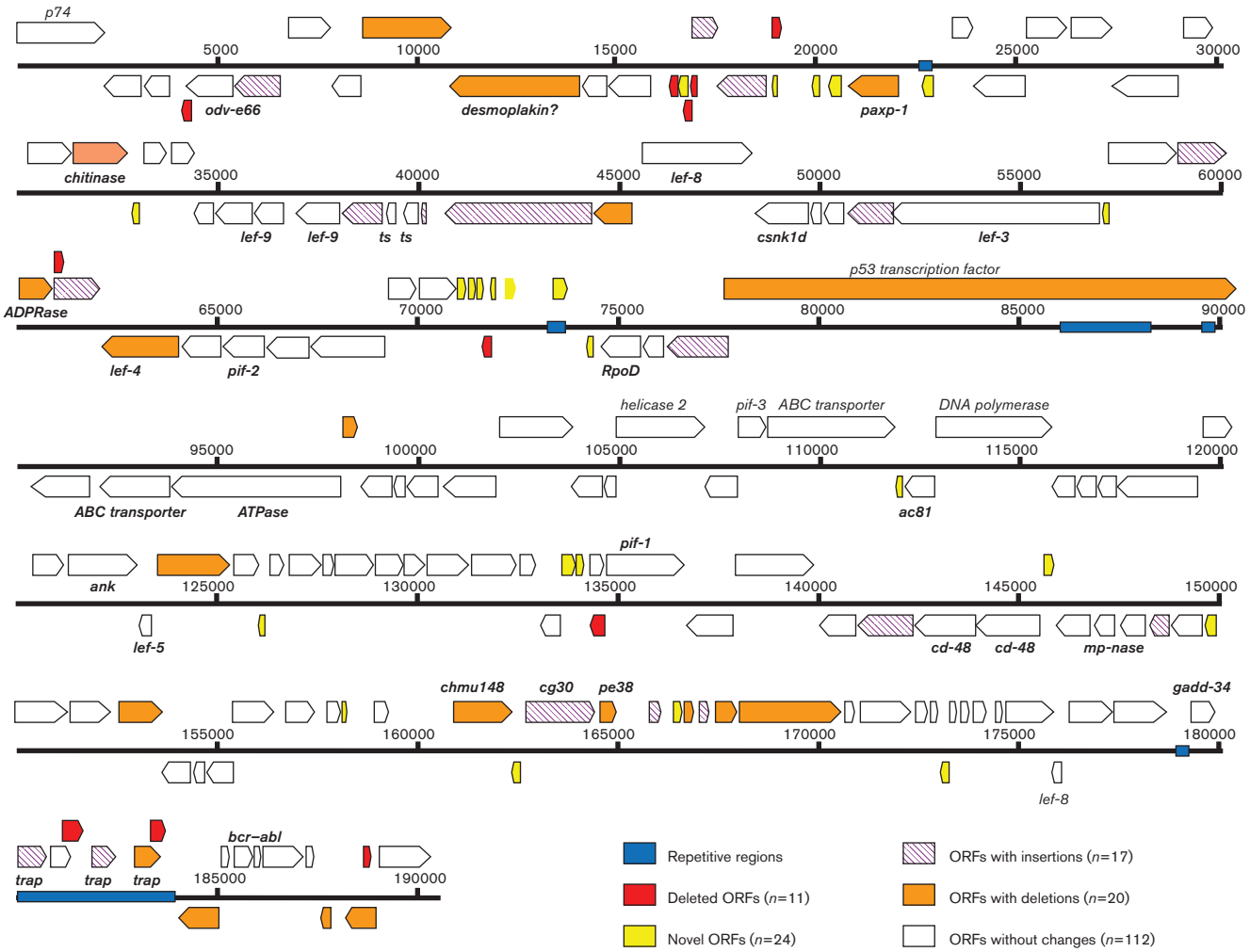

Fig. 1. Linear representation of the GpSGHV-Eth genome: The linearization is presented starting with the ATG initiating codon of *p74* (SGHV-Eth001). The arrows indicate the positions and orientations of the transcription potential of each ORF. In total, 112 ORFs in the GpSGHV-Eth genome were identical in length (amino acid residues) to the homologous ORFs in the GpSGHV-Uga genome; 37 ORFs had insertions and/or deletions, 11 ORFs were deleted and 24 ORFs were novel in the GpSGHV-Eth genome compared with the GpSGHV-Uga genome. The ORFs are colour-coded to show their homology to the corresponding ORFs in the GpSGHV-Uga genome. The abbreviations used are explained in the footnote to Table 1. The figure is drawn to scale.

baculovirus upstream late (T/G/A)TAAG transcription initiation motifs (Table 1).

Proteogenomic mapping of the GpSGHV-Eth genome

We authenticated the assignment of the GpSGHV-Eth ORFs by mapping transcripts and peptides onto the virus genomic sequence. After quality filtering of the transcriptome data, ~40 million paired reads remained, of which ~12.3 million mapped to the host (tsetse fly) and were therefore discarded from further analysis. The remaining reads mapped onto the viral genome and were used to construct 545 putative transcripts (represented by transcript isoforms). Of these, 431 transcripts were predicted to contain protein-coding regions and were used in the functional gene mapping; they mapped onto 141 of the 174 predicted GpSGHV-Eth ORFs (Table 1).

Similarly, filtering out reverse hits, contaminants and host-specific peptides from the liquid chromatography (LC)-MS/MS data resulted in 1314 unique viral peptides. These peptides mapped to 86 of the 174 predicted GpSGHV-Eth ORFs (Table 1). Of the 86 ORFs, 68 ORFs had ATG as the start codon, whilst 13 ORFs had either CTG or TTG as the start codon (see Table S2). Additionally, some of these ORFs contained mapped peptides on sequences upstream of the first methionine residues (Table S2). Non-AUG codons are mostly exhibited by

