Powerful Sequence Similarity Search Methods and In-Depth Manual Analyses Can Identify Remote Homologs in Many Apparently "Orphan" Viral Proteins

Durga B. Kuchibhatla,[a] Westley A. Sherman,[a] Betty Y. W. Chung,[b] Shelley Cook,[c] Georg Schneider,[a,d] Birgit Eisenhaber,[a] David G. Karlin[e,f]

Bioinformatics Institute (BII), A*STAR (Agency for Science, Technology and Research), Matrix, Singapore[a]; Department of Plant Sciences, University of Cambridge, Cambridge, United Kingdom[b]; Life Sciences—Parasites and Vectors Division, Natural History Museum, London, United Kingdom[c]; IST Austria (Institute of Science and Technology Austria), Klosterneuburg, Austria[d]; Department of Zoology, Oxford University, Oxford OX1 3PS, United Kingdom[e]; Division of Structural Biology, Oxford University, Oxford OX1 7BN, United Kingdom[f]

The genome sequences of new viruses often contain many "orphan" or "taxon-specific" proteins apparently lacking homologs. However, because viral proteins evolve very fast, commonly used sequence similarity detection methods such as BLAST may overlook homologs. We analyzed a data set of proteins from RNA viruses characterized as "genus specific" by BLAST. More powerful methods developed recently, such as HHblits or HHpred (available through web-based, user-friendly interfaces), could detect distant homologs of a quarter of these proteins, suggesting that these methods should be used to annotate viral genomes. In-depth manual analyses of a subset of the remaining sequences, guided by contextual information such as taxonomy, gene order, or domain cooccurrence, identified distant homologs of another third. Thus, a combination of powerful automated methods and manual analyses can uncover distant homologs of many proteins thought to be orphans. We expect these methodological results to be also applicable to cellular organisms, since they generally evolve much more slowly than RNA viruses. As an application, we reanalyzed the genome of a bee pathogen, *Chronic bee paralysis virus* (CBPV). We could identify homologs of most of its proteins thought to be orphans; in each case, identifying homologs provided functional clues. We discovered that CBPV encodes a domain homologous to the *Alphavirus* methyltransferase-guanylyltransferase; a putative membrane protein, SP24, with homologs in unrelated insect viruses and insect-transmitted plant viruses having different morphologies (cileviruses, higreviruses, bluerviruses, negeviruses); and a putative virion glycoprotein, ORF2, also found in negeviruses. SP24 and ORF2 are probably major structural components of the virions.

The detection of distant homologs of a protein has many applications. For example, it can provide clues to its function, guide the choice of substitutions for experimental studies, and facilitate three-dimensional (3D) structure determination (1). However, several sequence-based studies have reported that a significant fraction of viral proteins had no detectable homologs (2–4). These proteins have been called "orphans," "ORFans" (5), or more accurately, "taxonomically restricted" (6) to indicate that they have no detectable homologs outside a certain taxon.

Some proteins classified as taxonomically restricted are thought to be truly specific to a particular organism, which they may endow with unique functions (6, 7). However, given the high rate of evolution of viral proteins, especially in RNA viruses (8), some orphans may in fact be part of larger protein families whose other members have diverged in sequence beyond recognition (9, 10). In fact, studies of viral orphans have relied mainly on the BLAST (Basic Local Alignment Search Tool) program (11, 12) to identify homologs (2, 3), rather than on more recent, powerful methods based on sequence profiles, such as sequence-profile comparison (PSI-BLAST [11], HMMER3 [13]) or profile-profile comparison (HHpred [14], HHblits [15], FFAS [1], WebPRC [16]). In contrast to BLAST, which compares single sequences, these methods rely on the comparison of multiple-sequence alignments, encoded as sequence profiles. A sequence profile is a representation of a multiple-sequence alignment that contains information about which amino acids are "allowed" at each position of the alignment and with what probability (17). Comparing profiles is much more sensitive than comparing single sequences, because the profiles contain information about how the sequences can evolve and can thus identify faint similarities that remain after the sequences have evolved apart (18, 19).

In the first part of this study, we asked whether a combination of powerful automated methods and in-depth manual analysis could reveal overlooked homologs of viral proteins classified as "genus restricted" by BLAST. To answer this question, we set up an automated pipeline that could run various sequence similarity detection methods and analyze the taxonomic distribution of the homologs they identified.

In the second part, we applied these methods to the genome of a phylogenetically isolated virus, *Chronic bee paralysis virus* (CBPV), a pathogen of the honeybee, in which most of the open

reading frames (ORFs) were classified as orphans ([20]). We could find homologs of most of these ORFs and suggest putative functions for them. In particular, our results suggest that several insect and plant viruses that have different morphologies nevertheless have homologous structural proteins (SPs).

MATERIALS AND METHODS

Databases used. We ran BLAST ([11]) and PSI-BLAST ([11]) against the NCBI nonredundant (nr) database (1 April 2012 release). We ran HHsearch searches against version 26 of PFAM ([21]) and HHblits against its own database of UniProt sequence clusters, UniProt20 ([22]) (2 December 2011 release). We relied on the NCBI taxonomy (2/3 April 2012 release) to map UniProt identifiers (from the UniProt 21 March 2012 release) to the NCBI taxonomy.

Sequence similarity searches. For homology searches, we ran BLAST, PSI-BLAST, HHblits, and HHsearch with the following parameters: BLAST, executable blastall, version 2.2.23, E value cutoff of 10^{-3}, SEG low-complexity filtering enabled; PSI-BLAST, executable psiblast, version 2.2.26+, E value cutoff of 10^{-3}, low-complexity filtering enabled, 10 iterations maximum; HHblits and HHsearch, executable hhblits and hhsearch from HHsuite, version 2.0.13, E value of 10^{-3}, realignment using local maximum-accuracy algorithm enabled with a default maximum accuracy threshold parameter value of 0.35, four iterations maximum. We used the A3M multiple-sequence alignment generated by HHblits to run HHsearch.

ANNOTATOR environment. We used the ANNOTATOR web application ([23]) to run the search algorithms and calculate the raw taxonomy statistics. ANNOTATOR provides a convenient interface for running a comprehensive array of sequence analysis algorithms focused on protein function discovery. It is available at http://annotator.bii.a-star.edu.sg. The results were analyzed and tabulated by using a set of custom Perl scripts.

Constitution of a data set of viral proteins classified by BLAST as genus restricted. We adapted a data set of viral "ORFan" genes (species restricted) generously provided by YanBin Yin ([3]). To keep the size of the data set manageable, we focused on a subset composed of proteins from viruses with positive, single-stranded RNA viral genomes and for which BLAST detected no homologs in other genera by using the parameters and database described above. For the final data set, containing 351 sequences, see Table S1 in the supplemental material.

Identifying the taxonomic distribution of homologs. For each query, we compiled the taxonomic distribution of the hits retrieved by the similarity detection tools as follows. First, we collected all of the hits that had statistically significant similarity (E values of $\leq 10^{-3}$) to the query. For BLAST and PSI-BLAST, these hits correspond to single sequences and we retrieved their NCBI taxonomy. For HHblits, the hits correspond to clusters of protein sequences and we retrieved the NCBI taxonomy of all of the sequences listed in the clusters. For HHpred searches, the hits are "families" of protein sequences and we retrieved the taxonomy of all of the sequences from these families. Some PFAM families are grouped into "clans" ([24]). We retrieved the taxonomic distribution of all of the sequences from these clans. We discarded any nonviral sequence, which would have complicated the analysis without any benefit for our study (see Discussion). Second, having collected the taxonomic distribution of hits for each software, we counted the distinct taxons that were retrieved at the species, genus, and family ranks.

In some rare cases, virus species have no assigned genus or family. In these cases, we adjusted the taxonomic counts so that the counts for a particular query at a lower rank would be at least equal to the counts at a higher rank. For example, a query that had homologs in four viral families would also be counted as having homologs in four viral genera—even if the viral families in question did not have defined genera.

In-depth manual homology detection incorporating contextual information. To identify remote homologs missed by automated searches, we exploited "contextual" information, such as taxonomy, genome organization, and domain organization ([25–28]). Our procedure is similar to

that described previously ([29]) and consists of two steps, the detection of potential homologs and their validation. We first identified "straightforward" homologs of the query protein in the NCBI nr database (1 April 2012 release) by using HHpred ([14]), HHblits ([15]), and CSI-BLAST ([30], [31]) and selecting hits whose E values were below the cutoff of 10^{-3}. We then examined subsignificant hits (i.e., those with E values of $>10^{-3}$) up to an E value of 2,000, looking for viral proteins or domains that came from a virus taxonomically related to the query (or infecting similar hosts) and/or that occurred in the same position of the genome or of the viral polyprotein. Such subsignificant hits, which have weak similarity to the query protein and occur in a similar genomic context, constitute potential homologs. To validate these candidates, we gathered homologs of these subsignificant hits (as described above, i.e., with E values of $\leq 10^{-3}$) and used HHalign ([32]) to compare homologs of the query protein (obtained as described above) with homologs of the subsignificant hits. We considered an HHalign E value of $<10^{-5}$ to indicate homology between the subsignificant hit and the query. We performed additional checks, such as verifying that the secondary structure and function of the hits were compatible with those of the query. When we validated a potential homolog, we repeated the procedure after including it in query alignments (i.e., we performed iterative or "cascade" searches [33–35] until no new homologs were found).

