Molecular Evolution of Alphabaculovirus genomes: Evidence of Mutational bias and Natural selection

Puttatida Mahapattanakul
Queen Mary University of London
Pragun Rajbhandari
Mahidol University International College
Patsarin Rodpothong (patsarin.won@mahidol.edu)
Mahidol University International College https://orcid.org/0000-0002-1244-5192

Research article

Keywords: Codon usage, Baculoviruses, Mutational bias, Natural selection

Posted Date: March 3rd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-244707/v1

Molecular Evolution of Alphabaculovirus genomes: Evidence of Mutational bias and Natural selection

Puttatida Mahapattanakul[1], Pragun Rajbhandari[2] and Patsarin Rodpothong[2*]

[1] Queen Mary University of London, London, United Kingdom
[2] Mahidol International College, Mahidol University, Salaya, Nakhon-Pathom, Thailand
*Corresponding author email: patsarin.won@mahidol.edu

Keywords: Codon usage, Baculoviruses, Mutational bias, Natural selection

Abstract

Codon usage is a reflection of evolutionary adaptation to environmental pressure. The pattern of usage may be unique to species of viruses, genomes of the same species or genes within the same genome. Here we have analysed the overall nucleotide composition and the nucleotides at different codon positions in the genomes of 6 Alphabaculoviruses. Principle Component Analysis (PCA) based on Relative Synonymous Codon Usage (RSCU) of all Open Reading Frames (ORFs) was employed to investigate the pattern of the codon usage. The results suggest the Alphabaculovirus genomes, except that of Agrotis Ipsilon mNPV (AgipNPV), are predominantly under an influence of a neutral mutation that bias toward A/T. The majority of the ORFs, except those of the AgipNPV, cluster at the same location in the 2-dimensional PCA map with one prominent outlier that has been identified as a P6.9 gene. The six Alpha-baculovirus P6.9 genes have a high G/C content, dissimilar to the majority of the ORFs. The G/C content is found to be significantly high at the 2nd codon position, suggesting the influence of natural selection and perhaps reflecting its functional conservation in DNA packaging as well as its evolutionary relation to Protamine.

Introduction

The baculoviruses (family: *Baculoviridae*) are a group of large double-stranded DNA arthropod-specific viruses. They can be categorised into four genera; Alphabaculovirus, Betabaculovirus, Gammabaculovirus and Deltabaculovirus. Baculoviruses can also be classified into two types, nucleopolyhedroviruses (NPVs) and granuloviruses (GVs), based on their occlusion bodies (OBs) produced at the late stages of infection (Rohrmann 2019). The OB is an organized structure, composed of polyhedrin, which provides stability to virions embedded within, and is responsible for virus horizontal transmission among their insect hosts (Clem and Passarelli 2013, Sajjan and Hinchigeri 2016). Genera Alphabaculovirus, Gammabaculovirus and Deltabaculovirus consist of NPVs, infecting insects belonging to the orders Lepidoptera, Hymenoptera and Diptera (Herniou, Arif et al. 2012), while Betabaculovirus consists of GVs and only infects lepidopteran insects.

Genomes of baculoviruses range from 80–180 kbp in size, encoding 90-180 genes. Approximately 37 genes are conserved across different genera of baculoviruses and have been assigned as "Core genes", involving in viral DNA replication and packaging, transcription, architecture and assembly (Herniou, Olszewski et al. 2003, Herniou and Jehle 2007, van Oers and Vlak 2007, Miele, Garavaglia et al. 2011, Wang, Hou et al. 2018). Baculoviruses confer high degree of host specificity and insecticidal activity, thus various NPVs are being studied and developed as environmental-friendly biological pesticides that can be effectively used for pest management in agriculture and forestry (Szewczyk, Rabalski et al. 2009). Baculoviruses have also been used extensively in cell-expression system in the production of recombinant proteins (Kost, Condreay et al. 2005, Hitchman, Possee et al. 2009).

Evolution has imprinted its effect on nucleic acid sequences through various degrees of sequence homology, gene variants, types of noncoding sequences and codon usage pattern. Genes within the same genome may have their own evolutionary histories as they may originate from different ancestors or have been subjected to different environmental pressure. Frequencies of codon usage that code for the same amino acid have shown to be varied greatly between organisms, and between proteins within the same organism (Akashi 2001). Mutation (or synonymous mutation) and Natural selection have been suggested to be the two main forces that shape the pattern of codon usage bias within and between species (Duret 2002, Chamary, Parmley et al. 2006, Hershberg and Petrov 2008). The Mutation model states that codon usage bias arises from a bias in nucleotide composition, which in turn arises from a bias in the point mutation rate, or a bias in repair mechanism. For example, point mutations that favour the change from A to G and T to C may give rise to a GC-rich regions. It is deemed as “neutral” because these changes do not affect the amino acid sequence and thus, has no fitness advantage. In contrast, the Natural selection model suggests that synonymous

mutations would influence the fitness of an organism, such as accuracy and efficiency of translation, and therefore be promoted or repressed during evolution. The evolutionary driving force of codon usage bias has been studied in many viruses. Shackelton et. al. 2006 suggested that mutational pressure rather than natural selection is the main determinant of codon usage in vertebrate-infecting DNA viruses (Shackelton, Parrish et al. 2006). Jenkins & Holmes also suggested that mutation pressure is the most important determinant of the codon bias in human RNA viruses, but also proposed that translational selection may have some influence in shaping codon usage bias (Jenkins and Holmes 2003). Chen 2013 showed that 27% and 21% of total variation in the codon usage pattern could be attributed to mutational pressure, while 5% and 6% of total variation could be explained by natural selection for both DNA and RNA viruses, respectively (Chen 2013). Su et. al. 2009 demonstrated a positive correlation in codon usage preferences among RNA viruses that target the same host category, such as viruses infecting vertebrate hosts have different codon usage preferences to those of invertebrate viruses (Su, Lin et al. 2009). Codon usage has also been studied in nucleopolyhedroviruses through the sequence analyses of 6 genes, and the analyses showed that the patterns of codon usage were a direct function of the G+C content of the virus-encoded genes (Levin and Whittome 2000).

In this study, we would like to further explore the codon usage pattern of the nucleopolyhedrovirus (NPVs) genomes and evolutionary pressure that act on it using 6 Alphabaculoviruses as representatives of the NPVs. All Open-Reading-Frames (ORFs) in the Alphabaculovirus genomes were analysed. Principle-Component Analysis (PCA) was employed to cluster the ORFs, based on their Relative Synonymous Codon usage (RSCU). Nucleotide composition and nucleotides at different codon positions were also analysed.

Methods and Materials

Genome Sequence and Analysis

Complete genome sequences of baculoviruses and nudivirus were originally obtained from GenBank database (http://www.ncbi.nlm.nih.gov/genomes/VIRUSES/viruses.html). Viruses consist of 6 Alphabaculovirus (2 Tortricidae, 3 Noctuidae, 1 Bombycidae) and 1 nudivirus. There are in total of 950 ORFs used for calculating RSCU (Table 1) (Ayres, Howard et al. 1994, Gomi, Majima et al. 1999, Hyink, Dellow et al. 2002, Nakai, Goto et al. 2003, Yang, Lee et al. 2014, Noune and Hauxwell 2016). The nucleotide analyses were also performed using CAIcal server (http://genomes.urv.cat/CAIcal/)(Puigbo, Bravo et al. 2008). The results of the nucleotide composition analysis is in the Supplementary data 1.

Measures of Relative Synonymous Codon Usage (RSCU)

The relative synonymous codon usage (RSCU) score represents the frequency for which the codon is used relative to other synonymous codons, thus providing a metric for determining whether a mutation replaces a more common codon with a rarer codon or vice versa (Sharp and Li 1986). We use CAIcal server to calculate the RSCU (http://genomes.urv.es/CAIcal/). The relative synonymous codon usage (RSCU) is significant to the analysis of codon bias in terms of frequency. An important advantage of this index is its independence from amino acid composition bias. The RSCU value of each codon was calculated as follows:

$$RSCU = \frac{g_{ij}}{\frac{ni}{\sum_{j} g_{ij}}} ni$$

where the value is the observed number of the gth codon for the jth amino acid which has kinds of synonymous codons. Codons with higher (or lower) selected frequencies have higher (or lower) RSCU values. Hence, a frequent codon will have an RSCU > 1 and codons with RSCU < 1 are qualified as rare, which are the characteristics of a bias codon preference. The RSCU data of 950 genes is in the Supplementary data 2.

