Research

A potentially novel overlapping gene in the genomes of Israeli acute paralysis virus and its relatives

Niv Sabath*, Nicholas Price and Dan Graur

Address: Department of Biology and Biochemistry, University of Houston, Houston, TX 77204, USA

Email: Niv Sabath* - nsabath@uh.edu; Nicholas Price - price4890@gmail.com; Dan Graur - dgraur@uh.edu

* Corresponding author

Abstract

The Israeli acute paralysis virus (IAPV) is a honeybee-infecting virus that was found to be associated with colony collapse disorder. The IAPV genome contains two genes encoding a structural and a nonstructural polyprotein. We applied a recently developed method for the estimation of selection in overlapping genes to detect purifying selection and, hence, functionality. We provide evolutionary evidence for the existence of a functional overlapping gene, which is translated in the +1 reading frame of the structural polyprotein gene. Conserved orthologs of this putative gene, which we provisionally call *pog* (*p*redicted *o*verlapping *g*ene), were also found in the genomes of a monophyletic clade of dicistroviruses that includes IAPV, acute bee paralysis virus, Kashmir bee virus, and *Solenopsis invicta* (red imported fire ant) virus 1.

Background

Colony collapse disorder (CCD) is a syndrome characterized by the mass disappearance of honeybees from hives [1]. CCD imperils a global resource estimated at approximately $200 billion [2]. For example, it has been estimated that up to 35% of hives in the US may have been affected [3]. Many culprits have been suggested as causal factors of CCD, among them fungal, bacterial, and protozoan diseases, external and internal parasites, in-hive chemicals, agricultural insecticides, genetically modified crops, climatic factors, changed cultural practices, and the spread of cellular phones [1]. The Israeli acute paralysis virus (IAPV), a positive-strand RNA virus belonging to the family Dicistroviridae, was found to be strongly correlated with CCD [4]. It was first isolated in Israel [5], but was later found to have a worldwide distribution [4,6,7].

The genome of IAPV contains two long open reading frames (ORFs) separated by an intergenic region. The 5' ORF encodes a structural polyprotein; the 3' ORF encodes a non-structural polyprotein [5]. The non-structural polyprotein contains several signature sequences for helicase, protease, and RNA-dependent RNA polymerase [5]. The structural polyprotein, which is located downstream of the non-structural polyprotein, encodes two (and possibly more) capsid proteins.

Overlapping genes are easily missed by annotation programs [8], as evidenced by the fact that several overlapping genes were only detected by using the signatures of purifying selection [9-13]. Here, we apply a recently developed method for the detection of selection in overlapping reading frames [14] to the genome of IAPV and its relatives.

Results and Discussion

In the fourteen completely sequenced dicistroviral genomes (Table 1), we identified 43 same-strand overlap-

Table 1: A list of completely sequenced dicistroviruses used in this study

Name	Accession number
Israel acute paralysis virus (IAPV)	GenBank:NC_009025
Acute bee paralysis virus (ABPV)	GenBank:NC_002548
Kashmir bee virus (KBV)	GenBank:NC_004807
Solenopsis invicta virus (SINV-1)	GenBank:NC_006559
Black queen cell virus (BQCV)	GenBank:NC_003784
Cricket paralysis virus (CrPV)	GenBank:NC_003924
Homalodisca coagulata virus-1 (HoCV-1)	GenBank:NC_008029
Drosophila C virus (DCV)	GenBank:NC_001834
Aphid lethal paralysis virus (ALPV)	GenBank:NC_004365
Himetobi P virus (HiPV)	GenBank:NC_003782
Taura syndrome virus (TSV)	GenBank:NC_003005
Plautia stali intestine virus (PSIV)	GenBank:NC_003779
Triatoma virus (TrV)	GenBank:NC_003783
Rhopalosiphum padi virus (RhPV)	GenBank:NC_001874

ping ORFs of lengths equal or greater than 60 codons on the positive strand. Ten overlapping ORFs were found in concordant genomic locations in two or more genomes. The concordant overlapping ORFs were assigned to three orthologous clusters (Table 2). The overlapping ORFs in all three clusters are phase-1 overlaps, i.e., shifted by one nucleotide relative to the reading-frames of the known polyprotein genes. Two of the orthologous clusters overlap the gene encoding the nonstructural polyprotein and one overlaps the reading frame of the structural polyprotein. (In appendix 1, we present the results concerning the overlapping ORFs on the negative strand. We note, however, that dicistroviruses are not known to be ambisense [15].)