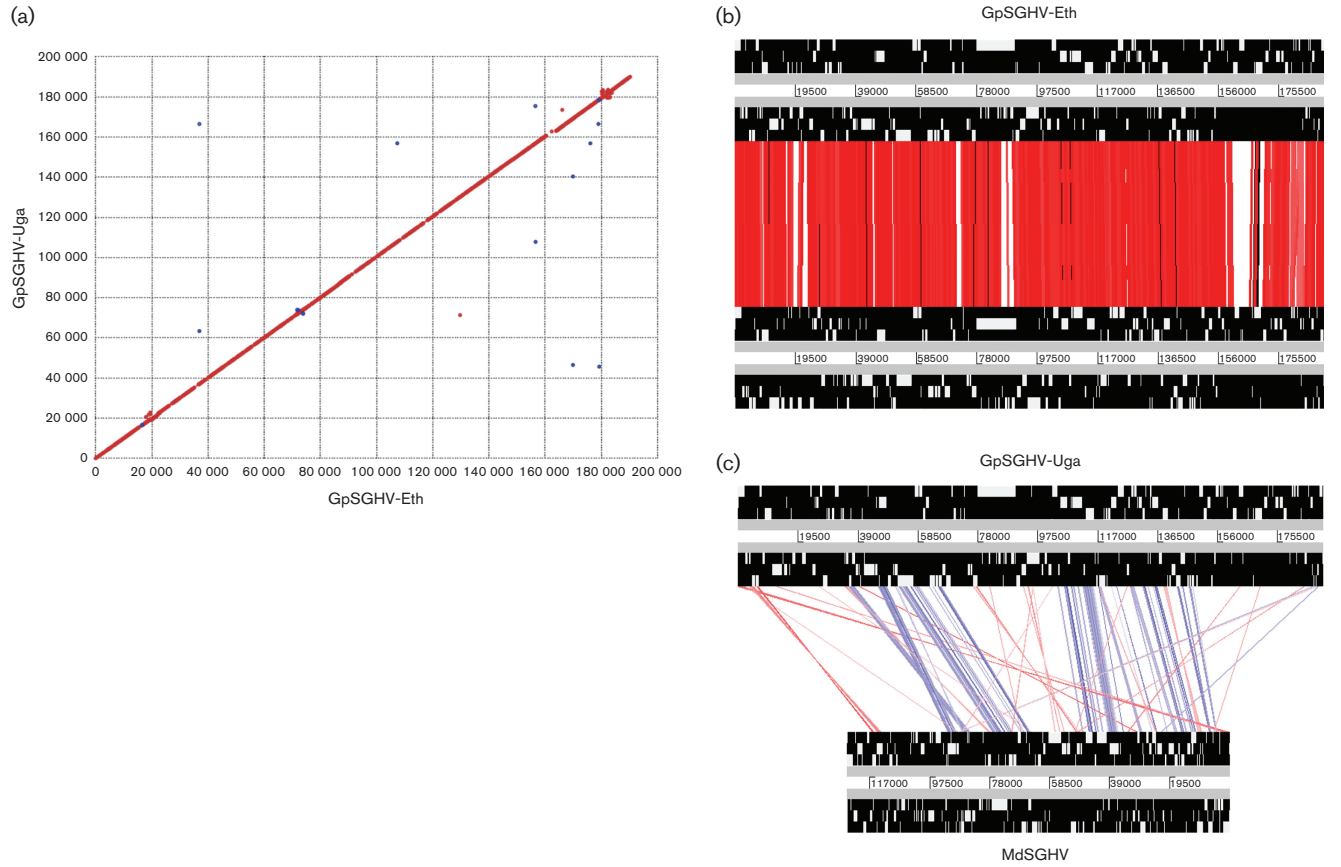

Fig. 2. Genome synteny visualization of the relationship between GpSGHV-Eth and GpSGHV-Uga. The positions of the genomes are shown between the thick black lines, which represent the dsDNA viral genomes. (a) The dot-plot was generated from the whole-genome DNA homology alignments between GpSGHV-Eth and GpSGHV-Uga genomic sequences. The red and blue colours indicate the sequence direction (reverse and forward). (b, c) Syntenic maps show the overall collinearity of GpSGHV-Eth and GpSGHV-Uga compared with *Musca domestica* salivary gland hypertrophy virus (MdSGHV). The red lines (b, c) indicate identity levels between the viruses, whilst the blue lines (c) indicate instances of inversions. The bands represented in the thick black lines do not necessarily indicate the ORFs, but rather the conserved genomic regions.

viral mRNAs encoding regulatory proteins (Corcelette *et al.*, 2000). Non-canonical translation initiation is a rare event that occurs in addition to initiation at a downstream AUG (Gupta *et al.*, 1994), possibly resulting in translation of multiple related proteins from a single ORF. Five of the 13 GpSGHV-Eth ORFs with non-AUG start codons were homologues to known regulatory proteins, i.e. SGHV-Eth008 (*MAL7P1.132*), SGHV-Eth033 (*lef-9*), SGHV-Eth036 (*ts*), SGHV-Eth117 (*cd-48*) and SGHV-Eth122 (*mp-nase*).

Overall, 45.4 % (n=79) of the 174 GpSGHV-Eth ORFs had both peptides and transcripts mapped onto their sequences, whilst only seven ORFs had peptide mapping only. Taking into account the proteomic datasets from GpSGHV-Uga (Kariithi *et al.*, 2010, 2011, 2013b, 2016) and the strong homologies between the two virus isolates (Table S1), additional homologous ORFs could be functionally annotated in the GpSGHV-Eth genome (Table 1). Furthermore, of these 79 ORFs, 68 had appropriately upstream positioned TATAA-like box and/or (G/T/A)TAAG transcriptional signals of early and late baculovirus genes, respectively (Chen *et al.*, 2013), whilst 70 ORFs had poly(A) signals. A total of 60 ORFs contained both the TATA-like box and poly(A) signals, and had mapped transcripts and peptides. Only one ORF (SGHV-Eth109) remained without any transcriptional signals. It is worth mentioning that all the ORFs harbouring the TAAG promoter motif were of the (G/T/A)TAAG type characteristic of baculovirus late promoters (Chen *et al.*, 2013; Rohrmann, 2013). This motif was present upstream of 80 % of the 45 ORFs encoding proteins identified in purified virus particles and thus assumed to correspond to late genes (see Table 1). The experimental validation by 5′–3′ RACE sequencing of whether these motifs correspond to functional transcription and polyadenylation signals is currently ongoing.

Functional elements in the GpSGHV-Eth genome

The peptides from the experimentally derived proteomics data helped us to identify potential functional ORFs in the GpSGHV-Eth genome and to improve the genome annotation. In particular, the GpSGHV-Eth genome contained homologues to 12 out of the 31 so-called core genes of baculoviruses and nudiviruses that are involved in five main processes of baculovirus infection, i.e. (i) replication, (ii) transcription, (iii) packaging, assembly and release, (iv) cell cycle arrest and/or interactions with host proteins, and (v) oral infectivity (Miele *et al.*, 2011). Of the 12 GpSGHV-Eth ORFs homologous to the core genes, nine contained both TATA-like box and poly(A) signals, and had both mapped transcripts and peptides, implying that they were functional in GpSGHV. The remaining three ORFs lacked peptide mapping only (Table 1). The homologues to the core genes in GpSGHV-Eth included one of the four baculovirus core genes involved in DNA repair and recombination, i.e. SGHV-Eth081, a homologue to helicase 2-like protein (*Spodoptera frugiperda* granulosis virus) (Wang *et al.*, 2007). Other homologues were five genes involved in transcription, i.e. SGHV-Eth052, SGHV-Eth095, SGHV-Eth041 and SGHV-Eth033/34, which are homologues to the transcription factors LEF-4, LEF-5, LEF-8 and LEF-9, respectively. Homologues to desmoplakin (Ac66) and to Ac81, which are involved in egress of baculovirus virions and in virus–host interactions at late infection stages (Miele *et al.*, 2011), respectively, were identified in GpSGHV (SGHV-Eth009 and SGHV-Eth086, respectively). Finally, five of the baculovirus *per os* infectivity factors (PIFs), i.e. SGHV-Eth001, SGHV-Eth054, SGHV-Eth112, SGHV-Eth083 and SGHV-Eth004, which are homologues to baculovirus P74 (= PIF-0), PIF-1, PIF-2, PIF-3 and ODV-e66 (Slack & Arif, 2007), respectively, were detected (Table 1). These homologies to the core genes suggest that hytrosaviruses share with baculoviruses and nudiviruses similar modes of entry and transcription of their late genes, which strongly supports the hypothesis that they are derived from a common ancestor (Jehle *et al.*, 2013). We also detected homologues to other viral genes that are potentially functional in GpSGHV, notably SGHV-Eth087 (homologue to herpesvirus DNA polymerase), SGHV-Eth095 (homologue to LEF-5 of *Culex nigripalpus* nucleopolyhedrovirus) and SGHV-Eth053 (homologue to a putative core protein encoded by ORF152 of *Melanoplus sanguinipes* entomopoxvirus) (Tables 1 and S2).

The expressed proteins provide experimental evidence that the viral genes are transcribed and translated to produce functional proteins. Nevertheless, some of the putative ORFs remained without peptide and/or transcript mapping. The lack of equivalence between the transcripts and peptides could be due to several reasons. The most important reasons are that not all mRNAs are actively translated at any particular time, and that the protein content is dependent on both synthesis of new proteins and degradation of existing proteins. Further, the intrinsic MS/MS peptide properties (e.g. abundance, ionization efficiency, solubility, etc.), incorrect assignment of peptides harbouring multiple coding mutations (Dresang *et al.*, 2011) and post-translational modifications may result in under-representation of the full protein repertoires. It is, however, worth mentioning that most of the ORFs without peptide or transcript mapping are of small size (<100 aa residues) and at least some of them could correspond to non-functional ORFs.

Genetic heterogeneity of the GpSGHV-Eth genome

Various insertions and deletion events were detected in the genome of GpSGHV-Eth (Tables 2 and S3). A 111 nt deletion was observed in the 89 508–89 648 nt region in 30 % of reads (Fig. 3a). This deletion was not detected in GpSGHV-Uga. The nucleotide variation in this region was within ORF SGHV-Eth069, which is homologous to the structural protein ORF147 of *Trichoplusia ni* ascovirus 2c (Cui *et al.*, 2007) (Table 1). Notably, compared with its SGHV-Uga062 homologue, ORF SGHV-Eth069 contains indels (Table S3); overall, the latter is 126 nt shorter than the former.

