BMC Genomics

RESEARCH ARTICLE Open Access

Genes in the terminal regions of orthopoxvirus genomes experience adaptive molecular evolution

David J Esteban* and Anne P Hutchinson

Abstract

Background: Orthopoxviruses are dsDNA viruses with large genomes, some encoding over 200 genes. Genes essential for viral replication are located in the center of the linear genome and genes encoding host response modifiers and other host interacting proteins are located in the terminal regions. The central portion of the genome is highly conserved, both in gene content and sequence, while the terminal regions are more diverse. In this study, we investigated the role of adaptive molecular evolution in poxvirus genes and the selective pressures that act on the different regions of the genome. The relative fixation rates of synonymous and non-synonymous mutations (the d_N/d_S ratio) are an indicator of the mechanism of evolution of sequences, and can be used to identify purifying, neutral, or diversifying selection acting on a gene. Like highly conserved residues, amino acids under diversifying selection may be functionally important. Many genes experiencing diversifying selection are involved in host-pathogen interactions, such as antigen-antibody interactions, or the "host-pathogen arms race."

Results: We analyzed 175 gene families from orthopoxviruses for evidence of diversifying selection. 79 genes were identified as experiencing diversifying selection, 25 with high confidence. Many of these genes are located in the terminal regions of the genome and function to modify the host response to infection or are virion-associated, indicating a greater role for diversifying selection in host-interacting genes. Of the 79 genes, 20 are of unknown function, and implicating diversifying selection as an important mechanism in their evolution may help characterize their function or identify important functional residues.

Conclusions: We conclude that diversifying selection is an important mechanism of orthopoxvirus evolution. Diversifying selection in poxviruses may be the result of interaction with host defense mechanisms.

Background

Poxviruses are a family of double stranded DNA viruses that infect diverse host species. The genus *Orthopoxvirus* includes Variola (the causative agent of smallpox), Vaccinia (the smallpox vaccine), Monkeypox, an emerging human pathogen, and Cowpox.

Poxviruses are among the largest and most complex of all animal viruses, some expressing over 200 genes [1]. A large fraction of the coding capacity of the genome is for processes essential for viral replication, such as virion assembly, transcription and replication. Unlike other DNA viruses, poxviruses replicate in the cytoplasm and therefore encode all genes necessary for DNA replication and transcription. These essential genes are highly conserved throughout the poxvirus family and in orthopoxviruses these core genes form a continuous block in the center of the linear genome [2,3].

Flanking the central region, the terminal regions of orthopoxvirus genomes show divergence among different genera, among species within a genus, and even among strains of the same species [4]. Many of these genes are non-essential for virus replication in cell culture but are virulence factors that mediate interactions with the host cell or immune system in their natural host. These include immune evasion genes that inhibit cytokines [5], inhibit the interferon response [6], or block apoptosis [7]. Many of these are host species specific, indicating adaptation to the specific host response to infection.

* Correspondence: daesteban@vassar.edu
Biology Department, Vassar College, 124 Raymond Ave, Poughkeepsie, NY, 12604, USA

Analysis of the content and organization of the orthopoxvirus genome implicates gene gain and loss as major mechanisms in their evolution [8]. Cowpox virus strain Brighton Red (CPXV-BR) has a "master set" of genes; all other orthopoxviruses have a smaller subset of those genes. Thus it is likely that as the orthopoxviruses evolved and diversified into different hosts, some genes were lost while those that were retained adapted to the specific host. The sequence conservation of the core genes may be the result of stringent structural or functional constraints on these core proteins. Host-response modifier genes in the terminal regions, however, may be more able to change and thus show greater sequence divergence. As such, we were interested in understanding the role of adaptive molecular evolution in poxvirus genes and the selective pressures that act on genes in different regions of the genome. Adaptive molecular evolution, or diversifying selection, is a key mechanism for species divergence and identifying proteins or specific residues experiencing diversifying selection may be important in understanding gene function.

Diversifying selection can be detected by measuring the ratio of non-synonymous/synonymous mutation fixation rates ($\omega = d_N/d_S$) [9-11]. Codons experience continual nucleotide substitutions as a result of errors in replication. Substitutions that result in synonymous mutations are selectively neutral. However, non-synonymous mutations can occur and become fixed as a result of selection. In purifying selection ($\omega < 1$), a higher rate of synonymous than non-synonymous mutations occurs, suggesting stringent structural or functional constraints on the amino acid. In neutral selection ($\omega = 1$) a change to a different amino acid has neither a positive nor negative effect on the protein. In diversifying selection ($\omega > 1$) the rate of fixation of non-synonymous mutations is greater than the rate of fixation of synonymous mutations, indicating adaptive molecular evolution.