We identified a strong signature of purifying selection in cluster A that contains overlapping ORFs from four genomes: IAPV, Acute bee paralysis virus (ABPV), Kashmir bee virus (KBV), and *Solenopsis invicta* virus 1 (SINV-1)

Table 2: Clusters of orthologous overlapping ORFs on the positive strand

Cluster	Virus	Start of ORF	End of ORF	Length (nucleotides)
A	IAPV	6589	6900	312
	ABPV	6513	6815	303
	KBV	6601	6909	309
	SINV-1	4382	4798	417
B	ABPV	5958	6227	270
	KBV	5974	6243	270
C	CrPV	2396	2614	219
	DCV	2216	2602	387
	HoCV-1	2377	2574	198
	PSIV	2333	2527	195

[16-18]. This ORF overlaps the 5' end of the structural polyprotein gene (Figure 1A). The detection of purifying selection is based on a method for the simultaneous estimation of selection intensities in overlapping genes [14]. To ascertain that each overlapping ORF is indeed subject to selection, we used the likelihood ratio test for two hierarchical models. In model 1, we assume no selection on the overlapping ORF. In model 2, the overlapping ORF is assumed to be under selection. If model 2 fits the data significantly better than model 1 ($p < 0.05$), then the overlapping ORF is predicted to be under selection and is most probably functional. The signature of selection was identified for the ORFs in the three bee viruses (IAPV, ABPV, and KBV). The protein product of the orthologous ORF in SINV-1 could not be tested for selection because the amino acid sequence identity between the ORF from SINV-1 and the ORFs from the three bee viruses (Table 3) is lower than the range of sequence identities for which the method can be applied (65-95%).

An additional indication for selection on these ORFs was obtained by comparing the degrees of conservation of the hypothetical protein sequences of the overlapping ORFs against the protein sequences of the known genes (structural and nonstructural polyproteins, Table 3). The degree of amino-acid conservation and, hence, sequence identity between orthologous protein-coding genes is influenced *ceteris paribus* by the intensity of purifying selection. If both overlapping genes are under similar strengths of selection, the amino-acid sequence identity of one pair of homologous genes would be similar to that of the overlapping pair. On the other hand, if a functional gene overlaps a non-functional ORF, the amino-acid identity between the hypothetical protein sequences of the non-functional ORFs would be much lower than that between the two homologous overlapping functional genes. We found that the degree of amino-acid conservation of the overlapping sequence identity between pairs of overlapping ORFs in cluster A is only slightly lower than that of the known gene (maximum of 12% difference between IAPV and SINV-1 in cluster A, Table 3). In contrast, the amino-acid sequence identity between ORF pairs in clusters B and C is much lower than that between the pairs of known genes (maximum of 44% difference between CrPV and DCV in cluster C, Table 3).

The signature of purifying selection on the ORFs in cluster A suggests that they may encode functional proteins. We provisionally term this gene *pog* (*p*redicted *o*verlapping *g*ene). In Figure 1, we show that *pog* is found in the genomes of four viruses that constitute a monophyletic clade, but not in any other dicistrovirid genome (Figure 1A). Its phylogenetic distribution suggests that *pog* originated before the divergence of SINV-1 from the three bee viruses. The phylogenetic distributions of the ORFs in

Table 3: Sequence conservation in comparisons of known orthologous proteins and orthologous products of overlapping ORFs.

Cluster	Genome pair		Identity of known proteins (%)	Identity of hypothetical product of overlapping ORFs (%)
A	IAPV	ABPV	80.2	74.8
	ABPV	KBV	79.3	75.6
	IAPV	KBV	77.4	72.5
	IAPV	SINV-1	42.7	30.3
	ABPV	SINV-1	41.6	32.6
	KBV	SINV-1	36.3	29.4
B	KBV	ABPV	87.7	52.3
C	CrPV	DCV	80.3	36.1
	HoCV-1	PSIV	64.3	40.0
	DCV	HoCV-1	56.4	28.8
	CrPV	HoCV-1	48.0	31.7
	DCV	PSIV	44.2	36.4
	CrPV	PSIV	35.7	25.0

clusters B and C (Figure 1B) are patchy. This patchiness is an additional indication that the overlapping ORFs in clusters B and C are spurious, i.e., non-functional.

