Sequence comparison between three geographically distinct *Spodoptera frugiperda* multiple nucleopolyhedrovirus isolates: Detecting positively selected genes

Oihane Simón, Leopoldo Palmá, Inés Beper, Delia Muñoz, Miguel López-Ferber, Primitivo Caballero, Trevor Williams

A R T I C L E   I N F O
Article history:
Received 28 July 2010
Accepted 10 January 2011
Available online 14 January 2011

Keywords:
*Spodoptera frugiperda* nucleopolyhedrovirus
Nicaraguan isolate
Genomic sequence
Positive selected genes

A B S T R A C T
The complete genomic sequence of a Nicaraguan plaque purified *Spodoptera frugiperda* nucleopolyhedrovirus (SfMNPV-B) was determined and compared to previously sequenced isolates from United States (SfMNPV-3AP2) and Brazil (SfMNPV-19). The genome of SfMNPV-B (132,954 bp) was 1623 bp and 389 bp larger than that of SfMNPV-3AP2 and SfMNPV-19, respectively. Genome size differences were mainly due to a deletion located in the SfMNPV-3AP2 egt region and small deletions and point mutations in SfMNPV-19. Nucleotide sequences were strongly conserved (99.35% identity) and a high degree of predicted amino acid sequence identity was observed. A total of 145 open reading frames (ORFs) were identified in SfMNPV-B, two of them (sf29 and sf710a) had not been previously identified in the SfMNPV-3AP2 and SfMNPV-19 genomes and one (sf57a) was absent in both these genomes. In addition, sf6 was not previously identified in the SfMNPV-19 genome. In contrast, SfMNPV-B and SfMNPV-19 both lacked sf129 that had been reported in SfMNPV-3AP2. In an effort to identify genes potentially involved in virulence or in determining population adaptations, selection pressure analysis was performed. Three ORFs were identified undergoing positive selection: sf49 (pif-3), sf57 (odv-e66b) and sf122 (unknown function). Strong selection for ODV envelope protein genes indicates that the initial infection process in the insect midgut is one critical point at which adaptation acts during the transmission of these viruses in geographically distant populations. The function of ORF sf122 is being examined.

1. Introduction

The fall armyworm, *Spodoptera frugiperda* (JE Smith) (Lepidoptera: Noctuidae) is a severe pest of maize, sorghum and rice in tropical and subtropical areas of the Americas. Isolates of *S. frugiperda* multiple nucleopolyhedrovirus (*SMNPV*) are known to cause natural epizootics of disease and have been recognized as potential control agents against this pest (Escribano et al., 1999; Fuxa, 1982; Shapiro et al., 1991). A Nicaraguan isolate (SMNPV-NIC) was selected for formulation and field trials in Honduras and Mexico (Williams et al., 1997), and has been extensively studied at the genomic and phenotypic levels (López-Ferber et al., 2003; Simón et al., 2004, 2005b, 2008). SMNPV-3AP2 is a single genotype isolate from Missouri (USA) with a fast killing phenotype that was tentatively attributed to its lack of the egt gene (Harrison et al., 2008). The SMNPV-19 isolate, purified from a single larval cadaver collected in Paramá State, southern Brazil, was selected from 22 SMNPV isolates collected in Brazil due to its high pathogenicity against fall armyworm larvae (Barreto et al., 2005; Wolff et al., 2008).

Adaptation of viruses to their local host populations has been described in various occasions. Comparative studies have reported that a Honduran population of *S. frugiperda* was most susceptible to a Central American isolate (Escribano et al., 1999), whereas insects from a Colombian population were most susceptible to a SMNPV isolate from Colombia (O. Simón, unpublished data). Similar findings have been observed by others working on different insect-nucleopolyhedrovirus systems elsewhere (Erlandson et al., 2007; Erlandson, 2009). This pattern is likely to be generated through processes of host–pathogen co-evolution, in which selection favors viruses with high transmissibility in the local host population. A consequence of this is that the choice of an isolate as the basis for a biological insecticide requires the assessment of a range of geographical isolates, including those from the region in which the biological control program is required.
A number of genes intervene in the species-specific virulence or host range of NPVs by affecting the ability to infect and replicate in cells and thereby influencing the dose required for host mortality, or the survival time of infected larvae from a given host population or species (Chen and Thiem, 1997; Chen et al., 1998; Popham et al., 1998). In most cases, species-specific effects of a gene on virulence or host range have been elucidated by ORF disruption or deletion. Eliminating or reducing virulence against one species following knock-out of a particular gene suggests that acquisition of new genes via recombination has the potential to shape NPV host range. However, very small changes such as individual amino acid replacements in particular viral genes have also proven influential in expanding host range or modifying virulence (Argaud et al., 1998; Kamita and Maeda, 1997; Yang, 1998). Additionally, nucleotide substitutions may lead to alterations in the activity of the encoded protein that facilitate adaptation to a new host or effectively overcome the defenses of a current host. Such mutations would confer a selective advantage and would be fixed in the population at a higher rate than silent or neutral substitutions (Harrison and Bonning, 2004).