In this study we analyzed a set of 175 orthopoxvirus gene families using CPXV-BR as a reference genome to determine the selective pressures acting upon them. We show that diversifying selection occurs most strongly in the terminal regions of the genome. The results of this study may help identify important functional regions in genes of known or unknown function, and may help categorize genes of unknown function as host-interacting genes.

Results

Gene families

The data set collected from the Viral Bioinformatics Resource Centre database contained 19,874 chordopoxvirus gene sequences. After the removal of identical sequences, 7758 unique sequences from 103 viruses remained (Additional file 1). Sequences were parsed into 380 gene families of which 214 were present in the CPXV-BR genome.

The CPXV-BR genome encodes 235 genes belonging to 214 gene families. Of the 214 gene families, 38 were excluded from the analysis (Additional file 2). 16 gene families (containing a total of 37 genes) were excluded because they were duplicated in the genome and therefore have the potential for differential evolutionary pressures on each paralog. 17 families with fewer than 6 sequences were excluded, since too few taxa in the dataset results in poor power to identify amino acid sites under diversifying selection [12]. Gene family size in the final dataset ranged from 6-69 sequences. 6 gene families were removed because of difficulty in generating a reliable alignment, fragmentation of the gene in CPXV, or excessive computational time required for analysis due to large family size and length of genes. The final dataset had a total of 175 gene families, comprising a total of 6058 sequences analyzed (Additional file 3). CPXV-BR is used as the reference strain for the purposes of referring to gene names, codon positions and genome organization, however, because the analysis is performed on gene families, not single gene sequences, the results apply to homologous genes of all poxviruses used in this study.

Treelength (S), measured as the number of substitutions per codon in the tree, can be used as a measure of sequence divergence in the gene family. Genes with treelengths that are too short ($S < 0.11$) are insufficiently divergent for this analysis [12]. All genes included in this analysis had $S > 0.11$. S values ranged from 0.225 to 43.02, with only three above 30 (Additional file 3). Genes found to be under diversifying selection (see below) were distributed throughout the range of treelengths.

The length of CPXV genes ranged from 42 to 1286 codons, with the majority having a length of less than 400 codons, and only 5 greater than 800 codons. Genes found to be under diversifying selection were distributed throughout the range (data not shown).

Diversifying selection

We wanted to determine which genes in poxvirus genomes showed adaptive molecular evolution (diversifying selection). To do this, we used site models of codon evolution to determine the rate of fixation of synonymous and non-synonymous mutations that can identify the presence of an amino acid site (codon) class under diversifying selection. Codon models use a statistical distribution to describe the variation in ω among sites, and make no assumptions about which sites are under diversifying selection. Site models permit differing d_N/d_S ratios at each codon, and because selective pressures are expected to vary across amino acids, this approach has high power to detect diversifying selection in genes [11]. Each gene family was analyzed for evidence of diversifying selection

by testing if the data fits a null model (which does not allow for sites under diversifying selection) better than an alternative model, which does allow for sites under diversifying selection. Two model comparisons were applied: M1a vs. M2a, M7 vs. M8, where M1a and M7 are null models and M2a and M8 are alternate models. A log likelihood ratio test was then applied to determine whether allowing for a site class with $\omega>1$ results in a significantly better fit of the model to the data.

Evidence for diversifying selection was considered to be strong where both comparisons were found to be significant at $p < 0.05$. 25 genes were identified by both models as having an amino acid site class experiencing diversifying selection (null models rejected, $p < 0.05$) (Table 1). With model M8 (a less conservative model) an additional 54 genes were identified as having sites experiencing diversifying selection (null model rejected, $p < 0.05$), for a total of 79 genes under diversifying selection (Additional file 4). All significant genes under model M2a were also significant under model M8, except one (Unknown YMTV120.5L).

Alternative models M2a and M8 allow for the existence of a codon site class under diversifying selection. The proportion of sites that fall into this site class provides insight into the selective pressure experienced by the gene. For each gene, the proportion of sites that are under diversifying selection was determined (Additional file 5). Among the genes for which the diversifying selection models showed a significantly better fit of the data, the proportion of sites under diversifying selection ranged from 0.1% to 14.6% (model M2a) and 0.1% to 29.5% (model M8). A gene does not necessarily need a large proportion of sites under diversifying selection to fit the alternative model better. In many genes, a large proportion of amino acids may be virtually invariable due to structural or functional constraints, while diversifying selection occurs at only a few key sites. For example, the F6L homolog (unknown function) has a small proportion of sites under diversifying selection (2.9%), while the mitochondrial associated apoptosis inhibitor has a large proportion of sites under diversifying selection (14.0%) and for both the null model was rejected.