Similarly, two repeat sequences, 27 nt each, occur two and three times, in 65 and 35 %, respectively, of the reads in the 183 153–183 354 nt region (Fig. 3b). The nucleotide variations in this region occur in ORF SGHV-Eth164, which was annotated as a trophozoite antigen-like protein and homologous to shrimp white spot syndrome virus ORF94 (Table 1). Compared with its homologous ORF SGHV-Uga148, SGHV-Eth164 has two deletion events encompassing a total of 57 nt (Table S3). However, this nucleotide region has more repeat sequences in GpSGHV-Uga than GpSGHV-Eth, making SGHV-Uga148 longer than SGHV-Eth164 by a total of 255 nt. It is, however, doubtful whether this ORF is expressed as it contains no transcriptional signals and remains without any peptide or transcript mapping (Table 1).

Likewise, 75 and 25 % of the reads in the 165 386 nt region contained two and one C repeats, respectively. This nucleotide region occurs between ORFs SGHV-Eth140 and SGHV-Eth141. Whereas the 967 nt between these two ORFs did not have any predicted ORF in the GpSGHV-Eth genome, the corresponding region in the GpSGHV-Uga genome (containing 1052 nt between ORFs SGHV-Uga126 and SGHV-Uga127) contained a predicted ORF (159 nt) which was not annotated during the sequencing of the GpSGHV-Uga genome (Abd-Alla *et al.*, 2008).

Such nucleotide deletions potentially result in a change at the ORF position or destroy the ORF if they are located in non-coding regions or coding regions, respectively. The apparent genetic heterogeneity observed within the GpSGHV-Eth genome could be attributed to the fact that the virus sample was isolated from HSGs collected from many symptomatically infected *G. pallidipes* individuals.

Table 2. Sequence polymorphisms in the GpSGHV-Eth genome: the table shows the positions with nucleotide polymorphisms in the genome detected from pyrosequencing reads

Genome position	A*	C*	G*	T*	Nucleotide sequence	Maximum†	Second‡	Percentage§	*P*‖
22760	2	2	2735	91	ACTTGGAAATAAG/TTCGCAATAAGTT	2735	91	3.33	1e–37
31111	5	2112	1	94	TTGTCGAATTTAC/TTTTTCTATTGCA	2112	94	4.45	2e–37
32379	51	1	2208	0	TCATTCAAAACAG/ACAATCGGCGCAA	2208	51	2.31	2e–24
62381	82	1	1533	1	AGCAAATCTTTCG/ACCACCACTAAAA	1533	82	5.35	4e–36
72054	72	3	1893	4	GGACAAACTTTCG/AGATAAGGCTACC	1893	72	3.80	4e–26
88953	11	3174	2	308	AGATGAATTATCC/TAAAGAGCAAGAA	3174	308	9.70	8e–128
101549	270	4	2228	1	GAAACTTCACTTG/ATAGCAATATTTA	2228	270	12.12	5e–120
101553	13	2219	2	106	CTTCACTTGTAGC/TAATATTTATTTG	2219	106	4.78	3e–36
130384	2	2140	0	100	AAAAAAATTACGC/TTGTGGACTAACA	2140	100	4.67	2e–46
137360	57	2701	0	1	CCTTTATATTATC/ATCCATTGAATAG	2701	57	2.11	2e–27
160377	0	182	2	2349	ACTGATCACACAT/CGTGTAGGGTGCC	2349	182	7.75	4e–85
163075	114	2	2400	3	GGATCATCCATTG/ACAACAAATTGGA	2400	114	4.75	5e–47
164395	11	3178	1	89	AAGCAAGAGATTC/TGTCATTTGAAAG	3178	89	2.80	2e–31
164892	2487	4	211	0	GCTGAGACATTCA/GTTCGCATTATTC	2487	211	8.48	4e–96
179546	2	1777	1	110	CAAAGTCCCAAGC/TAATTATTTTATG	1777	110	6.19	8e–48
183486	4	2981	4	220	ATAAAAAAGACGC/TAAATATGAATGA	2981	220	7.38	5e–92
185137	147	2603	1	4	GACATGGAAGAGC/AGATATGAATTAA	2603	147	5.65	1e–62

*Number of reads of each of the four bases at the stated position. Only reads matching exactly the flanking sequences in the column 'Nucleotide sequence' are included in the count.
†Count of the dominant residue.
‡Count of the second most frequently observed base.
§Percentage of all reads with the second most frequent base.
‖Probability that the second most frequent base occurs at this frequency or greater due to random sequence reading error calculated by the log-likelihood ratio test.

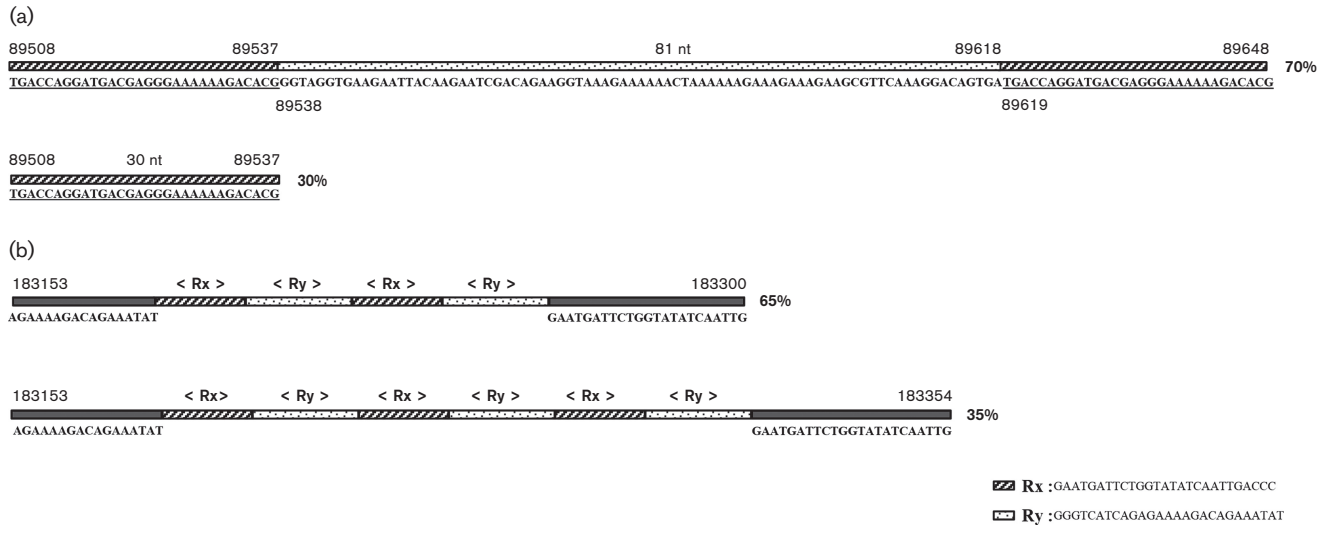

Fig. 3. Repeat status and genetic heterogeneity in the GpSGHV-Eth genome. (a) Approximately 30 and 70 % of the 89 508–89 648 nt region were covered by short repeats of 30 and 26 nt, respectively. (b) Two repeat sequences, 27 nt each, occur two and three times in 65 and 35 %, respectively, of the reads in the 183 153–183 354 nt region.

This notwithstanding, deletions and/or insertions are known to result in several virus strains, e.g. for baculoviruses (Barrera *et al.*, 2013; López-Ferber *et al.*, 2003; Virto *et al.*, 2014). Furthermore, such variation may play important roles in virus pathogenesis (Bernal *et al.*, 2013; Clavijo *et al.*, 2010; Simón *et al.*, 2004, 2005). The high error rate observed in high-throughput sequencing is well known and the observed single nucleotide polymorphisms (SNPs) could be due simply to read errors. To test this, the probability of obtaining the number of reads observed, assuming that read errors are random, was calculated by the log-likelihood ratio test (Sokal & Rohlf, 2012) and is presented in Table 2. All of the observed SNPs are highly significant.