Principal Component Analysis (PCA)

Principal component analysis (PCA) was carried out using BioVinci® program. The greatest variance represented by any projection of the data lies on the first coordinate, so called the first principal component (PC), the second greatest variance lies on the second PC, and so on. To minimize the effect of amino acid composition on codon usage, each coding sequence was represented as a 59 dimensional vector, and each dimension corresponds to the RSCU value of each sense codon, which only includes synonymous codons for a particular amino acid excluding the codons AUG, UGG, and the three stop codons.

Table 1. Genome sizes and ORFs of the 6 baculoviruses and 1 nudivirus

Virus	Genome Size(bp)	No. ORFs	Family	Reference	Accession no.
Adoxophyes Honmai NPV (AdhoNPV)	113,220	125	Tortricidae	Nakai (2003)	NC_004690
Epiphyas Postvittana NPV (EppoNPV)	118,584	136	Tortricidae	Hyink et al. (2002)	AY043265
Agrotis Ipsilon mNPV (AgipNVP)	155,122	163	Noctuidae	Harrison (unpublished)	NC_011345
Autographa Californica MNPV (AcMNPV)	133,894	155	Noctuidae	Ayers et al. (1994)	NC_001623
Helicoverpa armigera NPV (HearNPV)	136,740	113	Noctuidae	Noune and Hauxwell (2016)	KJ909666
Bombyx mori NPV (BmNPV)	128,413	143	Bombycidae	Gomi et al. (1999)	NC_001962
Penaeus monodon nudivirus (PmNV)	119,638	115	-	Yang et al. (2014)	NC_024692

Results

Nucleotide composition analysis

The overall nucleotide composition and the frequency of the nucleotides at the synonymous third codon position of 6 Alphabaculovirus genomes and 1 Nudivirus were analysed. Penaeus monodon nudivirus *(*PmNV) is used as a control for virus of a different family *Nudiviridae*. PmNV also produces an occlusion body, similar to baculovirus, and is the causative agent of spherical baculovirosis in shrimp (*Penaeus monodon*) (Yang, Lee et al. 2014).

The mean values of the nucleotide composition are presented in Table 2. In all species, except Agrotis Ipsilon mNPV (AgipNPV), the A+T content ranges from 58.5%-65.2%, in which the genome of PmNV contains the highest percentage of A+T content, compared to the other Alphabaculoviruses. The genome of AgipNPV shows an approximately equal percentage of the A+T and G+C contents at 50.93% and 49.07%, respectively.

The mean values of the nucleotide composition at the third codon position was also investigated. The results revealed that all viruses, except AgipNPV, prefer A or T at the third codon position (Table 3). The A3+T3 ranges from 52.46%-65.89% with PmNV contains the highest percentage of A3+T3 content. The A3+T3 and G3+C3 contents of AgipNPV are 37% and 63%, respectively, indicating that the AgipNPV prefers G and C at the third codon position.

Principal Component Analysis of Viral RSCU

The RSCU of 950 viral ORFs, belonging to 6 Alphabaculovirus and 1 nudivirus were calculated and subjected to the principal component analysis (PCA). The two-dimensional PCA of all the 7 viruses are shown in Figure 1. Each coloured dots represent individual ORFs. The statistical result confirms the validity of the test, in which principal component 1 (PC1) explains 65% of the total data variance, followed by principal component 2 (PC2) which explains 5% of the total variance (Table 4). Both PCs explain in total 70% of the data variance.

There are 3 distinct clusters in the PCA plot (Figure 1); cluster 1 is the ORFs from genomes of AgipNPV (yellow dots), locating in the upper-left area of the plot, cluster 2 is the ORFs of PmNV (red dots), locating in the lower-left area of the plot, and cluster 3 is the ORFs of the rest of the Alphabaculoviruses that located between cluster 1 and 2. There are also some ORFs that do not cluster, but disperse around the main clusters. Some can be identified as outliers because they positioned in the far-right area of the plot. We analysed the ORFs further by plotting the ORFs of baculoviruses that infect the same family of insects, i.e. AgipNPV, AcMNPV and HearNPV infect insects of the family *Noctuidae* (Figure 2) and AdhoNPV and EppoNPV infect insects of the family *Tortricidae* (Figure 3). The plot of AcMNPV and HearNPV ORFs reveals a tight clustering pattern, by which many overlap one another, while the plot of AdhoNPV and EppoNPV ORFs form a loose cluster. The PCA plot of BmNPV and AcMNPV that infect different families of insects was also processed (Figure 4). Interestingly, majority of the ORFs overlap one another forming a tight cluster in a single location, despite the fact that the two infect different families of insects. The majority of the core genes, such as Helicase and DNA polymerase are found within the main cluster, where majority of the ORFs are present. In all the plots, the outliers on the far right of the plot were identified as either hypothetical or P6.9 genes. The one-dimensional PCA of all the 7 viruses were plotted to further emphasise the outliers. The results are consistent with the two-dimensional PCA, in which the outliers are either hypothetical or P6.9 genes in all plots (Figure 5). Interestingly, the outlier of the nudivirus has also been identified as P6.9 gene.

Table 2. The mean values of the nucleotide composition.

Virus	%A	%C	%T	%G
Adoxophyes Honmai NPV (AdhoNPV)	34.9	17.65	28.96	18.44
Epiphyas Postvittana NPV (EppoNPV)	31.19	20.45	27.3	21.05
Agrotis Ipsilon mNPV (AgipNVP)	27.76	25.11	23.17	23.97
Autographa Californica MNPV (AcMNPV)	32.18	20.3	26.83	20.68
Helicoverpa armigera NPV (HearNPV)	32.94	19.78	27.84	19.45
Bombyx mori NPV (BmNPV)	32.54	20.27	26.72	20.48
Penaeus monodon nudivirus (PmNV)	34.84	16.99	30.33	17.84

Table 3. The mean values of the nucleotide composition at third codon position.

Virus	%G3+C3	%A3+T3
Adoxophyes Honmai NPV (AdhoNPV)	38.66	61.33
Epiphyas Postvittana NPV (EppoNPV)	44.83	55.16
Agrotis Ipsilon mNPV (AgipNVP)	62.99	37.00
Autographa Californica MNPV (AcMNPV)	47.54	52.46
Helicoverpa armigera NPV (HearNPV)	42.47	57.53
Bombyx mori NPV (BmNPV)	47.47	52.52
Penaeus monodon nudivirus (PmNV)	34.10	65.89

Table 4. Statistical Test of PCA.

Component	Standard Deviation	Proportion of Variance
PC1	6.18684299949783	0.64876
PC2	1.7418816087488	0.05143

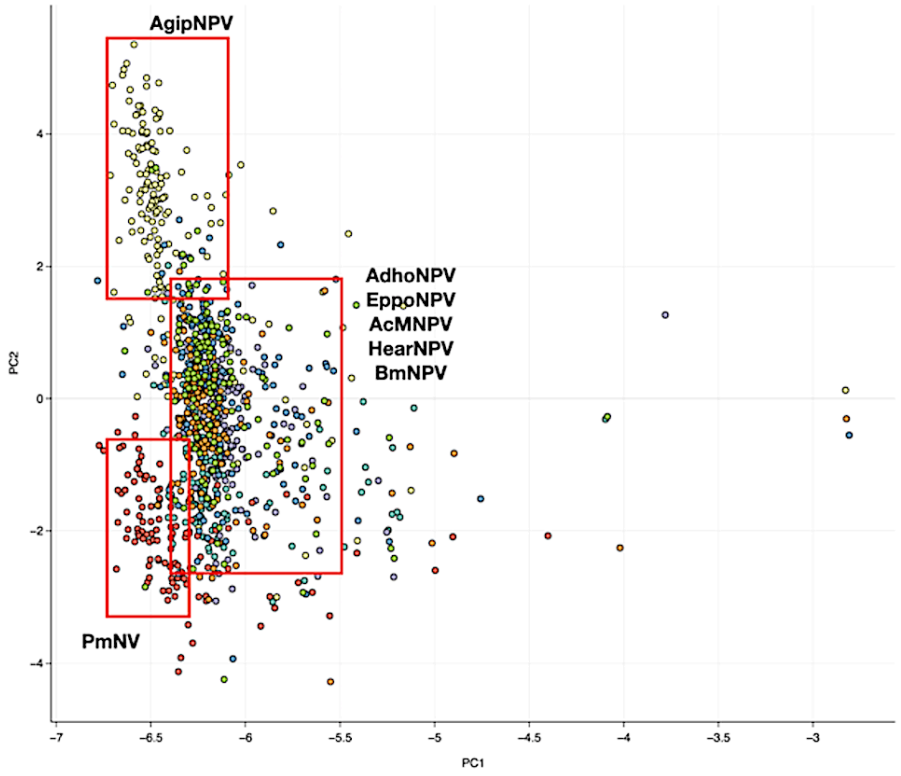

Fig. 1. Two-dimensional plot of the ORFs from six baculoviruses; AdhoNPV (mint green), AgipNPV (yellow), EppoNPV (grey), AcMNPV (blue), HearNPV (orange), BmNPV (light green), and one nudivirus, PmNV (red).