The availability of the complete sequences of isolates from the same virus species from different geographical areas offers the possibility of another kind of analysis, based on a bioinformatics approach. In the present study, the complete sequence of an SfMNPV-NIC genotype was compared with those of SfMNPV-3AP2 and SfMNPV-19. Maximum-likelihood models of codon substitutions were used to test for selection on ORFs. The non-synonymous to synonymous substitution rate ratio ($\omega = d_\text{a}/d_\text{s}$) provides a sensitive measure of selection on amino acid sequences. We reasoned that comparative analyses of SfMNPV genomes from geographically distinct sources could highlight genetic differences involved in the marked phenotypic variation in virulence that has been reported in these isolates, and thereby focus attention on the key genes involved in traits that are considered desirable for the development of effective bioinsecticidal products.

2. Materials and methods

2.1. Virus

The Nicaraguan SfMNPV isolate (SfMNPV-NIC) was initially collected from larval cadavers collected from a maize field in Nicaragua. The isolate was propagated in S. frugiperda larvae reared in the laboratory and subjected to an uncertain number of passages. SfMNPV-NIC genotypic variants were isolated by plaque purification (Simón et al., 2004) and the predominant genotype that possessed the largest genome (SfMNPV-B) was selected for bacmid construction (Simón et al., 2008a).

Ocular clones (OBs) obtained from larvae infected with SfMNPV-B were extracted filtered through cheesecloth, washed twice with 0.1% SDS and once with 0.1 M NaCl and finally resuspended in bidistilled water. OB suspensions were quantified using a bacterial counting chamber and stored at 4 °C until used.

2.2. Viral DNA isolation and cloning

SfMNPV-B bacmid DNA was sequenced to ensure that the nucleotide sequence originated from a single genotype and was not the result of contamination from a persistent infection present in the insect colony (Simón et al., 2010). The complete SfMNPV-B genome was cloned in a pBACe3.6 vector that includes a chloramphenicol resistance gene (Frengen et al., 1999). The pUC19 plasmid was removed from pBACe3.6 by treatment with BamHI. pBluescript KS-I with a modified polylinker that included the Ascl restriction site, was inserted using T4 DNA ligase (New England Biolabs, Ipswich, MA). To include the Ascl restriction site in pBluescript KS-I polylinker, the polylinker was amplified by PCR using two specific primers. One of the primers included a SacI and an Ascl restriction site sequence, and a 20 nt homolog to the right arm of the polylinker, whereas the second primer included a 20 nt homolog to the left arm of the polylinker, and Ascl and KpnI restriction site sequences. The PCR product was double digested with SacI and KpnI and cloned into pBluescript KS-I previously digested with these two enzymes. Once the modified pBACe3.6 was obtained, this and the SfMNPV-B DNA purified by CsCl were digested with Ascl and ligated overnight at 16 °C using the T4 DNA ligase (Simón et al., 2008a). The ligation reaction was dialyzed for 4 h against TE buffer. After dialysis, the ligation was used to transform DH10B GeneHogs electrocompetent cells. Transformed cells were incubated at 37 °C for 1 h in SOC medium and colonies were selected in the presence of chloramphenicol. Colonies were selected on chloramphenicol medium and bacmid DNAs were purified by alkaline lysis. DNA was digested with PstI for comparison with the restriction profiles of each of the Nicaraguan SfMNPV genotypes (Simón et al., 2004). The bacmid with the complete genotype was selected for sequencing. Alkaline lysis and CsCl gradient purified bacmid DNA was used for sequencing (Sistemas Genómicos S.L., Paterna, Valencia, Spain).