Repetitive regions in the GpSGHV-Eth genome

We detected several repetitive regions with head-to-tail tandem repeat sequences (TRSs) and one inverted repeat sequence in the GpSGHV-Eth genome (Fig. 4). These repeat elements were found to be distributed throughout the GpSGHV-Eth genome, representing ~3 % of the genomic sequence. The lengths of the repetitive regions ranged from 171 to 556 nt and consisted of 78 TRS minifragments. Most of these TRS segments were highly homologous to each other and clustered in two genomic regions (86 000–88 000 and 179 000–183 000 nt). The sizes of the TRS varied from 52 to 246 nt and the number of TRSs per repetitive region varied from 2.7 to 14.5. Within the same repetitive region, the identity of TRS was >80 %, but the sequence identity amongst the different repetitive region varied from 21.2 to 96.2 %. TRS9, 11 and 13, and 10, 12 and 14 shared >80 % sequence identities amongst each other. Occurrence of repeat sequences at multiple locations along the genomes has been reported in baculoviruses (Cochran & Faulkner, 1983), nudiviruses (Wang *et al.*, 2011) and whispovirus (Syed Musthaq *et al.*, 2006). Potentially, the repetitive elements may serve as regulators of viral gene expression (Schnitzler *et al.*, 1987).

Comparison between the GpSGHV-Eth and GpSGHV-Uga genomes

We observed a strong collinearity between GpSGHV-Eth and GpSGHV-Uga genomic sequences (Fig. 2), which was corroborated by BLAST searches in that only 24 of the 174 GpSGHV-Eth ORFs remained without any hits to GpSGHV-Uga ORFs (Table 1). The major differences between the genomes of the two viruses are depicted in Figs 1 and 5. Compared with the GpSGHV-Uga genome, the GpSGHV-Eth genome had an insertion of ~500 nt in the 20 000 nt region, which is immediately followed by a deletion of ~600 nt. Similarly, a short insertion was observed in the 72 000 nt region and a long deletion (~400 nt) was observed in the 87 000 nt region. Several other insertions of a combined length of almost 1250 nt were observed in the 150 000–161 000 nt region, followed by several deletions in the 162 000–1 90 000 nt region (see summary in Fig. 5). Combined, these insertions and deletions make the GpSGHV-Eth genome 249 nt longer and more ORF-dense than the GpSGHV-Uga genome. Taken together, our analyses of the two virus genomes led us to omit 10 small ORFs (four in the 16 000–18 000 nt region and six dispersed along the sequence) reported in the GpSGHV-Uga genome. On the same note, we included 24 new small ORFs (five in the 16 000–22 000 nt region and 19 dispersed along the sequence) in the GpSGHV-Eth genome (Fig. 1).

We then determined which specific ORFs contained deletions and/or insertions. We found that, compared with GpSGHV-Uga, a total of 37 ORFs in the GpSGHV-Eth ORFs contained several nucleotide insertions and/or deletions, of which 17 and 20 ORFs had insertions and deletions, respectively (Table S1; compare with Fig. 1). As can be observed in Fig. 1, the ORFs with insertions and deletions appeared to occur in clusters. For instance, the 5000–25 000 nt region contains three ORFs with insertions (SGHV-Eth005, SGHV-Eth013 and SGHV-Eth014) and three ORFs with deletions (SGHV-Eth008, SGHV-Eth009 and SGHV-Eth018). Interestingly, this region also contained four ORFs that lack homologues in GpSGHV-Eth (SGHV-Uga013, SGHV-Uga14, SGHV-Uga015 and SGHV-Uga018) and five novel ORFs (SGHV-Eth012, SGHV-Eth015, SGHV-Eth016 and SGHV-Eth019) (Fig. 1). Another notable region is the 70 000-90 000 nt region, which contains two ORFs with insertions (SGHV-Eth064 and SGHV-Eth068), a high-molecular-mass ORF with deletions (SGHV-Eth069) and seven novel ORFs (SGHV-Eth059–SGHV-Eth065). Similarly, the 150 000–170 000 nt region contains three ORFs with insertions, six ORFs with deletions and three novel ORFs (Fig. 1). The 162 000–190 000 nt region contains six ORFs with insertions, nine ORFs with deletions, three novel ORFs and two missing ORFs (Fig. 1). Finally, except for ORF SGHV-Eth073 (with deletion) and a small novel ORF (SGHV-Eth085), the remaining 24 ORFs in the 90 000–123 000 nt region did not have any variations when comparing the two virus genomes.

Putative novel ORFs in the GpSGHV-Eth genome

A total of 24 ORFs in GpSGHV-Eth had no homologues to any known gene, including to GpSGHV-Uga (Table 1), implying that they are novel. However, notable motifs/domains were detected in some of the novel ORFs. For instance, ORF SGHV-Eth016 contained an NAD-specific glutamate dehydrogenase (GluDH) motif, whilst ORF SGHV-Eth028 contained the subtelomic variant ORF and repetitive interspersed family (Rif/Stevor) domain. Stevor and Rif family proteins are implicated in the regulation of antigenic variations and gene duplication events in some pathogens (Joannin *et al.*, 2008; Niang *et al.*, 2009), but their role in viruses is not well studied. ORF SGHV-Eth110 contained a viral small hydrophobic protein (v-SHP) domain, which is a retention signal for

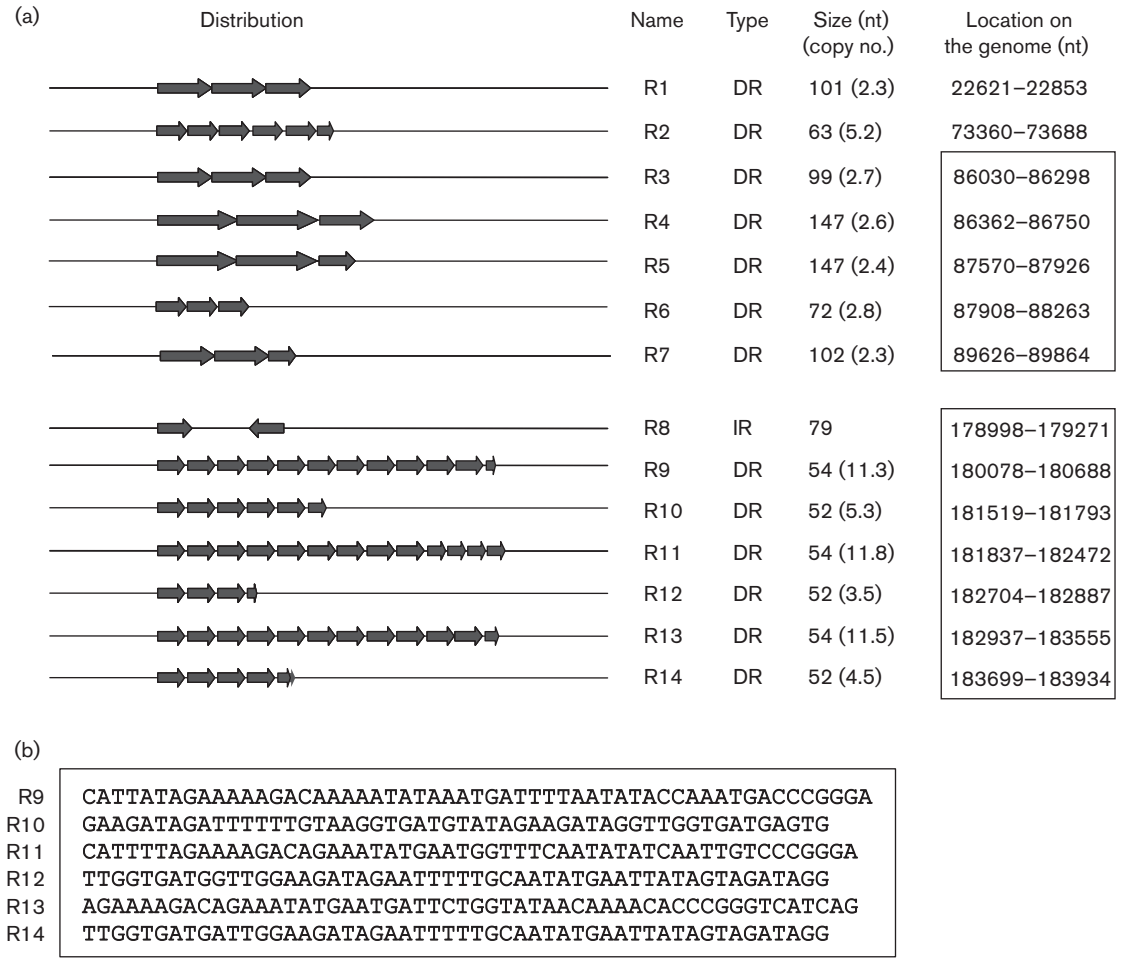

Fig. 4. Loci of the GpSGHV-Eth genome with repeating elements. (a) The arrows represent the repeat core elements in the indicated directions. The name, type, size, copy number and genomic location of the repeat are indicated on the right. DR, direct repeat; IR, indirect repeat. Parameters: length >50 nt, $n>2$. (b) Alignment of the consensus nucleotide sequences of the direct repeats R9, R10, R11, R12, R13 and R14.

intracellular microvesicles (McCarthy & Theilmann, 2008) potentially involved in virus docking. ORF SGHV-Eth155 contained a repeat-associated mysterious proteins (RAMPs) domain – a protospacer sequence targeted by Cas nucleases to cleave viral genomes (Heler *et al.*, 2015). Other notable domains detected in the putative novel ORFs are shown in Table 1. Out of the 24 novel ORFs, 14 contained both putative transcriptional signals [TATA-like box and poly(A) signals]. Four of these ORFs (SGHV-Eth060, SGHV-Eth061, SGHV-Eth109 and SGHV-Eth110) showed both transcripts and peptides mapping onto their sequences, implying that these ORFs are functional.