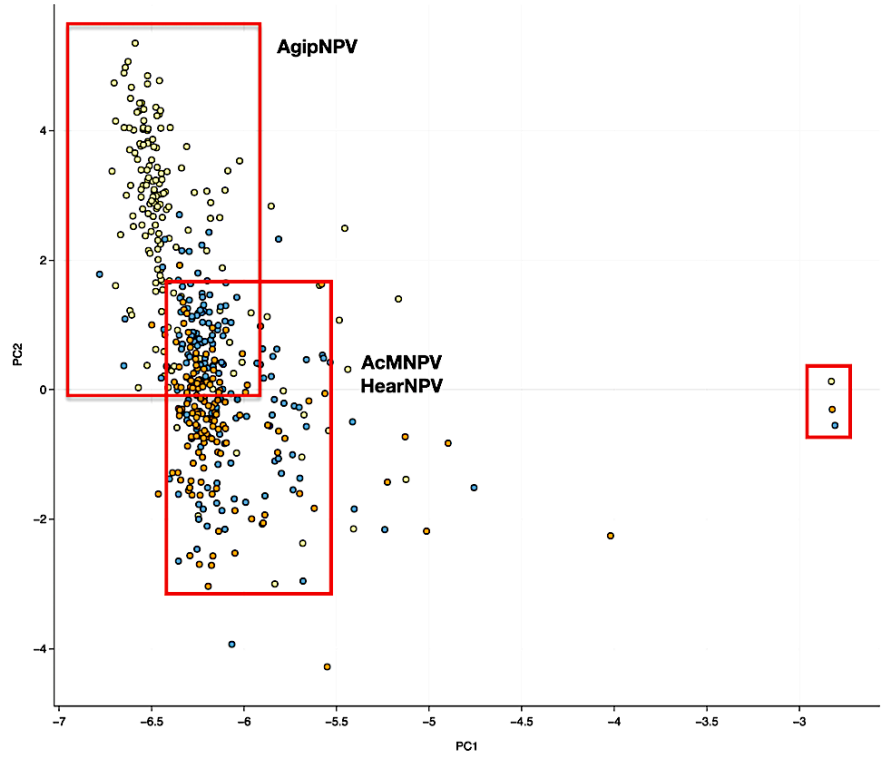

Fig. 2. Two-dimensional plot of the ORFs from AgipNPV (yellow), AcMNPV (blue), HearNPV (orange). The three baculoviruses infect insects of the family *Noctuidae*. Red square identifies a distinct clustering of AgipNPV ORFs, while ORFs of AcMNPV and HearNPV form a tight cluster with many overlapping positions. Small red square indicates P6.9 gene.

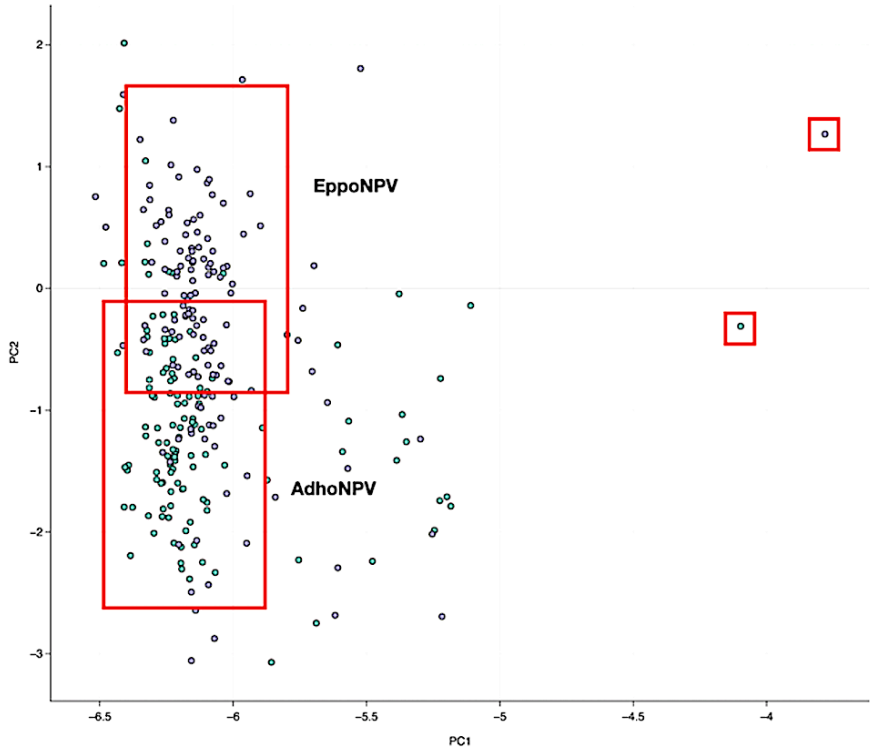

Fig. 3. Two-dimensional plot of the ORFs from AdhoNPV (mint green) and EppoNPV (grey). The two baculoviruses infection insects in the family *Tortricidae*. Loose-clustering pattern is shown. Small red squares indicate P6.9 gene.

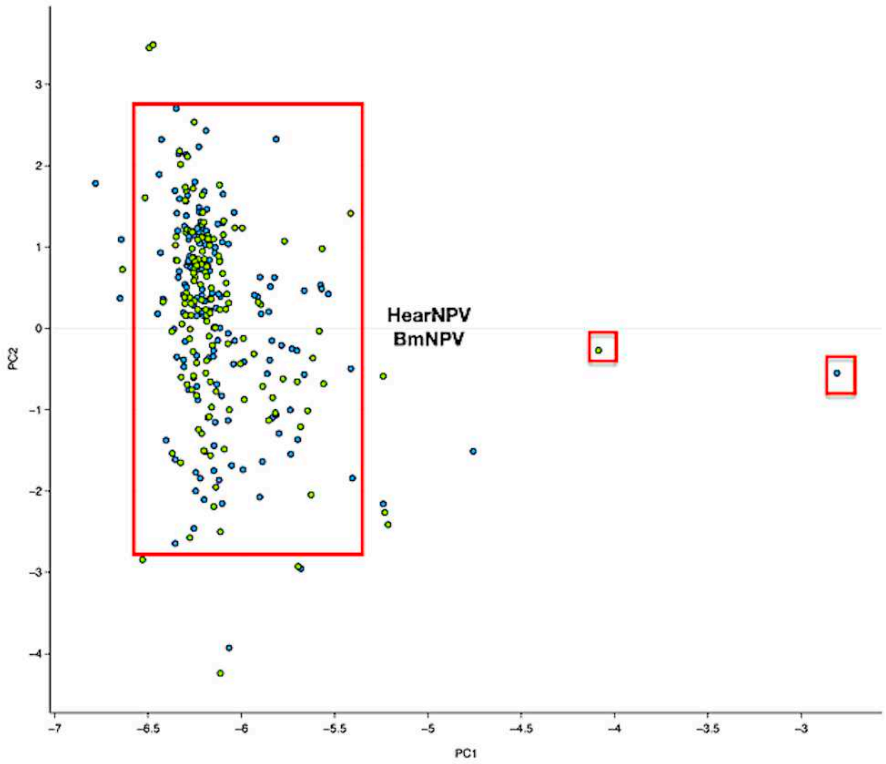

Fig. 4. Two-dimensional plot of the ORFs from BmNPV (green) and AcMNPV (blue). BmNPV infect insects of the family *Bombicidae,* while AcMNPV infect insects of the family *Noctuidae*. Tight-clustering pattern has shown. Small red squares indicate P6.9 gene.