2.3. DNA sequencing and sequence analysis

Shotgun sequencing was performed using a genomic library of SfMNPV-B bacmid DNA constructed into a sequencing vector. Sequence information was generated from 2304 reactions performed on 1152 clones. The depth of sequence coverage across the genome was 4–8x. Sequencing reactions were set up using the ABI Prism Big Dye Terminator Cycle Sequencing Reaction kit on a 9600 of PE model thermocycler. The reaction products were loaded in an automated DNA sequencer ABI PRISM. The DNA and deduced amino acid sequences were compared against the updated GenBank/EMBL, SWISS-PROT and PIR databases using BLASTn, (Altschul et al., 1990; Pearlson, 1990). SfMNPV-3AP2 (accession number EF035042) and SfMNPV-19 (accession number EU258200) genomic sequences were obtained from the GenBank database and published sources. Sequence alignments and gene-parity plots were performed using NCBI BLAST alignment tools (Altschul et al., 1990) to examine genome organization and order of homologous ORFs. SignalP (Bendtsen et al., 2004), TargetP (Emanuelsson et al., 2000) and Virus-Ploc tools (Chou and Shen, 2008) were also used to determine the nature of selected proteins of unknown function.

2.4. Analysis of selection pressures

PAML software package version 4 (Yang, 2007; http://abacusgene.ucl.ac.uk/software/paml.html) was used to estimate positive selection acting on SfMNPV ORFs and to infer amino acid sites subject to this process. This software uses a maximum-likelihood approach to determine the numbers of non-synonymous (amino acid changing) substitutions per non-synonymous site ($d_\text{a}$) and of synonymous (silent) substitutions per synonymous site ($d_\text{s}$). The ratio of $d_\text{a}$ to $d_\text{s}$, $\omega$, is a measure of the magnitude of selection acting on a gene. Genes with a value of $\omega < 1$ are undergoing negative or purifying selection, in which non-synonymous mutations are eliminated at a faster rate than synonymous mutations because of their deleterious influence on fitness. Finally, genes for which $\omega > 1$ are undergoing positive or diversifying selection in which non-synonymous mutations are fixed at a faster rate than synonymous mutations as they positively influence fitness. To achieve this, sequences were examined for the presence...
of indels causing a frameshift and the affected codons were removed from the alignments; indels that did not result in a frameshift were not eliminated. Sequence data were then fitted to codon-based substitution models that allow \( \omega \) to vary among sites, by using the CODEML application implemented in PAML. In the simplest model M0 (one ratio), it is assumed that the \( \omega \) ratio is an average over all sites. The “nearly neutral” model (M1) allows for conserved sites when \( 0 < \omega < 1 \) and completely neutral sites when \( \omega = 1 \). The “positive selection model” (M2) adds a third class to M1 in which \( \omega \) can take values > 1. Model M3 (discrete) has three classes with proportions \( p_0, p_1 \) and \( p_2 \) and \( \omega_0, \omega_1 \) and \( \omega_2 \) values estimated from the data. Model M7 (\( \beta \) and \( \omega \)) adds an extra class of sites to the M7 model in which \( \omega \) can be greater than one. Models M0 and M1 are nested with models M2 and M3, and model M7 is nested with model M8. Nested models can prove particularly useful since log-likelihood values can be compared using a likelihood ratio test. The level of significance was calculated as follows: the difference between the log-likelihood values for two models was doubled and the resulting value was compared with a \( \chi^2 \) distribution with the number of degrees of freedom calculated from the difference in the number of parameters between the models. The comparison provides a \( P \) value that allows the null hypothesis (no positive selection, integrated in models M0, M1 and M7) to be accepted or rejected. Positive selection can be only inferred when models M2, M3 or M8 indicate codons with a \( \omega \) ratio > 1 and the likelihood ratio test of positive selection is significant at \( P < 0.05 \). Finally, a Bayesian Empirical Bayes (BEB) approach was used to identify the sites that potentially are under positive selection in the protein. Thus, identification of genes with these values represents persuasive evidence of adaptive evolution.

