

Poxvirus Orthologous Clusters: toward Defining the Minimum Essential Poxvirus Genome

Chris Upton,* Stephanie Slack, Arwen L. Hunter, Angelika Ehlers, and Rachel L. Roper

Department of Biochemistry and Microbiology, University of Victoria, Victoria, British Columbia, Canada

Received 17 December 2002/Accepted 26 March 2003

Increasingly complex bioinformatic analysis is necessitated by the plethora of sequence information currently available. A total of 21 poxvirus genomes have now been completely sequenced and annotated, and many more genomes will be available in the next few years. First, we describe the creation of a database of continuously corrected and updated genome sequences and an easy-to-use and extremely powerful suite of software tools for the analysis of genomes, genes, and proteins. These tools are available free to all researchers and, in most cases, alleviate the need for using multiple Internet sites for analysis. Further, we describe the use of these programs to identify conserved families of genes (poxvirus orthologous clusters) and have named the software suite POCs, which is available at www.poxvirus.org. Using POCs, we have identified a set of 49 absolutely conserved gene families—those which are conserved between the highly diverged families of insect-infecting entomopoxviruses and vertebrate-infecting chordopoxviruses. An additional set of 41 gene families conserved in chordopoxviruses was also identified. Thus, 90 genes are completely conserved in chordopoxviruses and comprise the minimum essential genome, and these will make excellent drug, antibody, vaccine, and detection targets. Finally, we describe the use of these tools to identify necessary annotation and sequencing updates in poxvirus genomes. For example, using POCs, we identified 19 genes that were widely conserved in poxviruses but missing from the vaccinia virus strain Tian Tan 1998 GenBank file. We have reannotated and resequenced fragments of this genome and verified that these genes are conserved in Tian Tan. The results for poxvirus genes and genomes are discussed in light of evolutionary processes.

Poxviruses are large double-stranded DNA viruses with genomes ranging from 130 to 380 kbp (40). They are highly successful pathogens known to infect a tremendous variety of animals, including insects, reptiles, birds, and over 30 mammals. They can be spread by aerosol, direct contact, and insects. The most infamous member of the family *Poxviridae* is *Variola virus*, the causative agent of smallpox, which caused millions of deaths before its eradication from the natural environment. Today, there still remain multiple poxvirus threats to humans, including the use of smallpox as a bioterrorism weapon in a now largely unvaccinated population. Furthermore, there exist related animal poxviruses, including monkeypox virus, tanapoxvirus, Yaba-like disease virus, and cowpox virus, which can infect humans and cause morbidity (16, 19, 23–25, 33, 39). Monkey poxvirus is of particular concern because it causes high human mortality and can spread from human to human (31). Additionally, there is an emerging poxvirus infection, cantagalo, in Brazil, which has apparently evolved from the locally used smallpox vaccine strain (vaccinia virus [VV]) (17). Another member of the poxvirus family, molluscum contagiosum virus (MCV), occurs commonly in humans (39% of a population over 50 years old tested seropositive [32]) but causes significant disease only in immunocompromised individuals and rarely in children. MCV may cause long-standing disfiguring infections, mostly on the skin of the face (66). As the number of transplant recipients and human immunodeficiency

virus-infected individuals has risen, morbidity due to MCV infections has also increased.

The existence of multiple poxviruses that can infect humans raises the possibility of the evolution of a new smallpox-like virus through host gene acquisitions or interviral recombination events. If the new virus retained the ability to infect animals, however, then its eradication would be unlikely due to the natural animal reservoir of infection. In addition to the importance of poxvirus pathogens, multiple attenuated poxviruses are being used as vectors for clinical purposes, including cancer treatment, vaccines for human immunodeficiency virus, cytomegalovirus, and measles virus, and a successful rabies virus vaccine for feral animals (1, 9, 10, 29, 36, 42, 55).

The present study includes an analysis of 21 poxvirus genomes that have been completely sequenced (Table 1)—19 members of the subfamily *Chordopoxvirinae* and 2 members of the subfamily *Entomopoxvirinae*. With this wealth of sequence information, it is possible to move from the laborious and slow techniques of single-gene functional analysis to a global comprehension of poxvirus genes. The development of new bioinformatic tools is required for these large-scale analyses. As part of the Poxvirus Bioinformatics Resource (PBR; www.poxvirus.org) funded by the Canadian Protein Engineering Network Centre of Excellence and the U.S. National Institutes of Health, our group and collaborators have developed the Viral Genome Organizer (62), the Virus Genome Database (27), and Poxvirus Orthologous Clusters (POCs) (21). POCs, which is the successor of the Virus Genome Database, is an MySQL database containing sequenced poxvirus genomes and a software suite, with graphics, designed for users to search for and analyze poxvirus genes, promoters, and gene or protein homologs (orthologs) in related viruses. In addition, POCs en-

* Corresponding author. Mailing address: Department of Biochemistry and Microbiology, University of Victoria, Ring Rd., Petch Bldg., Rm. 150, Victoria, British Columbia V8P 5C2, Canada. Phone: (250) 721-6507. Fax: (250) 721-8855. E-mail: cupton@uvic.ca.

TABLE 1. Poxvirus genomes

Genome	Abbreviation	GenBank accession no.	Reference or source
Chordopoxviruses			
Vaccinia virus (Copenhagen)	VV-Cop	M35027	26
Vaccinia virus (modified vaccinia virus Ankara)	VV-MVA	U94848	8
Vaccinia virus (Tian Tan)	VV-Tan	AF095689	Unpublished
Variola major virus (Bangladesh-1975)	VAR-Bang	L22579	37
Variola major virus (India-1967)	VAR-Ind	NC_001611	49, 52
Variola minor virus (Garcia-1966)	VAR-Gar	NC_000900	51
Fowlpox virus (virulent challenge virus)	FPV-V	AF198100	3
Molluscum contagiosum virus (subtype 1)	MCV	U60315	47
Myxoma virus (Lausanne)	MYX	NC_001132.2	15
Shope rabbit fibroma virus (Kasza)	SFV	AF170722	63
Lumpy skin disease virus (Neethling 2490)	LSDV	NC_003027	59
Monkeypox virus (Zaire)	MPV	AF380138	50
Yaba-like disease virus (Smith)	YLDV	NC_002642	35
Swinepox virus (Nebraska 17077-99)	SPV	AF410153	4
Cowpox virus (Brighton Red)	CPV	AF482758	Unpublished
Camelpox virus (Kazakhstan M-96)	CMLV	AF438165	5
Ectromelia virus (Moscow)	EVM	AF012825	Unpublished
Rabbitpox virus (Utrecht)	RPV	To be submitted	Unpublished
Sheepox virus (Turkey; TU-V02127)	ShPV	NC_004002	60
Entomopoxviruses			
<i>Melanoplus sanguinipes</i> (Tucson)	MsEPV	NC_001993	2
<i>Amsacta moorei</i> (Moyer)	AmEPV	NC_002520	11

ables users to analyze the likelihood that an open reading frame (ORF) encodes an expressed protein and allows searches for unique viral genes, genes missing from particular viral genomes, or genes present in a user-defined subset of viruses. These tools are available to all researchers at no cost. We have developed and used these computer applications to group genes into families and have identified genes that are most highly conserved in the family *Poxviridae*. Thus, this analysis represents the first step toward identifying the minimum poxvirus genome and also identifies less-well-conserved genes which may be involved in host-specific virulence. Finally, we have used data analysis by POCs to predict necessary sequencing updates in the VV strain Tian Tan (VV-Tan) GenBank file and have confirmed the hypothesis by DNA sequence analysis.

MATERIALS AND METHODS

POCs software and database. POCs is a JAVA client-server application which accesses a curated sequence query language (SQL) database containing all complete poxvirus genomes (see Ehlers et al. [21] for a technical description of the software). MySQL was chosen because it is an open source (www.mysql.com), and versions are available for many different computer platforms. The database and JAVA server reside on a computer at PBR (www.poxvirus.org). The user downloads the free client software from PBR, and it runs on the user's computer (Windows, Macintosh, and UNIX platforms), connecting to PBR via the Internet. POCs has been developed as a JAVA Web Start application (<http://java.sun.com/products/javawebstart/>) because each time the client software is run, it checks for new versions on the PBR website and automatically upgrades the software without intervention by the user. POCs automatically groups orthologous genes into families based on BLASTP scores and allows assessment by a human database curator. Most importantly, the software has a user-friendly interface permitting complex SQL queries to retrieve interesting groups of DNA and protein sequences as well as gene families for subsequent interrogation by a variety of integrated tools: BLASTP, BLASTX, TBLASTN, PSIBLAST, Jalview (a JAVA graphic multiple-sequence-alignment viewer and editor available at <http://www.compbio.dundee.ac.uk/>), JDotter (a JAVA-based user interface in POCs for viewing Dotter results [54]), Laj (local alignment in JAVA [46]), and NAP (nucleotide-to-amino-acid alignment [30]).