Detection of homologs of CBPV ORFs. To detect homologs of ORF1, we obtained the following tools from their web servers and used them with default parameters: HHpred ([14]) (http://toolkit.tuebingen.mpg.de/hhpred), FFAS ([1]) (http://ffas.sanfordburnham.org/ffas-cgi/cgi/ffas.pl), and WebPRC ([16]) (http://www.ibi.vu.nl/programs/prcwww). We used PROMALS ([36]) to compare the secondary structure of ORF1 with that of known methyltranferase-guanylyltransferases (MTase-GTases).

We used contextual information coupled with sequence similarity searches as described above to detect homologs of ORF2 and ORF3. We obtained CSI-BLAST ([30], [31]) from its web server (http://toolkit.tuebingen.mpg.de/cs_blast#) and used it with five iterations and most of the default parameters (inclusion cutoff E value of 10^{-3}, low-complexity filter not enabled, nr database). However, to examine as many subsignificant hits as possible, even extremely weak ones, we set the maximum E value reported to 2,000 (the default value is 10) and the maximum number of sequences reported to 2,000 (the default is 100).

We used ANNIE ([37]; http://annie.bii.a-star.edu.sg) to predict the structural properties of all of the ORFs of CBPV; MetaPrDOS ([38]) for disorder prediction, respecting the principles described in reference [39]; Composition Profiler ([40]) for sequence composition analyses; and LOMETS ([41]) for fold recognition.

Prediction of TM segments in CBPV ORF2 and ORF3. To predict the number and locations of transmembrane (TM) segments of ORF2 and ORF3, we relied on two complementary approaches. On the one hand, for each virus, we compared the predictions of multiple programs for a single sequence (vertical approach). We considered the robustness of the prediction to be proportional to the number of predictors that detected a given TM segment. The predictors used were those applied and displayed by ANNIE ([37]). On the other hand, we compared the prediction of a single program for several homologs (horizontal approach) by using TM-coffee ([42]), which predicts TM segments in multiple sequences via HMMTOP ([43]). We considered the robustness of the prediction to be proportional to the number of sequences that contained a given predicted TM segment.

Finally, we used Phobius ([44]) to predict signal peptides and the topology of membrane proteins and TMSOC ([45]) to distinguish "complex" from "simple" anchor-type TM segments liable to give spurious hits in similarity searches.

RESULTS

An intuitive understanding of the power of similarity search methods can be provided by the genetic distance at which they can detect homologs. However, the genetic distances between viral

TABLE 1 Capacities of the methods tested to detect homologs at different taxonomic depths[a]

Algorithm	% of sequences for which homologs were found with the following taxonomic distribution:		
	At most 1 genus	>1 genus	>1 family
BLAST	100	0	0
PSI-BLAST	94	6	2.6
HHblits	81.5	18.5	8.3
HHpred	80.1	19.9	14.2
All combined	74.6	25.4	14.2

[a] The total for each row can be >100% because "beyond genus level" includes "beyond family level." Likewise, the total proportion of the different algorithms in each column can be greater than the value of the cell "all methods combined" because for some proteins, distant homologs were detected by several algorithms. Percentages were calculated from a total of 351 sequences.

proteins often cannot be easily established by standard phylogenetic analyses, owing to their high rate of evolution (particularly in RNA viruses). Instead, we relied on the precomputed viral taxonomy as a proxy for genetic distances, since recent studies suggest that viral taxonomy reflects sequence-based phylogeny well (46–48) (despite being based on additional factors such as antigenicity).

We compiled a data set of 351 proteins from positive-sense, single-stranded RNA viruses that BLAST classifies as genus restricted by adapting a previously published data set identified as ORFans on the basis of BLAST searches (3) (see Table S1 in the supplemental material).

Profile-profile methods find more distant homologs of 25% of the proteins classified as genus restricted by BLAST. We first ran several automated sequence similarity search programs on this data set and compared the taxonomic depths at which they detected homologs. For a given program, a protein was deemed homologous to the query if the program reported a statistically significant sequence similarity between them, i.e., with an E value smaller than the cutoff of 10^{-3}. The programs we compared included the widely used BLAST and PSI-BLAST programs and the more recent profile-profile comparison methods HHpred (14) and HHblits (15).

If the proteins in the data set were truly genus restricted, then no algorithm would detect homologs in other genera. This is not what we observed. Table 1 and Fig. 1 show that all of the methods, aside from BLAST, detect homologs in other genera for at least some of the proteins in the data set. For example, HHpred detected distant homologs (i.e., in more than one genus) for about 20% of the sequences and very distant homologs (i.e., in more than one family) for about 14% of the sequences. Methods based on profile-profile comparison, i.e., HHpred and HHblits, detected many more distant homologs than PSI-BLAST (Fig. 1 and Table 1). Among these methods, HHpred found about as many distant homologs as HHblits but markedly more very distant homologs. This is probably due to the fact that contrary to HHblits, HHpred relies on a database of protein profiles built with human supervision (PFAM); in particular, PFAM clans often incorporate other information than merely sequence data, such as 3D structure or function (24). When combining results from PSI-BLAST, HHblits, and HHpred (right side of Fig. 1), in total, >25% of the 89 proteins in the data set had distant homologs and >14% had very distant homologs. Thus, many of the proteins characterized

as genus restricted by BLAST and thought to be ORFans (3) are actually members of protein families with a wide taxonomic distribution.

For the proteins for which homologs were detected in more than one genus by at least one program and their taxonomic distribution according to each method, see Table S2 in the supplemental material. To corroborate these results, we examined in-depth 18 proteins (one-fifth) selected at random from among the 89 proteins found to have distant homologs (see Materials and Methods). We found that all were correct (not shown). We also ran the same calculations with a more stringent E value cutoff (10^{-5}) and obtained qualitatively similar results (not shown), confirming that they are not an artifact due to the detection of false positives.

Manual analysis using biological context reveals markedly more distant homologs than automated methods. Even the profile-profile methods described above could not find distant homologs of 75% of the proteins in the data set with the standard significance cutoff (E = 10^{-3}). However, more distant homologs can sometimes be detected by examining subsignificant hits (i.e., those with E values of $>10^{-3}$) and using "extrinsic" or "contextual" sequence-based information that goes beyond simple sequence similarity (such as taxonomy, gene order, or domain organization) (25). We randomly selected 10 such proteins (Table 2) and analyzed them in depth manually (see Materials and Methods). This allowed the detection of more homologs in half of the cases (5 out of 10) listed in the top half of Table 2 (compare columns 4 and 5). The evidence supporting homology is described in the last column of Table 2.

Of these five proteins, three have homologs in more than one genus, including two that have homologs in more than one family. The 6K2 protein of *Cucumber vein yellowing virus* is homologous to the 6K2 proteins of several genera in the *Potyviridae* family (Table 2). The coat protein of *Strawberry latent ringspot virus* and that of *Mushroom bacilliform virus* are each predicted to have a jellyroll fold (49) and to have homologs in >40 families (Table 2). Experiments confirmed that they function as a capsid (50, 51).

Intriguingly, four of the five other genus-restricted proteins belong to the same family, *Closteroviridae*. The fact that this family
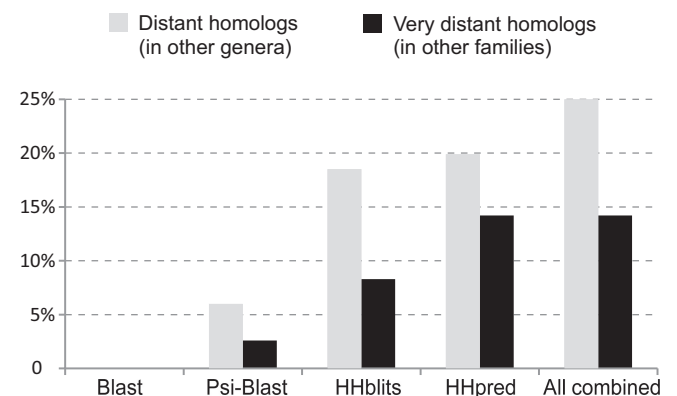

FIG 1 Capacities of the methods tested to detect homologs at different taxonomic depths. Shown are the proportions of proteins classified as genus restricted by BLAST and found by the different similarity search methods to have homologs beyond the genus level and beyond the family level. Precise values are in Table 1, columns 2 and 3.