Using the CPXV-BR genome as a reference for gene positions within the genome, we noted a distinct pattern of genes with a high proportion of sites under diversifying selection being located closer to the termini of the genome, as well as a tendency for significant genes (model M2a) to be located in the terminal regions (Figure 1). A centrally located core set of 90 genes is present in all chordopoxvirues [2]. A chi square test revealed that genes experiencing diversifying selection identified by model M2a were significantly more likely to be found among the 85 terminally located less conserved genes ($\chi^2 = 12.56$, df = 1, $p < 0.05$). The 79 genes identified by model M8 however were proportionately distributed between the core and less conserved genes ($\chi^2 = 0.05$, df = 1, $p > 0.05$).

The non-synonymous/synonymous rate ratio ($\omega = d_N/d_S$) is an important indicator of selective pressure at the protein level. The d_N/d_S ratio of a gene is calculated as the average of the ω of each codon in the gene. A higher d_N/d_S ratio suggests a greater role for diversifying selection in the molecular evolution of the gene. Again we note that genes with high d_N/d_S ratios are located in the termini of the genome (Figure 2). Some significant genes have an average d_N/d_S ratio less than 1, indicating that diversifying selection has occurred in these genes, but not sufficiently to cause the average to be greater than 1. These genes may have only a small proportion of amino acids under diversifying selection or a larger proportion that experiences weaker diversifying selection.

The d_N/d_S ratio of the site class under diversifying selection ($\omega 2$) was also determined (Additional file 5). This value is a measurement of the intensity of diversifying selection within that site class. For example, Semaphorin (CPXV-BR-182) experiences very strong diversifying selection ($\omega 2 = 10.417$) in only a small proportion of total sites (2.7%). The F7L homolog (CPXV-BR-055, unknown function) also has a site class experiencing diversifying selection ($\omega 2 = 8.343$), however this slightly weaker selection is spread over a greater proportion of sites (12.9%). For this reason, diversifying selection cannot be identified solely on the basis of quantity of sites experiencing diversifying pressure along the gene, but also the strength of the pressure acting on these genes. A few genes have $\omega 2 > 1$ but the data do not fit the diversifying selection model significantly better than the null model, indicating that a site class under diversifying selection likely does not exist for these genes.

Function and localization of genes under diversifying selection

The 79 genes identified were broadly categorized by function and localization (Table 1, Additional file 4 and Figure 3) based on annotation in the Viral Bioinformatics Resource Center curated genes database [13] and several reports on poxvirus genomes [1,14,15]. The largest category is that of genes of unknown function (20 genes), but many host response modifiers or host-range genes (16 genes) were also found. Enzymes and structural components were also identified. Classified by localization, 36 of the proteins under diversifying selection have been found to be virion associated (Figure 3B and 3C). Localization in the virion was determined from two proteomics studies [16,17] and previous literature [1]. All of the proteins of unknown function, and most (81%) of the host response modifiers are not found in the virion. 18 out of 25 (72%) of the enzymes involved

Table 1 Genes under diversifying selection identified by both models

ORF Number	Gene Family Name	Category	VACV-Cop ORF	Specific Function	Virion
021	EGF_Growth factor	Host Response Modifier	C11R	EGF homolog with mitogenic activity	No
024	IL_18_BP (Bsh_D7L)	Host Response Modifier	n/a	Inhibition of IL-18 activity	No
035	Kelch_like_ (Cop_C2L)	Host Response Modifier	C2L	Unknown	No
036	Unknown (Cop_C1L)	Unknown	C1L	Unknown	No
046	Unknown (Cop_K7R)	Entry/Exit	K7R	inhibitor of pattern recognition receptor signalling pathway	No
048	Apoptosis inhibitor (mitochondrial associated)	Host Response Modifier	F1L	localizes to mitochondria and inhibits apoptosis	No
054	Unknown (Cop_F6L)	Unknown	F6L	Unknown	No
055	Unknown (Cop_F7L)	Unknown	F7L	Unknown	No
056	Cytoplasmic protein (Cop_F8L)	Unknown	F8L	Cytoplasmic localization, unknown function	No
063	Unknown (Cop_F14L)	Unknown	F14L	Unknown	No
129	Carbonic anhydrase Virion	Entry/Exit	D8L	Binds Chondroitin Sulfate on cell surface	Yes
140	Core_protein (Cop_A4L)	Structure/Assembly	A4L	associated with virion core and membranes	Yes
145	Membrane protein (Cop_A9L)	Structure/Assembly	A9L	located on virion membrane, essential for morphogenesis	Yes
162	RNA pol 132 (RPO132)	RNA metabolism	A24R	Component of RNA polymerase	Yes
171	Unknown (YMTV_120.5L)	Unknown	A30.5L	Unknown	No
172	Unknown (Cop_A31R)	Host Response Modifier	A31R	Unknown	Yes
182	Semaphorin	Host Response Modifier	A39R	Induces IL-6 and IL-8 secretion from monocytes	No
191	Unknown (Cop_A47L)	Unknown	A47L	Unknown	No
192	Thymidylate kinase	DNA metabolism	A48R	Thymidylate kinase activity	No
200	Hemagglutinin	Host Response Modifier	A56R	prevents cell-cell fusion, present on EEV and cell surface	Yes
203	Schlafen (Cop_B2R)	Host Response Modifier	B2R	Virulence factor in CMLV	No
204	Ankyrin (Cop_B4R)	Unknown	B4R	Unknown	No
212	Ser_Thr_Kinase (Cop_B12R)	Unknown	B12R	Unknown	No
215	IL_1_beta receptor	Host Response Modifier	B15R	Inhibits IL-1 beta activity	No
216	Unknown (Cop_B17L)	Unknown	B17R	Unknown	No