3. Results and discussion

3.1. General characteristics of the Nicaraguan SfMNPV genome

The complete genome of the SfMNPV-B plaque isolate comprised 132,954 bp, approximately 4.2 kb larger than the genome size estimated for this isolate by REN analysis (Simón et al., 2005a), and 1617 and 389 bp larger than the SfMNPV-3AP2 and SfMNPV-19 genomes, respectively. These size differences among geographically distinct SfMNPV isolates were accounted for almost entirely by a deletion of 1428 bp located in the egt region of the SfMNPV-3AP2 genome (Harrison et al., 2008) and by a series of point mutations and short deletions or insertions that resulted in the addition or deletion of codons in the SfMNPV-19 genome (Wolff et al., 2008). The SfMNPV-B genome sequence has a G+C content of 40.30% compared to 40.20% and 40.26% for SfMNPV-3AP2 and SfMNPV-19, respectively (Harrison et al., 2008; Wolff et al., 2008). This value is also comparable to those of other group II NPVs. Deletion of the SfMNPV-B egt gene had no effect on OB infectivity (Simón et al., 2005a), whereas the egt and sf27 genes of the SfMNPV-B genome are 525 and 168 codons respectively (Supplemental material, Table S1), in agreement with findings for the Brazilian isolate (Tumilasci et al., 2003; Wolff et al., 2008). The baculovirus egt encodes an ecdysteroid UDP-glucosyltransferase (EGT) that alters the hormonal balance of the host larva, preventing the larval molt and ensuring the continuation of insect feeding and weight gain (ÖReilly, 1995; ÖReilly and Miller, 1991). The continued growth of the infected insect is reflected in an increased production of viral OBs. Deletion of egt from the viral genome has been shown to increase the speed of kill of the virus with a corresponding reduction in crop damage (Chen et al., 2000; ÖReilly, 1995; Slavicek et al., 1999; Treacy et al., 1997). Harrison et al. (2008) suggested that the deletion of most of the egt ORF in SfMNPV-3AP2 provided a potential explanation for its rapid speed of kill in bioassays. However, minority genotypes present in the Nicaraguan SfMNPV isolate that present deletions that do not encompass egt, among other genes, proved to be less virulent than deleted genotypes that carried this gene intact, possibly due to the absence of genes located elsewhere (Simón et al., 2004, 2005b). This deletion also affected the sf27 ORF of SfMNPV-3AP2 (Table S1), an ORF that occurs only in group II NPVs (Harrison et al., 2008; Ijkel et al., 1999; Jakubowska et al., 2006; Li et al., 2002). The presence of a conserved early motif suggests that sf27 is an early transcribed gene but the expression and function of this ORF have yet to be characterized.

The SfMNPV-3AP2 and SfMNPV-19 genomes both contain a 310 bp deletion with conserved endpoints in the sf57–sf58 region. This deletion includes the final ten codons of sf57 (odv-e66a), residues 683 to 691 in the complete intergenic region that contains the polyadenylation signals, and the complete unique ORF sf57a of SfMNPV-B (Fig. 1). The ortholog of sf57 in SeMNPV is 723 codons in length. sf57 is a paralog of ac46 from Autographa californica MNPV (AcMNPV) that encodes an ODV envelope protein, ODV-E66, that directs reporter proteins to the ODV envelope (Braunagel et al., 2004; Hong et al., 1997). Two copies of odv-e66 are present in some members of group II NPVs, odv-e66a (sf57) and odv-e66b (sf115), whereas just one copy is found in group I NPVs (Ayers et al., 1994; Harrison et al., 2008; Ijkel et al., 1999; Jakubowska et al., 2006; Li et al., 2002); based on sequence identity values, sf115 was considered to be the ortholog of ac46. The biological effects of the presence of two copies of this gene in some NPV genomes have not been studied. This deletion also affects the sf57a ORF of unknown function. This ORF is absent in both SfMNPV-3AP2 and SfMNPV-19 (Table S1, Fig. 1). The expression and function of this ORF are currently the subject of study.

Two insertions were found in the SfMNPV-B genome. The longest insertion, 74 bp in length was located in the intergenic region between sf85 ORF (unique to SfMNPV) and the sf86 gene (p26a), located on opposite strands. Neither the start codon of sf85, the promoter, or the stop codon of sf86 were affected, suggesting that the insertion is unlikely to substantially affect the biological activity of the virus. The second insertion in the SfMNPV-B genome, 61 bp in length, starts 437 nt upstream from the ATG start codon of sf131, resulting in a putative product of 299 amino acids, compared to 279 in the other two isolates (Table S1) (Harrison et al., 2008; Wolff et al., 2008) or 278 residues in SeMNPV (Ijkel et al., 1999). sf131 is the ortholog of ac136, that encodes the P26 protein. Acp26 is a non-essential gene for AcMNPV replication in vitro (Rodems and Friesen, 1993) or in vivo (Simón et al., 2008b), and its deletion had no substantial effects on the infectivity of AcMNPV OBs. SfMNPV isolates possess two copies of this gene as do other members of group II NPVs. Deletion of the SfMNPV-B p26b gene had no effect on OB infectivity (Simón et al., 2008b).
Within the SfMNPV-19 genome a deletion of 186 bp located in the 5' region of the \emph{dutpase} gene (sf53) was detected when compared with the SfMNPV-B and SfMNPV-3AP2 genomes. In SfMNPV-19 the \emph{dutpase} gene consisted of 124 aa, which is 21 aa shorter than that in SfMNPV-B and SfMNPV-3AP2 (sf54) (Table S1). Wolff et al. (2008) indicated that the non-coding region located immediately upstream from the SfMNPV-19 \emph{dutpase} gene was poorly conserved and suggestive of a mutation hotspot.