Sequencing of VV-Tan genes. VV-Tan DNA was kindly provided by Joe Esposito, Centers for Disease Control and Prevention, Atlanta, Ga. (CDC), and was isolated from a passaged VV-Tan strain generously provided to CDC by a smallpox vaccine producer in People's Republic of China. Each VV-Tan gene to be sequenced was first PCR amplified with the following reaction mixture to make a final volume of 100 μ l: 34 ng of VV-Tan DNA, 0.1 μ M forward and reverse primers, 0.1 mM deoxynucleoside triphosphates (GibcoBRL), PCR buffer, *Taq* polymerase, and double-distilled H₂O. The reaction mixture was heated to 94°C for 2 min and subjected to 30 cycles of 94°C for 1 min, 45°C for 1 min, and 72°C for 2 min. PCR products were purified by using a Qiagen QIAquick PCR kit and sequenced by using a LI-COR, Inc., 4200 global edition DNA sequencer with nested primers (see below). Sequences were assembled and analyzed (minimum redundancy of two) by using Dear-Staden software [18].

Nucleotide sequence accession numbers. Gene and primer sequences are available under GenBank accession numbers AY188507, AY188508, AY188509, AY188510, AY188511, AY188512, AY188513, AY188514, and AY188515.

RESULTS

Database and software. The first step in the development of POCs was the insertion of all of the available 21 poxvirus genomes (Table 1), with a total of 4,197 predicted genes, into the MySQL database. Since the database imports GenBank files, it contains a number of ORFs that are likely nonfunctional; the number of such ORFs per virus varies depending on the stringency applied to ORF prediction by the original authors. This database is continuously updated and corrected, in response to new publications and feedback from poxvirus researchers. For example, annotation for the virulence gene corresponding to VV strain Copenhagen (VV-Cop) A14.5L (12) has been added to seven viruses in POCs, but this gene remains unannotated for these viruses in the GenBank files. Similarly, the VV-Cop G5.5R RNA polymerase 7-kDa subunit (7) and the VV-Cop A2.5L redox protein (48) are listed in the POCs database but not in the GenBank files describing the genomic sequence. We have also extensively updated and annotated the POCs VV-Tan files, which were deposited in GenBank but

never published. The development of this database gives poxvirus researchers a single source for updated genomic information at www.poxvirus.org. Should users of the database wish to customize their own databases on a local server, they can request authorized access and install the administrative program. This system allows users to add new genomes or genes to their own databases and group genes into families according to their own guidelines.

We have designed a suite of tools in which the user interface with the database is specifically designed for molecular virologists with little computer and no SQL experience. POCs enables users to quickly perform a great variety of complex queries for genes or proteins from a single virus, subsets of viruses, or all viruses. The user can perform searches for sets of genes based on the nucleotide content or sequence, pI, gene size, and/or codon usage (alone or in any combination) by using the "Sequence Query" window. Data on individual genes and proteins (e.g., predicted molecular weight, pI, and amino acid and nucleotide contents and sequences) from any virus can also be viewed by using this window. Protein hydrophobicity plotting is available by three methods under the "Analysis" pull-down menu (28, 34, 43). The nucleotide composition and predicted amino acid sequence can be used to aid in predicting whether an ORF is authentic (61). In addition, the 100-bp upstream sequence is included in the Sequence query interface to allow for promoter analysis.

Comparisons of poxvirus DNA sequences and/or protein sequences from one or more viruses can be performed with the National Center for Biotechnology Information BLAST programs TBLASTN, BLASTX, BLASTP, and PSIBLAST (6), which have been integrated into the POCs software. TBLASTN compares a protein sequence to six-frame translations of DNA sequences in the database; BLASTX compares a translated DNA sequence to protein sequences in the database; BLASTP compares the protein sequence to protein sequences in the database; and PSIBLAST does iterative protein-protein sequence searches and often uncovers distant homologies not found by other methods. These BLAST programs allow the user to search for orthologous genes or proteins and to identify ORFs that were not originally annotated by the sequencing group. The programs are also very useful for the detection of possible errors in reported gene sequences by allowing the entire region of an ORF or fragmented ORFs to be viewed and determining whether there is a nucleotide change with conservation of the downstream sequence. If the downstream sequence is conserved intact, it is possible that the reported nucleotide change is an error, whereas truly fragmented genes accumulate further errors over time because there is no longer any selective pressure for sequence conservation (see the discussion of VV-Tan sequencing below).

Comparisons of DNA or protein sequences can also be made in POCs with Jalview (<http://www.compbio.dundee.ac.uk/>), JDotter (54), Laj (46), and NAP (30). Jalview is a multiple-sequence-alignment program that allows the user to manually edit multiple alignments (which is often necessary) after either local or remote preliminary alignment by CLUSTAL W (57) and T-Coffee (41). JDotter is an alignment program that compares two gene or two protein sequences and displays similar regions in a dot plot via an interactive graphic interface. Laj also generates a dot plot-like picture of two

aligned sequences but actually displays a series of local alignments generated by BLAST. These gapped local alignments provide useful information regarding the conservation of DNA sequences. Laj is best at comparing two genomic sequences, whereas JDotter is more useful for comparing smaller regions, i.e., specific genes or proteins. NAP greatly simplifies searching for frameshifts or for insertions or deletions in a DNA sequence because it generates nucleotide-amino acid alignments that allow the user to compare a DNA sequence from one virus to a predicted protein sequence from a different virus and shows point mutations and insertions or deletions.

These tools enable the user to compare poxvirus genomes and determine whether genes are fragmented. In addition, POCs aids in predicting whether fragmented genes are likely to produce a truncated but functional protein (depending on the size, conservation, and potential promoter sequences).

Analysis with these tools in POCs is fast and straightforward because the POCs software is devoted to poxvirus genomes. In addition, POCs is able to manage spliced genes; therefore, the software may be used for other large viral genomes, such as those of herpesviruses and baculoviruses.

Gene family organization. We created a POCs database of gene families by grouping related genes based on similarities in BLASTP results. The families were named based on the predicted or known functions of the constituent proteins, as described in poxvirus sequencing studies (5), *Fields Virology* (40), and original research (12, 14, 38, 53, 56, 58). Gene families were assigned initially by performing a BLASTP analysis of each protein against every protein in every poxvirus. We proceeded based on the expectation that there should be a large set of "essential" genes that are present in each and every poxvirus genome and relatively few families that contain multiple genes from a single genome. Therefore, in automatically generating these families, it was desirable that in most cases, each "essential gene" family (e.g., DNA polymerase) contain one gene from each virus, resulting in families with 21 genes, one from each of the 21 genomes. We performed a large number of trials with different expect (E) values to create the largest number of families meeting this criterion (data not shown). These trials indicated that the largest number of such families was generated by using an E-value of 10^{-17} , and this value was used as the basis for family designation. Thus, the gene families that we have created primarily represent groups of orthologs (hence, the name POCs, for poxvirus orthologous clusters); however, some families contain paralogs (related genes in the same genome). In future versions of the database, we intend to implement a system to annotate the ortholog-paralog relationships.

Each POCs gene family was then manually inspected (using the various tools built into POCs) to verify correct assignment of the poxvirus orthologs. This was accomplished by analyzing every family with the TBLASTN, BLASTX, BLASTP, PSIBLAST, Jalview, and NAP tools. The creation of the families made it simple to identify genes that were conserved in all but a few virus genomes, and further investigation of this group of genes identified errors in GenBank files. During this analysis, an annotation error was found in the fowlpox virus (FPV) GenBank file for the rifampin resistance gene, FPV-050. The GenBank file listed the gene ORF as a fragment between bp 52647 and 52856 (210 bp). However, POCs TBLASTN analysis

of related poxvirus rifampin resistance genes indicated that the FPV gene started at bp 52914 and stopped at bp 54569 (1,656 bp). Thus, in the original GenBank annotation, it appeared that FPV did not contain a functional counterpart of this gene, suggesting that it is not essential for poxvirus replication. However, with POCs we have shown that this gene is predicted to be full length and functional in FPV and is therefore present in every poxvirus genome sequenced to date. A similar error was found for the myxoma virus (MYX) gene 077L, where TBLASTN analysis showed the ORF as being located between bp 75602 and 75171 (432 bp) and the GenBank file reported the ORF as being located between bp 75735 and 75556 (180 bp). These errors have been corrected in the POCs database and were reported to the authors for correction of the GenBank files. It is extremely important that these errors are detected and corrected; otherwise, they are propagated throughout other public databases and may influence experiments that depend on database sequence information. Thus, the above examples highlight the utility of the POCs software package for scientific analysis because this database is updated and corrected by poxvirus researchers.