in DNA or RNA metabolism, as well as the majority of proteins used to assemble or form the structure of the virion (87%) are found in the virion. Focusing on the 25 genes identified by both models, host-response modifiers and genes of unknown function make up the largest proportions, 36% and 40%, respectively (Figure 3A). Of the 6 remaining genes, 5 are found in the virion (Thymidine kinase is not found in the virion).

Identification of specific residues under diversifying selection

For each gene identified as experiencing diversifying selection, the posterior probabilities of each site falling into the three categories were calculated using a Bayes empirical Bayes (BEB) analysis [18]. For each gene, every codon was assigned to either the purifying, neutral or diversifying site classes. Two examples are shown: the Interleukin-18 binding protein (IL-18BP) (Figure 4A) and the mitochondrial associated apoptosis inhibitor (Figure 4B). As with most genes, most codons of the IL-18BP gene are under purifying selection, with a small proportion under neutral or diversifying selection. For example, the posterior probabilities for site W124 are 0.000, 0.057, and 0.942 for purifying, neutral and diversifying selection, thus the site is highly likely to be under diversifying selection. A total of 8 specific sites are

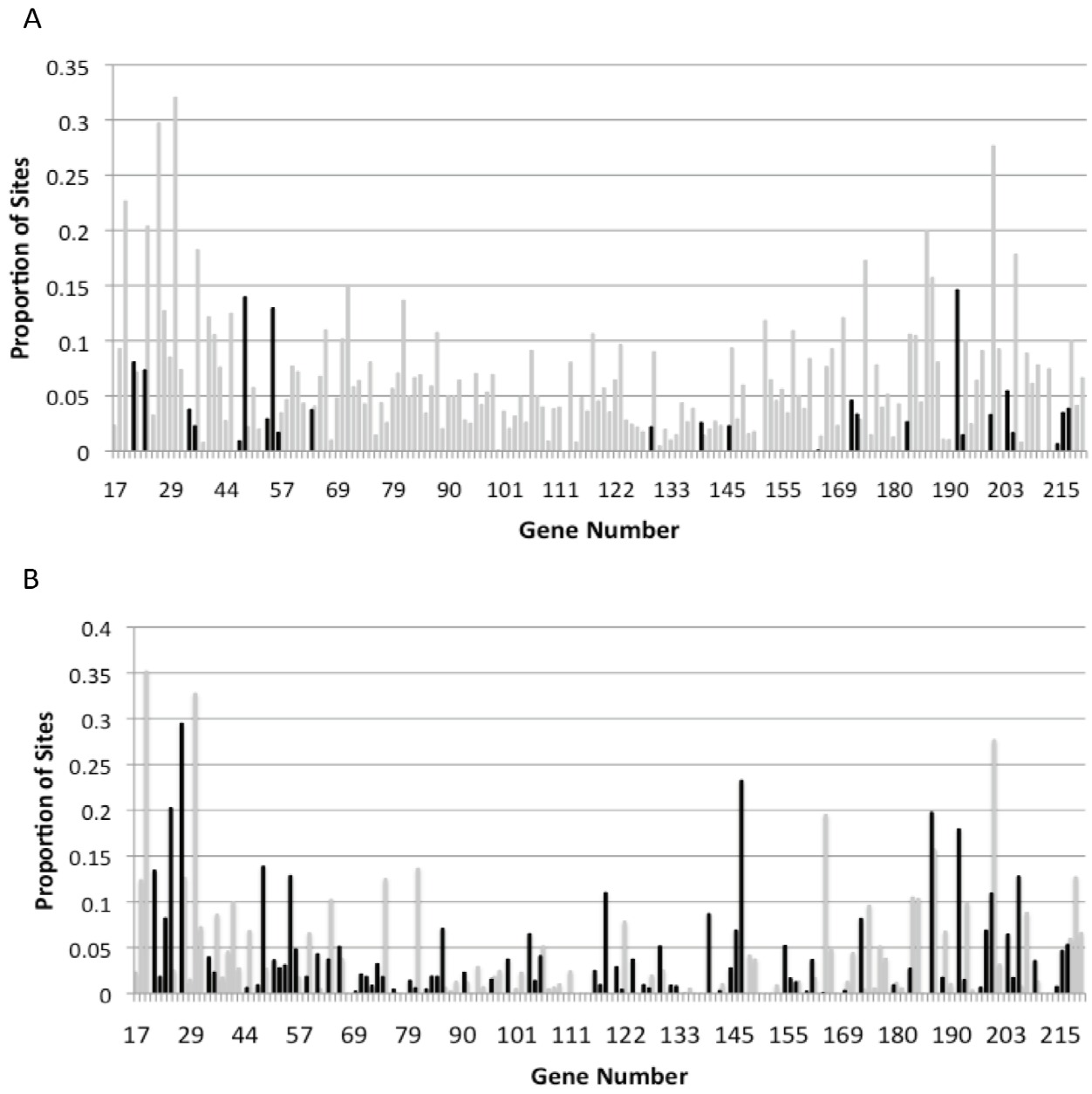

Figure 1 Proportion of sites under diversifying selection in each gene in the CPXV genome . Black: Genes with sites under diversifying selection (Genes for which the alternative model, allowing positively sites with $\omega > 1$, fits the data better than the null model ($p < 0.05$)). White: Genes without sites under diversifying selection. A) Proportions determined using model M2a and LRT against M1a. B) Proportions determined using model M8 and LRT against M7.

predicted to be under diversifying selection in IL-18BP, and 7 sites are under diversifying selection in the highly conserved C-terminal domain of the mitochondrial associated apoptosis inhibitor, and an additional 2 are in the N-terminal region. For other genes, the specific residues with a high posterior probability of falling into the diversifying selection site class are given in Additional file 6 and Additional file 7.

Discussion

In this study we analyzed a large set of poxvirus genes with the purpose of identifying genes experiencing adaptive molecular evolution. Of 235 genes in the Cowpox virus genome, 175 were analyzed using phylogenetic analysis of maximum likelihood (PAML) and subsequently analyzed by Bayes empirical Bayes to determine the probability of each codon falling into the three