Finally, a deletion of 192 bp was detected in \emph{hr8} located immediately after the \emph{sf128} gene (\emph{lef-6}) in the SfMNPV-B genome. The effect of this deletion is not known. Both the number and length of \emph{hrs} regions are variable among the different baculovirus genomes that have been sequenced to date. The number of \emph{hrs} varies between six in SeMNPV (Ijkel et al., 1999), five in \emph{Agrotis segetum} NPV (Ag-seNPV) (Jakubowska et al., 2006) and four in MacoNPV (Li et al., 2002). The biological significance of this variability in \emph{hr} number is not known. As \emph{hrs} are possible origins of DNA replication and enhancers of gene expression (Possee and Rohrmann, 1997) this deletion could influence DNA replication and gene expression in SfMNPV-B. A total of eight \emph{hr} regions consisting of one to four imperfect palindromic repeats with conserved repeat size of around 44 nt were detected in the SfMNPV-B genome dispersed along the genome in intergenic regions (Table S1). This pattern concurs with the distribution of \emph{hrs} found in SfMNPV-3AP2 (Harrison et al., 2008) and SfMNPV-19 (Wolff et al., 2008).

Other point mutations and short deletions or insertions that resulted in additional or elimination of codons were identified among the different genomes, the biological implications of which are currently being investigated.

### 3.2. Comparison of SfMNPV ORFs

The principal differences between isolates at the genome level were due to differences in gene content between genomes. A total of 145 ORFs were identified in the SfMNPV-B genome, defined as methionine-initiated ORFs encoding putative proteins of more than 50 aa and with minimal overlapping with adjacent ORFs (Table S1). Among these ORFs, 141 are common to the three genome sequences. To examine the genome organization, the order of homologous ORFs between all the isolates was compared using gene-parity plot analysis. The gene arrangement and orientation was almost completely collinear between all the isolates, whereas minimal differences were observed in gene content (Fig. 1). SfMNPV-B presented three ORFs (sf39a, sf57a and sf110a) that were not detected in the genomes of SfMNPV-3AP2 (Fig. 1A) and SfMNPV-19 (Fig. 1B). These three ORFs did not overlap adjacent ORFs. The sf57a gene encodes a protein comprising 53 aa and is included in the 310 bp deletion present in the SfMNPV-3AP2 and SfMNPV-19 genomes. In contrast, we did not observe major inter-genomic differences (deletions or insertions) at the nucleotide level in the sf39a or sf110a regions that resulted in the deletion of these genes in SfMNPV-3AP2 or SfMNPV-19. However, both sf39a and sf110a were located downstream and upstream from homologous regions \emph{hr4} and \emph{hr7}, respectively. Harrison et al. (2008) and
Wolff et al. (2008) considered both regions as hrs, whereas in the SfMNPV-B genome two methionine-initiated ORFs larger than 50 aa were identified. In contrast, the SfMNPV-B and SfMNPV-19 genomes lacked the SfMNPV-3AP2 ORF129 homolog. Harrison et al. (2008) identified sf129 comprising 122 codons between nt 117,291 and 117,659 in polyhedrin sense orientation. SfMNPV-3AP2 encodes a 94 aa polypeptide, located between nt 117,372 and 117,656 in anti-polyhedrin sense orientation, that completely overlaps the sf129 ORF. When the sf129–sf130 region was compared between both genomes, nucleotide sequence variation was observed between nt 117,345 and 117,350 in SfMNPV-3AP2 and nt 118,934 and 118,340 in SfMNPV-B (Fig. 2). This comprises an AAATTT sequence present in the SfMNPV-3AP2 genome, whereas a TATGCGC sequence is present in SfMNPV-B. The latter was followed by four nucleotides and a stop codon four that reduced sf129 ORF to 21 aa, leading us to reject sf129 as a valid ORF. In addition, sf129 of SfMNPV-3AP2 has no homologs in other baculovirus genomes, whereas sf130 had sequence homology with an ORF of unknown function that is present in most baculoviruses; SeMNPV se128 (Ijkel et al., 1999), Spodoptera litura MNPV splt28 (Pang et al., 2001) or AcMNPV ac29 (Ayres et al., 1994), suggesting that the sf129 ORF described by Harrison et al. (2008) may not be a true ORF based on the degree of overlap with sf130. Finally, the sf6 ORF is present in the genomes of SfMNPV-B, SfMNPV-3AP2 and SfMNPV-19, although this appears to have been overlooked in the latter, but no differences in nucleotide sequences were observed among the different genomes in this region. All the other ORFs were conserved between all genomes and showed a high degree of predicted amino acid sequence conservation. In general, nucleotide sequence identities were at least 99.5% with few gaps.