Since MCV and the entomopoxviruses have diverged significantly from the chordopoxviruses, we searched for more distantly related orthologs in these viruses. Conserved gene families that did not contain MCV or entomopoxvirus gene members were used to specifically search these genome sequences for orthologs. If an MCV or entomopoxvirus BLASTP hit was found, regardless of the E-value, the MCV or entomopoxvirus gene was then used to search against the POCs database. If gene homology to additional chordopoxvirus genes in the same family was found and the entomopoxvirus or MCV gene had higher homology to this family than to other chordopoxvirus gene families, the gene was included in the poxvirus gene family. Manual alignment analysis, considering conservation of the most highly conserved amino acid residues in the family and hydrophobic regions, was used as the final criterion for family designation. The criterion for inclusion of a gene in a family, called the "family assignment," is available from the "Family View" window in POCs (BLASTP or manual), so that users will know how the gene was assigned.

We have identified a number of entomopoxvirus genes (including homologs of the VV-Cop A9L, E6R, H3L, and L5R genes) that we have manually placed in the larger orthopoxvirus families (Table 2), despite the fact that they have BLASTP E-values above 10^{-17} . Figure 1 shows an example of one such alignment. Several of the more diverse members of the POCs family containing VV-Cop L5R (putative membrane protein) were aligned by PEPTOOL, available at PBR (www.poxvirus.org). We have manually included the entomopoxviruses in this family based on a significant number of absolutely conserved amino acids, including two cysteines; a large well-conserved hydrophobic domain; and similarity in gene length. The two entomopoxvirus proteins shown in Fig. 1 are 43.6% identical, but they are no more than 24% identical to other proteins in this alignment. Some highly conserved gene families (Table 3) do not currently contain entomopoxvirus members. For example, the family containing the VV-Cop A20R gene does not include entomopoxvirus genes, although this family is expected to encode a very important DNA polymerase processivity factor. A BLASTP analysis of entomopoxviruses uncovered sev-

eral different entomopoxvirus genes with E-values below 1 (one hit was 10^{-7}); however, further BLASTP analysis of these hits showed that the identified entomopoxvirus genes had higher homology to individual members of several other gene families than to the A20R gene family. Furthermore, PSI-BLAST did not reveal any orthologs, and analysis of multiple alignments did not demonstrate the conservation of amino acid residues that were conserved in all other members of the family. These results do not mean that there is no VV-Cop A20R homolog in entomopoxviruses but rather that it cannot be recognized by these similarity search methods. While these family designation criteria are subjective in nature and may or may not prove to be functionally relevant, we have nonetheless attempted to classify MCV and entomopoxvirus genes into suitable families when possible. These designations will provide a scaffold for testing of hypotheses, and the families will be updated as biochemical and functional data become available.

Using POCs software, we have identified 768 gene families; however, many contain only a single unique gene and some contain multiple paralogs. A total of 342 families contain genes conserved in at least two different viral genomes. However, 42 of these families contain only small ORFs, predicted to encode proteins smaller than 60 amino acids and overlapping larger genes, from the very closely related VV-Cop and VV-Tan. We believe that most of these ORFs are unlikely to encode functional proteins (61, 62). Thus, we submit that there are currently 300 bona fide gene families.

Gene family analysis. In addition to querying the database for a specific gene, it is also possible to query for family information. Within POCs, the user may access the family of any gene and immediately view all poxvirus orthologs from either the Sequence Query or the "Gene Family Analyzer" window. The POCs Gene Family Analyzer interface also allows the user to perform several complex queries. The user can search for the family containing a specific gene (by family name, family number, or ORF designation in any virus), families conserved in a certain number of virus genomes, families containing a certain number of genes, or families that contain or do not contain genes from a particular viral genome. Any of these queries will result in a table with links to all of the data for any requested family or gene. From these queries, the user can compare genes that are present in one virus but not in another, compare genes within a family, and search for fragmented genes. For example, a query to retrieve families that contain genes from MYX and Shope rabbit fibroma virus (SFV) but not swinepox virus (SPV) and lumpy skin disease virus (LSDV) takes only a few seconds to perform. This search identifies 13 gene families and lists all of the viruses and all of the genes that are members in the 13 families. The advantage of displaying the data in this format and detail is that the user is informed as to whether these families have been identified because (i) the genes in the family are present only in SFV and MYX or (ii) the genes are present in many viruses but absent from SPV and LSDV. Similarly, it is possible to search for poxvirus genes required for infecting mammalian hosts as opposed to avian hosts. A query to identify genes present in viruses that infect mammalian hosts (Yaba-like disease virus, GenBank LSDV, monkey poxvirus, SPV, camel poxvirus, VV-Cop, variola virus [Ind, Bang, and Gar], SFV, MYX, and MCV) but not present in FPV results in the software retrieving three families (fami-

TABLE 2. Completely conserved gene families

VV-Cop ORF	Family names of 49 poxvirus conserved genes	Family identification no.	Function ^a
A1L	Late transcription factor 2 (VLTF-2)	1153	T
A2L	Late transcription factor 3 (VLTF-3)	1228	T
A3L	P4b precursor	1072	M
A5R	RNA polymerase subunit 19 (RPO19)	1225	T
A7L	Early transcription factor—large (VETF-I)	914	T
A9L	Late virion membrane protein (MP), essential	1218	M
A10L	P4a precursor	1750	M
A11R	Unknown	1217	U
A16L	Unknown soluble-myristylated	887	U
A18R	DNA helicase, transcription	896	T
A21L	Unknown	1202	U
A22R	Holliday junction resolvase	1201	R
A23R	Intermediate transcription factor 3—large (VITF-3)	946	T
A24R	RNA polymerase subunit 132 (RPO132)	880	T
A28L	Unknown predicted signal peptide	1099	U
A29L	RNA polymerase subunit 35 (RPO35)	1197	T
A32L	ATPase-DNA packaging protein	1192	M
D1R	mRNA capping enzyme large subunit	1109	T
D4R	Uracil-DNA glycosylase	1130	R
D5R	NTPase, DNA replication	950	R
D6R	Early transcription factor—small (VETF-s), morphogenesis	1027	T, M
D7R	RNA polymerase subunit 18 (RPO18)	1141	T
D10R	Nucleophosphohydrolase-pyrophosphohydrolase downregulator	1143	T
D11L	Nucleophosphohydrolase I (NPH-I), virion	1093	T, M
D12L	mRNA capping enzyme small subunit, VITF	1151	T
D13L	Rifampin resistance MP, morphogenesis	883	M
E1L	Poly(A) polymerase large subunit	939	T
E6R	Unknown	1253	U
E9L	DNA polymerase	910	R
E10R	Redox-EVR-1, morphogenesis	900	M
F9L	Unknown predicted MP	1001	U
F10L	Serine-threonine kinase, morphogenesis	1065	M
G1L	Protease morphogenesis	1275	M
G5R	Unknown	1017	U
G6R	Unknown	1131	U
G9R	Myristylated protein	957	U
H2R	Unknown	1116	U
H3L	Intracellular mature virus morphogenesis viral protein (VP55)	1269	M
H4L	RNA polymerase-associated protein (RAP94)	1695	T
H6R	Topoisomerase type I	908	R
I7L	Virion core protease	1015	M
I8R	RNA helicase, NPH-II	1104	T
J3R	Poly(A) polymerase small subunit VP39	895	T
J5L	Late MP, essential	1777	M
J6R	RNA polymerase subunit 147 (RPO147)	1040	T
L1R	Myristylated MP virion	1044	M
L3L	Unknown	1285	U
L4R	Core packaging transcription	1283	T, M
L5R	Unknown predicted MP	1511	U

^a T, transcription; M, morphogenesis; U, unknown; R, replication.

lies 1309 and 1537, of unknown functions and containing VV-Cop F16L and VV-Cop A37R, respectively; and family 1541, containing the VV-Cop A33R envelope protein) (45). This result suggests that these three genes may be specifically important in poxvirus infection of mammalian hosts but not avian hosts.