TABLE 2 In-depth analysis, using contextual information, of a random subset of proteins classified as genus restricted by all automated methods

Accession no.	Protein name	Taxonomic position (family, genus, species)	Taxonomic distribution found by automated methods	Taxonomic distribution found by in-depth manual analysis	Type of protein (reference)	Evidence
YP_227375	Large coat protein	*Secoviridae*, unclassified, *Strawberry latent ringspot virus*	2 species	>40 families	Contains two domains of capsids with a jellyroll fold (PFAM clan Viral_ssRNA_CP) (49)	Marginal HHipred hits (E = 0.28 for region 241–308 and E = 7.5 for 43–159 region) to PFAM family RhV, HHalign comparison with RhV alignment (E = 5 × 10^{-2}), functional confirmation (50)
NP_042511	Coat protein	*Barnaviridae, Barnavirus, Mushroom bacilliform virus*	1 species	>40 families	Capsid with a jellyroll fold (PFAM clan Viral_ssRNA_CP) (49)	Marginal HHipred hit (E = 0.2) to PFAM family Viral_Coat for 67–181 region, HHalign comparison with Viral_coat (E = 1.4 × 10^{-3}), functional confirmation (51)
YP_308882	6K2	*Potyviridae, Ipomovirus, Cucumber vein yellowing virus*	4 species	4 genera in *Potyviridae* family (*Ipomovirus, Poacevirus, Tritimovirus, Macluravirus*)	6K2 (87)	Subsignificant CSI-BLAST hits to other 6K2 proteins (also located between CI protein and VPg protein), significant HHalign comparison between full-length 6K2 of *Ipomovirus* and 6K2 of *Tritimovirus, Poacevirus,* and *Macluravirus* (E = 2.9 × 10^{-7})
NP_776026	Putative matrix protein M	*Flaviviridae, Flavivirus, Tamana bat virus*	1 species	At least 1 whole genus (*Flavivirus*), may also be homologous to matrix protein of related genus *Pegivirus*	Membrane protein (52)	Subsignificant CSI-BLAST hits of M proteins of flaviviruses to *Tamana bat virus* M protein, which has identical domain position within polyprotein (between C and E proteins), significant HHalign comparison between 53–159 region of M of *Tamana bat virus* and M of other flaviviruses (E = 6 × 10^{-6})
NP_778215	VPg	Unclassified, *Sobemovirus, Turnip rosette virus*	1 species	1 whole genus (*Sobemovirus*)	Genome-linked protein VPg (88)	CSI-BLAST finds full-length significant matches to VPg of many sobemoviruses, further iterative sequence searches identify as homologs VPg of all other sobemoviruses, all have same position within 2a/2b polyprotein (downstream of serine protease domain)
YP_293702	P9	*Closteroviridae, Crinivirus, Tomato chlorosis virus*	1 genus	1 genus	P9 (53)	
NP_619694	Hypothetical protein p34	*Closteroviridae, Crinivirus, Lettuce infectious yellows virus*	1 species	2 species[a]	Endonuclease	HHipred hit to RNase Dicer[a] (PDB[b] accession no. 3c4b, E = 2.10^{-4})
NP_851570	p22	*Closteroviridae, Crinivirus, Cucurbit yellow stunting disorder virus*	1 species	1 species		
YP_293697	p5	*Closteroviridae, Crinivirus, Tomato chlorosis virus*	1 species	1 species		
YP_053926	Hypothetical peptide	*Comoviridae, Nepovirus* (subgroup A), *Tobacco ringspot virus*	1 species	1 species		

[a] p34 has homologs in only two viral species but is homologous to a vast family of RNases from cellular organisms and thus most probably originated by horizontal transfer, which is beyond the scope of this study (see Materials and Methods and Discussion).

[b] PDB, Protein Data Bank.

contains numerous proteins that lack identifiable orthologs has been noted previously (53).

In summary, 3 of 10 proteins that all of the automated methods found to be genus restricted actually have distant homologs detectable by manual sequence analysis. This value cannot be simply extrapolated to the whole data set because of the small size of the subset examined. Nevertheless, it is clear that numerous proteins classified as genus restricted by BLAST and initially thought to be ORFans (3) actually have more distant homologs. Therefore, methods relying on profile-profile comparison should be used in addition to BLAST and PSI-BLAST to annotate viral genomes.

Case study: in-depth analysis of the genome of CBPV. We applied a combination of automated profile-profile methods and manual examination of subsignificant hits to analyze the genome of a phylogenetically isolated virus, CBPV (proposed genus, *Chroparavirus* [P. Blanchard, personal communication]), that induces paralysis in the honeybee *Apis mellifera* (20). The first genome segment (RNA1) encodes three ORFs (all accession numbers are in Table 3). ORF1 and ORF3 are thought to give rise to a fusion protein, ORF1-ORF3, by a frameshift (20). ORF2 overlaps ORF1; its expression is unproven. PSI-BLAST detected significant similarity between ORF3 and viral RNA-dependent RNA polymerases (RdRPs) but could detect no homolog of ORF1 or ORF2 (20). The second genome segment (RNA2) of CBPV contains three ORFs (ORF1 to ORF3) for which PSI-BLAST could detect no homolog either (20).

ORF1 of CBPV RNA1 is homologous to the *Alphavirus* MTase-GTase. HHpred reported a statistically significant hit (E = 4.6×10^{-4}) between aa 132 to 325 of RNA1 ORF1 and the first 231 aa of the PFAM family Vmethyltransf, corresponding to the MTase-GTase of the *Alphavirus* supergroup (54). Two recently discovered viruses infecting bees have an organization similar to that of CBPV, *Lake Sinai virus 1* and *Lake Sinai virus 2* (55). Their first genome segment is also composed of an ORF1 with significant similarity to that of CBPV, followed by an ORF encoding the RdRP. We aligned the ORF1 of CBPV and those of the Lake Sinai viruses and submitted the alignment to HHpred. HHpred reported a longer match between this alignment (corresponding to aa 131 to 338 of CBPV) and almost the entire Vmethyltransf domain, albeit with a lower E value (E = 0.002), marginally under the threshold of significance. The strictly conserved histidine of the MTase-GTase of the *Alphavirus* superfamily (54) is also conserved in the ORF1 of the bee viruses (aa 158 in CBPV ORF1). In addition, the predicted secondary-structure elements of the putative MTase-GTase of CBPV matched that of the *Alphavirus* supergroup (not shown). Thus, we conclude that the aa 131 to 338 region of CBPV ORF1 contains a domain homologous to the MTase-GTase of the *Alphavirus* superfamily. This prediction is coherent with the facts that the genome of CBPV is capped (20) and that the MTase-GTase is generally found at the N terminus of the viral replicase (54). Finally, we could find no homolog of RNA1 ORF2.

ORF3 of CBPV RNA2 is a putative virion membrane protein found in various insect and plant viruses. We could detect no homologs of ORF2 and ORF3 of RNA2 by using HHblits or HHpred, but these programs would not detect sequences deposited very recently, since they rely on databases that are not updated daily. We therefore used CSI-BLAST (an improved version of PSI-BLAST [30, 31]), which searches the up-to-date NCBI nr database. We present first the analysis of ORF3 and then that of ORF2.

TABLE 3 accession numbers of ORFs of chronic bee viruses and homologous ORFs

Genus and species or products and host species	Protein name	Accession no.
Chroparavirus[a]		
Chronic bee paralysis virus (CBPV)	ORF1 from RNA1	YP_001911136.1
	ORF2 from RNA2	YP_001911140.1
	ORF3 from RNA2 (SP24)	YP_001911141.1
Anopheline-associated C virus (AACV)	ORF1	Being submitted
	ORF2	Being submitted
Sinaivirus[b]		
Lake Sinai virus 1	ORF1	AEH26192.1
Lake Sinai virus 2	ORF1	AEH26187
Cilevirus		
Citrus Leprosis virus C	p24	ABC75826.1
Citrus leprosis virus cytoplasmic type 2	p24	AGE82891.1
Blunervirus,[c] *Blueberry necrotic ring blotch virus*	p24	YP_004901704.1
Higrevirus, Hibiscus green spot virus	p23	AER13452.1
Negevirus (Negev group)		
Negev virus	ORF2	AFI24682.1
	ORF3	AFI24674.1
Ngewotan virus	ORF2	AFY98073.1
Ngewotan virus	ORF3	AFY98074.1
Piura virus	ORF2	AFI24679.1
Piura virus	ORF3	AFI24680.1
Loreto virus	ORF3	AFI24694.1
Loreto virus	ORF2	AFI24692.1
Negevirus (Santana group)		
Santana virus	ORF2	AFI24676.1
Dezidougou	ORF2	AFI24670.1
Cellular proteins		
Drosophila melanogaster[d]	IP15837p	ABC86319.1
Glossina morsitans[d]	Hypothetical nonconserved protein	ADD20599.1

[a] Proposed genus (P. Blanchard, personal communication).
[b] Proposed genus (this study).
[c] Proposed genus (89).
[d] May be an endogenous viral protein (see text).