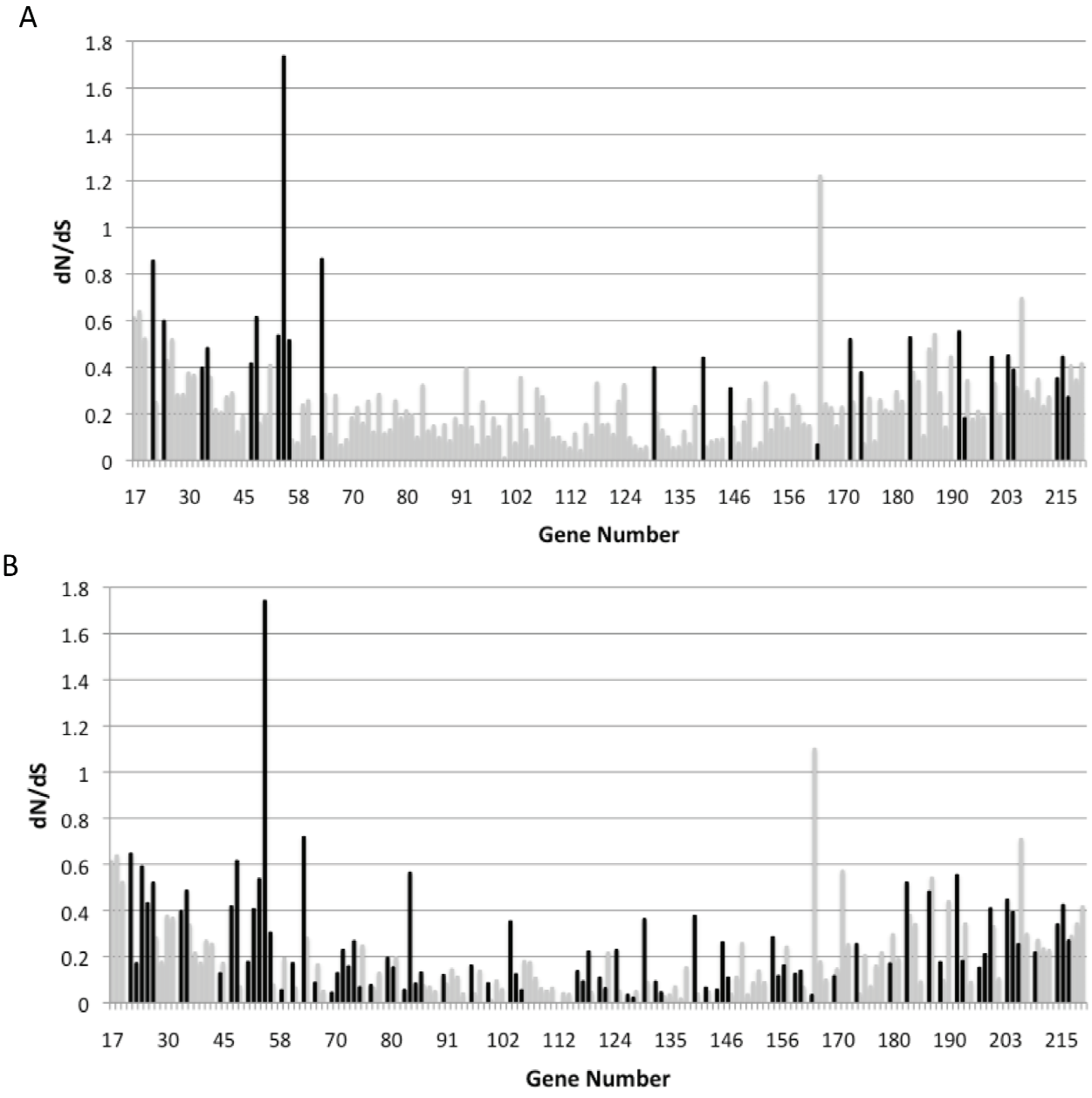

Figure 2 d_N/d_S (ω) ratios of genes in the CPXV genome . The ratio for the gene is the average ω of all codons in the gene. Black: Genes with sites under diversifying selection (Genes for which the alternative model, allowing positively sites with ω>1, fits the data better than the null model ($p < 0.05$)). White: Genes without sites under diversifying selection. A) Ratios determined using model M2a, with statistical significance determined by comparison with M1a. B) Ratios determined using model M8, with statistical significance determined by comparison with M7.

possible site classes of purifying, neutral or diversifying selection. We identified 79 genes under diversifying selection, representing 45% of the analyzed genes in the genome. Of those, 25 (14% of genes analyzed) were