The fall armyworm is a migratory pest that is distributed from Argentina to the United States, with a correspondingly high likelihood of virus dispersal along the insect’s migratory routes. This may explain the high similarity found among the different SfMNPV genomes from these areas; high levels of gene conservation were observed between SfMNPV-3AP2 and SfMNPV-19 despite the geographical separation of their points of isolation. SfMNPV-3AP2 purified from a wild-type population is a single genotype with a deletion in the egt region. Harrison et al. (2008) also reported the presence of different genotypes within the wild-type population that presented similar deletions to those found in the SfMNPV-NIC isolate (Fig. 3). Genotypes that lacked oral infectivity and egt minus genotypes were present in both isolates (Harrison et al., 2008; Simón et al., 2004). It seems that the SfMNPV isolate from Missouri is genotypically structured in a similar way to the SfMNPV-NIC isolate. Indeed, the genotypic structure of the SfMNPV-NIC population appears to maximize the likelihood of transmission (López-Ferber et al., 2003; Simón et al., 2005b, 2008c). Experiments are in progress to determine the similarity at the population level among the different SfMNPVs and its importance in the ecology of the virus.

3.3. Analysis of selection on SfMNPV isolates genes

Selection analysis (Yang, 1998) performed on vertebrate virus genes has identified positively selected sites in regions involved in host immune response and receptor binding (Holmes et al., 2002; Twiddy et al., 2002; Woelk et al., 2001; Woelk and Holmes, 2001). Analysis of variable selection performed among amino acid sites in HIV genes (Yang et al., 2003), Trypanosoma brucei essential

Fig. 2. Sequence comparison between sf129–sf130 gene regions of SfMNPV-3AP2 and SfMNPV-B genomes.
surface proteins (Emes and Yang, 2008) and M proteins of influenza A virus (Furuse et al., 2009) revealed sites involved in the adaptive evolution of these intracellular parasites. This technique could help to identify baculovirus genes involved in species-specific virulence and host range, or those involved in adaptation to new host species. Alternatively, positively selected sites could also be involved in selection for improved physical stability or transmission. In the present study, pairwise comparisons and models that allow changes in the value of \( \omega \) among codon sites were used to evaluate selection on genes from geographically distinct SfMNPV genotypes.

Following the pairwise approach in PAML, the average \( \omega \)-value was 0.1743 suggesting that negative selection has prevailed during the evolution of most of the ORFs present in SfMNPV genomes (data not shown). When comparing genes among different viruses, overall negative selection values have also been observed (Harrison and Bonning, 2003, 2004). Nevertheless, it is difficult to detect positive selection using a pairwise approach mainly because a large number of amino acids in a protein remains invariant due to strong functional constraints (\( \omega \) close to 0), requiring that this kind of analysis be complemented with models that allow for heterogeneous \( \omega \) values among different codon sites (Harrison and Bonning, 2003). In this second analysis, ORF alignments were fitted to six models, M0, M1, M2, M3, M7 and M8 (Harrison and Bonning, 2003, 2004). Models M2, M3 and M8 identified positively selected sites in several ORFs. However, the null hypothesis models (M0, M1 and M7) were rejected only for sf49, sf57 and sf122. To rule out sequencing errors that could have generated anomalous results in the selection analysis, these three ORFs were subjected to re-sequencing. For this, sf49, sf57 and sf122 were amplified from SfMNPV-B DNA with a high fidelity Taq Polymerase (Takara), cloned into pUC19 and re-sequenced using standard primers. The sequences and alignments of the positively selected ORFs from the three SfMNPV genomes showed no evidence of sequencing errors following re-sequencing (Fig. 4). The sf49 and sf57 genes have been previously identified as pif-3 and odv-e66a, respectively, whereas no function or protein product has been demonstrated for the sf122 ORF. Homologs of sf122 have been found in some, but not all members, of group II NPVs; SeMNPV (se121) (Ijkel et al., 1999), Agrotis ipsilon NPV (agip140) (Harrison, 2008), Ag-seNPV (agse133) (Jakubowska et al., 2006) and SpltNPV (splt126) (Pang et al., 2001).