CSI-BLAST on ORF3 from CBPV RNA2 detected protein p24 of *Blueberry necrotic ring blotch virus* (56) with a marginal E value (E = 0.13) but also a weaker similarity (E = 15) to ORF3 from *Negev virus*, the type species of *Negevirus*, a new genus of viruses infecting insects (57). Aligning these proteins with CBPV ORF3 and resubmitting the alignment to CSI-BLAST gave significant hits to several other viral or (apparently) cellular proteins (Table 3 and Fig. 2), i.e., p24 of *Citrus leprosis virus C* (58–60), p23 of the related *Hibiscus green spot virus* (61), and proteins of insects. The latter probably corresponds to sequences of endogenous viruses integrated into *Drosophila* or *Glossita* genomes (62), since a
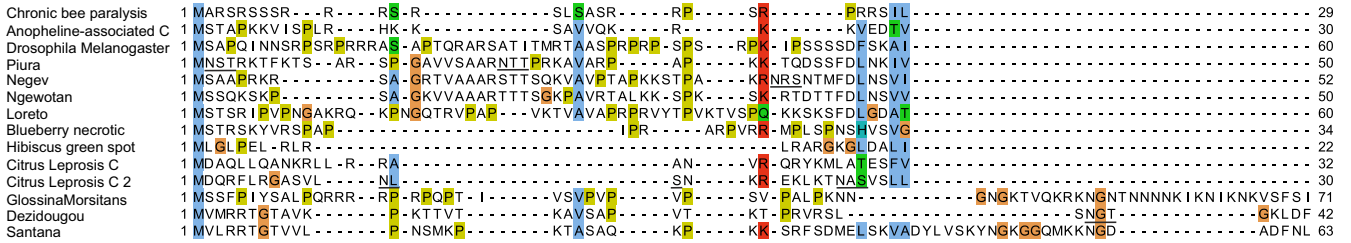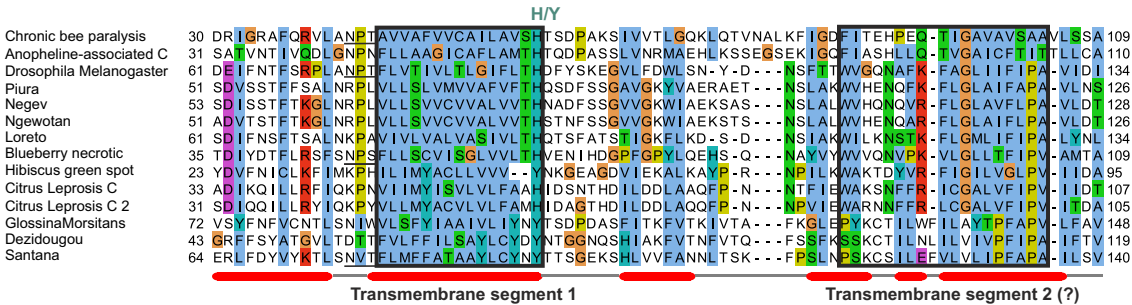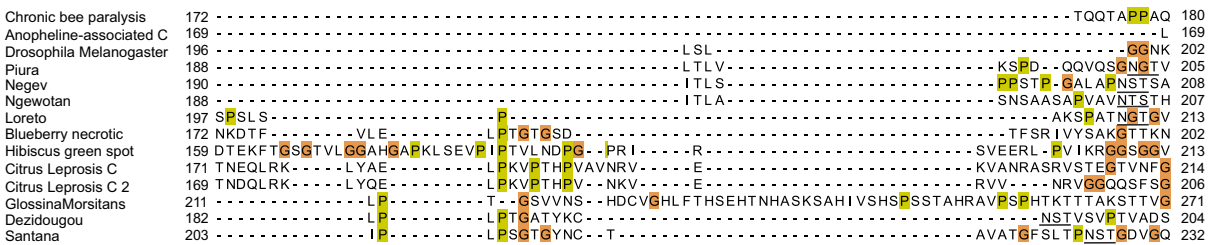
FIG 2 The SP24 family of virion membrane proteins of insect viruses. The boundaries of the predicted TM segments are approximate. We assumed that the topology of SP24 was conserved in all of the viruses, but in chroparaviruses (first two sequences), TM segment 2 is less hydrophobic and thus may be simply membrane associated, which would generate a different overall topology. The N-and C-terminal regions have no detectable sequence similarity and are presented only for information; whether they are homologous is unknown. Predicted N-glycosylation sites are indicated for N- and C-terminal regions only. Actual N-glycosylation can occur only if these regions are on the outside the virion, which we cannot reliably predict (see text).

tblastn search confirmed their presence in these genomes. To confirm the homology, we aligned ORF3 of CBPV with ORF3 of another recently discovered *Chroparavirus*, *Anopheline-associated C virus* (AACV; [90]) and compared their alignment to that of the other proteins described above. The two groups had highly significant similarity (HHalign E value of 6×10^{-14}) between the regions corresponding to aa 27 to 175 of CBPV ORF3 and aa 50 to 194 of *Negev virus* ORF3, confirming the homology. One cautionary note is that ORF3 and its homologs contain several TM segments (see below), which could give spurious hits in similarity searches if they were of a simple, anchor-like type ([63]). However, the program TMSOC ([45]) indicated that all of the TM segments of
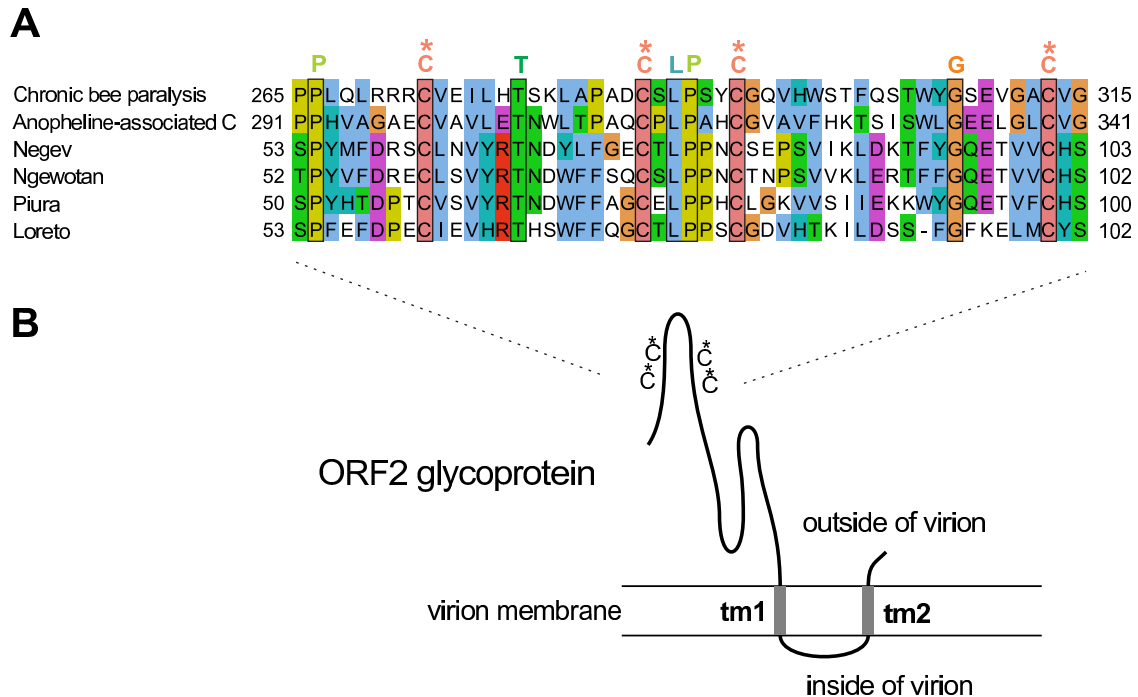
A

B

FIG 3 The viral ORF2 glycoproteins of chroparaviruses and negeviruses. (A) Alignment of the cysteine-rich region of ORF2 sequences of chroparaviruses and negeviruses. Conventions are the same as in Fig. 2. The conserved cysteines, predicted to form disulfide bridges, are indicated by an asterisk. (B) Predicted organization of ORF2. We make no claim to accurately predict disulfide connectivity. Other disulfide bridges are likely to occur elsewhere in ORF2 but are not conserved across taxons (see text). tm1 and tm2, TM segments 1 and 2, respectively.