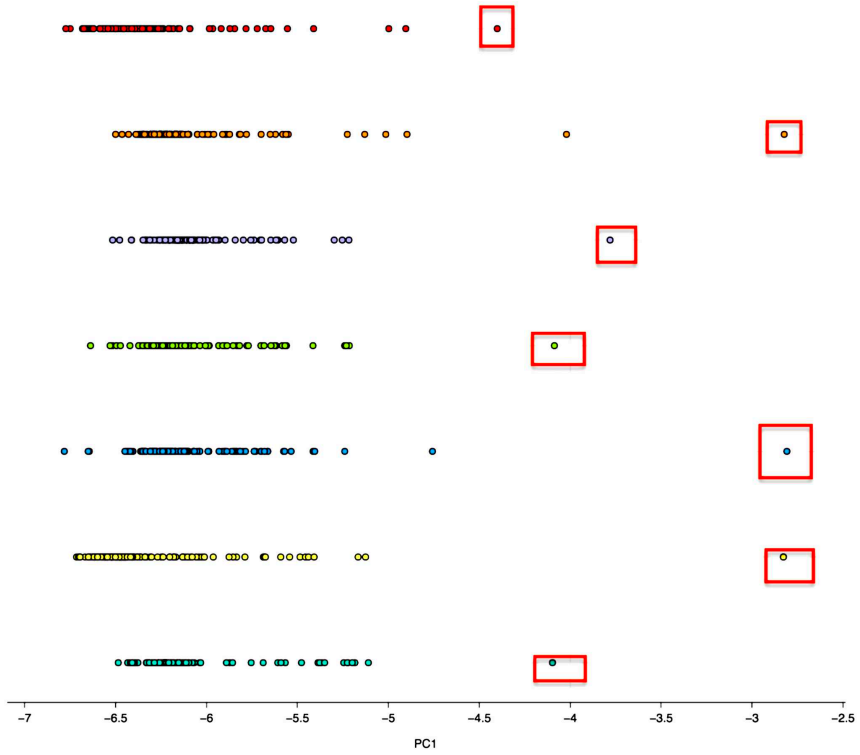

Fig. 5. One-dimensional Plot of ORFs from all seven viruses. AdhoNPV (mint green), AgipNPV (yellow), EppoNPV (grey), AcMNPV (blue), HearNPV (orange), BmNPV (light green), and one nudivirus, *PmNV* (red). P6.9 genes (outlier) are in red squares.

Codon usage of P6.9`

It is consistent in all PCA plots that one of the outliers has been identified as P6.9 gene, which is one of the core genes that is present in all baculoviruses. Therefore, we would like to analyse the codon usage of P6.9 gene further.

The Alphabaculovirus P6.9 genes use 17 different amino acids and 40 codons. The amino acids used are Phenylalanine, Leucine, Valine, Serine, Proline, Threonine, Alanine, Tyrosine, Histidine, Glutamine, Asparagine, Lysine, Aspartic acid, Glutamic acid, Arginine and Glycine (Supplementary data 2). All P6.9 genes, except that of AgipNPV, uses amino acids and codons ranging from 7-11 different amino acids and 17-25 codons, respectively (Table 6). The AgipNPV P6.9 gene uses a more diverse set of amino acids and codons, 17 amino acids and 35 codons respectively. We have categorised the most preferred codon as RSCU ≥ 2 and least preferred codon as RSCU < 1. All baculoviruses have 5-11 preferred codons, in which some of the codons are used exclusively to code for specific amino acids. For example, TTA is used exclusively for a Leucine in the AdhoNPV P6.9 gene (RSCU = 6), GTC for Valine in EppoNVP (RSCU = 4) and GCC for Alanine in BmNPV (RSCU = 4). The degree of codon usage bias appears to be higher in AdhoNPV, EppoNPV, AcMNPV, HearNPV and BmNPV compared to AgipNPV as a higher proportion of codons has RSCU ≥ 2.

Amino acid sequence alignment of the six P6.9 genes shows the evidence of either deletions or insertions, which indicates by the alignment gaps (Figure 6). All sequences are Arginine-riched, in which this amino acid contributes to approximately 35-44% of the sequences. The second highest is Serine, which is present between 12-23%. Sequences of HearNPV and AgipNPV also have a high percentage of Glycine, 32% and 20%, respectively.

The G+C content of P6.9 gene

The baculovirus P6.9 gene has been annotated as a protamine-like gene, and the encoded protein plays an important role in condensing the viral genome into the nucleocapsid. Protamine-like genes have also been identified in insects, thus we explore sequence relationship, focusing on the G+C content, between the baculovirus P6.9 and host insect protamine-like genes (Table 7). *Bombus bifarius* belongs to the Order Hymenoptera, *Drosophila melanogaster* belongs to the Order Diptera, and the rest of the insect species belongs to the Order Lepidoptera. Baculovirus Helicase gene is also used as a representative of the baculovirus core genes located within the main clusters in the 2-dimensional PCA plots.

The overall %G+C of the P6.9 gene is consistently high across the 6 baculoviruses, ranging from 56-67%, while that of the Helicase is lower, ranging between 50-34% (Table 7). The Protamine-like genes exhibits a more diverse %G+C, ranging between 48-68%. The %G+C at the three different codon positions in the P6.9 gene shows an interesting pattern, by which the %G2+C2 establishes an outstanding high value between 80-94%, compared to the other codon positions (Table 7). This pattern is not observed in the rest of the genes analysed. The %G2+C2 of the Helicase gene establishes the lowest value, ranging between 25-30%, when comparing to the other 2 codon positions. The %G2+C2 and %G3+C3 are comparable in the insect protamine-like genes, with an exception in *Papilio machaon* protamine-like gene that shows 92% G3+C3.

Table 5. Summary of a number of amino acid and codon usage in the 6 baculovirus P6.9 genes.

Baculoviruses	**No. of Amino acids and codons**	**No. of codons with RSCU ≥ 2 (most preferred)**	**No. of codons with 2 > RSCU ≥ 1**	**No. of codons with RSCU < 1**
AdhoNPV	11, 25	8	10	7
EppoNPV	11, 21	11	8	2
AgipNPV	**17, 35**	7	18	10
AcMNPV	7, 17	6	6	5
HearNPV	8, 19	5	8	6
BmNPV	10, 20	9	6	5

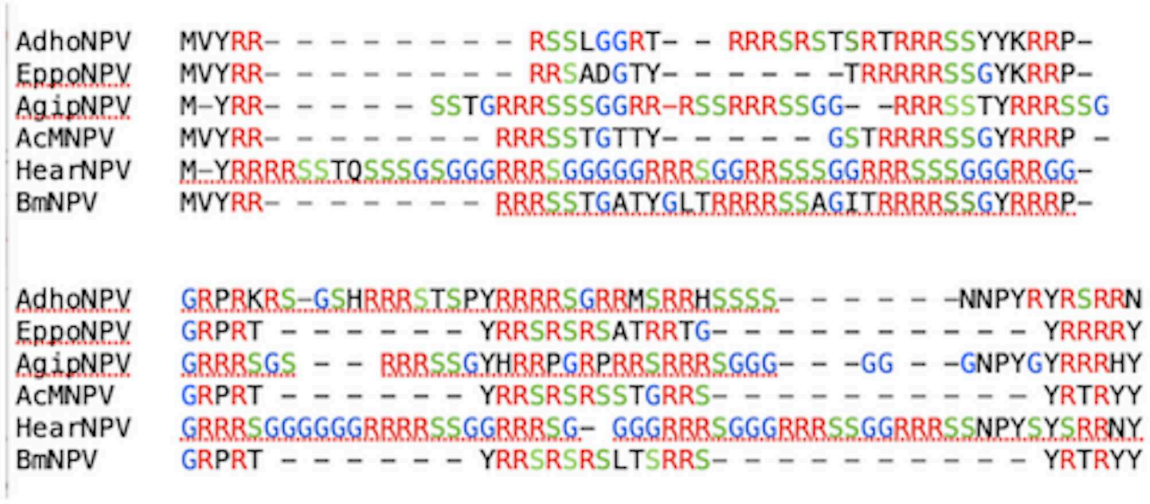

Fig. 6. Sequence alignment of P6.9 genes from AdhoNPV, EppoNPV, AgipNPV, AcMNPV, HearNPV and BmNPV. Arginine (Red), Serine (Green) and Glucine (blue).

Table 6. Overall % G+C content and % G+C content at the three codon positions.