Further analysis of genes and families may be easily done with POCs. By selecting the POCs page displaying information on these retrieved families, it is possible to rapidly evaluate all of the members of the family. For instance, by selecting the VV-Cop A37R family, it is immediately apparent that this family contains genes whose products are truncated to 68 amino acids in the three published variola virus sequences,

compared to about 270 amino acids in most other chordopoxviruses. This truncation suggests that A37R is not required for infection in mammals; however, experimental data are required to determine the functionality of fragmented proteins and proteins of different sizes. POCs provides the most comprehensive suite of tools in one location for the analysis of poxvirus genomes, genes, and proteins and can provide invaluable assistance in the formation and investigation of experimental hypotheses.

Conserved gene families. Of the 300 families generated by the POCs database, there are 49 families that are conserved in all 21 poxvirus genomes (Table 2), suggesting that these genes are essential in the poxvirus life cycle. As might be expected,

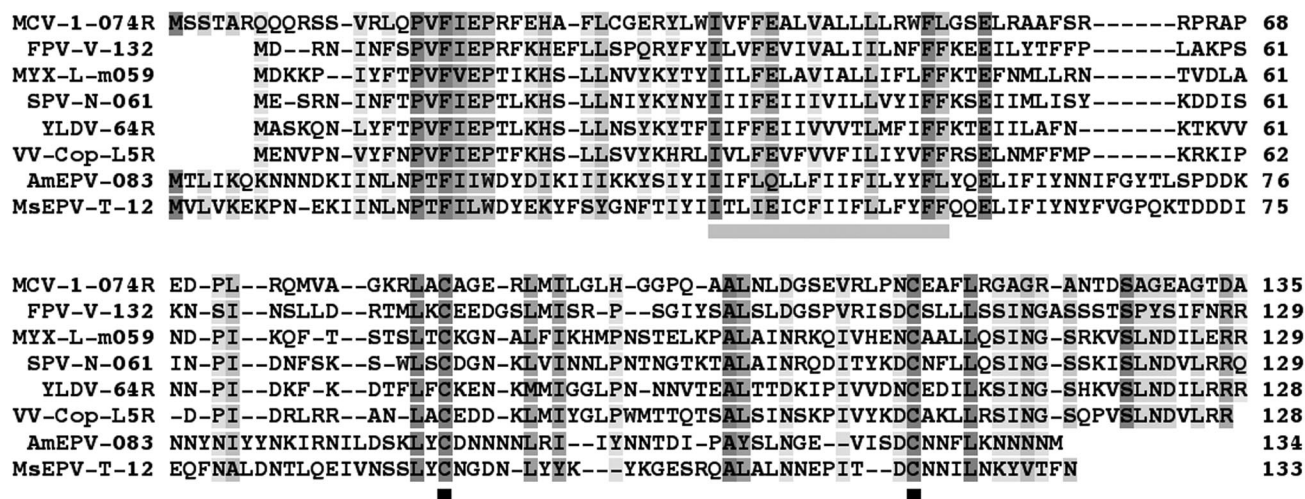


FIG. 1. Alignment of entomopoxvirus protein sequences with six widely diverged chordopoxvirus protein sequences in the VV-Cop L5R family. Shading shows the most highly conserved regions of the proteins (darker shading indicates more conservation). The grey bar shows a well-conserved hydrophobic domain, and black squares indicate cysteines that are conserved in every protein in this family. YLDV, Yaba-like disease virus; AmEPV, *Amsacta moorei* entomopoxvirus.

many of the completely conserved genes are known to be involved in DNA replication and transcription (25 of 49), and 15 of the gene family products are associated with virions, virion assembly, or maturation. Notably, 12 of the putative essential conserved genes are of unknown function, thus highlighting genes that require functional characterization.

A total of 41 additional gene families are conserved only among the chordopoxviruses (Table 3), without clear orthologs in the entomopoxviruses. Eleven of these chordopoxvirus conserved gene families are responsible for replication and transcription; 17 families are associated with virions, virion morphogenesis, or egress; and 13 families have unknown functions. The tyrosine-serine phosphatase (VV-Cop H1L) is included in this list, although it is present in 20 of 21 viruses and missing only from *Melanoplus sanguinipes* entomopoxvirus (MsEPV). A total of 90 genes are completely conserved in the chordopoxviruses, and we hypothesize that these genes comprise the minimum essential chordopoxvirus genome. Since VV-MVA is a highly passaged and attenuated virus, we hypothesized that it might be missing some genes that are conserved in all natural pathogens; however, no genes were found to be conserved in all chordopoxviruses but absent from VV-MVA.

Interestingly, two gene families, those for deoxyuridine triphosphatase (VV-Cop F2L) and thymidine kinase (VV-Cop J2R), were found to be conserved in all poxviruses except for MCV and MsEPV. Both of these genes are involved in nucleotide metabolism, and it has been hypothesized that MCV does not require these two enzymes because it replicates slowly in rapidly dividing skin cells, which are expected to contain a high concentration of DNA nucleotide precursors (47). Similarly, MsEPV infection begins in the midgut, which also contains a preponderance of rapidly dividing cells (2). The VV-Cop B1R serine-threonine kinase family contains genes from 20 of the 21 viruses and is missing only from MCV. As such, this family is not listed in either the completely conserved or the chordopoxvirus conserved families (Tables 2 and 3). The reasons why such genes would be so highly conserved in insect

poxviruses and chordopoxviruses but missing from MCV are unknown; however, the unique life cycle and slow replication of MCV may again hold the answer. The serine protease inhibitor (SPI) family genes are also represented in all chordopoxviruses except for MCV. The skewed AT genome contents of MCV and the entomopoxviruses (36 and 82%, respectively, compared to 66.6% for the orthopoxvirus VV-Cop) confound homology searching. However, we have manually searched for homologs and evaluated the best matches as described above, regardless of E-values.

It has long been noted that essential conserved genes tend to be located in the central regions of genomes. Figure 2 shows the first systematic analysis of conserved gene locations and confirms that all 49 conserved putative essential poxvirus genes and the 41 conserved chordopoxvirus genes from Tables 2 and 3, respectively, are located in the central region of the VV-Cop genome. The terminal regions of the genome contain the majority of the virulence and host range genes. These genes are not as highly conserved among the orthopoxviruses as the genes in the central region of the genome. Genome maps and information on all of the gene families (including those not listed in the tables presented here) are available at www.poxvirus.org.

Gene fragments and multigene families. One difficulty in interpreting DNA sequences lies in the prediction of a "real" gene, one that encodes a functional protein. Some genes are much larger in one virus than in others because of a mutation causing the gain or loss of a stop codon. If there is a downstream methionine and the conserved nucleotide sequence continues for 180 nucleotides (60 amino acids) after the introduction of a stop codon, computer analysis will reveal two (or more) ORFs that are homologous to one larger ORF in another poxvirus. Any ORF that is smaller in one virus than another may be considered a gene fragment. Since it is not possible to correctly predict the ORF length required for a functional protein or whether two gene fragments can produce two small proteins that, together, may still be functional, these

TABLE 3. Gene families conserved in chordopoxviruses

W-Cop ORF	Family names of 41 chordopoxvirus conserved genes	Family identification no.	Function ^a
A2.5L	Thioredoxin-like protein	1552	M
A4L	Core protein	1580	M
A8R	Intermediate transcription factor 3—small (VITF-3)	1758	T
A6L	Unknown	1224	U
A12L	Structural protein	1216	M
A13L	Virion membrane protein (MP)	1575	M
A14L	Intracellular mature virus (IMV) phosphorylated MP	1323	M
A14.5L	IMV MP, virulence factor	1547	M
A15L	Unknown	1546	U
A17L	IMV phosphorylated MP	1206	M
A19L	Unknown	1545	U
A20R	DNA polymerase processivity factor	1203	R
A30L	Virion morphogenesis	1561	M
A34R	Extracellular enveloped virion glycoprotein	1540	M
D2L	Structural protein	1351	M
D3R	Structural protein	1350	M
D9R	<i>mutT</i> motif, nucleoside triphosphate pyrophosphohydrolase	1142	U
E2L	Unknown	1251	U
E4L	RNA polymerase subunit 30 (RPO30), VITF-1	1252	T
E8R	Endoplasmic reticulum-localized MP	1254	U
F12L	Actin tail, microtubule	1145	M
F13L	Phospholipase extracellular enveloped virion	1146	M
F15L	Unknown	1148	U
F17R	DNA-binding phosphoprotein	1150	R
G2R	Late transcription factor (VLTF)	1277	T
G3L	Unknown	1521	U
G4L	Glutaredoxin 2	1279	M
G5.5R	RNA polymerase subunit 7 (RPO7)	1133	T
G7L	Structural protein	1776	M
G8R	Late transcription factor 1 (VLTF-1)	1287	T
H1L	Tyrosine-serine phosphatase	1270	M
H5R	Late transcription factor 4 (VLTF-4)	1353	T
H7R	Unknown	1260	U
I1L	DNA-binding protein	1263	R
I2L	Unknown	1598	U
I3L	DNA-binding phosphoprotein	1265	R
I5L	Unknown VP13	1367	U
I6L	Unknown	1268	U
J1R	Virion	1273	M
J4R	RNA polymerase subunit 22 (RPO22)	1272	T
L2R	Unknown	1584	U