Selection pressures acting on GpSGHV ORFs

We estimated numbers and rates of synonymous and non-synonymous substitutions in the ORF sequences of GpSGHV-Eth in comparison with the homologous ORFs in GpSGHV-Uga (Table S4). Mutation and selection have different effects on silent (d_S) and amino acid replacement (d_N) substitutions. The d_S/d_N ratios amongst sites provide insights into the functional constraints at different amino acid sites and are useful in detection of sites under positive selection (Nielsen & Yang, 1998). A $d_S/d_N<1.0$ or $d_N/d_S>1.0$ is considered to be a convincing indicator of genes that are under purifying or positive selection pressure, respectively (Kreitman & Akashi, 1995). Based on this criterion, a total of 21 ORFs were considered to be under positive selection pressure. Of the 21 ORFs under positive selection, 11 ORFs had significant homologies to known viral genes. These include three homologues to baculovirus genes (SGHV-Eth136, SGHV-Eth139 and SGHV-Eth140), four homologues to entomopoxvirus genes (SGHV-Eth009, SGHV-Eth035, SGHV-Eth096 and SGHV-Eth116), and homologues to nudivirus (SGHV-Eth052), ascovirus (SGHV-Eth069), mimivirus (SGHV-Eth68) and nimavirus (SGHV-Eth164) genes (Table 1). Two ORFs (SGHV-Eth049 and SGHV-Eth064)

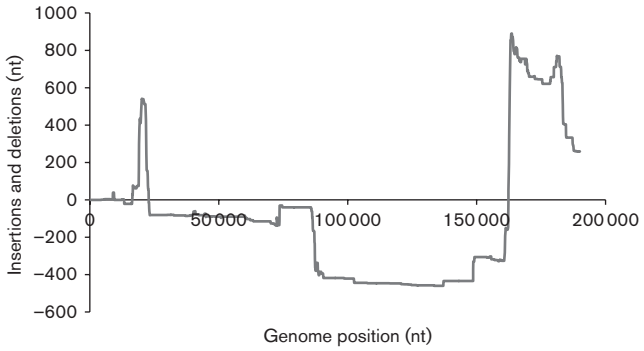

Fig. 5. Inserted and deleted nucleotides in the GpSGHV-Eth genome compared with the GpSGHV-Uga genome. Positive and negative numbers indicate insertions and deletions, respectively, in the two virus genomes.

were homologous to known cellular genes, whilst the GpSGHV envelope and tegument proteins encoded by ORFs SGHV-Eth040 and SGHV-Eth123 were annotated to be HSP90-like ATPase and regulatory proteins, respectively (Table 1). The remaining six ORFs (i.e. SGHV-Eth051, SGHV-Eth108, SGHV-Eth141, SGHV-Eth153, SGHV-Eth165 and SGHV-Eth171) remained without homologies to known genes. It is thought that genes encoding virulence factors undergo intensified episodes of positive selection to maintain or improve the advantage of pathogens over their hosts (adaptive evolution) (van der Ende *et al.*, 1998; Wolf *et al.*, 2006). In this case, some of the notable GpSGHV genes under positive selection pressure include SGHV-Eth139 and SGHV-Eth140 (homologues to baculovirus *cg30* and *pe38* involved in maximum production of occlusion bodies and apoptosis, respectively), SGHV-Eth052 (homologue to the nudivirus very late gene transcription factor *lef-4*), and SGHV-Eth069 (homologue to the ascovirus 2c ORF147) (Table 1). Negative selection occurs in conservative viral genes that are maintained under structural and/or functional constraints to avoid reproductive and other fitness disadvantages (Doi, 1991; Wilson *et al.*, 1977). Knowledge of viral genes under positive selection pressure is critical in the identification of genes involved in virulence/pathogenesis without prior knowledge of the exact mechanism(s) that govern virulence and pathogenesis. Our analysis may clarify the driving force of *Hytrosaviridae* evolution.

Differential pathogenesis of GpSGHV-Eth and GpSGHV-Uga

The differential pathogenesis of GpSGHV in different tsetse colonies could be attributed to the nucleotide variations described above or to the susceptibility of the host itself to virus infection. It has been demonstrated that insect host-encoded factors have an impact on virus pathogenesis, e.g. in the baculoviruses (Asser-Kaiser *et al.*, 2010, 2011). It should be noted that whereas the Ugandan *G. pallidipes* fly population has been maintained at the IPCL for many generations ($\geq$30 years), the Ethiopian *G. pallidipes* population is quite recently colonized. Consequently, the prolonged exposure of the Ugandan *G. pallidipes* population to GpSGHV infections could have resulted in establishment of a virus–host equilibrium and co-existence. Alternatively, colony handling and feeding regimes practised in different mass production facilities could influence expression of SGH, as recently reported by Abd-Alla *et al.* (2013).

CONCLUSIONS

Based on our data, we draw the following conclusions. (i) With a size of 190 291 nt and encoding 174 putative ORFs, the GpSGHV-Eth genome is 98.1 % identical to the reference genome of GpSGHV-Uga. (ii) The combined proteogenomic and transcriptomic mapping significantly improved the annotation of the GpSGHV-Eth genome, allowing the detection of 141 transcripts and 86 proteins. Of the 86 ORFs that were proteogenomically mapped on the virus genome, 13 ORFs appear to contain non-canonical start codons; four of the 13 ORFs had peptides mapping upstream of the first methionine residues. (iii) Of the 132 ORFs that contained transcriptional signals [TATA-like box promoter elements and poly(A) signals], 60 ORFs had both mapped transcripts and peptides, implying that these ORFs are translated into functional viral proteins. Compared with the GpSGHV-Uga genome, where only 47 ORFs could be annotated, 114 ORFs could be annotated in the GpSGHV-Eth genome, including 61 structural proteins. (iv) Amongst the potentially functional elements in the GpSGHV-Eth genome were homologues to the core proteins of the baculoviruses, nudiviruses and herpesvirus that are important in various virus replication cycles. (v) Over 10 % (n=21) of the GpSGHV genes are potentially under positive selection pressure, implying that these genes may have evolved new functions in GpSGHV-Eth compared with the reference GpSGHV-Uga. (vi) As the characteristics of virulence are often due to the synergistic effects of various genomic loci, the cohesive activities of the genetic heterogeneity presented here potentially result in the differential pathogenesis of the two GpSGHV isolates. In addition, it is possible that *G. pallidipes* colonies maintained under different insectary conditions may vary in their susceptibility to GpSGHV infections and may therefore influence the occurrence of overt SGH symptoms. Our data provide a foundation for future investigations of the *Hytrosaviridae* family of insect viruses.

METHODS

Isolation, purification, extraction and sequencing of viral genomic DNA. *G. pallidipes* flies were obtained from a tsetse production facility at the National Institute for Control and Eradication of Tsetse Fly and Trypanosomosis (NICETT), Addis Ababa, Ethiopia.