SP24 are complex, i.e., carry significant evolutionary information, and thus that the similarity is not spurious.

CBPV is the only virus for which functional information about ORF3 is available; it is thought to be an SP of the virion (20, 64, 65). We named ORF3 and its homologs SP24, owing to their size (20 to 24 kDa). Figure 2 presents an alignment of SP24 proteins. They are composed of an N-terminal region of 20 to 70 aa with low sequence complexity that is enriched in basic residues and predicted to be disordered in most species, a central region containing several predicted TM segments, and a C terminus predicted to be disordered and rich in P, S, and T. Therefore, SP24 is probably an integral membrane protein of the virion.

To predict the topology of SP24, we first tried to predict its number of TM segments and then to predict which region was inside or outside the virion. We used two complementary approaches to assess the robustness of TM segment predictions (see Materials and Methods). Overall, the predictions (data not shown) suggested that SP24 may contain four TM segments (Fig. 2). However, it is difficult to accurately predict TM segments in a multipass membrane protein, even when using consensus approaches (for instance, in a recent study of the NS2A protein of *Dengue virus*, a region predicted by all of the predictors to span the membrane was, in fact, found experimentally to be only membrane associated [66]), and therefore, only experiments can settle the matter.

Even assuming that there were four TM segments, we could not reliably predict which parts of SP24 were internal or external to the virion, since Phobius (44) gave discordant results for different homologs. Nevertheless, we made two observations. (i) The loop before TM segment 4 contains positively charged residues

(R/K), and the very C terminus of SP24 contains potential N-glycosylation sites in negeviruses (Fig. 2). This suggests that the C terminus is on the outside of the virion. (ii) The N terminus of SP24 is basic, which would allow it to bind the viral RNA, if the N terminus were in the interior of the virion. Hypotheses i and ii are not simultaneously possible if there are four TM segments, since in that case both the N and C termini of SP24 would necessarily be on the same side of the membrane. However, it is possible that one of the four putative TM segments is, in fact, only membrane associated; a candidate would be segment 2, which is less hydrophobic, particularly in chroparaviruses (Fig. 2).

ORF2 is probably a virion glycoprotein, detected in several insect viruses. We next examined ORF2 of CBPV RNA2. CSI-BLAST reported a very weak hit (E = 691) to ORF2 of *Piura virus*, a *Negevirus*. This ORF2 has significant sequence similarity to the ORF2 of other negeviruses, except *Santana virus* and *Dezidougou virus*, which form a separate clade (we will call it the Santana group). A comparison of ORF2 of CBPV and AACV with ORF2 of negeviruses (with the Santana group excluded) confirmed that they were homologous (HHalign E value of 6×10^{-7} between the regions corresponding to aa 265 to 315 of CBPV ORF2 and aa 53 to 103 of *Negev virus* ORF3). We could not find other homologs of ORF2, even by in-depth examination of the genome of the insect viruses encoding SP24. Finally, we could not find homologs of the remaining ORF (ORF1) of CBPV RNA2.

The region of similarity between the ORF2 sequences of CBPV and negeviruses corresponds to 50 aa in the N-terminal or central part of ORF2, which contains nine conserved residues, including four cysteines (Fig. 3A), predicted to form disulfide bridges by Metaldetector (67). Outside of this region, we could detect no
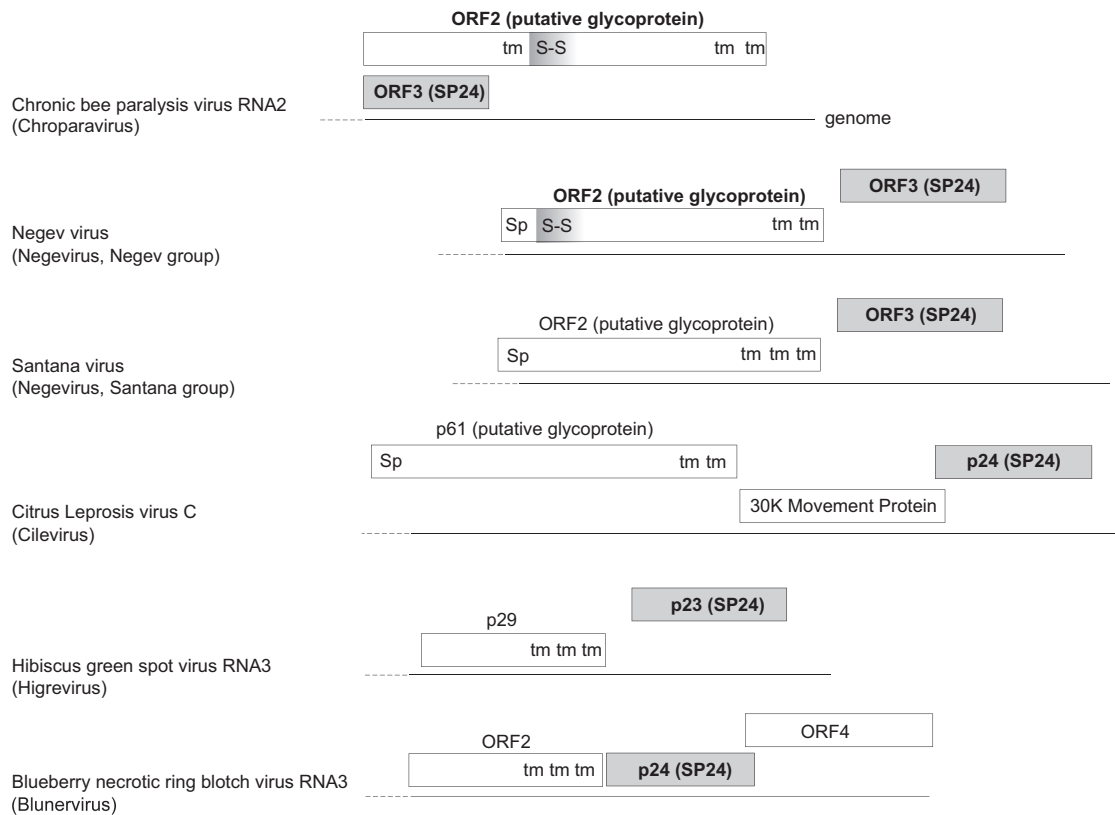
FIG 4 Comparison of SP24 and the ORF2 glycoprotein in different viruses. The genomic contexts of SP24 in insect viruses are shown. Genomes and proteins are approximately to scale. The names of proteins that have significant sequence similarity (and are thus demonstrably homologs) are in bold.

further sequence or secondary-structure similarity between ORF2 of CBPV and that of negeviruses. However, they are similar in organization, being composed of a predicted TM segment (or a signal peptide for negeviruses) 30 to 40 aa upstream of the conserved cysteine-rich region, followed by a variable region of about 200 aa, and two or three predicted C-terminal TM segments. In all of the viruses, ORF2 contains predicted N-glycosylation sites (not shown) and other cysteines conserved only in closely related species, which may form other disulfide bridges. Thus, ORF2 has all of the features of a virion glycoprotein. A speculative model of its topology is presented in Fig. 3B.

Figure 4 presents a summary of the organization of ORF2 and ORF3 and of their genomic context in different viruses. The long (200-aa) N-terminal extension of ORF2 (predicted to be disordered) that overlaps ORF3 only in chroparaviruses (Fig. 4, top) probably originated by overprinting (29, 68, 69) in their common ancestor. Since the predicted membrane segment of CBPV ORF2 occurs in the same position as the signal peptide of *Negevirus* ORF2 (compare the first two viruses in Fig. 4), it may be cleaved to give rise to the same topology.

Finally, we searched for putative glycoproteins analogous to ORF2 in other viruses that encode SP24. *Cilevirus* p61 and ORF2 of the Santana group of negeviruses have features similar to those

of ORF2 and may be their functional equivalent (Fig. 4). *Blunervirus* and *Higrevirus* also encode proteins with predicted C-terminal TM segments (Fig. 4), but they are considerably shorter than ORF2 or p61.