^a See Table 2, footnote a.

fragmented ORFs are included in the POCs database as real genes. Gene fragments are immediately visible in POCs when viewing a family because the molecular weight is listed and two adjacent genes may be listed for a single virus. The presence of gene fragments has resulted in several families, such as the ankyrin family, that contain more than one gene from a viral genome. A full analysis of fragments is beyond the scope of this study, and biochemical analysis will be required to elucidate their functionality. Currently, the POCs database counts gene fragments as real genes (as long as they were annotated in the GenBank file), and the gene fragments are included in the total gene count; however, annotations to allow users to distinguish the gene fragments from complete full-length genes in families are being developed for POCs so that fragments can be viewed and analyzed but not included in the overall gene count.

We have worked from the principle that each genome contributes a single gene to each essential gene family. Poxvirus genomes do, however, contain some multigene families, resulting in families that contain more than one gene from an indi-

vidual viral genome. These are true nonfragmented (full-length) genes, but it is not clear whether they have evolved by gene duplication to produce paralogs or represent multiple gene acquisition events. One such family is the SPI family, which contains 47 genes from 18 genomes. It is known that there are at least three types of SPIs; however, when BLASTP analysis is carried out for each SPI protein, all other known SPI proteins are found by the search with an E-value of less than 10^{-17} . This situation results in the database grouping all SPI genes into one family. Since all three types of SPIs tend to be distantly but equally related, it is difficult to reliably separate the SPI proteins into three families; therefore, we have chosen to place all of the SPI proteins in a single family for the time being. Additional information, such as gene location (synteny), will be used in future work to aid in the classification of orthologs and paralogs.

VV-Tan sequencing and analysis. Analysis of the families indicated that 19 genes in the VV-Tan 1998 GenBank file had significant changes relative to conserved genes in variola virus strains and VV-Cop. Thirteen of these gene families were

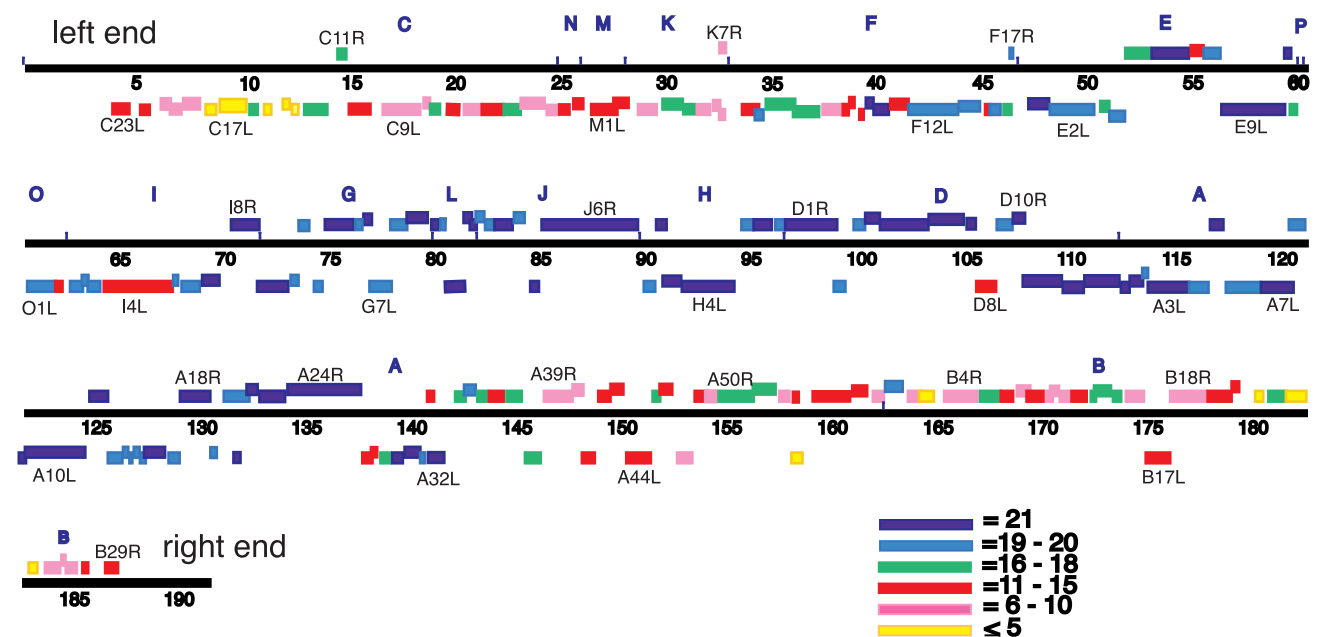


FIG. 2. *Hind*III restriction map of VV-Cop with fragments A to P and genome positions numbered from left to right. Colored bars indicate ORFs and the number of viruses in which an ortholog is conserved. Dark blue bars indicate the 49 genes conserved in all poxviruses (Table 2), and lighter blue bars indicate genes conserved in at least 19 of the 21 genomes, including the 41 genes listed in Table 3. ORFs transcribed rightward are shown above the line, and ORFs transcribed leftward are shown below the line. A few of the larger ORFs are labeled for easy reference.

missing genes from VV-Tan. The TBLASTN function built into POCs permitted a rapid analysis of corresponding VV and variola virus genes to find the homologous DNA region in VV-Tan. Of the 13 genes missing from the GenBank file, we found 10 (TD4R, TL5R, TA37L, TF15L, TG9R, TA62R, TA65R, TB20R, TA63R, and TK5L) in VV-Tan that were listed under “miscellaneous features” and were not fully annotated in this submitted but unpublished sequence. These genes have been added to the POCs database but are still not annotated in the 1998 GenBank file and are therefore unlikely to be represented in any other public database (Table 4). The remaining 3 of the 13 missing VV-Tan genes (corresponding to VV-Cop A2.5L [thioredoxin], VV-Cop F12L [actin or microtubule-associated protein], and VV-Cop G6R [unknown function]) were reported in the GenBank file to have highly conserved DNA sequences over the full length compared to the VV-Cop genes. However, the VV-Tan genes in the GenBank file were reported to have nucleotide insertions or deletions causing premature stop codons (Table 4). Since these three genes are conserved in all chordopoxviruses (Tables 2 and 3), we decided to resequence these genes with modern techniques and equipment. We sequenced the regions of the reported frameshifts in the VV-Tan genes. Analysis of our sequence data (Table 4) showed that these genes were in fact not truncated in the genome of the separately passaged VV-Tan strain from CDC.

In addition to the 13 genes described above, 5 gene families were found to each contain two adjacent VV-Tan ORFs, and 1 gene family was found to contain one truncated VV-Tan ORF (Table 4). Thus, the VV-Tan genes from these six families had been reported in 1998 to be fragments of the larger corresponding VV-Cop and variola virus genes. Since all of

these families are highly conserved, being present in at least 15 of the analyzed genomes, we sequenced the VV-Tan genome around the reported mutations. Analysis of our data revealed that these genes were intact in our VV-Tan DNA (Table 4). When there were sequence differences between the VV-Cop and VV-Tan GenBank files (excluding the areas of frameshift mutations reported for VV-Tan), our sequence matched the VV-Tan GenBank sequence twice as often as the VV-Cop sequence, confirming that the genomic DNA that we sequenced was indeed VV-Tan. However, when we compared the sequence of our separately passaged isolate of VV-Tan to the sequence in the 1998 GenBank file, we found a 1.46% difference in the sequences, about 1 difference per 68 nucleotides. The VV-Tan sequences have been updated in POCs and were included in the evaluations of data for Tables 2 and 3 describing conserved genes. With these updates to VV-Tan, there are now six gene families that contain VV-Cop genes but that do not contain any VV-Tan gene; however, these families are not highly conserved and appear to contain gene fragments. Conversely, VV-Tan genes are found in six families in which there is no VV-Cop gene, but most of these families are not conserved and also contain apparent gene fragments. The updated VV-Tan genome is available in POCs; however, researchers must be aware that only nine gene sequences have been updated for the VV-Tan genome.