The virus sample was purified from HSGs (dissected from naturally infected 4-week-old males) and the viral DNA sequenced as described by Abd-Alla *et al.* (2008), with slight modifications. Briefly, after extraction from purified virus suspension, viral DNA was released by Sarkosyl/proteinase K treatment (Qiagen), followed by phenol/chloroform extraction. Based on the genomic sequence of the reference GpSGHV genome, 136 primer pairs were designed and used to amplify 1500 bp amplicons covering the entire virus genomic sequence (Abd-Alla *et al.*, 2007). Each of the PCR products was then purified using a High Pure PCR Purification kit (Roche Biochemicals). The respective PCR primers were used to sequence the PCR products from both ends by the Sanger method (MWG-Biotech). In certain cases, PCR products were cloned into pGEM-T/pGEM T-Easy vector systems (Promega), and then sequenced using T7 and SP6 universal primers according to standard protocols. To cover the sequence of the AT-rich regions of the virus genomic sequence, 10 μg intact DNA was subjected to pyrophosphate-based sequencing (pyrosequencing) (454 Life Sciences) according to Margulies *et al.* (2005). Repeat regions were resolved by DNA sequencing at Macrogen using the HiSeq 2000 platform (Illumina). Sequencing reads were trimmed by Trimmomatic tools version 0.36 (Bolger *et al.*, 2014), and assembled by SeqMan (Lasergene version 7.0; Dnastar) and Vector NTI version 9.0 (Invitrogen) packages. Sequence assembly was validated using a set of routines as described by Parker & Parker (2008).

Protein extraction, fractionation and MS/MS data acquisition. Preparation of *G. pallidipes* salivary gland proteins and MS analyses were performed as described by Kariithi *et al.* (2013b). Briefly, proteins were extracted from HSG extracts, fractionated using 12 % SDS-PAGE and subjected to in-gel trypsin digestion. The resultant tryptic peptides were analysed by LC coupled to electrospray ionization and high-accuracy LTQ-Orbitrap XL tandem MS (LC-MS/MS). The MS/MS spectra were searched against *Glossina* hytrosavirus and tsetse fly protein databases downloaded from UniProt (version May 2014), a database containing sequences of common contaminants and a decoy database created by reversing the database sequences. The MS searches were performed using MaxQuant version 1.3.0.5 (Cox & Mann, 2008) with the Andromeda search engine (Cox *et al.*, 2011). The default false discovery rate of ≤ 0.01 was used at protein and peptide levels during the searches. Any peptide hits to the decoy sequences and hits with modified peptides only were excluded from further analyses.

Deep sequencing of RNAs, assembly and annotation of viral transcriptome. Total RNA was extracted from HSGs (obtained from 4-week-old males as described above) using TRIzol reagent (Invitrogen) and treated with DNase I to remove residual DNA contaminants. RNA samples were purified using a RNeasy Plus Mini kit (Qiagen). Then, RNA libraries were constructed using a TruSeq RNA Sample Prep kit version 2 (Illumina), and the pooled libraries sequenced using the Illumina HiSeq2000 platform in high-throughput mode. To map the RNA-Seq reads onto the GpSGHV genomic sequence, residual Illumina sequencing adapters, low-quality read pairs with average Phred +33 quality scores <30 and low-quality bases (<30) on the ends of the reads were removed using EA-Utils software (Aronesty, 2013). To ascertain from which part of the GpSGHV genome the sequencing reads occurred, the paired reads were aligned to the virus genome sequence using Bowtie 2 (Langmead *et al.*, 2009). Estimation of the virus gene expression levels from the RNA-Seq data was done using EDGE-pro (Magoc *et al.*, 2013).

Prediction, validation and annotation of virus ORFs. The composition and features of the virus genome were analysed as described by Abd-Alla *et al.* (2008), with modifications. The modifications were that, in addition to selection of ORFs containing ≥ 50 aa and with minimal overlaps (<100 nt), the coding regions of the ORFs predicted in the GpSGHV genome were delineated by mapping RNA-Seq transcripts and LC-MS/MS peptides onto the virus genome (Armengaud, 2009). For the transcript mapping, the transcripts were mapped on to the virus genome sequence translated into all six reading frames using BioEdit local BLAST alignment. For proteogenomic mapping, the validated unique peptides derived from the LC-MS/MS spectral matches were mapped back to the virus genome (translated in all six reading frames) using the Proteogenomic Mapping Pipeline (PMP) (Sanders *et al.*, 2011). Briefly, three files were inputted into the PMP: a FASTA file containing validated unique MS/MS peptides, a FASTA file of the GpSGHV genome and a text file containing the National Center for Biotechnology Information (NCBI) genetic code (*genetic_code_table*). The output file was used to analyse the expressed protein sequence tags created after the peptides had mapped to the virus nucleotide sequence (Sanders *et al.*, 2011). Gene ontology annotation of the predicted ORFs was performed using Blast2GO version 3.0.4 (Conesa *et al.*, 2005). Protein domain analyses were performed using various databases, including Pfam (Finn *et al.*, 2008), InterPro (Mitchell *et al.*, 2015) and the NCBI conserved domain database (Marchler-Bauer *et al.*, 2015).

Comparative analysis and genomic synteny of viral genomes. After the annotation, the genomic sequence of GpSGHV-Eth was compared with that of the reference GpSGHV-Uga. For this, the predicted GpSGHV-Eth ORFs were used as query sequences to BLAST against the nr (non-redundant) NCBI protein database. The Artemis comparison tool (Carver *et al.*, 2005) was used to analyse the gene synteny of GpSGHV-Eth compared with GpSGHV-Uga and the closely related hytrosavirus of the house fly, *Musca domestica* salivary gland hypertrophy virus (Garcia-Maruniak *et al.*, 2009). The occurrence of insertions, deletions and SNPs in the protein-coding regions of the GpSGHV-Eth genomic sequences compared with those of GpSGHV-Uga was investigated by aligning the nucleotide sequences of the homologous ORFs using MegAlign (Dnastar). The pair-wise alignments were visualized within MegAlign and descriptions of the nucleotide variations were based on a previously described nomenclature system (den Dunnen & Antonarakis, 2001).

Analysis of positive and negative selection pressures. The occurrence of selection pressures in the protein-coding regions of the GpSGHV-Eth genome in comparison with the GpSGHV-Uga genome was assessed by analysing the rates of non-synonymous substitution per synonymous site (d_N) versus synonymous substitution per synonymous site (d_S). For this, codon-aligned nucleotide sequence sets of GpSGHV-Eth and GpSGHV-Uga ORFs were analysed using SNAP version 2.1.1 (Korber, 2000). SNAP computes the nucleotide substitution differences (codon by codon) between pairs of homologous sequences based on the Nei–Gojobori method (Nei & Gojobori, 1986). Evidence for the occurrence of negative or positive selection pressures is indicated when $d_S/d_N > 1.0$ or $d_N/d_S > 1.0$, respectively.

Nucleotide/protein sequence databases and accession numbers. The accession numbers for the ORFs presented in this paper are from the GenBank, UniProt, or PIR databases. The GpSGHV-Eth genome sequence has been deposited in GenBank.

ACKNOWLEDGEMENTS

This work was funded by the Joint FAO/IAEA Division of Nuclear Techniques in Food and Agriculture, IAEA, Austria. The authors acknowledge Ms Irene Meki, Mr Moges Hidoto, and the staff members of NICETT and IPCL for technical assistance.

REFERENCES

Abd-Alla, A., Bossin, H., Cousserans, F., Parker, A., Bergoin, M. & Robinson, A. (2007). Development of a non-destructive PCR method for detection of the salivary gland hypertrophy virus (SGHV) in tsetse flies. *J Virol Methods* **139**, 143–149.

Abd-Alla, A. M. M., Cousserans, F., Parker, A. G., Jehle, J. A., Parker, N. J., Vlak, J. M., Robinson, A. S. & Bergoin, M. (2008). Genome analysis of a *Glossina pallidipes* salivary gland hypertrophy virus reveals a novel, large, double-stranded circular DNA virus. *J Virol* **82**, 4595–4611.

Abd-Alla, A. M. M., Kariithi, H. M., Parker, A. G., Robinson, A. S., Kiflom, M., Bergoin, M. & Vreysen, M. J. B. (2010). Dynamics of the salivary gland hypertrophy virus in laboratory colonies of *Glossina pallidipes* (Diptera: *Glossinidae*). *Virus Res* **150**, 103–110.