Putative roles of SP24 and ORF2 in light of experimental data. The viruses that encode an SP24 homolog have a variety of morphologies (Table 4), spherical (*Negevirus*), ellipsoidal (*Chroparavirus*), and short, bacilliform (*Cilevirus*, *Higrevirus*). Interestingly, these three groups of morphologies correlate with the phylogeny of the viral replicative enzymes. Since CBPV SP24 and ORF2 are thought to be SPs (20, 64, 65), they are probably the main membrane virion proteins. The fact that SP24 and ORF2 are both also encoded by negeviruses suggests that they may interact. How they would underlie a different morphology in these viruses is unclear; we note, however, that treatment of CBPV virions with acids or bases gave rise to a nearly spherical morphology similar to that of negeviruses (65). It is unclear whether chroparaviruses and cileviruses have enveloped virions. Chroparaviruses are regularly cited as being nonenveloped (e.g., reference 70), though to the best of our knowledge, this is not firmly proven. Cileviruses are also regularly cited as being nonenveloped (e.g., reference 71), but earlier reports described them as enveloped (72, 73). Obviously, if

TABLE 4 virion morphology of viruses encoding an SP24 matrix protein

Genus	Type species	Morphology	Host(s)	Reference(s)
Chroparavirus (proposed)	*Chronic bee paralysis virus*	Ellipsoidal, different populations (220 by 41, 54, or 64 nm), treatment with acid or base solutions results in more rounded, apparently empty shells (20–30 by 20–50 nm)	Insects (bees, mosquitoes)	65
Cilevirus	*Citrus Leprosis virus C*	Short, membrane-bound, enveloped, bacilliform particles (40–50 by 80–120 nm)	Plants (citrus), insects (erythrophyte mites)	60, 72, 73
Higrevirus	*Hibiscus green spot virus*	Short, bacilliform particles (30 by 50 nm)	Plant (hibiscus), insects (erythrophyte mites)	61
Blunervirus (proposed)	*Blueberry necrotic ring blotch virus*	Unknown	Plant (blueberry), probably transmitted by erythrophyte mites	
Negevirus	*Negev virus*	Spherical, enveloped particles (45–55 nm)	Insects (mosquitoes, phlebotomine sand flies)	57

our hypotheses regarding the role of SP24 are correct and if its function is conserved, then these viruses must be enveloped.

Most plant viruses encode a capsid protein that gives them a flexuous, icosahedral, or tubular morphology (74). Thus, SP24 is probably a new type of SP of plant and insect viruses. Its predicted topology is reminiscent of the *Coronaviridae* M protein, which contains three TM segments and a membrane-associated region (75) and forms spherical particles.

In conclusion, a combination of automated profile-profile methods and in-depth manual analysis allowed the detection of remote homologs and gave functional clues about most of the ORFs of a phylogenetically isolated virus.

DISCUSSION

Does the similarity among SP24, ORF2, and other viral proteins come from homology? Significant sequence similarity is widely considered evidence of homology, because there is no imperious constraint on protein sequences that would make convergent evolution likely (i.e., very different sequences can perform the same function or adopt the same structure [76]). However, this is not rigorously applicable to regions with low sequence complexity, since E values have been calibrated on globular proteins. In addition, convergent evolution should always be considered if there is no plausible mechanism by which two proteins could have evolved by common descent.

Several lines of argument strongly suggest that the similarity among SP24, ORF2, and other viral proteins is due to homologous descent. (i) The region similar among ORF2 proteins is most likely globular (Fig. 3), and thus, the caveat above does not apply. (ii) The TM segments of SP24 are predicted to carry significant evolutionary information, unlike simple "anchor" TM segments (45). (iii) The fact that both ORF2 and ORF3 of chroparaviruses have significant similarity to those of negeviruses considerably strengthens the homology hypothesis, since it seems difficult to envision why convergent evolution would have occurred twice. (iv) There is a plausible mechanism to explain homology, i.e., horizontal transfer between similar organisms (RNA viruses) that infect similar hosts (insects and plants).

Manual analyses incorporating contextual information are an indispensable complement of automated searches on viral proteins. Our study shows that making use of the biological context and examining search results far beyond the threshold of statistical significance allows the detection of homologs even in phylogenetically isolated viruses. Such approaches are successful for

three main reasons. (i) Gene and domain order are often conserved in viruses, at least within the same family. For instance, the order of domains is mostly conserved in the *Flaviviridae* polyprotein (77), as is the order of genes in the *Coronaviridae* genome (78). (ii) Because RNA viruses have very few genes, a weak hit to a protein from a related virus conveys strong information (contrary to, for instance, a weak hit from a human protein to another). (iii) Though it remains difficult to detect very distant homologs, it has recently become easier to validate candidate homologs by pairwise profile-profile comparison (32).

The results presented here suggest that the "limits of homology detection" (79) are far from having been reached for viral proteins, despite their fast evolution. In particular, proteins that have strictly conserved residues owing to catalytic activity (such as the presumed MTase-GTase of CBPV) are expected to retain detectable sequence similarity over long distances.

A cautionary note: checks to perform when doing manual analyses. Some well-established checks that will avoid many false positives in sequence similarity searches include (17, 39, 68, 80) (i) excluding regions with low sequence complexity, coiled coils, disordered regions (by using, for instance, ANNIE [37]), and simple TM segments (45); (ii) comparing the lengths of the query and the hit (viral proteins rarely change dramatically in length, unlike eukaryotic proteins, except in specific cases like polyproteins), their functions, and their secondary structures; (iii) using a relatively stringent cutoff for the validation step (e.g., HHalign E values of $<10^{-5}$) of candidate homologs identified in the detection step; and (iv) waiting for the sequence of new, divergent viruses if there remains a doubt over a prediction; they will often settle the matter.

Limitations of our study and comparison with previous studies. A limitation of our work is that we only considered viral homologs of the proteins of the ORFan data set, because cellular homologs often correspond to horizontal transfer (81), for instance, isolated cases of endogenous viruses (82, 83). The limitations of BLAST have already been noted in archaeal viruses (84) and mimiviruses (85).

Implications for the study of ORFans. In conclusion, we suspect that our results are applicable to all organisms and not only viruses. BLAST can reliably identify ORFans in mammalian genomes (86) owing to their low rate of evolution. However, as the phylogenetic distance increases (for instance, when comparing vertebrates and invertebrates), homologs are expected to be increasingly difficult to detect. Tautz et al. wrote in a recent review "we are still missing a systematic study that uses PSI-BLAST-based

searches to provide a reliable estimate of orphan gene affiliation to the known protein folds" (9). We agree but suggest that more powerful tools should also be used, such as profile-profile comparison or fold recognition.