DISCUSSION

We have described here the development of a comprehensive, curated, poxvirus genome database and extensive software designed specifically for poxvirus gene, protein, and genome analysis by molecular biologists. Using POCs, we have

TABLE 4. Annotation and sequence updates to VV-Tan genes

Family name	VV-Tan gene	VV-Cop gene	VV-Tan POCs new start	VV-Tan POCs new stop	No. of amino acids	GenBank problem ^a	Conservation ^b
Genes not originally annotated							
Ankyrin	TB20R	B20R	179955	181796	613	Not annotated	13/12
Kelch ring canal	TA65R	A55R	160131	161825	564	Not annotated	19/14
Phospholipase D-like	TK5L	K4L	28424	27150	424	Not annotated	9/9
Putative membrane protein	TL5R	L5R	79474	79860	128	Not annotated	21/21
Putative signal peptide	TA37L	A28L	140396	139956	146	Not annotated	21/21
Unknown (Cop A51R)	TA62R	A51R	157363	158367	334	Not annotated	16/16
Unknown (Cop A52R)	TA63R	A52R	158437	159009	190	Not annotated	9/9
Unknown (Cop F15L)	TF15L	F15L	42506	42030	158	Not annotated	19/19
Uracil DNA glycosylase	TD4R	D4R	97157	97813	218	Not annotated	21/21
Late transcription factor 1 (VLTf-1)	TG9R	G8R	74770	75552	260	Not annotated	19/19
Thioredoxin-like	TA2.5.1L	A2.5L	110849	110619	76	AY188514	19/19
Actin tail, microtubule	TF12.1L	F12L	40360	38453	635	AY188508	19/19
Unknown (Cop G6R)	TG7.1R ^c	G6R	73162	73659	165	AY188510	21/21
Genes originally annotated as fragments							
α -Amanitin sensitivity	TN2.1L	N2L	22363	21836	175	AY188507	17/15
Extracellular envelope virus glycoprotein	TA43R/TA42R ^d	A33R	143013	143570	185	AY188515	18/18
Interferon resistance, inhibitor of protein kinase PKR	TE3L/TE4L ^d	E3L	47921	47349	190	AY188509	17/17
Nucleophosphohydrolase I (NPH-I)	TD13L/TD12L ^d	D11L	106867	104972	631	AY188513	21/21
NPH-PPH downregulator	TD11R/TD10R ^d	D10R	104225	104971	248	AY188512	21/21
RNA polymerase-associated protein (RAP94)	TH5L/TH4L ^d	H4L	88871	91255	794	AY188511	21/21

^a The new GenBank accession number is given for each updated sequence.

^b Number of genes/number of viruses in family.

^c Listed in GenBank as TF7R due to an apparent typographical error.

^d Listed in GenBank as two ORFs.

identified both completely conserved and chordopoxvirus conserved gene families. Together, they describe the natural minimum essential chordopoxvirus genome. It should be noted that not all of these genes are essential for replication *in vitro*. For example, the VV-Cop F13L and A34R genes can be experimentally deleted, resulting in an attenuated but replication-competent virus (13, 20, 44, 65). The fact that viruses missing these genes have not been identified in nature may nonetheless indicate that the genes are essential for the survival and spread of a poxvirus in host populations.

We have used gene conservation in POCs to identify areas for annotation and sequencing updates in available genomes. Notably, the 1998 GenBank VV-Tan genome file has been updated. POCs is helpful in this process because it identifies genes that are absolutely or highly conserved in a certain group of viruses and therefore flags genomes of closely related viruses for which a truncation or fragmentation has been reported. These areas may be sequenced to verify accuracy.

Table 3 lists 41 genes conserved in all chordopoxviruses but not entomopoxviruses. A number of the chordopoxvirus conserved genes have what might be predicted to be essential functions in the virus life cycle, including transcription, replication, virion formation, and egress. It is possible that there remain unrecognized, highly diverged orthologs in AT-rich entomopoxvirus genomes. Alternatively, insect cells may possess various complementing factors that are absent from vertebrate hosts. In the completely conserved orthologs, genes encoding transcriptional and replicative functions are more common than genes involved in morphogenesis (at a 5:3 ratio). However, in the chordopoxvirus conserved group, this pattern is reversed, with relatively more genes being involved in mor-

phogenesis than in RNA or DNA synthesis (at a ratio of 11:17). This pattern suggests that replication and transcription are more broadly conserved across genera than morphogenesis. This finding may be expected because morphogenesis and egress are likely more host specific, since the processes rely substantially on host membranes and host proteins.

The large number of highly conserved genes (24) without functional characterizations highlights areas for future experimental research. It should be noted that gene families were first constructed by using BLAST E-values and that these scores are dependent on the length of the query sequence (gene size). Therefore, it is possible that smaller genes were less likely to be identified as family members due to this computational bias. The concept of bias is supported by the fact that the average molecular masses of proteins in the completely conserved and chordopoxvirus conserved gene families are 49 and 25 kDa, respectively. Larger entomopoxvirus genes may have been more easily recognized as family members than smaller genes. However, the molecular mass range in completely conserved genes is 11 to 147 kDa, and that in chordopoxvirus conserved genes is 6 to 86 kDa, so both Tables 2 and 3 contain a range of gene sizes. Furthermore, each family was manually assessed by comparing conserved amino acid patterns, and even very small and distantly related orthologs (e.g., the 8.9-kDa entomopoxvirus protein member of the A9L family) were assigned to families. An additional reason that entomopoxvirus orthologs may be difficult to identify is that entomopoxviruses may have acquired genes independently from other sources. A precedent for this situation appears to exist in the uracil DNA glycosylase (UDG) family. The entomopoxvirus UDG is more similar to bacterial and herpesvirus

UDGs than to the chordopoxvirus UDG, suggesting that it was acquired independently. Hence, there are several reasons why entomopoxvirus orthologs may not be recognized.

Our ability to reliably predict whether an ORF will encode a functional protein remains poor. Many gene fragments have been reported for poxvirus genomes, and these may or may not be functional. Gene fusion events resulting in multidomain, composite proteins comprise a major pathway of evolution; therefore, the presence of a fragmented or smaller gene in one genome and a much larger or fused gene in another genome does not mean that the smaller or fragmented gene is non-functional (22, 64, 67). One study comparing 23 genomes of eukaryotes, eubacteria, and archaeobacteria identified 7,224 domains present both as independent genes and as gene fusions (22). Increased regulatory efficiency is believed to provide a selective advantage for gene fusion events. It is also known that genes may be lost through mutations in the absence of selective pressure, and gene fission events are also possible (8, 22). Various POCs tools may be used to evaluate the likelihood of an ORF actually encoding a protein; these include analyses of the upstream promoter region and nucleotide and amino acid compositions.

The POCs database will be updated regularly, and we will also respond to user requests for changes and new annotations. The database and family identifications should be very useful for antipoxvirus drug development research. The most broad-spectrum antipoxvirus therapies should be directed against the 90 completely or chordopoxvirus conserved gene products. The absolute conservation of the genes for these products identifies them as both essential and broadly conserved; thus, they are the ideal therapeutic targets. Their conservation indicates that they are under strict selective pressure. Since these genes are more mutationally constrained by their functional requirements than are nonessential genes, the use of the conserved proteins encoded by these genes as drug targets would provide the best chances of reducing the development of drug-resistant strains. In addition, vaccines and antibody therapeutic agents might also be directed against these conserved target proteins, which would be the most likely targets to provide cross-protection between virus species. Finally, these conserved genes would be ideal for the detection of poxviruses in environmental or clinical samples, as all known poxviruses infecting vertebrates share these genes. If a poxvirus were identified, then further testing would indicate what species was present. Thus, POCs software and POCs gene families should provide a useful tool for poxvirus researchers.