identified with both models, indicating high confidence. Thus we identify diversifying selection as an important mechanism of evolution in poxviruses.

Analysis of genes spanning the CPXV genome revealed that diversifying selection is a more important mechanism of molecular evolution in the genome's terminal regions than in the central genes. This may be due to diversifying pressures applied by host interactions. Among different orthopoxvirus species and even strains, terminal regions are more diverse both in gene

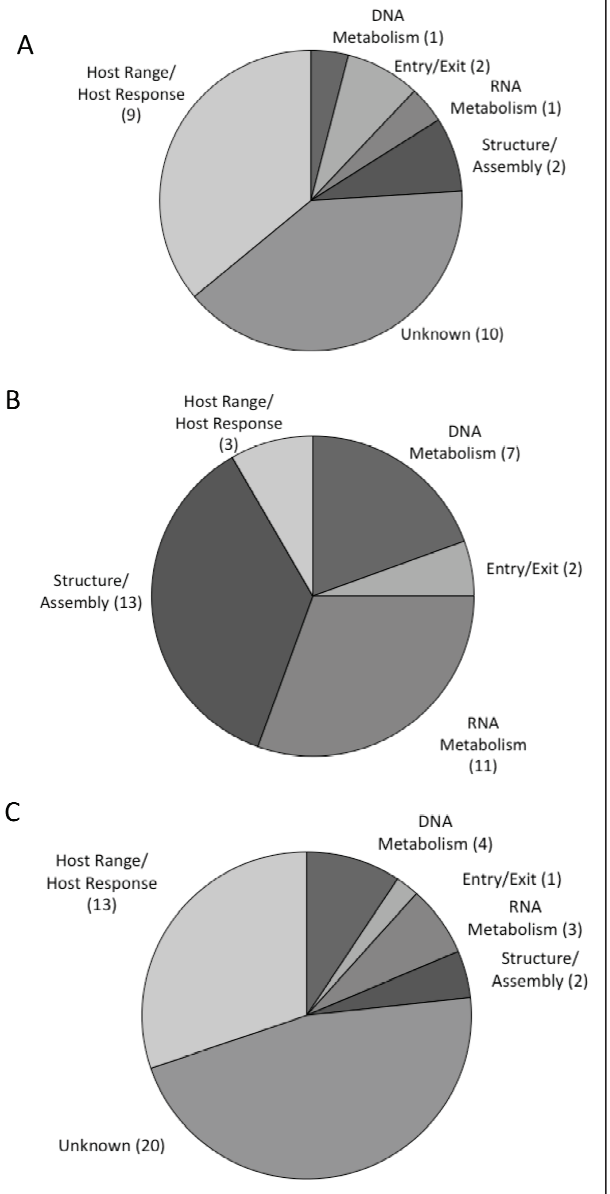

Figure 3 Functional categories of the genes under diversifying selection . A) Functions of the 25 genes identified by both models. B) Functions of 43 non-virion-associated proteins identified by model M8, B) Functions of 36 virion-associated proteins identified by model M8.