The evidence of positive selection for pif-3 was stronger than that of odv-e66a but weaker than that of sf122. Estimation of \( \omega = 0.150 \) as an average across all sites (M0) suggested that pif-3 was being subjected to purifying selection (Table 1). However, analyses using models that allow \( \omega \) to vary among sites, indicated which sites of the pif-3 gene were subjected to differing selection intensities. Moreover, the likelihood ratio test suggested rejection of the null hypothesis for each of the three comparisons that were performed (Table 2). Models M2, M3 and M8 consistently identified a positively selected site (61C). The \( \omega \)-values were very high for all three models in approximately 0.5% of the codon sites (\( \omega > 150 \)), indicating a very strong degree of positive selection for

![Fig. 3. Schematic representation of the gene order in the genotypes present within the Nicaraguan isolate of SfMNPV. The deletions and the genes absent in each genotype are indicated. The genes and ORFs found in this variable region are indicated below the arrows in the complete SfMNPV-B genome that point to their directions of transcription.](image-url)
Furthermore, using Bayesian Empirical Bayes (BEB) analysis, site 61 was identified in M2 and M8, as a positively selected site with probabilities greater than 90% but lower than 95% (Table 3).

The pif-3 gene, along with other ODV envelope protein genes p74, pif-1, pif-2, pif-4 and odv-e56 is among the 30 core genes that are found in all baculoviruses sequenced to date (Braunagel and Summers, 2007; Fang et al., 2009; Harrison et al., 2010). These genes encode factors that are essential for per os infection (Fang et al., 2006, 2009; Faulkner et al., 1997; Harrison et al., 2010; Kikhnno et al., 2002; Pijlman et al., 2003). Peroral infection occurs following a two step process (Horton and Burand, 1993): beginning with ODV binding to specific host cell receptors (Haas-Stapleton et al., 2004; Slack et al., 2008), followed by fusion of the cell and ODV membranes, although the mechanisms by which this occurs remain unclear. Recently, it has been shown that PIF-1, PIF-2 and PIF-3, but not P74 are essential for the formation of a stable complex on the AcMNPV ODV surface, and that P74 is associated with this complex that appears to have an essential function in virus entry (Peng et al., 2010). PIF-3 could therefore be influential in determining host range or infectivity at the level of midgut cell binding and internalization of ODVs.

Fig. 4. Multiple sequence alignment of the deduced proteins with positively selected sites from the three SMNPV genomes; (A) PIF-3, (B) ODV-E66A, and (C) SF122. Amino acid sequences were obtained by translating the nucleotide sequences from the codon alignments. Sites at which positive selection is inferred following removal of indels (BEB analysis) with at least 90% of posterior probabilities are indicated in bold type and gray shading (see Table 3).
For odv-e66α the evidence of positive selection was also strong. The average ω value (0.33816) indicates that purifying selection dominated the evolution of the odv-e66α. Model M3 (discrete) suggests a proportion of sites (~1.8%) under weaker diversifying selection with ω2 = 0.15748 and a smaller proportion (~0.7%) under stronger diversifying selection with ω2 = 59.6440. Under Model M2 and M8, the estimates suggest that ~1.6% of sites are evolving by positive selection with ω values ~38 (Table 1). All models that allow for positive selection sites reject the null hypothesis in the likelihood ratio tests (Table 2). BEB analysis of models M2 and M8 identified positively selected sites consisting of the same five residues with probabilities greater than 95% (Table 3).