There are two other public domain resources that have made attempts to cluster poxvirus proteins into families. Like POCs, Clusters of Related Viral Proteins at the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/crp_start.html) and the Virus Database at University College London (http://www.biochem.ucl.ac.uk/bsm/virus_database/VIDA.html) both use BLASTP alignments for clustering. These programs, however, are not dedicated to poxviruses, and although they attempt to provide a resource for all viruses, we have found that the families that they create are not complete. In addition, these databases do not provide the rich variety of database queries or analysis tools that are available in POCs.

ACKNOWLEDGMENTS

This work was supported by a Natural Science and Engineering Research Council discovery grant (Canada) and NIH grant AI48653-02.

We thank Graeme Roch, Melissa Da Silva, Monika Fazekas, Ryan Brody, David Meeuwis, and Ross Gibbs for expert technical assistance and discussions and Geoff Smith and Bart Hazes for helpful discussions regarding family assignments.

REFERENCES

- Adler, S. P., S. A. Plotkin, E. Gonczol, M. Cadoz, C. Meric, J. B. Wang, P. Dellamonica, A. M. Best, J. Zahradnik, S. Pincus, K. Berencsi, W. I. Cox, and Z. Gyulai. 1999. A canarypox vector expressing cytomegalovirus (CMV) glycoprotein B primes for antibody responses to a live attenuated CMV vaccine (Towne). *J. Infect. Dis.* **180**:843–846.
- Afonso, C. L., E. R. Tulman, Z. Lu, E. Oma, G. F. Kutish, and D. L. Rock. 1999. The genome of *Melanoplus sanguinipes* entomopoxvirus. *J. Virol.* **73**:533–552.
- Afonso, C. L., E. R. Tulman, Z. Lu, L. Zsak, G. F. Kutish, and D. L. Rock. 2000. The genome of fowlpox virus. *J. Virol.* **74**:3815–3831.
- Afonso, C. L., E. R. Tulman, Z. Lu, L. Zsak, F. A. Osorio, C. Balinsky, G. F. Kutish, and D. L. Rock. 2002. The genome of swinepox virus. *J. Virol.* **76**:783–790.
- Afonso, C. L., E. R. Tulman, Z. Lu, L. Zsak, N. T. Sandybaev, U. Z. Kerembekova, V. L. Zaitsev, G. F. Kutish, and D. L. Rock. 2002. The genome of camelpox virus. *Virology* **295**:1–9.
- Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**:3389–3402.
- Amegadzie, B. Y., B. Y. Ahn, and B. Moss. 1992. Characterization of a 7-kilodalton subunit of vaccinia virus DNA-dependent RNA polymerase with structural similarities to the smallest subunit of eukaryotic RNA polymerase II. *J. Virol.* **66**:3003–3010.
- Antoine, G., F. Scheiffinger, F. Dörner, and F. G. Falkner. 1998. The complete genomic sequence of the modified vaccinia Ankara strain: comparison with other orthopoxviruses. *Virology* **244**:365–396.
- Artois, M., K. M. Charlton, N. D. Tolson, G. A. Casey, M. K. Knowles, and J. B. Campbell. 1990. Vaccinia recombinant virus expressing the rabies virus glycoprotein: safety and efficacy trials in Canadian wildlife. *Can. J. Vet. Res.* **54**:504–507.
- Barouch, D. H., S. Santra, M. J. Kuroda, J. E. Schmitz, R. Plishka, A. Buckler-White, A. E. Gaitan, R. Zin, J. H. Nam, L. S. Wyatt, M. A. Lifton, C. E. Nickerson, B. Moss, D. C. Montefiori, V. M. Hirsch, and N. L. Letvin. 2001. Reduction of simian-human immunodeficiency virus 89.6P viremia in rhesus monkeys by recombinant modified vaccinia virus Ankara vaccination. *J. Virol.* **75**:5151–5158.
- Bawden, A. L., K. J. Glassberg, J. Diggins, R. Shaw, W. Farmerie, and R. W. Moyer. 2000. Complete genomic sequence of the Amsacta moorei entomopoxvirus: analysis and comparison with other poxviruses. *Virology* **274**:120–139.
- Betakova, T., E. J. Wolfe, and B. Moss. 2000. The vaccinia virus A14.5L gene encodes a hydrophobic 53-amino-acid virion membrane protein that enhances virulence in mice and is conserved among vertebrate poxviruses. *J. Virol.* **74**:4085–4092.
- Blasco, R., and B. Moss. 1991. Extracellular vaccinia virus formation and cell-to-cell virus transmission are prevented by deletion of the gene encoding the 37,000-dalton outer envelope protein. *J. Virol.* **65**:5910–5920.
- Boulanger, D., P. Green, T. Smith, C. P. Czerny, and M. A. Skinner. 1998. The 131-amino-acid repeat region of the essential 39-kilodalton core protein of fowlpox virus FP9, equivalent to vaccinia virus A4L protein, is nonessential and highly immunogenic. *J. Virol.* **72**:170–179.
- Cameron, C., S. Hota-Mitchell, L. Chen, J. Barrett, J. X. Cao, C. Macaulay, D. Willer, D. Evans, and G. McFadden. 1999. The complete DNA sequence of myxoma virus. *Virology* **264**:298–318.
- Crandell, R. A., H. W. Casey, and W. B. Brumlow. 1969. Studies of a newly recognized poxvirus of monkeys. *J. Infect. Dis.* **119**:80–88.
- Damaso, C. R., J. J. Esposito, R. C. Condit, and N. Moussatche. 2000. An emergent poxvirus from humans and cattle in Rio de Janeiro State: Cantagalo virus may derive from Brazilian smallpox vaccine. *Virology* **277**:439–449.
- Dear, S., and R. Staden. 1991. A sequence assembly and editing program for efficient management of large projects. *Nucleic Acids Res.* **19**:3907–3911.
- Downie, A. W. 1972. The epidemiology of tanapox and Yaba virus infections. *J. Med. Microbiol.* **5**:14.
- Duncan, S. A., and G. L. Smith. 1992. Identification and characterization of an extracellular envelope glycoprotein affecting vaccinia virus egress. *J. Virol.* **66**:1610–1621.
- Ehlers, A., J. Osborne, S. Slack, R. L. Roper, and C. Upton. 2002. Poxvirus orthologous clusters (POCs). *Bioinformatics* **18**:1544–1545.
- Enright, A. J., and C. A. Ouzounis. 2001. Functional associations of proteins