ACKNOWLEDGMENTS

REFERENCES

1. **Jaroszewski L, Li Z, Cai XH, Weber C, Godzik A.** 2011. FFAS server: novel features and applications. Nucleic Acids Res. **39**:W38–W44.
2. **Boyer M, Gimenez G, Suzan-Monti M, Raoult D.** 2010. Classification and determination of possible origins of ORFans through analysis of nucleocytoplasmic large DNA viruses. Intervirology **53**:310–320.
3. **Yin Y, Fischer D.** 2008. Identification and investigation of ORFans in the viral world. BMC Genomics **9**:24. http://dx.doi.org/10.1186/1471-2164-9-24.
4. **Fancello L, Raoult D, Desnues C.** 2012. Computational tools for viral metagenomics and their application in clinical research. Virology **434**:162–174.
5. **Fischer D, Eisenberg D.** 1999. Finding families for genomic ORFans. Bioinformatics **15**:759–762.
6. **Wilson GA, Bertrand N, Patel Y, Hughes JB, Feil EJ, Field D.** 2005. Orphans as taxonomically restricted and ecologically important genes. Microbiology **151**:2499–2501.
7. **Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TCG.** 2009. More than just orphans: are taxonomically-restricted genes important in evolution? Trends Genet. **25**:404–413.
8. **Sanjuán R, Nebot MR, Chirico N, Mansky LM, Belshaw R.** 2010. Viral mutation rates. J. Virol. **84**:9733–9748.
9. **Tautz D, Domazet-Loso T.** 2011. The evolutionary origin of orphan genes. Nat. Rev. Genet. **12**:692–702.
10. **Wissler L, Gadau J, Simola DF, Helmkampf M, Bornberg-Bauer E.** 2013. Mechanisms and dynamics of orphan gene emergence in insect genomes. Genome Biol. Evol. **5**:439–455.
11. **Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**:3389–3402.
12. **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ.** 1990. Basic local alignment search tool. J. Mol. Biol. **215**:403–410.
13. **Finn RD, Clements J, Eddy SR.** 2011. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. **39**:W29–W37.
14. **Hildebrand A, Remmert M, Biegert A, Söding J.** 2009. Fast and accurate automatic structure prediction with HHpred. Proteins 77(Suppl 9):128–132.
15. **Remmert M, Biegert A, Hauser A, Söding J.** 2012. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat. Methods **9**:173–175.
16. **Brandt BW, Heringa J.** 2009. webPRC: the Profile Comparer for alignment-based searching of public domain databases. Nucleic Acids Res. **37**:W48–W52.
17. **Söding J, Remmert M.** 2011. Protein sequence comparison and fold recognition: progress and good-practice benchmarking. Curr. Opin. Struct. Biol. **21**:404–411.
18. **Dunbrack RL.** 2006. Sequence comparison and protein structure prediction. Curr. Opin. Struct. Biol. **16**:374–384.
19. **Karlin D, Belshaw R.** 2012. Detecting remote sequence homology in disordered proteins: discovery of conserved motifs in the N-termini of Mononegavirales phosphoproteins. PLoS One **7**:e31719. http://dx.doi.org/10.1371/journal.pone.0031719.
20. **Olivier V, Blanchard P, Chaouch S, Lallemand P, Schurr F, Celle O, Dubois E, Tordo N, Thiery R, Houlgatte R, Ribiere M.** 2008. Molecular characterisation and phylogenetic analysis of chronic bee paralysis virus, a honey bee virus. Virus Res. **132**:59–68.
21. **Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer EL, Eddy SR, Bateman A, Finn RD.** 2012. The Pfam protein families database. Nucleic Acids Res. **40**:D290–D301.
22. **UniProt Consortium.** 2012. Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res. **40**:D71–D75.
23. **Schneider G, Sherman W, Kuchibhatla D, Ooi H, Sirota F, Maurer-Stroh S, Eisenhaber B, Eisenhaber F.** 2012. Protein-sequence-structure-function-network links discovered with the ANNOTATOR software suite: application to ELYS/Mel-28, p 111–144. *In* Trajanoski Z (ed), Computational medicine: tools and challenges. Springer, New York, NY.
24. **Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer EL, Bateman A.** 2006. Pfam: clans, web tools and services. Nucleic Acids Res. **34**:D247–D251.
25. **Aravind L.** 2000. Guilt by association: contextual information in genome analysis. Genome Res. **10**:1074–1077.
26. **Ochoa A, Llinas M, Singh M.** 2011. Using context to improve protein domain identification. BMC Bioinformatics **12**:90. http://dx.doi.org/10.1186/1471-2105-12-90.
27. **Terrapon N, Gascuel O, Marechal E, Breehelin L.** 2009. Detection of new protein domains using co-occurrence: application to Plasmodium falciparum. Bioinformatics **25**:3077–3083.
28. **Huynen M, Snel B, Lathe W, III, Bork P.** 2000. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. Genome Res. **10**:1204–1210.
29. **Pavesi A, Magiorkinis G, Karlin DG.** 2013. Viral proteins originated de novo by overprinting can be identified by codon usage: application to the "gene nursery" of deltaretroviruses. PLoS Comput. Biol. **9**:e1003162. http://dx.doi.org/10.1371/journal.pcbi.1003162.
30. **Biegert A, Söding J.** 2009. Sequence context-specific profiles for homology searching. Proc. Natl. Acad. Sci. U. S. A. **106**:3770–3775.
31. **Angermüller C, Biegert A, Söding J.** 2012. Discriminative modelling of context-specific amino acid substitution probabilities. Bioinformatics **28**:3240–3247.
32. **Biegert A, Mayer C, Remmert M, Söding J, Lupas AN.** 2006. The MPI Bioinformatics Toolkit for protein sequence analysis. Nucleic Acids Res. **34**:W335–W339.
33. **Schneider G, Neuberger G, Wildpaner M, Tian S, Berezovsky I, Eisenhaber F.** 2006. Application of a sensitive collection heuristic for very large protein families: evolutionary relationship between adipose triglyceride lipase (ATGL) and classic mammalian lipases. BMC Bioinformatics **7**:164. http://dx.doi.org/10.1186/1471-2105-7-164.
34. **Kaushik S, Mutt E, Chellappan A, Sankaran S, Srinivasan N, Sowdhamini R.** 2013. Improved detection of remote homologues using Cascade PSI-BLAST: influence of neighbouring protein families on sequence coverage. PLoS One **8**:e56449. http://dx.doi.org/10.1371/journal.pone.0056449.
35. **Sandhya S, Chakrabarti S, Abhinandan KR, Sowdhamini R, Srinivasan N.** 2005. Assessment of a rigorous transitive profile based search method to detect remotely similar proteins. J. Biomol. Struct. Dyn. **23**:283–298.
36. **Pei JM, Kim BH, Tang M, Grishin NV.** 2007. PROMALS web server for accurate multiple protein sequence alignments. Nucleic Acids Res. **35**:W649–W652.
37. **Ooi HS, Kwo CY, Wildpaner M, Sirota FL, Eisenhaber B, Maurer-Stroh S, Wong WC, Schleiffer A, Eisenhaber F, Schneider G.** 2009. ANNIE: integrated de novo protein sequence annotation. Nucleic Acids Res. **37**:W435–W440.
38. **Ishida T, Kinoshita K.** 2008. Prediction of disordered regions in proteins based on the meta approach. Bioinformatics **24**:1344–1348.
39. **Ferron F, Longhi S, Canard B, Karlin D.** 2006. A practical overview of protein disorder prediction methods. Proteins **65**:1–14.
40. **Vacic V, Uversky VN, Dunker AK, Lonardi S.** 2007. Composition Profiler: a tool for discovery and visualization of amino acid composition differences. BMC Bioinformatics **8**:211. http://dx.doi.org/10.1186/1471-2105-8-211.
41. **Wu S, Zhang Y.** 2007. LOMETS: a local meta-threading-server for protein structure prediction. Nucleic Acids Res. **35**:3375–3382.
42. **Chang JM, Di Tommaso P, Taly JF, Notredame C.** 2012. Accurate multiple-sequence alignment of transmembrane proteins with PSI-Coffee. BMC Bioinformatics **13**(Suppl 4):S1. http://dx.doi.org/10.1186/1471-2105-13-S4-S1.
43. **Tusnády GE, Simon I.** 2001. The HMMTOP transmembrane topology prediction server. Bioinformatics **17**:849–850.
44. **Käll L, Krogh A, Sonnhammer ELL.** 2007. Advantages of combined

45. **Wong WC, Maurer-Stroh S, Schneider G, Eisenhaber F.** 2012. Transmembrane helix: simple or complex. Nucleic Acids Res. **40:**W370–W375.

46. **Lauber C, Gorbalenya AE.** 2012. Toward genetics-based virus taxonomy: comparative analysis of a genetics-based classification and the taxonomy of picornaviruses. J. Virol. **86:**3905–3915.

47. **Lauber C, Gorbalenya AE.** 2012. Partitioning the genetic diversity of a virus family: approach and evaluation through a case study of picornaviruses. J. Virol. **86:**3890–3904.

48. **Yu C, Hernandez T, Zheng H, Yau SC, Huang HH, He RL, Yang J, Yau SS.** 2013. Real time classification of viruses in 12 dimensions. PLoS One **8:**e64328. http://dx.doi.org/10.1371/journal.pone.0064328.

49. **Dolja VV, Koonin EV.** 1991. Phylogeny of capsid proteins of small icosahedral RNA plant-viruses. J. Gen. Virol. **72:**1481–1486.

50. **Kreiah S, Strunk G, Cooper JI.** 1994. Sequence-analysis and location of capsid proteins within Rna-2 of strawberry latent ringspot virus. J. Gen. Virol. **75:**2527–2532.

51. **Revill PA, Davidson AD, Wright PJ.** 1994. The nucleotide sequence and genome organization of mushroom bacilliform virus: a single-stranded RNA virus of Agaricus bisporus (Lange) Imbach. Virology **202:**904–911.

52. **Li L, Lok SM, Yu IM, Zhang Y, Kuhn RJ, Chen J, Rossmann MG.** 2008. The flavivirus precursor membrane-envelope protein complex: structure and maturation. Science **319:**1830–1834.

53. **Dolja VV, Kreuze JF, Valkonen JP.** 2006. Comparative and functional genomics of closteroviruses. Virus Res. **117:**38–51.

54. **Rozanov MN, Koonin EV, Gorbalenya AE.** 1992. Conservation of the putative methyltransferase domain: a hallmark of the 'Sindbis-like' supergroup of positive-strand RNA viruses. J. Gen. Virol. **73(Pt 8):**2129–2134.

55. **Runckel C, Flenniken ML, Engel JC, Ruby JG, Ganem D, Andino R, DeRisi JL.** 2011. Temporal analysis of the honey bee microbiome reveals four novel viruses and seasonal prevalence of known viruses, Nosema, and Crithidia. PLoS One **6:**e20656. http://dx.doi.org/10.1371/journal.pone.0020656.

56. **Martin RR, Quito-Avila DF, Cline WO, Harmon PF, Brannen PM.** 2011. Blueberry necrotic ring blotch, a new blueberry disease caused by a virus. Phytopathology **101:**S115–S115.