Organisms	Genes	Overall % G+C	%G1+C1	%G2+C2	%G3+C3
AdhoNPV	P6.9	57	34	**80**	60
EppoNPV	P6.9	63	55	**82**	57
AgipNPV	P6.9	67	51	**89**	63
AcMNPV	P6.9	56	41	**85**	46
HearNPV	P6.9	64	57	**94**	44
BmNPV	P6.9	56	44	**83**	44
AdhoNPV	Helicase	**34**	36	25	42
EppoNPV	Helicase	**39**	41	30	47
AgipNPV	Helicase	**50**	45	30	75
AcMNPV	Helicase	**41**	39	28	55
HearNPV	Helicase	**37**	40	30	42
BmNPV	Helicase	**39**	38	27	52
Bombus bifarius (bumble bee)	Protamine-like	48	44	52	49
Pieris rapae (white and yellow butterfly)	Protamine-like	56	55	62	51
Papilio machaon (Swallow tail butterfly)	Protamine-like	68	56	56	92
Amyelois transitella (monotypic snout moth)	Protamine-like	68	63	66	74
Helicoverpa armigera (Cotton bollworm)	Protamine 2-like	55	54	53	59

Discussion

The overall nucleotide composition of the 5 baculoviruses (AdhoNPV, EppoNPV, AcMNPV, HearNPV and BmNPV) genomes suggests that the Alphabaculoviruses may prefer AT-rich genomes. This observation is consistent with the analysis of the Third codon-position that also prefers A or T. The genome of PmNV shows the highest percentage of A+T content and at the Third codon-position, compared to the 5 baculoviruses. Since the percentage of A T C G nucleotide composition correlates with the percentage of A T G C at the Third codon position and mutations at this codon position is subjected to the codon redundancy and wobble pairing, any changes at this position do not affect the amino acid coded, thus it is a reflection of mutational bias in the genome. This is suggests that the codon usage in these five alpha-baculoviruses are predominantly under an influence of a neutral mutation that biases toward A/T. The mutational bias towards A/T is perhaps due to the high rate of G/C to A/T transitions. However, the analysis of AgipNPV genome appears to be different to the other Alphabaculovirus genomes. The nucleotide composition reveals an equal usage of A/T and G/C, but the third codon-position analysis showed that this virus prefers G/C at this position. This suggests that the A/T content is mostly found at either the First or Second codon-position. Since the preferred nucleotides at the Third codon-position does not correlate with the overall nucleotide composition, and changes at the

First and Second codon-positions affect the coded amino acid. This may reflect an influence of natural selection on the AgipNPV genome and the usage of codons. Natural selection acts on the nucleotide content of genome when the percentage of A/T or G/C affects its fitness and survival. For example, Auewarakul 2004 showed that the G/C content directly affects the viral codon adaptation index and codon usage preference, which plays a key role in predicting the efficiency of viral gene expression in the host cells (Auewarakul 2005). The G/C content also plays an important role in the adaptation to the host environment as shown in the study by Brown (2007) that Herpes Simplex Virus-1 (HSV-1) uses its high G/C content to protect itself from the insertion of an AT-rich retrotransposon (L1) abundantly found in the brain (Brown 2007).

Principle Component Analysis (PCA) has shown that the ORFs from the genomes of AdhoNPV, EppoNPV, AcMNPV, HearNPV and BmNPV are clustering at a similar location, reflecting similarities in the Relative Synonymous Codon Usage (RSCU) patterns and thus, the same evolutionary force that drive the usage of codon. However, the codon usage pattern is not a reflection of insect host specificity as shown that the ORFs of AcMNPV and BmNPV that infect 2 different hosts show a tight clustering pattern, while the ORFs of AcMNPV and AgipNPV that infect the same host show two distinct clusterings. The clustering of the ORFs and the RSCU patterns are likely determined by the overall nucleotide composition and perhaps nucleotides at the different codon position mentioned above. We further looked at the outliers that appear in the PCA plots from all the genomes tested, by which they have been identified as the Protamine-like genes (P6.9). It is interesting that the P6.9 gene is also an outlier in the Nudivirus, which is a different family of viruses. This distinct characteristic of P6.9 is perhaps a reflection of its distinct function, especially in the occlusion-forming viruses.

The sequences of P6.9 gene from the 6 baculoviruses were analysed further. The analysis of the six P6.9 gene sequences shows no similarity in the codon usage preference. The P6.9 genes of different Alphabaculoviruses have their own preferred codons, with AgipNPV uses the most diverse sets of codons. Some codons are exclusively used to code for specific amino acids, i.e. codon with RSCU = 6 or RSCU = 4. This indicates a strong positive selection on those codons, perhaps relating to the abundance of tRNA and thus, the translation efficiency. The G+C content of the P6.9 genes shows a distinct pattern compared to that of the genes in the main cluster, represented by the Helicase genes. The P6.9 genes have a high G+C content, similar to insect protamine-like genes, while the Helicase genes have a low G+C content. The G+C content of the P6.9 genes is significantly high at the second codon position, which coincides with the high percentage of arginine, serine and glycine in the sequences. These amino acids have G or C at their second codon position. Thus, the high percentage of these 3 amino acids is likely to contribute to the high percentage of G+C content, especially at the second codon position, observed in the sequences. A high content of arginine and its positively-charged property, which is also known to be a characteristic of a protamine gene, has been selected for its ability to compact DNA to a very high density (Brewer, Corzett et al. 1999, DeRouchey, Hoover et al. 2013). Tight DNA packing has also been proposed to prevent DNA damage from radical as well as to inactivate the gene. Therefore, the high percentage of G+C at the second codon position in P6.9 is likely to reflect its function in DNA packaging. This is an evidence of natural selection that acts on both the codon usage and the nucleotide composition of a gene.

In conclusion, we have shown that different genes within the same genome may subject to different types of evolutionary pressure. The evidence is shown in the overall nucleotide composition and G+C content at different codon positions. In addition, homologous P6.9 genes in different baculoviruses may have different codon usage pattern, but their overall nucleotide composition may be similar because they perform the same functions and subject to the same evolutionary pressure.

Declarations

Ethics approval and consent to participate: Not applicable

Consent for publication: Not applicable

Competing interests: The authors declare that they have no competing interests

Funding: No funding was received for conducting this study.

Authors' contributions: P. Mahapattanakul acquired the sequence data, performed the RSCU, PCA and nucleotide composition analyses, Interpret the results and draft the manuscript. P. Rajbhandari acquired the sequence data, performed the P6.9 sequence analyses. P. Rodpothong designed the work, performed the P6.9 sequence analyses, Interpret the results and finalise the manuscript.

Acknowledgements: Thank you Prof. Prasert Auewarakul for valuable discussion on the codon usage pattern of the virus.

References

Akashi, H. (2001). "Gene expression and molecular evolution." Curr Opin Genet Dev **11**(6): 660-666.

Auewarakul, P. (2005). "Composition bias and genome polarity of RNA viruses." Virus Res **109**(1): 33-37.

Ayres, M. D., S. C. Howard, J. Kuzio, M. Lopez-Ferber and R. D. Possee (1994). "The complete DNA sequence of Autographa californica nuclear polyhedrosis virus." Virology **202**(2): 586-605.

Brewer, L. R., M. Corzett and R. Balhorn (1999). "Protamine-induced condensation and decondensation of the same DNA molecule." Science **286**(5437): 120-123.

Brown, J. C. (2007). "High G+C Content of Herpes Simplex Virus DNA: Proposed Role in Protection Against Retrotransposon Insertion." Open Biochemistry Journal **1**: 33-42.

Chamary, J. V., J. L. Parmley and L. D. Hurst (2006). "Hearing silence: non-neutral evolution at synonymous sites in mammals." Nat Rev Genet **7**(2): 98-108.

Chen, Y. (2013). "A comparison of synonymous codon usage bias patterns in DNA and RNA virus genomes: quantifying the relative importance of mutational pressure and natural selection." Biomed Res Int **2013**: 406342.

Clem, R. J. and A. L. Passarelli (2013). "Baculoviruses: sophisticated pathogens of insects." PLoS Pathog **9**(11): e1003729.

DeRouchey, J., B. Hoover and D. C. Rau (2013). "A comparison of DNA compaction by arginine and lysine peptides: a physical basis for arginine rich protamines." Biochemistry **52**(17): 3000-3009.

Duret, L. (2002). "Evolution of synonymous codon usage in metazoans." Curr Opin Genet Dev **12**(6): 640-649.

Gomi, S., K. Majima and S. Maeda (1999). "Sequence analysis of the genome of Bombyx mori nucleopolyhedrovirus." J Gen Virol **80 (Pt 5)**: 1323-1337.

Herniou, E. A., B. M. Arif and J. J. Becnel (2012). Family Baculoviridae. Virus Taxonomy, Ninth Report of the International Committee on Taxonomy of Viruses. A. M. Q. King, M. J. Adams, E. B. Carstens and E. J. Lefkowitz. Amsterdam, Elsevier Academic Press**:** 163-173.