Abd-Alla, A. M. M., Parker, A. G., Vreysen, M. J. B. & Bergoin, M. (2011). Tsetse salivary gland hypertrophy virus: hope or hindrance for tsetse control? *PLoS Negl Trop Dis* **5**, e1220.

Abd-Alla, A. M. M., Kariithi, H. M., Mohamed, A. H., Lapiz, E., Parker, A. G. & Vreysen, M. J. B. (2013). Managing hytrosavirus infections in *Glossina pallidipes* colonies: feeding regime affects the prevalence of salivary gland hypertrophy syndrome. *PLoS One* **8**, e61875.

Abd-Alla, A. M. M., Marin, C., Parker, A. G. & Vreysen, M. J. B. (2014). Antiviral drug valacyclovir treatment combined with a clean feeding system enhances the suppression of salivary gland hypertrophy in laboratory colonies of *Glossina pallidipes*. *Parasit Vectors* **7**, 214.

Armengaud, J. (2009). A perfect genome annotation is within reach with the proteomics and genomics alliance. *Curr Opin Microbiol* **12**, 292–300.

Aronesty, E. (2013). Comparison of sequencing utility programs. *Open Bioinform J* **7**, 1–8.

Asser-Kaiser, S., Heckel, D. G. & Jehle, J. A. (2010). Sex linkage of CpGV resistance in a heterogeneous field strain of the codling moth *Cydia pomonella* (L.). *J Invertebr Pathol* **103**, 59–64.

Asser-Kaiser, S., Radtke, P., El-Salamouny, S., Winstanley, D. & Jehle, J. A. (2011). Baculovirus resistance in codling moth (*Cydia pomonella* L.) caused by early block of virus replication. *Virology* **410**, 360–367.

Barrera, G., Williams, T., Villamizar, L., Caballero, P. & Simón, O. (2013). Deletion genotypes reduce occlusion body potency but increase occlusion body production in a Colombian *Spodoptera frugiperda* nucleopolyhedrovirus population. *PLoS One* **8**, e77271.

Barrett, M. P., Vincent, I. M., Burchmore, R. J., Kazibwe, A. J. & Matovu, E. (2011). Drug resistance in human African trypanosomiasis. *Future Microbiol* **6**, 1037–1047.

Bernal, A., Williams, T., Muñoz, D., Caballero, P. & Simón, O. (2013). Complete genome sequences of five *Chrysodeixis chalcites* nucleopolyhedrovirus genotypes from a Canary Islands isolate. *Genome Announc* **1**, e00873–e00813.

Bolger, A. M., Lohse, M. & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120.

Boucias, D. G., Kariithi, H. M., Bourtzis, K., Schneider, D. I., Kelley, K., Miller, W. J., Parker, A. G. & Abd-Alla, A. M. M. (2013). Transgenerational transmission of the *Glossina pallidipes* hytrosavirus depends on the presence of a functional symbiome. *PLoS One* **8**, e61150.

Carver, T. J., Rutherford, K. M., Berriman, M., Rajandream, M. A., Barrell, B. G. & Parkhill, J. (2005). ACT: the Artemis comparison tool. *Bioinformatics* **21**, 3422–3423.

Chen, Y. R., Zhong, S., Fei, Z., Hashimoto, Y., Xiang, J. Z., Zhang, S. & Blissard, G. W. (2013). The transcriptome of the baculovirus *Autographa californica* multiple nucleopolyhedrovirus in *Trichoplusia ni* cells. *J Virol* **87**, 6391–6405.

Clavijo, G., Williams, T., Muñoz, D., Caballero, P. & López-Ferber, M. (2010). Mixed genotype transmission bodies and virions contribute to the maintenance of diversity in an insect virus. *Proc Biol Sci* **277**, 943–951.

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

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M. & Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676.

Corcelette, S., Massé, T. & Madjar, J. J. (2000). Initiation of translation by non-AUG codons in human T-cell lymphotropic virus type I mRNA encoding both Rex and Tax regulatory proteins. *Nucleic Acids Res* **28**, 1625–1634.

Cox, J. & Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **26**, 1367–1372.

Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V. & Mann, M. (2011). Andromeda: a peptide search engine integrated into the MaxQuant environment. *J Proteome Res* **10**, 1794–1805.

Cui, L., Cheng, X., Li, L. & Li, J. (2007). Identification of *Trichoplusia ni* ascovirus 2c virion structural proteins. *J Gen Virol* **88**, 2194–2197.

den Dunnen, J. T. & Antonarakis, S. E. (2001). Nomenclature for the description of human sequence variations. *Hum Genet* **109**, 121–124.

Doi, H. (1991). Importance of purine and pyrimidine content of local nucleotide sequences (six bases long) for evolution of the human immunodeficiency virus type 1. *Proc Natl Acad Sci U S A* **88**, 9282–9286.

Dresang, L. R., Teuton, J. R., Feng, H., Jacobs, J. M., Camp, D. G. II, Purvine, S. O., Gritsenko, M. A., Li, Z., Smith, R. D. & other authors (2011). Coupled transcriptome and proteome analysis of human lymphotropic tumor viruses: insights on the detection and discovery of viral genes. *BMC Genomics* **12**, 625.

Feldmann, U., Dyck, V. A., Mattioli, R. C. & Jannin, J. (2005). Potential impact of tsetse fly control involving the sterile insect technique. In *Sterile Insect Technique. Principles and Practice in Area-Wide Integrated Pest Management*, pp. 701–723. Edited by V. A. Dyck, J. Hendrichs & A. S. Robinson. Dordrecht: Springer.

Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H. R., Ceric, G., Forslund, K., Eddy, S. R. & other authors (2008). The Pfam protein families database. *Nucleic Acids Res* **36** (Database), D281–D288.

Garcia-Maruniak, A., Abd-Alla, A. M. M., Salem, T. Z., Parker, A. G., Lietze, V. U., van Oers, M. M., Maruniak, J. E., Kim, W., Burand, J. P. & other authors (2009). Two viruses that cause salivary gland hypertrophy in *Glossina pallidipes* and *Musca domestica* are related and form a distinct phylogenetic clade. *J Gen Virol* **90**, 334–346.

Gupta, K. C., Ono, E., Ariztia, E. V. & Inaba, M. (1994). Translation initiation from non-AUG codons in COS1 cells is mRNA species dependent. *Biochem Biophys Res Commun* **201**, 567–573.

Heler, R., Samai, P., Modell, J. W., Weiner, C., Goldberg, G. W., Bikard, D. & Marraffini, L. A. (2015). Cas9 specifies functional viral targets during CRISPR-Cas adaptation. *Nature* **519**, 199–202.

Jaenson, T. G. T. (1978a). Mating behaviour of *Glossina pallidipes* Austen (Diptera, Glossinidae): genetic differences in copulation time between allopatric populations. *Entomol Exp Appl* **24**, 100–108.

Jaenson, T. G. T. (1978b). *Reproductive biology of the tsetse Glossina pallidipes Austen (Diptera, Glossinidae) with special reference to mating behaviour*. PhD thesis, Uppsala University, Uppsala, Sweden.

Jehle, J. A., Abd-Alla, A. M. & Wang, Y. (2013). Phylogeny and evolution of *Hytrosaviridae*. *J Invertebr Pathol* **112** (Suppl), S62–S67.

Joannin, N., Abhiman, S., Sonnhammer, E. L. & Wahlgren, M. (2008). Sub-grouping and sub-functionalization of the RIFIN multi-copy protein family. *BMC Genomics* **9**, 19.

Jordan, A. M. (1986). *Trypanosomiasis Control and African Rural Development*. London: Longman Higher Education.

Kariithi, H. M., Ince, I. A., Boeren, S., Vervoort, J., Bergoin, M., van Oers, M. M., Abd-Alla, A. M. M. & Vlak, J. M. (2010). Proteomic analysis of *Glossina pallidipes* salivary gland hypertrophy virus virions for immune intervention in tsetse fly colonies. *J Gen Virol* **91**, 3065–3074.

Kariithi, H. M., Ince, I. A., Boeren, S., Abd-Alla, A. M. M., Parker, A. G., Aksoy, S., Vlak, J. M. & van Oers, M. M. (2011). The salivary secretome of the tsetse fly *Glossina pallidipes* (Diptera: Glossinidae) infected by salivary gland hypertrophy virus. *PLoS Negl Trop Dis* **5**, e1371.