The odv-e66α is a late gene that encodes for an ODV envelope protein, that redirects protein trafficking to the ODV envelope (Braunagel et al., 2004; Hong et al., 1997). One interesting feature is the presence of two homologs of odv-e66 (sf57 and sf115) in SIMNPV genomes. sf115 appears to be the ortholog of ac66 characterized as odv-e66, whereas sf57 is the paralog that showed the highest degree of homology to odv-e66 (AB009613) from Leucania separate NPV with 64% identity (ID) and 76% similarity (Sim). Two copies of this gene are also found in other members of group II NPVs such as MacoMNPV, SeMNPV and Orgyia leucostigma NPV, whereas a single copy is present in AcMNPV, SplitNPV, AgseNPV or Helicoverpa armigera SNPV. The position of these two genes in relation to nearby ORFs is conserved in these NPVs. It is probable that the two copies of the odv-e66 gene were acquired independently and the first copy originated from a source that was more closely related to LesoNPV and Lymnaea dispar MNPV than to the ortholog of ac66 (sf115). The presence of two copies of odv-e66 with both late and early baculovirus promoters may be related to the two forms of odv-e66 found in mature AcMNPV ODVs (Jikelo et al., 1999; Li et al., 2002). The biological significance of the presence of these two copies in some members of baculoviruses is unknown. In addition, the first copy (sf57) is positively selected which is suggestive of diversifying selection, whereas the sf115 copy did not show this diversify. Such directional selection may cause alterations in the activity of the encoded protein, facilitating the adaptation to a new host species and/or overcoming host defenses.

Finally, analysis of the M2, M3 and M8 models for ORF sf122 resulted in extremely high ω values that are usually obtained when
d_0 tends to zero. The average ω value (~0.152) in M0 indicates relatively strong purifying selection acting on sf122 (Table 1). M0, M1 and M7 models were rejected by likelihood ratio tests (Table 2). BEB analysis of M2, and M8 identified three sites as being positively selected, but only one (96E) with probabilities greater than 95% (Table 3).

sf122 encodes a putative protein 139 aa in length of unknown function that is present in some members of group II NPVs. The sf122 transcripts were classified as late transcribed by RT-PCR (data not shown), suggesting that this is a functional late gene. BLAST searches revealed homologies to SeMNPV SE121 (49% ID and 69% Sim), SpltNPV AGSE133 (38% ID and 61% Sim) and AgipNPV AGIP140 (33% ID and 62% Sim). Sequence comparison revealed that the SF122 protein presented local areas of homology in different regions along the proteins of other organisms, suggesting that those regions could be related to specific protein domains. SF122 is a small putative protein of around 16.7 kDa, without conserved domains. SF122 TargetP analyses demonstrated that the amino acid sequence contained a cytoplasmic membrane localization signal supported by a reliability class value of RC = 3. SF122 SignalP analysis using gram-negative and gram-positive networks, identified this protein as a non-secreted protein. Moreover, virus-Ploc indicated the inner layer of the cytoplasmic membrane as the likely localization site of the putative protein. Further studies will be performed to determine the function of this protein in SIMNPV.

In summary, the likelihood ratio test rejected the null model in each of the three comparisons for the three positively selected genes: M0 (one ratio) against M3 (discrete), M1 (neutral) against M2 (selection), and M7 (β) against M8 (β and α) (Table 2). This suggests diversifying selection for pif-3, odv-e66a and sf122. The pif-3 and odv-e66a genes encode proteins involved in primary infection, whereas no function has yet been described for the late transcribed gene. The primary infection is the first barrier to diversification for these viruses. Diversifying selection may favor alterations in the activity of the encoded protein, facilitating the adaptation to a new host species or overcoming the immune response in the novel host. Experiments are in progress to determine the possible function of these three proteins, particularly the unknown protein SF122. The three geographical SIMNPV genotypes differ significantly in a number of biological activities that have direct relevance to their use as biological insecticides (Harrison et al., 2008; Wolff et al., 2008; Simón et al., 2004).

The probability of rejecting the null hypothesis (neutral selection) when it is wrong, leading us to accept that the alternative hypothesis, positive selection, is likely correct. This probability decreases in accuracy when a low number of sequences are analyzed (Anisimova et al., 2001). It may be expected that from the analysis of only three genome sequences of the same virus species, as reported here, we were only able to reject the null hypothesis models in three instances for which the strength of selection was very high. Similar results were obtained when comparing two closely related sequences (AcMNPV, RoMNPV), for which positively selected sites were detected in only two genes (Harrison and Bonning, 2003). Selection analysis that includes sequences from other NPV genomes may help to identify a greater diversity of viral genes that are subjected to positive selection.

Acknowledgments

We thank N. Gorria (Universidad Pública de Navarra, Pamplona, Spain) for technical assistance. This work was supported by the Grant AGL200507905-C03-01. TW acknowledges support from CONACYt infrastructure Grant 10-110/162/10.

Appendix A. Supplementary material


References