An examination of the DNA alignment of *pog* (Figures 2) reveals a conservation of the first potential start codon (ATG or CTG) in the +1 reading frame in three out of the four viral genomes (IAPV, ABPV, and SINV-1). As seen in Figure 3, this conservation cannot be explained by constraints on the overlapping polyprotein, in which the corresponding site is variable and encodes different amino acids (His, Asn, and Pro, in IAPV, ABPV, and SINV-1, respectively). We note, however, that we did not find a conserved Kozak consensus sequence [19] upstream of the potential initiation site. This situation is similar to that described in [13].

A protein motif search resulted in several matches, all with a weak score. Two patterns were found in all four proteins: (1) a signature of rhodopsin-like GPCRs (G protein-coupled receptors), and (2) a protein kinase C phosphorylation site (Figure 3). Prediction of the secondary structures [20] suggests that the proteins contain two conserved helix domains, separated by 3-5 residues (except for SINV-1, in which one long domain is predicted), at the C-terminus (Figure 3). A search for transmembrane topology [21] indicates that the longer helix may be a transmembranal segment (Figure 3). Although viruses often use GPCRs to exploit the host immune system through molecular mimicry [22-25], the lengths of the proteins encoded by *pog* are shorter than the average virus-encoded GPCR. Therefore, these proteins may have a different function.

Conclusion

In this note, we provide evolutionary evidence (purifying selection) for the existence of a functional overlapping gene, *pog*, in the genomes of IAPV, ABPV, KBV, and SINV-1. To our knowledge, this putative gene, whose coding region overlaps the structural polyprotein, has not been described in the literature before.

Methods

Sequence Data, Processing, and Analysis

Fourteen completely sequenced dicistrovirid genomes were obtained from NCBI (Table 1). Each genome was scanned for the presence of overlapping ORFs. We used BLASTP [26] with the protein sequences of the known genes to identify matches of orthologous overlapping ORFs (E value < 10^{-6}). Matching overlapping ORFs were assigned into clusters. Within each cluster, we aligned the amino-acid orthologs by using the sequences of the known genes as references. If alignment length of the overlapping sequence exceeded 60 amino-acids, and if the amino-acid sequence identity among the hypothetical genes within a cluster was higher than 65%, we tested for selection on the hypothetical gene (see below).

We aligned the protein sequences of the two polyproteins with CLUSTAW [27] as implemented in the MEGA package [28]. Alignment quality was confirmed using HoT [29]. We reconstructed two phylogenetic trees (one for each polyprotein) by applying the neighbor joining method [30], as implemented in the MEGA package [28]. Trees were rooted by the mid-point rooting method [31] and confidence of each branch was estimated by bootstrap with 1000 replications.