Kariithi, H. M., Ahmadi, M., Parker, A. G., Franz, G., Ros, V. I. D., Haq, I., Elashry, A. M., Vlak, J. M., Bergoin, M. & other authors (2013a). Prevalence and genetic variation of salivary gland hypertrophy virus in wild populations of the tsetse fly *Glossina pallidipes* from southern and eastern Africa. *J Invertebr Pathol* **112** (Suppl), S123–S132.

Kariithi, H. M., van Lent, J. W., Boeren, S., Abd-Alla, A. M., İnce, I. A., van Oers, M. M. & Vlak, J. M. (2013b). Correlation between structure, protein composition, morphogenesis and cytopathology of *Glossina pallidipes* salivary gland hypertrophy virus. *J Gen Virol* **94**, 193–208.

Kariithi, H. M., van Oers, M. M., Vlak, J. M., Vreysen, M. J., Parker, A. G. & Abd-Alla, A. M. (2013c). Virology, epidemiology and pathology of *Glossina* hytrosavirus, and its control prospects in laboratory colonies of the tsetse fly, *Glossina pallidipes* (Diptera; Glossinidae). *Insects* **4**, 287–319.

Kariithi, H. M., Ince, A. I., Boeren, S., Murungi, E. K., Meki, I. K., Otieno, E. A., Nyanjom, S. R. G., van Oers, M. M., Vlak, J. M. & other authors (2016). Comparative analysis of salivary gland proteomes of two *Glossina* species that exhibit differential hytrosavirus pathologies. *Front Microbiol* **7**, 89.

Korber, B. T. M. (2000). HIV signature and sequence variation analysis. In *Computational Analysis of HIV Molecular Sequences*, pp. 55–72. Edited by A. G. Rodrigo & G. H. Learn. Dordrecht: Kluwer Academic.

Kreitman, M. & Akashi, H. (1995). Molecular evidence for natural selection. *Annu Rev Ecol Syst* **26**, 403–422.

Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25.

López-Ferber, M., Simón, O., Williams, T. & Caballero, P. (2003). Defective or effective? Mutualistic interactions between virus genotypes. *Proc Biol Sci* **270**, 2249–2255.

Magoc, T., Wood, D. & Salzberg, S. L. (2013). EDGE-pro: estimated degree of gene expression in prokaryotic genomes. *Evol Bioinform Online* **9**, 127–136.

Marchler-Bauer, A., Derbyshire, M. K., Gonzales, N. R., Lu, S., Chitsaz, F., Geer, L. Y., Geer, R. C., He, J., Gwadz, M. & other authors (2015). CDD: NCBI's conserved domain database. *Nucleic Acids Res* **43** (D1), D222–D226.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J. & other authors (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380.

McCarthy, C. B. & Theilmann, D. A. (2008). AcMNPV *ac143* (*odv-e18*) is essential for mediating budded virus production and is the 30th baculovirus core gene. *Virology* **375**, 277–291.

Miele, S. A., Garavaglia, M. J., Belaich, M. N. & Ghiringhelli, P. D. (2011). Baculovirus: molecular insights on their diversity and conservation. *Int J Evol Biol* **2011**, 379424.

Mitchell, A., Chang, H. Y., Daugherty, L., Fraser, M., Hunter, S., López, R., McAnulla, C., McMenamin, C., Nuka, G. & other authors (2015). InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res* **43** (D1), D213–D221.

Nei, M. & Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* **3**, 418–426.

Nesvizhskii, A. I. (2014). Proteogenomics: concepts, applications and computational strategies. *Nat Methods* **11**, 1114–1125.

Niang, M., Yan Yam, X. & Preiser, P. R. (2009). The *Plasmodium falciparum* STEVOR multigene family mediates antigenic variation of the infected erythrocyte. *PLoS Pathog* **5**, e1000307.

Nielsen, R. & Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**, 929–936.

Parker, N. J. & Parker, A. G. (2008). Simple tools for assembling and searching high-density picolitre pyrophosphate sequence data. *Source Code Biol Med* **3**, 5.

Rohrmann, G. F. (2013). The AcMNPV genome: gene content, conservation, and function. In *Baculovirus Molecular Biology*, 3rd edn, pp. 153–193. Bethesda, MD: National Center for Biotechnology Information.

Sanders, W. S., Wang, N., Bridges, S. M., Malone, B. M., Dandass, Y. S., McCarthy, F. M., Nanduri, B., Lawrence, M. L. & Burgess, S. C. (2011). The proteogenomic mapping tool. *BMC Bioinformatics* **12**, 115.

Schnitzler, P., Delius, H., Scholz, J., Touray, M., Orth, E. & Darai, G. (1987). Identification and nucleotide sequence analysis of the repetitive DNA element in the genome of fish lymphocystis disease virus. *Virology* **161**, 570–578.

Schofield, C. J. & Kabayo, J. P. (2008). Trypanosomiasis vector control in Africa and Latin America. *Parasit Vectors* **1**, 24.

Simón, O., Williams, T., López-Ferber, M. & Caballero, P. (2004). Genetic structure of a *Spodoptera frugiperda* nucleopolyhedrovirus population: high prevalence of deletion genotypes. *Appl Environ Microbiol* **70**, 5579–5588.

Simón, O., Williams, T., López-Ferber, M. & Caballero, P. (2005). Functional importance of deletion mutant genotypes in an insect nucleopolyhedrovirus population. *Appl Environ Microbiol* **71**, 4254–4262.

Slack, J. & Arif, B. M. (2007). The baculoviruses occlusion-derived virus: virion structure and function. *Adv Virus Res* **69**, 99–165.

Sokal, R. R. & Rohlf, F. J. (2012). *Biometry: The Principles and Practice of Statistics in Biological Research*. New York, NY: Freeman.

Steelman, C. D. (1976). Effects of external and internal arthropod parasites on domestic livestock production. *Annu Rev Entomol* **21**, 155–178.

Syed Musthaq, S., Sudhakaran, R., Ishaq Ahmed, V. P., Balasubramanian, G. & Sahul Hameed, A. S. (2006). Variability in the tandem repetitive DNA sequences of white spot syndrome virus (WSSV) genome and suitability of VP28 gene to detect different isolates of WSSV from India. *Aquaculture* **256**, 34–41.

van der Ende, A., Pan, Z. J., Bart, A., van der Hulst, R. W. M., Feller, M., Xiao, S. D., Tytgat, G. N. J. & Dankert, J. (1998). *cagA*-positive

Helicobacter pylori populations in China and The Netherlands are distinct. *Infect Immun* **66**, 1822–1826.

Virto, C., Navarro, D., Tellez, M. M., Herrero, S., Williams, T., Murillo, R. & Caballero, P. (2014). Natural populations of *Spodoptera exigua* are infected by multiple viruses that are transmitted to their offspring. *J Invertebr Pathol* **122**, 22–27.

Vreysen, M. J. B., Seck, M. T., Sall, B. & Bouyer, J. (2013). Tsetse flies: their biology and control using area-wide integrated pest management approaches. *J Invertebr Pathol* **112** (Suppl), S15–S25.

Wang, Y., Kleespies, R. G., Huger, A. M. & Jehle, J. A. (2007). The genome of *Gryllus bimaculatus* nudivirus indicates an ancient diversification of baculovirus-related nonoccluded nudiviruses of insects. *J Virol* **81**, 5395–5406.

Wang, Y., Bininda-Emonds, O. R., van Oers, M. M., Vlak, J. M. & Jehle, J. A. (2011). The genome of *Oryctes rhinoceros* nudivirus provides novel insight into the evolution of nuclear arthropod-specific large circular double-stranded DNA viruses. *Virus Genes* **42**, 444–456.

Whitnall, A. B. M. (1934). The trypanosome infections of *Glossina pallidipes* in the Umfolosi Game Reserve, Zululand. *Onderstepoort J Vet Sci Anim Ind* **2**, 2–21.

Wilson, A. C., Carlson, S. S. & White, T. J. (1977). Biochemical evolution. *Annu Rev Biochem* **46**, 573–639.

Wolf, Y. I., Viboud, C., Holmes, E. C., Koonin, E. V. & Lipman, D. J. (2006). Long intervals of stasis punctuated by bursts of positive selection in the seasonal evolution of influenza A virus. *Biol Direct* **1**, 34.