content and gene sequences [2]. Many terminally located genes are host response modifiers that directly interact with components of the host immune response or cellular response to infection. Grouping the genes identified in this study into broad functional categories (Figure 3), we identified several host response modifiers and host range genes that have sites under diversifying selection. These may be sites that are involved in host-specific interactions, demonstrating adaptation to the virus's particular host. Previous studies have also demonstrated diversifying selection in poxvirus host response modifying genes [19-21]. The current study uses the largest dataset providing the most comprehensive analysis of diversifying selection in poxvirus genomes. Among the host response modifiers identified were several secreted immunomodulators (IL-18 binding protein, IL-1 beta receptor and IFN gamma receptor) that are known to contribute to virulence [22-24]. Genes that modulate the cellular response to infection were also identified, including the mitochondrial associated apoptosis inhibitor, an inhibitor of the protein kinase R (PKR) response to double stranded RNA, a ubiquitin ligase, and an inhibitor of Toll-like receptor signaling [25-28].

In addition to host response modifying genes, we also showed, using model M8, that genes involved in viral replication and virion structure experience diversifying selection, possibly due to their protein products being packaged in the virion. A proteomics study [17] identified 75 viral proteins in the VACV virion, 18 of which are present at abundances greater than 1% of the weight of the virion. Of those, 6 were identified in the current study, including the major core protein (A4), the most abundant protein in the virion by weight. Further, the immune response to poxvirus infection induces the formation of neutralizing antibodies to several virion membrane proteins [29,30]. We identified 3 known major targets of neutralizing antibodies: an immature virion membrane protein (A7, CPXV-BR-154), carbonic anhydrase (D8, CPXV-BR-129) and an enveloped virion protein (B5, CPXV-BR-205). Thus detection of diversifying selection is not limited to host response modifying genes but more broadly to genes whose products interact with the host, such as major antigens.

Designation of families in the VOCs database was based on a BLAST expect value of 10^{-17}, thus families are likely to be composed of orthologs [2]. Horizontal gene transfer (HGT) is recognized as a factor in evolution of poxvirus genomes. It is possible that some of the genes within a gene family may not be orthologs if they arose through multiple independent horizontal gene transfer events. Several lines of evidence are needed to demonstrate HGT, including phylogenetic clustering of the gene in taxa unrelated to the genome under study. The origins of most poxvirus genes are unknown, although most chordopoxvirus genes show greater similarity to eukaryotic genes than to other viral genes [31]. Full phylogenies of each poxvirus gene family are needed to determine if multiple horizontal gene transfer events have occurred within the family. Such research is underway and will be valuable in further interpretation of the current data.