- in entire genomes by means of exhaustive detection of gene fusions. *Genome Biol.* **2**:34.1–34.7.
23. Espana, C. 1971. Review of some outbreaks of viral disease in captive nonhuman primates. *Lab. Anim. Sci.* **21**:1023–1031.
 24. Espana, C., M. A. Brayton, and B. H. Ruebner. 1971. Electron microscopy of the Tana poxvirus. *Exp. Mol. Pathol.* **15**:34–42.
 25. Esposito, J., and F. Fenner. 2001. Poxviruses, p. 2885–2921. *In* D. M. Knipe, P. M. Howley, D. E. Griffin, R. A. Lamb, M. A. Martin, B. Roizman, and S. E. Straus (ed.), *Fields virology*, 4th ed., vol. 2. Lippincott Williams & Wilkins, Philadelphia, Pa.
 26. Goebel, S. J., G. P. Johnson, M. E. Perkins, S. W. Davis, J. P. Winslow, and E. Paoletti. 1990. The complete DNA sequence of vaccinia virus. *Virology* **179**:247–266.
 27. Hiscock, D., and C. Upton. 2000. Viral Genome DataBase: storing and analyzing genes and proteins from complete viral genomes. *Bioinformatics* **16**:484–485.
 28. Hopp, T. P., and K. R. Woods. 1981. Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl. Acad. Sci. USA* **78**:3824.
 29. Hu, Y., J. Lee, J. A. McCart, H. Xu, B. Moss, H. R. Alexander, and D. L. Bartlett. 2001. Yaba-like disease virus: an alternative replicating poxvirus vector for cancer gene therapy. *J. Virol.* **75**:10300–10308.
 30. Huang, X., and J. Zhang. 1996. Methods for comparing a DNA sequence with a protein sequence. *Comput. Appl. Biosci.* **12**:497–506.
 31. Hutin, Y. J., R. J. Williams, P. Malfait, R. Pebody, V. N. Loparev, S. L. Ropp, M. Rodriguez, J. C. Knight, F. K. Tshioko, A. S. Khan, M. V. Szczeniowski, and J. J. Esposito. 2001. Outbreak of human monkeypox, Democratic Republic of Congo, 1996 to 1997. *Emerg. Infect. Dis.* **7**:434–438.
 32. Konya, J., and C. H. Thompson. 1999. Molluscum contagiosum virus: antibody responses in persons with clinical lesions and seroepidemiology in a representative Australian population. *J. Infect. Dis.* **179**:701–704.
 33. Kupper, J. L., H. W. Casey, and D. K. Johnson. 1970. Experimental Yaba and benign epidermal monkey pox in rhesus monkeys. *Lab. Anim. Care* **20**:979–988.
 34. Kyte, J., and R. F. Doolittle. 1982. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**:105–132.
 35. Lee, H. J., K. Essani, and G. L. Smith. 2001. The genome sequence of Yaba-like disease virus, a yatapoxvirus. *Virology* **281**:170–192.
 36. Masson, E., V. Bruyere-Masson, P. Vuillaume, S. Lemoyne, and M. Aubert. 1999. Rabies oral vaccination of foxes during the summer with the VRG vaccine bait. *Vet. Res.* **30**:595–605.
 37. Massung, R. F., L. I. Liu, J. Qi, J. C. Knight, T. E. Yuran, A. R. Kerlavage, J. M. Parsons, J. C. Venter, and J. J. Esposito. 1994. Analysis of the complete genome of smallpox variola major virus strain Bangladesh-1975. *Virology* **201**:215–240.
 38. McCreith, S., T. Holtzman, B. Moss, and S. Fields. 2000. Genome-wide analysis of vaccinia virus protein-protein interactions. *Proc. Natl. Acad. Sci. USA* **97**:4879–4884.
 39. McNulty, W. P., W. C. Lobitz, F. Hu, C. A. Maruppo, and A. S. Hall. 1968. A pox disease in monkeys transmitted to man. *Arch. Dermatol.* **97**:286–293.
 40. Moss, B. 2001. Poxviruses, p. 2849–2884. *In* D. M. Knipe, P. M. Howley, D. E. Griffin, R. A. Lamb, M. A. Martin, B. Roizman, and S. E. Straus (ed.), *Fields virology*, 4th ed., vol. 2. Lippincott Williams & Wilkins, Philadelphia, Pa.
 41. Notredame, C., D. G. Higgins, and J. Heringa. 2000. T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**:205–217.
 42. Ourmanov, I., M. Biliska, V. M. Hirsch, and D. C. Montefiori. 2000. Recombinant modified vaccinia virus Ankara expressing the surface gp120 of simian immunodeficiency virus (SIV) primes for a rapid neutralizing antibody response to SIV infection in macaques. *J. Virol.* **74**:2960–2965.
 43. Parker, J. M. R., D. Guo, and R. S. Hodges. 1986. New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites. *Biochemistry* **25**:5425–5432.
 44. Roper, R. L., and B. Moss. 1999. Envelope formation is blocked by mutation of a sequence related to the HKD phospholipid metabolism motif in the vaccinia virus F13L protein. *J. Virol.* **73**:1108–1117.
 45. Roper, R. L., L. G. Payne, and B. Moss. 1996. Extracellular vaccinia virus envelope glycoprotein encoded by the A33R gene. *J. Virol.* **70**:3753–3762.
 46. Schwartz, S., Z. Zhang, K. A. Frazer, A. Smit, C. Riemer, J. Bouck, R. Gibbs, R. Hardison, and W. Miller. 2000. PipMaker—a web server for aligning two genomic DNA sequences. *Genome Res.* **10**:577–586.
 47. Senkevich, T. G., E. V. Koonin, J. J. Bugert, G. Darai, and B. Moss. 1997. The genome of molluscum contagiosum virus: analysis and comparison with other poxviruses. *Virology* **233**:19–42.
 48. Senkevich, T. G., C. L. White, A. Weisberg, J. A. Granek, E. J. Wolffe, E. V. Koonin, and B. Moss. 2002. Expression of the vaccinia virus A2.5L redox protein is required for virion morphogenesis. *Virology* **300**:296–303.
 49. Shchelkunov, S. N., A. V. Totmenin, I. V. Babkin, P. F. Safronov, V. V. Gutorov, S. G. Pozdnaikov, V. M. Blinov, S. M. Resenchuk, and L. S. Sandakhchiev. 1996. Study of the structure-activity organization of the smallpox viral genome. V. Sequencing and analysis of the nucleotide sequence of the left terminus of the India-1967 strain genome. *Mol. Biol. (Mosc)* **30**:595–612.
 50. Shchelkunov, S. N., A. V. Totmenin, I. V. Babkin, P. F. Safronov, O. I. Ryazankina, N. A. Petrov, V. V. Gutorov, E. A. Uvarova, M. V. Mikheev, J. R. Sisler, J. J. Esposito, P. B. Jahrling, B. Moss, and L. S. Sandakhchiev. 2001. Human monkeypox and smallpox viruses: genomic comparison. *FEBS Lett.* **509**:66–70.
 51. Shchelkunov, S. N., A. V. Totmenin, V. N. Loparev, P. F. Safronov, V. V. Gutorov, V. E. Chizhikov, J. C. Knight, J. M. Parsons, R. F. Massung, and J. J. Esposito. 2000. Alastrim smallpox variola minor virus genome DNA sequences. *Virology* **266**:361–386.
 52. Shchelkunov, S. N., A. V. Totmenin, and L. S. Sandakhchiev. 1996. Analysis of the nucleotide sequence of 23.8 kbp from the left terminus of the genome of variola major virus strain India-1967. *Virus Res.* **40**:169–183.
 53. Shuman, S., S. S. Broyles, and B. Moss. 1987. Purification and characterization of a transcription termination factor from vaccinia virions. *J. Biol. Chem.* **262**:12372–12380.
 54. Sonnhammer, E. L., and R. Durbin. 1995. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* **167**:GC1–GC10.
 55. Stittelaar, K. J., L. S. Wyatt, R. L. de Swart, H. W. Vos, J. Groen, G. van Amerongen, R. S. van Binnendijk, S. Rozenblatt, B. Moss, and A. D. Osterhaus. 2000. Protective immunity in macaques vaccinated with a modified vaccinia virus Ankara-based measles virus vaccine in the presence of passively acquired antibodies. *J. Virol.* **74**:4236–4243.
 56. Takahashi, T., M. Oie, and Y. Ichihashi. 1994. N-terminal amino acid sequences of vaccinia virus structural proteins. *Virology* **202**:844–852.
 57. Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**:4673–4680.
 58. Tolonen, N., L. Doglio, S. Schleich, and J. Krijnse Locker. 2001. Vaccinia virus DNA replication occurs in endoplasmic reticulum-enclosed cytoplasmic mini-nuclei. *Mol. Biol. Cell* **12**:2031–2046.
 59. Tulman, E. R., C. L. Afonso, Z. Lu, L. Zsak, G. F. Kutish, and D. L. Rock. 2001. Genome of lumpy skin disease virus. *J. Virol.* **75**:7122–7130.
 60. Tulman, E. R., C. L. Afonso, Z. Lu, L. Zsak, J. H. Sur, N. T. Sandybaev, U. Z. Kerembekova, V. L. Zaitsev, G. F. Kutish, and D. L. Rock. 2002. The genomes of sheeppox and goatpox viruses. *J. Virol.* **76**:6054–6061.
 61. Upton, C. 2000. Screening predicted coding regions in poxvirus genomes. *Virus Genes* **20**:159–164.
 62. Upton, C., D. Hogg, D. Perrin, M. Boone, and N. L. Harris. 2000. Viral genome organizer: a system for analyzing complete viral genomes. *Virus Res.* **70**:55–64.
 63. Willer, D. O., G. McFadden, and D. H. Evans. 1999. The complete genome sequence of shope (rabbit) fibroma virus. *Virology* **264**:319–343.
 64. Wolf, Y. I., A. S. Kondrashov, and E. V. Koonin. 2000. Interkingdom gene fusions. *Genome Biol.* **1**:13.1–13.13.
 65. Wolffe, E. J., E. Katz, A. Weisberg, and B. Moss. 1997. The A34R glycoprotein gene is required for induction of specialized actin-containing microvilli and efficient cell-to-cell transmission of vaccinia virus. *J. Virol.* **71**:3904–3915.
 66. Xiang, Y., and B. Moss. 2001. Correspondence of the functional epitopes of poxvirus and human interleukin-18-binding proteins. *J. Virol.* **75**:9947–9954.
 67. Yanai, I., Y. I. Wolf, and E. V. Koonin. 2002. Evolution of gene fusions: horizontal transfer versus independent events. *Genome Biol.* **3**:24.1–24.13.