Herniou, E. A. and J. A. Jehle (2007). "Baculovirus phylogeny and evolution." Curr Drug Targets **8**(10): 1043-1050.

Herniou, E. A., J. A. Olszewski, J. S. Cory and D. R. O'Reilly (2003). "The genome sequence and evolution of baculoviruses." Annu Rev Entomol **48**: 211-234.

Hershberg, R. and D. A. Petrov (2008). "Selection on codon bias." Annu Rev Genet **42**: 287-299.

Hitchman, R. B., R. D. Possee and L. A. King (2009). "Baculovirus expression systems for recombinant protein production in insect cells." Recent Pat Biotechnol **3**(1): 46-54.

Hyink, O., R. A. Dellow, M. J. Olsen, K. M. B. Caradoc-Davies, K. Drake, E. A. Herniou, J. S. Cory, D. R. O'Reilly and V. K. Ward (2002). "Whole genome analysis of the Epiphyas postvittana nucleopolyhedrovirus." J Gen Virol **83**(Pt 4): 957-971.

Jenkins, G. M. and E. C. Holmes (2003). "The extent of codon usage bias in human RNA viruses and its evolutionary origin." Virus Res **92**(1): 1-7.

Kost, T. A., J. P. Condreay and D. L. Jarvis (2005). "Baculovirus as versatile vectors for protein expression in insect and mammalian cells." Nature Biotechnology **23**(5): 567-575.

Levin, D. B. and B. Whittome (2000). "Codon usage in nucleopolyhedroviruses." J Gen Virol **81**(Pt 9): 2313-2325.

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

Nakai, M., C. Goto, W. Kang, M. Shikata, T. Luque and Y. Kunimi (2003). "Genome sequence and organization of a nucleopolyhedrovirus isolated from the smaller tea tortrix, Adoxophyes honmai." Virology **316**(1): 171-183.

Noune, C. and C. Hauxwell (2016). "Complete Genome Sequences of Seven Helicoverpa armigera SNPV-AC53-Derived Strains." Genome Announc **4**(3).

Puigbo, P., I. G. Bravo and S. Garcia-Vallve (2008). "CAIcal: a combined set of tools to assess codon usage adaptation." Biol Direct **3**: 38.

Rohrmann, G. F. (2019). Baculovirus Molecular Biology. th. Bethesda (MD).

Sajjan, D. B. and S. B. Hinchigeri (2016). "Structural Organization of Baculovirus Occlusion Bodies and Protective Role of Multilayered Polyhedron Envelope Protein." Food Environ Virol **8**(1): 86-100.

Shackelton, L. A., C. R. Parrish and E. C. Holmes (2006). "Evolutionary basis of codon usage and nucleotide composition bias in vertebrate DNA viruses." J Mol Evol **62**(5): 551-563.

Sharp, P. M. and W. H. Li (1986). "An evolutionary perspective on synonymous codon usage in unicellular organisms." J Mol Evol **24**(1-2): 28-38.

Su, M. W., H. M. Lin, H. S. Yuan and W. C. Chu (2009). "Categorizing host-dependent RNA viruses by principal component analysis of their codon usage preferences." J Comput Biol **16**(11): 1539-1547.

Szewczyk, B., L. Rabalski, E. Krol, W. Sihler and M. L. de Souza (2009). "Baculovirus biopesticides – a safe alternative to chemical protection of plants." Journal of Biopesticides **2**: 209-216.

van Oers, M. M. and J. M. Vlak (2007). "Baculovirus genomics." Curr Drug Targets **8**(10): 1051-1068.

Wang, J., D. Hou, Q. Wang, W. Kuang, L. Zhang, J. Li, S. Shen, F. Deng, H. Wang, Z. Hu and M. Wang (2018). "Genome analysis of a novel Group I alphabaculovirus obtained from Oxyplax ochracea." PLoS One **13**(2): e0192279.

Yang, Y. T., D. Y. Lee, Y. Wang, J. M. Hu, W. H. Li, J. H. Leu, G. D. Chang, H. M. Ke, S. T. Kang, S. S. Lin, G. H. Kou and C. F. Lo (2014). "The genome and occlusion bodies of marine Penaeus monodon nudivirus (PmNV, also known as MBV and PemoNPV) suggest that it should be assigned to a new nudivirus genus that is distinct from the terrestrial nudiviruses." BMC Genomics **15**: 628.

Figures

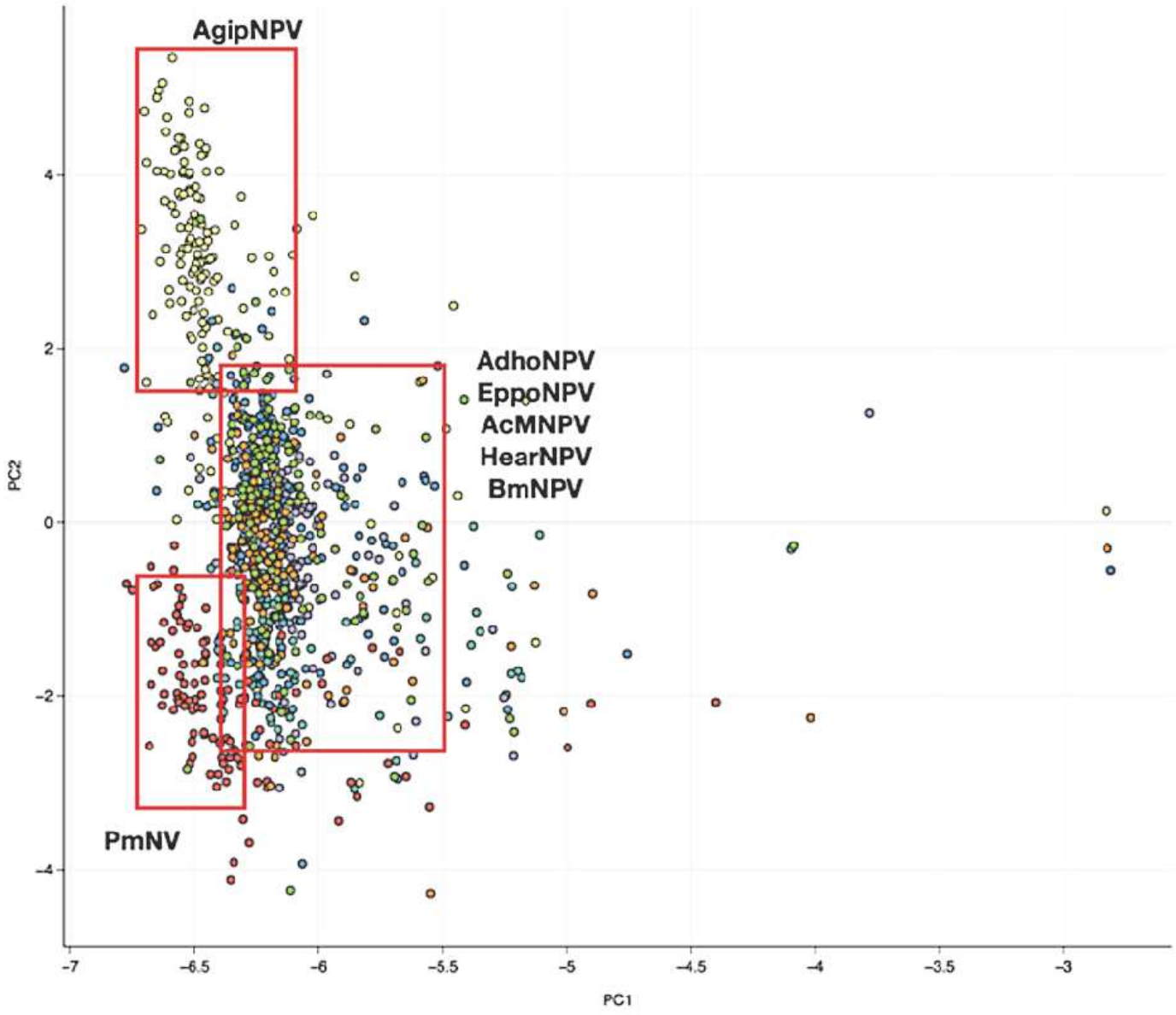

Figure 1

Two-dimensional plot o f t h e ORF s f r om s i x baculoviruses; AdhoNPV (mint gr e en) , AgipNPV (yellow), EppoNPV (grey), AcMNPV (blue), HearNPV (orange), BmNPV (light green), and one nudivirus, PmNV (red).