Detection of Selection in Overlapping Genes

We used the method of Sabath et al. [14] for the simultaneous estimation of selection intensities in overlapping genes. This method uses a maximum-likelihood framework to fit a Markov model of codon substitution to data
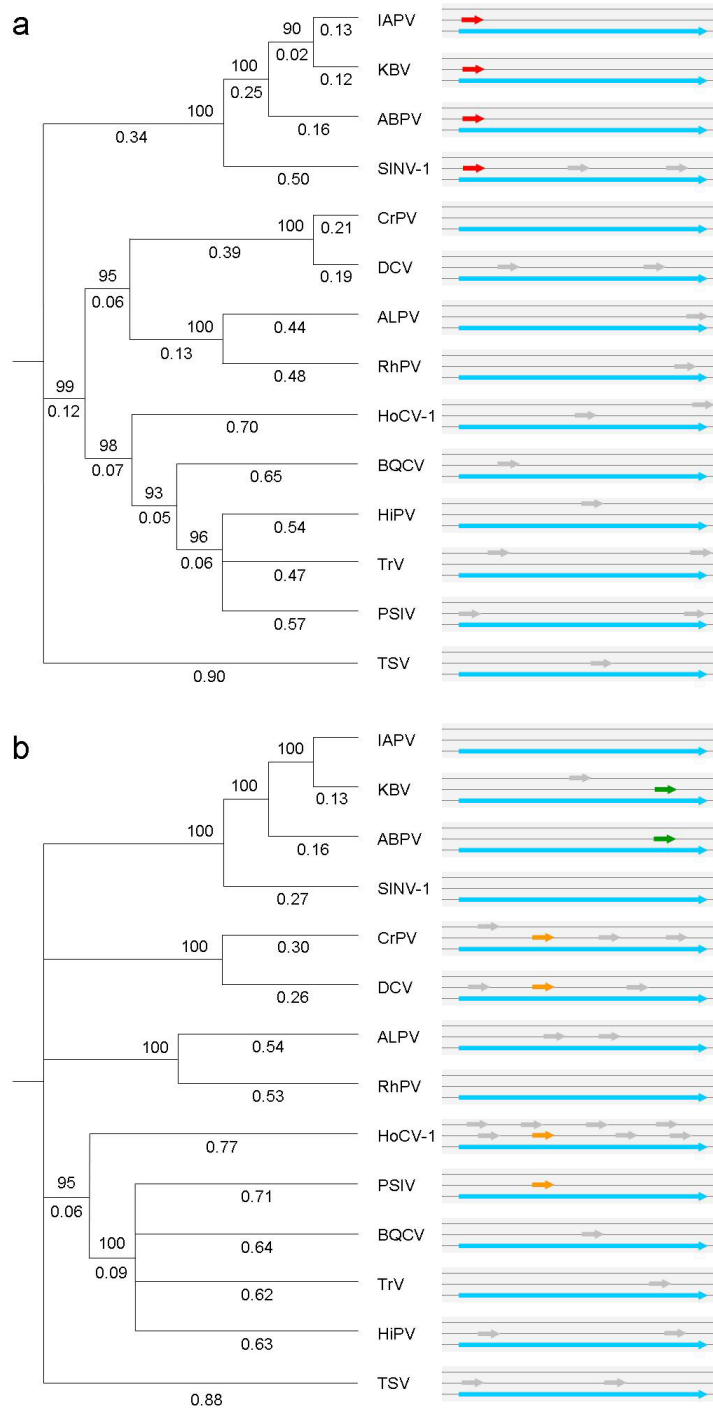
Figure 1
Phylogenetic trees and schematic representation of the dicistrovirid genomes (a. structural polyprotein; b. non-structural polyprotein). Trees were inferred using the neighbor joining method [30] and rooted by the mid-point rooting method [31]. Numbers above and below the branches are bootstrap values (1000 replications) and branch lengths (amino-acid substitutions per site), respectively. Phylogenetic analyses were conducted with MEGA [28]. The approximate locations and sizes of the known genes (blue), overlapping hypothetical genes (red, green, and orange), and singlet ORFs (gray) are noted in the three reading frames.

```
              1    2    3    4    5    6    7    8    9    10   11   12   13   14   15   16   17   18   19   20
IAPV        gaa  cag  ctg  tac  tgg  gca  gtt  aca  gca  gtc  gta_ tgg  taa  cac  atg  cgg  cgt  tcc  gaa  ata

ABPV        gaa  cag  cta  tat  tgg  gta  gtt  gta  gca  gtt  gta  ttc  aaa_ tga  atg  cag  cgt  tcc  gaa  ata

KBV         aaa  ccg  cta  tat  cgg  gta  gct  ata  gca  gtc  gga  tag  taa  tat  atc  cgg  cgt  ttc  gaa  ata

SINV-1      tag  cag  tca  gga_ tgt  cat  tct  ggc  gtt  ccg  aaa  tac  cca  aac  ctg  ctc  aat  caa  aca  atg


              21   22   23   24   25   26   27   28   29   30   31   32   33   34   35   36   37   38   39   40
IAPV        cca  tgc  ctg  gcg  att  cac  aac  aag  aaa  gca  ata  ctc  cca  acg  tac  aca  ata  cgg  aac  tcg

ABPV        tca  tac  ctg  ccg  atc  ---  ---  aag  aaa  caa  ata  ctt  cca  acg  tac  ata  ata  cgc  aac  tcg

KBV         cca  tac  ctg  ctg  ata  ---  acc  aag  aaa  acg  att  cta  cca  atg  tac  ata  aca  cga  aac  tcg

SINV-1      cga  ata  ctt  ttg  aga  cga  aaa  cgg  caa  caa  cct  ctg  ctt  ccc  acg  cac  aat  cgg  aac  tta
```