57. **Vasilakis N, Forrester NL, Palacios G, Nasar F, Savji N, Rossi SL, Guzman H, Wood TG, Popov V, Gorchakov R, Gonzalez AV, Haddow AD, Watts DM, da Rosa AP, Weaver SC, Lipkin WI, Tesh RB.** 2013. Negevirus: a proposed new taxon of insect-specific viruses with wide geographic distribution. J. Virol. **87:**2475–2488.

58. **Locali-Fabris EC, Freitas-Astua J, Souza AA, Takita MA, Astua-Monge G, Antonioli-Luizon R, Rodrigues V, Targon ML, Machado MA.** 2006. Complete nucleotide sequence, genomic organization and phylogenetic analysis of citrus leprosis virus cytoplasmic type. J. Gen. Virol. **87:**2721–2729.

59. **Pascon RC, Kitajima JP, Breton MC, Assumpcao L, Greggio C, Zanca AS, Okura VK, Alegria MC, Camargo ME, Silva GG, Cardozo JC, Vallim MA, Franco SF, Silva VH, Jordao H, Jr, Oliveira F, Giachetto PF, Ferrari F, Aguilar-Vildoso CI, Franchiscini FJ, Silva JM, Arruda P, Ferro JA, Reinach F, da Silva AC.** 2006. The complete nucleotide sequence and genomic organization of citrus leprosis associated virus, cytoplasmatic type (CiLV-C). Virus Genes **32:**289–298.

60. **Roy A, Choudhary N, Leon G, Shao J, Govindarajulu A, Achor D, Wei G, Picton DD, Levy L, Nakhla MK, Hartung JS, Brlansky RH.** 2013. A novel virus of the genus Cilevirus causing symptoms similar to citrus leprosis. Phytopathology **103:**488–500.

61. **Melzer MJ, Sether DM, Borth WB, Hu JS.** 2012. Characterization of a virus infecting Citrus volkameriana with citrus leprosis-like symptoms. Phytopathology **102:**122–127.

62. **Cui J, Holmes EC.** 2012. Endogenous RNA viruses of plants in insect genomes. Virology **427:**77–79.

63. **Wong WC, Maurer-Stroh S, Eisenhaber F.** 2011. Not all transmembrane helices are born equal: towards the extension of the sequence homology concept to membrane proteins. Biol. Direct **6:**57. http://dx.doi.org/10.1186/1745-6150-6-57.

64. **Ribiere M, Faucon JP, Pepin M.** 2000. Detection of chronic honey bee (Apis mellifera L.) paralysis virus infection: application to a field survey. Apidologie **31:**567–577.

65. **Bailey L, Gibbs AJ, Woods RD.** 1968. The purification and properties of chronic bee paralysis virus. J. Gen. Virol. **2:**251–260.

66. **Xie X, Gayen S, Kang C, Yuan Z, Shi PY.** 2013. Membrane topology and function of dengue virus NS2A protein. J. Virol. **87:**4609–4622.

67. **Passerini A, Lippi M, Frasconi P.** 2011. MetalDetector v2.0: predicting the geometry of metal binding sites from protein sequence. Nucleic Acids Res. **39:**W288–W292.

68. **Rancurel C, Khosravi M, Dunker AK, Romero PR, Karlin D.** 2009. Overlapping genes produce proteins with unusual sequence properties and offer insight into de novo protein creation. J. Virol. **83:**10719–10736.

69. **Keese PK, Gibbs A.** 1992. Origins of genes: "big bang" or continuous creation? Proc. Natl. Acad. Sci. U. S. A. **89:**9489–9493.

70. **Chevin A, Schurr F, Blanchard P, Thiery R, Ribiere M.** 2012. Experimental infection of the honeybee (Apis mellifera L.) with the chronic bee paralysis virus (CBPV): infectivity of naked CBPV RNAs. Virus Res. **167:**173–178.

71. **Locali-Fabris EC, Freitas-Astúa J, Machado MA.** 2012. Genus Cilevirus, p 1169–1172. In King AM, Lefkowitz E, Adams MJ, Carstens EB (ed), Virus taxonomy—ninth report of the International Committee on Taxonomy of Viruses. Elsevier Academic Press, New York, NY.

72. **Collariccio A, Lovisolo O, Chagas CM, Galleti SR, Rossetti V, Kitajima EW.** 1995. Mechanical transmission and ultrastructural aspects of citrus leprosis disease. Fitopatol. Bras. **20:**208–213.

73. **Kitajima EW, Rosillo MA, Portillo MM, Müller GW, Costa AS.** 1974. Electron microscopy of leaf tissues of orange trees infected by "lepra explosiva" in Argentina. Fitopatologia (Lima) **9:**55–56.

74. **Masson P, Hulo C, De Castro E, Bitter H, Gruenbaum L, Essioux L, Bougueleret L, Xenarios I, Le Mercier P.** 2013. ViralZone: recent updates to the virus knowledge resource. Nucleic Acids Res. **41:**D579–D583.

75. **Voss D, Pfefferle S, Drosten C, Stevermann L, Traggiai E, Lanzavecchia A, Becker S.** 2009. Studies on membrane topology, N-glycosylation and functionality of SARS-CoV membrane protein. Virol. J. **6:**79. http://dx.doi.org/10.1186/1743-422X-6-79.

76. **Theobald DL, Miller C.** 2010. Membrane transport proteins: surprises in structural sameness. Nat. Struct. Mol. Biol. **17:**2–3.

77. **Simmonds P, Becher P, Collett MS, Gould EA, Heinz FX, Meyers G, Monath T, Pletnev A, Rice CM, Stiasny K, Thiel H-J, Weiner A, Bukh J.** 2011. Flaviviridae, p 1003–1020. In King AM, Lefkowitz E, Adams MJ, Carstens EB (ed), Virus taxonomy—ninth report of the International Committee on Taxonomy of Viruses. Elsevier Academic Press, New York, NY.

78. **Brian DA, Baric RS.** 2005. Coronavirus genome structure and replication. Curr. Top. Microbiol. Immunol. **287:**1–30.

79. **Pearson WR, Sierk ML.** 2005. The limits of protein sequence comparison? Curr. Opin. Struct. Biol. **15:**254–260.

80. **Iyer LM, Aravind L, Bork P, Hofmann K, Mushegian AR, Zhulin IB, Koonin EV.** 2001. Quoderat demonstrandum? The mystery of experimental validation of apparently erroneous computational analyses of protein sequences. Genome Biol. **2:**RESEARCH0051.

81. **Koonin EV, Wolf YI.** 2012. Evolution of microbes and viruses: a paradigm shift in evolutionary biology? Front. Cell Infect. Microbiol. **2:**119. http://dx.doi.org/10.3389/fcimb.2012.00119.

82. **Aswad A, Katzourakis A.** 2012. Paleovirology and virally derived immunity. Trends Ecol. Evol. **27:**627–636.

83. **Patel MR, Emerman M, Malik HS.** 2011. Paleovirology—ghosts and gifts of viruses past. Curr. Opin. Virol. **1:**304–309.

84. **Prangishvili D, Garrett RA, Koonin EV.** 2006. Evolutionary genomics of archaeal viruses: unique viral genomes in the third domain of life. Virus Res. **117:**52–67.

85. **Saini HK, Fischer D.** 2007. Structural and functional insights into mimivirus ORFans. BMC Genomics **8:**115. http://dx.doi.org/10.1186/1471-2164-8-115.

86. **Albà MM, Castresana J.** 2007. On homology searches by protein Blast and the characterization of the age of genes. BMC Evol. Biol. **7:**53. http://dx.doi.org/10.1186/1471-2148-7-53.

87. **Adams MJ, Zerbini FM, French R, Rabenstein F, Stenger DC, Valkonen JPT.** 2012. Potyviridae, p 1069–1089. In King AMQ, Lefkowitz E, Adams MJ, Carstens EB, (ed), Virus taxonomy—ninth report of the International Committee on Taxonomy of Viruses. Elsevier Academic Press, New York, NY.

88. **Olspert A, Arike L, Peil L, Truve E.** 2011. Sobemovirus RNA linked to VPg over a threonine residue. FEBS Lett. **585:**2979–2985.

89. **Quito-Avila DF, Brannen PM, Cline WO, Harmon PF, Martin RR.** 2013. Genetic characterization of blueberry necrotic ring blotch virus, a novel RNA virus with unique genetic features. J. Gen. Virol. **94:**1426–1434.

90. **Cook S, Chung BYW, Bass D, Moureau G, Tang S, McAlister E, Culverwell CL, Glucksman E, Wang H, Brown TDK, Gould EA, Harbach RE, De Lamballerie X, Firth AE.** Novel virus discovery and genome reconstruction from field RNA samples reveals highly divergent viruses in dipteran hosts. PLoS One, in press.