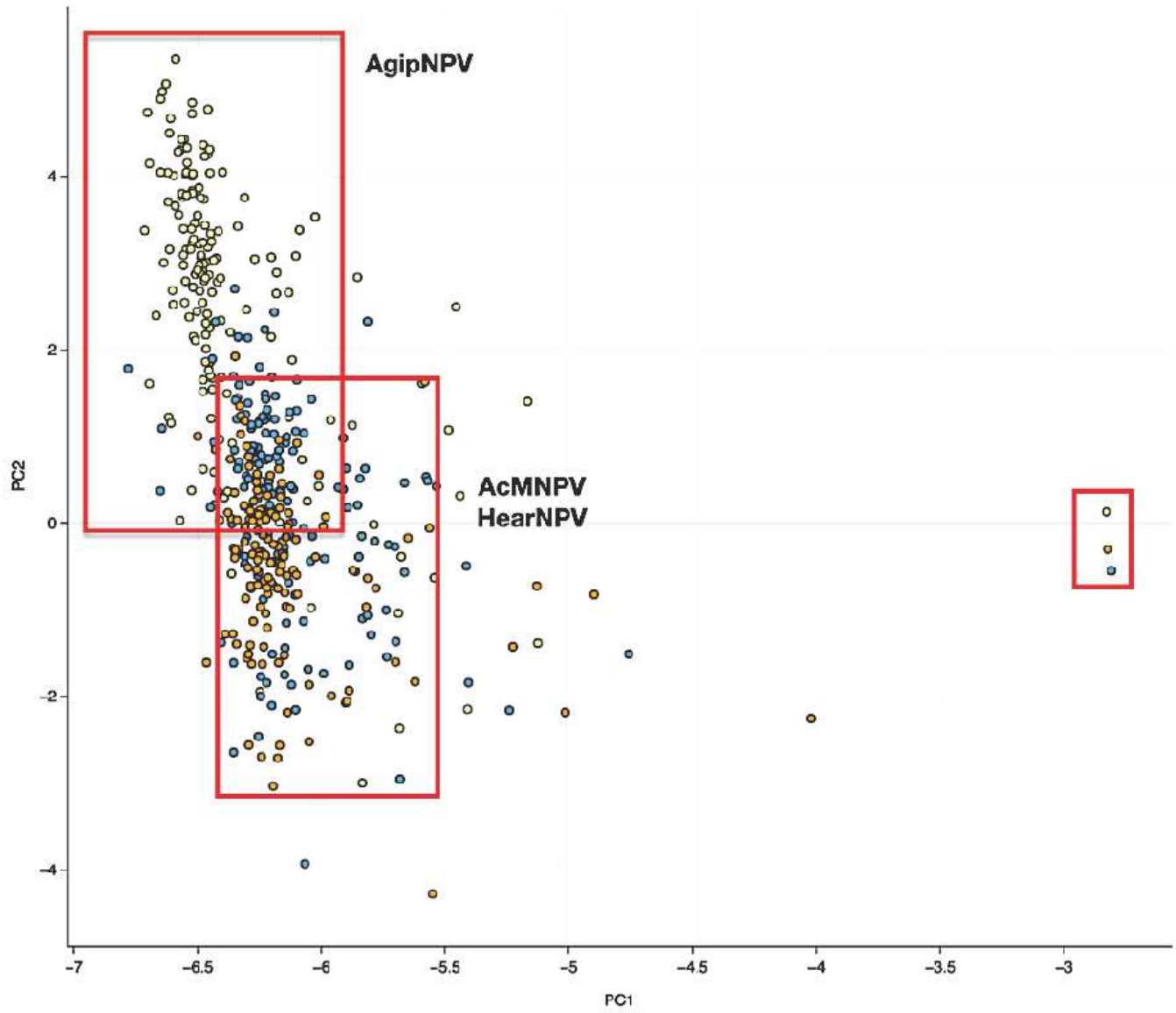

Figure 2

Two-dimensional plot of the ORFs from AgipNPV (yellow), AcMNPV (blue), HearNPV (orange). The three baculoviruses infect insects of the family Noctuidae. Red square identifies a distinct clustering of AgipNPV ORFs, while ORFs of AcMNPV and HearNPV form a tight cluster wi t h ma n y o v e r l a p p i n g positions. Small red square indicates P6.9 gene.

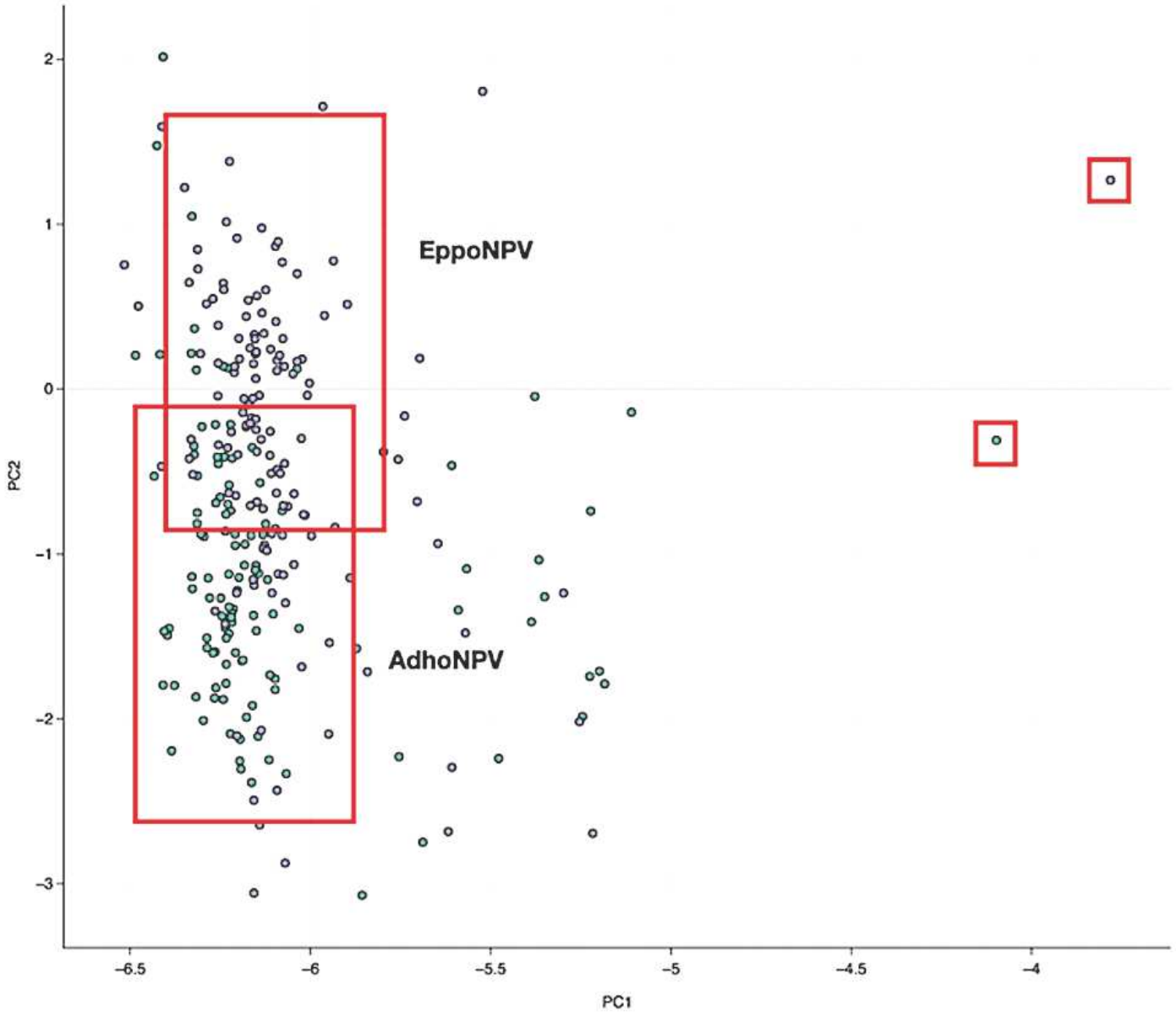

Figure 3

Two-dimensional plot of the ORFs from AdhoNPV (mint green) and EppoNPV (grey). The two baculoviruses infection insects in the family Tortricidae. Loose-clustering pattern is shown. Small red squares indicate P6.9 gene.