Figure 2
Codon alignment of the 5' overlap region between the structural polyprotein and the hypothetical gene. The alignment is shown in the reading frame of the hypothetical gene. The annotated initiation site of the polyproteins is underlined. The first potential initiation site (AUG or CUG) of the hypothetical genes is marked in red. The last stop codon at the +1 reading frames is marked in green.

```
IAPV         GTAVLGSYSSRMVTHAAFRNTMPGDSQQESNTPNVHNTELASSTSENSVETQEITTFHDV   60
ABPV         GTAILGSCSSCIQMNAAFRNIIPADQ--ETNTSNVHNTQLASTSEENSVETEQITTFHDV   58
KBV          ETAISGSYSSRIVIYPAFRNTIPADN-QENDSTNVHNTKLASTSAENAIEKEQITTFHDV   59
SINV-1       IAVRMSFWRSEIPKPAQSNNANTFETKTATTSASHAQSELSETTPENSLTRQELTVFHDV   60


IAPV    +1   EQLYWAVTAVVW*HMRRSEIPCLAIHNKKAILPTYTIRNSLRPLVKTRLRPKKSQPFMMW
ABPV    +1   EQLYWVVVAVVFK*MQRSEISYLPI--KKQILPTYIIRNSRRPLKKTQLKRNKSPPFMMW
KBV     +1   KPLYRVAIAVG**YIRRFEIPYLLI-TKKTILPMYITRNSRRPQRRMPLRRNKSPPFMMW
SINV-1  +1   *QSGCHSGVPKYPNLLNQTMRILLRRKRQQPLLPTHNRNLARRPQKIPLPDKNSQFSMML


IAPV         ETPNRIDTPMAQDTSSARNMDDTHSIIQFLQRPVLIDNIEIIAGTTADANKPLSRYV---  117
ABPV         ETPNRINTPMAQDTSSARSMDDTHSIIQFLQRPVLIDHIEVIAGSTADDNKPLNRYV---  115
KBV          ETPNRIDTPMAQDTSSARSMDDTHSIIQFLQRPVLIDNIEIVAGTTADNNTALSRYV---  116
SINV-1       EQPRVALPIAPQTTSSLAKLDSTATIVDFLSRTVVLDQFELVQGESNDNHKPLNAATFKD  120


IAPV    +1   KLQIGSIPPWLRILHRLGTWMIRTVLFSFYSAPFSLTTLRSLLEQRPMQTNPLADM*---
ABPV    +1   KLQIGSIPPWLKTLHRLGAWMIRTVLFSFYNAPYSLTTLRSLLDQQQMITNPSIDM*---
KBV     +1   KLQIGSIPPWLRILHRLGAWMIRTVLFSFYNAPFSLTTLRLLQEQLPITTQHSVDM*---
SINV-1  +1   NNLASLFQLLRKRLALLLSLILQRQLWIFFLELLSSINSSLFKVNQTITTNPLTQQLLKT
```

Figure 3
The amino-acid alignment of the overlap region between the structural polyprotein and the hypothetical gene (+1 reading frame). The annotated initiation site of the polyproteins is marked in blue. The first potential initiation site (AUG or CUG) of the hypothetical genes is marked in red. The last stop codon at the +1 reading frames is marked in green. Transmembranal helixes predicted by MEMSAT [21] are marked in blue. Conserved protein kinase C phosphorylation sites predicted through My-Hits server http://hits.isb-sib.ch/cgi-bin/PFSCAN are marked in yellow.

from two aligned homologous overlapping sequences. To predict functionality of an ORF that overlaps a known gene, we modified an existing approach for predicting functionality in non-overlapping genes [32]. Given two aligned orthologous overlapping sequences, we estimate the likelihood of two hierarchical models. In model 1, there is no selection on the ORF. In model 2, the ORF is assumed to be under selection. The likelihood-ratio test is used to test whether model 2 fits the data significantly better than model 1, in which case, the ORF is predicted to be under selection and most probably functional.

Motifs
We looked for motifs within the inferred protein sequences encoded by the overlapping ORF by using the motif search server http://motif.genome.jp/ and the My-Hits server http://hits.isb-sib.ch/cgi-bin/PFSCAN with the following motif databases: PRINTS [33], PROSITE [34], and Pfam [35]. We used PSIPRED [20] to predict secondary structure, and MEMSAT [21] to predict transmembrane protein topology.

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

Authors' contributions
NS carried out the analysis and wrote the draft manuscript. NP performed the motif search. DG and NP contributed to the interpretation of the results and the final version.