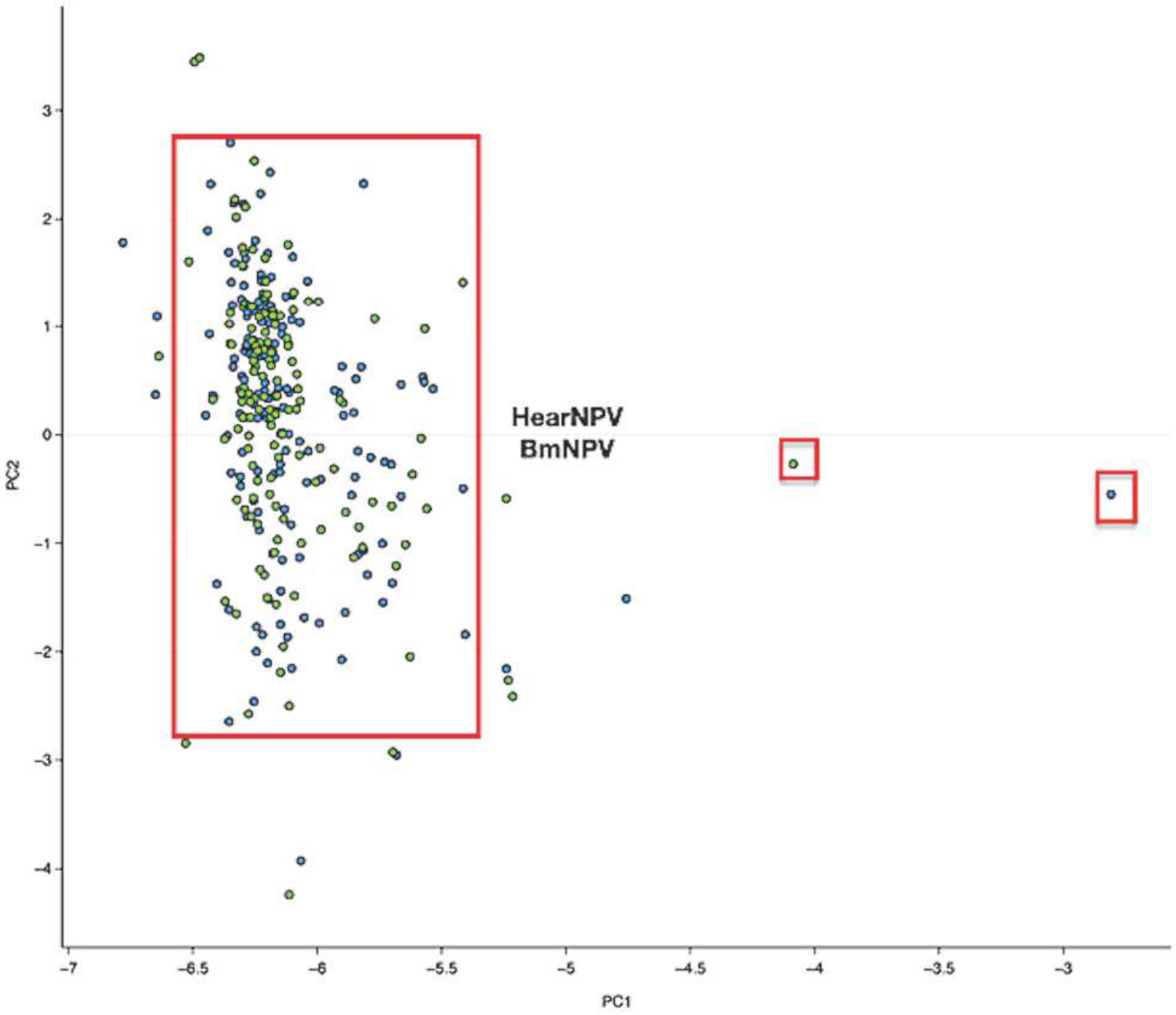

Figure 4

Two-dimensional plot of the ORFs from BmNPV (green) and AcMNPV (blue). BmNPV infect insects of the family Bombicidae, while AcMNPV infect insects of the family Noctuidae. Tight-clustering pattern has shown. Small red squares indicate P6.9 gene.

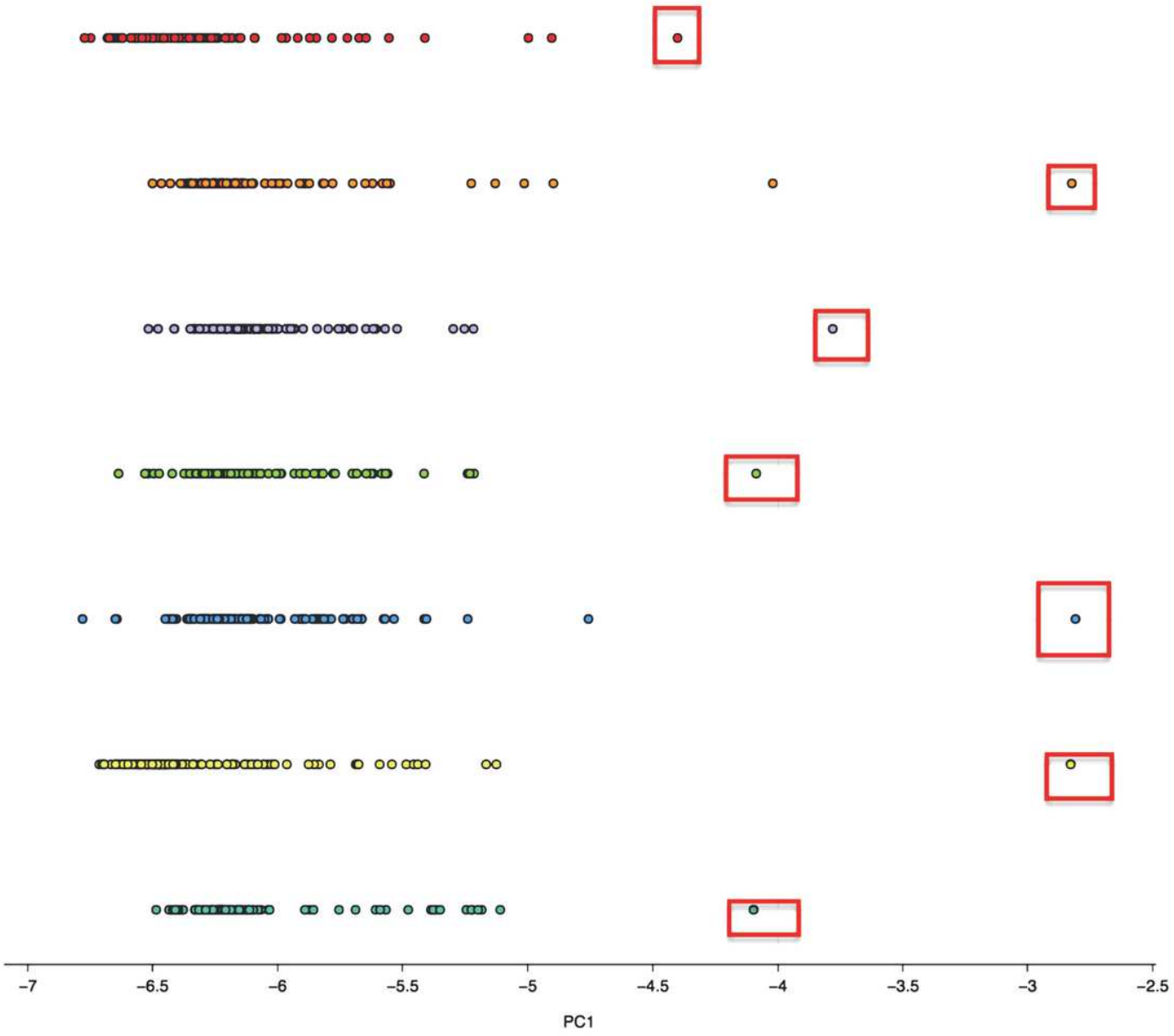

Figure 5

One-dimensional Plot of ORFs from all seven viruses. AdhoNPV (mint green), AgipNPV (y e l l ow) , Ep p oNPV (g r e y) , AcMNPV (b l u e) , He a rNPV (orange), BmNPV (light green), and one nudivirus, PmNV (red). P6.9 genes (outlier) are in red squares.

Figure 6

Sequence alignment of P6.9 genes from AdhoNPV, EppoNPV, AgipNPV, AcMNPV, HearNPV and BmNPV. Arginine (Red), Serine (Green) and Glucine (blue).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarydata1.xlsx
- Supplementarydata2.xlsx