All authors have read and approved the manuscript.

Appendix 1
Overlapping ORFs on the negative strand
In the fourteen completely sequenced dicistroviruse genomes (Table 1), we identified 240 overlapping ORFs of length equal or greater than 60 codons on the negative strand. Of the 240 ORFs, 113 were found in concordant genomic locations in two or more genomes. The concordant overlapping ORFs were assigned into 29 clusters (Additional file 1). There are 9, 1, and 19 clusters in phase 0, 1, and 2, respectively. The cluster size ranges from 2 to 9. In two clusters, 5 and 10, both in phase 2, there is a weak signature of selection. However, this signature seems to be a false positive, which was driven by the unique structure of opposite-strand phase-2 overlap (Additional file 2). In this structure, codon positions one and two of one gene match codon positions two and one of the overlapping gene. This structure leads to a situation where most changes are either synonymous or nonsynonymous in both overlapping genes and occasionally, to false signal of purifying selection on the overlapping ORF. In addition, one of the clusters (cluster 10) does not constitute a monophyletic clade, and is, therefore, unlikely to be func-

tional. We therefore conclude that dicistroviruses most probably do not encode proteins on the negative strand.

Additional material

> ### Additional file 1
> *Clusters of orthologous overlapping ORFs on the negative strands of dicistrovirid genomes.*
> Click here for file
> [http://www.biomedcentral.com/content/supplementary/1743-422X-6-144-S1.DOC]
>
> ### Additional file 2
> *The corresponding codon positions of overlapping genes in opposite-strand phase-2. First and second codon positions, in which ~5% and 0% of the changes are synonymous, are marked in red. Third codon positions, in which ~70% of the changes are synonymous, are marked in blue.*
> Click here for file
> [http://www.biomedcentral.com/content/supplementary/1743-422X-6-144-S2.PPT]

References
1.　Oldroyd BP: **What's killing American honey bees?** *PLoS Biol* 2007, **5**:e168.
2.　Gallai N, Salles J-M, Settele J, Vaissière BE: **Economic valuation of the vulnerability of world agriculture confronted with pollinator decline.** *Ecological Economics* 2009, **68**:810-821.
3.　van Engelsdorp D, Hayes J Jr, Underwood RM, Pettis J: **A survey of honey bee colony losses in the U.S., fall 2007 to spring 2008.** *PLoS ONE* 2008, **3**:e4071.
4.　Cox-Foster DL, Conlan S, Holmes EC, Palacios G, Evans JD, Moran NA, Quan PL, Briese T, Hornig M, Geiser DM, *et al.*: **A metagenomic survey of microbes in honey bee colony collapse disorder.** *Science* 2007, **318**:283-287.
5.　Maori E, Lavi S, Mozes-Koch R, Gantman Y, Peretz Y, Edelbaum O, Tanne E, Sela I: **Isolation and characterization of Israeli acute paralysis virus, a dicistrovirus affecting honeybees in Israel: evidence for diversity due to intra- and inter-species recombination.** *J Gen Virol* 2007, **88**:3428-3438.
6.　Blanchard P, Schurr F, Celle O, Cougoule N, Drajnudel P, Thiery R, Faucon JP, Ribiere M: **First detection of Israeli acute paralysis virus (IAPV) in France, a dicistrovirus affecting honeybees (*Apis mellifera*).** *J Invertebr Pathol* 2008, **99**:348-350.
7.　Palacios G, Hui J, Quan PL, Kalkstein A, Honkavuori KS, Bussetti AV, Conlan S, Evans J, Chen YP, vanEngelsdorp D, *et al.*: **Genetic analysis of Israel acute paralysis virus: distinct clusters are circulating in the United States.** *J Virol* 2008, **82**:6209-6217.
8.　Chen W, Calvo PA, Malide D, Gibbs J, Schubert U, Bacik I, Basta S, O'Neill R, Schickli J, Palese P, *et al.*: **A novel influenza A virus mitochondrial protein that induces cell death.** *Nat Med* 2001, **7**:1306-1312.
9.　Chung BY, Miller WA, Atkins JF, Firth AE: **An overlapping essential gene in the Potyviridae.** *Proc Natl Acad Sci USA* 2008, **105**:5897-5902.
10.　Firth AE: **Bioinformatic analysis suggests that the Orbivirus VP6 cistron encodes an overlapping gene.** *Virol J* 2008, **5**:48.
11.　Firth AE, Atkins JF: **Bioinformatic analysis suggests that the Cypovirus 1 major core protein cistron harbours an overlapping gene.** *Virol J* 2008, **5**:62.

12. Firth AE, Atkins JF: **Bioinformatic analysis suggests that a conserved ORF in the waikaviruses encodes an overlapping gene.** *Arch Virol* 2008, **153:**1379-1383.
13. Firth AE, Atkins JF: **Analysis of the coding potential of the partially overlapping 3' ORF in segment 5 of the plant fijiviruses.** *Virol J* 2009, **6:**32.
14. Sabath N, Landan G, Graur D: **A method for the simultaneous estimation of selection intensities in overlapping genes.** *PLoS ONE* 2008, **3:**e3996.
15. Nguyen M, Haenni AL: **Expression strategies of ambisense viruses.** *Virus Res* 2003, **93:**141-150.
16. de Miranda JR, Drebot M, Tyler S, Shen M, Cameron CE, Stoltz DB, Camazine SM: **Complete nucleotide sequence of Kashmir bee virus and comparison with acute bee paralysis virus.** *J Gen Virol* 2004, **85:**2263-2270.
17. Govan VA, Leat N, Allsopp M, Davison S: **Analysis of the complete genome sequence of acute bee paralysis virus shows that it belongs to the novel group of insect-infecting RNA viruses.** *Virology* 2000, **277:**457-463.
18. Valles SM, Strong CA, Dang PM, Hunter WB, Pereira RM, Oi DH, Shapiro AM, Williams DF: **A picorna-like virus from the red imported fire ant, *Solenopsis invicta*: initial discovery, genome sequence, and characterization.** *Virology* 2004, **328:**151-157.
19. Kozak M: **Comparison of initiation of protein synthesis in procaryotes, eucaryotes, and organelles.** *Microbiol Rev* 1983, **47:**1-45.
20. McGuffin LJ, Bryson K, Jones DT: **The PSIPRED protein structure prediction server.** *Bioinformatics* 2000, **16:**404-405.
21. Jones DT: **Improving the accuracy of transmembrane protein topology prediction using evolutionary information.** *Bioinformatics* 2007, **23:**538-544.
22. Murphy PM: **Viral exploitation and subversion of the immune system through chemokine mimicry.** *Nat Immunol* 2001, **2:**116-122.
23. Lalani AS, McFadden G: **Evasion and exploitation of chemokines by viruses.** *Cytokine Growth Factor Rev* 1999, **10:**219-233.
24. McLysaght A, Baldi PF, Gaut BS: **Extensive gene gain associated with adaptive evolution of poxviruses.** *Proc Natl Acad Sci USA* 2003, **100:**15655-15660.
25. Hughes AL, Friedman R: **Genome-wide survey for genes horizontally transferred from cellular organisms to baculoviruses.** *Mol Biol Evol* 2003, **20:**979-987.
26. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.
27. Thompson JD, Gibson TJ, Higgins DG: **Multiple sequence alignment using ClustalW and ClustalX.** *Curr Protoc Bioinformatics* 2002, **Chapter 2(Unit 2):**3.
28. Kumar S, Nei M, Dudley J, Tamura K: **MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences.** *Brief Bioinform* 2008, **9:**299-306.
29. Landan G, Graur D: **Heads or tails: a simple reliability check for multiple sequence alignments.** *Mol Biol Evol* 2007, **24:**1380-1383.
30. Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Mol Biol Evol* 1987, **4:**406-425.
31. Farris JS: **Estimating phylogenetic trees from distance matrices.** *Am Nat* 1972, **106:**645-668.
32. Nekrutenko A, Makova KD, Li WH: **The K(A)/K(S) ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study.** *Genome Res* 2002, **12:**198-202.
33. Attwood TK, Blythe MJ, Flower DR, Gaulton A, Mabey JE, Maudling N, McGregor L, Mitchell AL, Moulton G, Paine K, Scordis P: **PRINTS and PRINTS-S shed light on protein ancestry.** *Nucleic Acids Res* 2002, **30:**239-241.
34. Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux PS, Pagni M, Sigrist CJ: **The PROSITE database.** *Nucleic Acids Res* 2006, **34:**D227-230.
35. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL, Bateman A: **The Pfam protein families database.** *Nucleic Acids Res* 2008, **36:**D281-288.