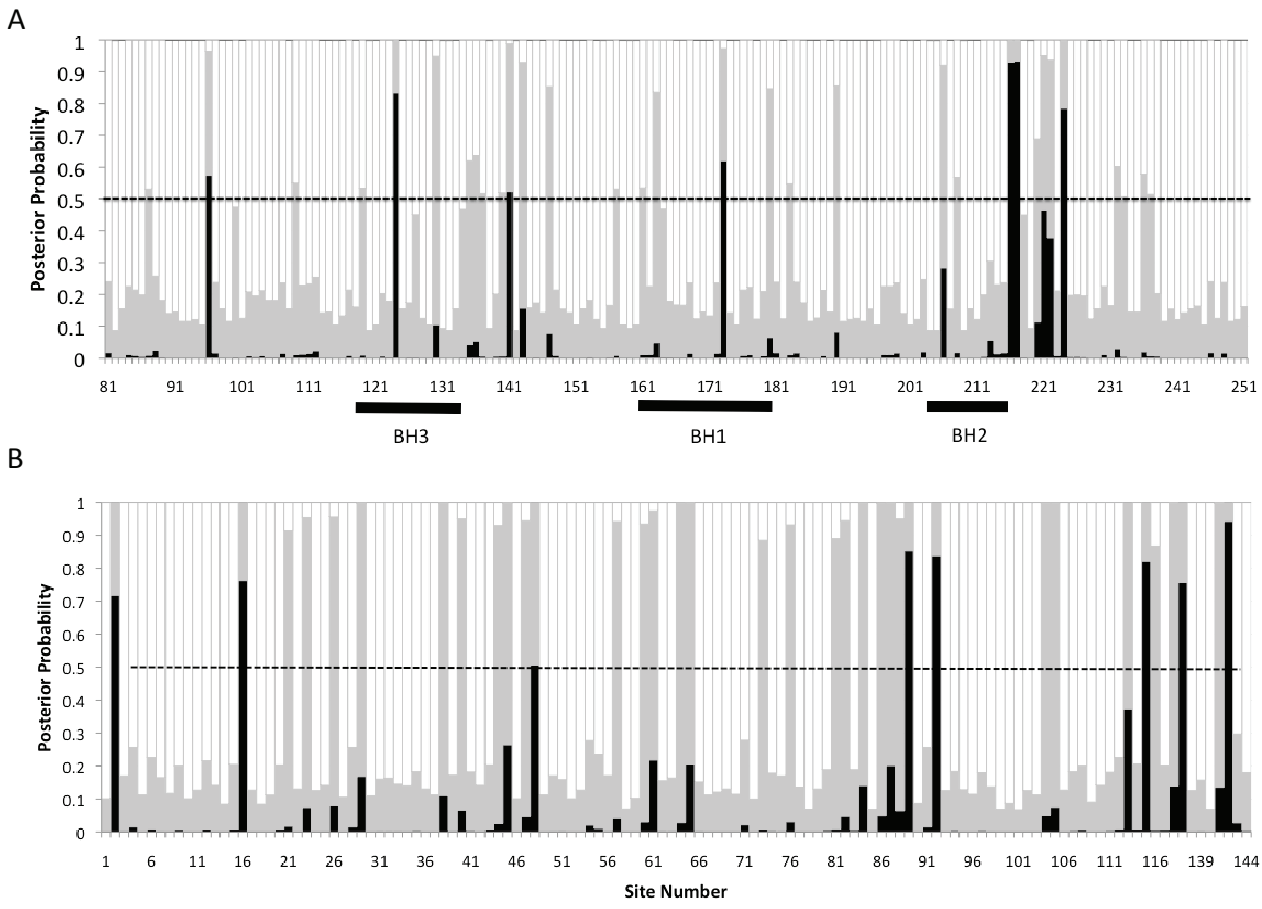

Figure 4 Posterior probability of site classes for amino acids in two host response modifier genes . Shown for each site is the posterior probability of falling into the purifying (white), neutral (grey), or diversifying (black) class. Posterior probabilities were determined by Bayes empirical Bayes analysis. Dashed line at 0.5 shows the threshold for a site to be classified as diversifying. A) IL-18 binding protein. B) mitochondrial associated apoptosis inhibitor. Only the conserved C-terminal domain is shown, and the locations of the BH domains are indicated below the axis.

Like highly conserved amino acids, variable positions under diversifying selection may indicate functionally or structurally important positions. Computational identification of amino acids under diversifying selection has revealed important functional or antigenic sites in several studies [32,33]. One of the genes identified in this study was the Interleukin-18 binding protein (IL-18BP), which was previously shown to attenuate the immune response in mice [22]. The crystal structure of the Ectromelia virus (ECTV) IL-18 binding protein was recently determined [34]. Bayes empirical Bayes analysis demonstrated that most residues are under purifying selection, while some are neutral and a few are under diversifying selection (Figure 4). The crystal structure and previous mutagenesis studies [35] identified contact residues important in binding the ligand, IL-18. Most of these residues were found to be under purifying selection, supporting a requirement for conservation of these residues to maintain function. Interestingly, 2 residues shown through crystallography to be contacts in the binding interface were also identified as being under diversifying selection (D48 and I115 in CPXV-BR, E48 and L115 in ECTV). Both of these residues interact with binding site C on human IL-18 [34]. Evidence of diversifying selection in these positions may suggest a role for these residues in adaptation to IL-18 of the specific host species of the virus.

Another important host response modifying gene that we identified is the mitochondrial associated apoptosis inhibitor (F1L in VACV-Cop). Apoptosis is an important cellular response to infection that serves to limit viral replication through removal of infected cells. F1L inhibits apoptosis through binding to the pro-apoptotic protein Bak [36,37] thereby inhibiting the permeabilization of the mitochondrial membrane, a critical step in apoptosis. Interaction with Bak is via Bcl2-like homology (BH) domains that are highly divergent but nonetheless form characteristic BH domain folds [37]. Among the

poxviruses, the C-terminus (containing the BH domains) of the F1L family is highly conserved. There are 7 residues under diversifying selection located in this domain (Figure 4B). Of those, 2 residues identified in this study are located in the BH domains. BH3 and BH1, primarily located on alpha helices α2 and α5, make up the BH3 binding pocket responsible for binding Bak [37]. One residue (A173 CPXV) identified in this study, corresponding to A144 in the F1L homolog of Modified Vaccinia Ankara (MVA), is located in the binding pocket and was shown by mutagenesis to increase binding affinity if mutated to phenylalanine. M124 in CPXV (I95 in MVA) is located in α2 and therefore could be involved in ligand binding or in determining the shape of the pocket. Another 3 sites under diversifying selection are located immediately C-terminal to BH2. Overall, the capacity of the Bayes empirical Bayes analysis to identify residues known to be important in protein function, such as in the IL-18BP and mitochondrial associated apoptosis inhibitor, suggests that it may be valuable to test other predicted sites across the genome for their role in protein function.

The identification of host-interacting genes in poxviruses as ones experiencing adaptive molecular evolution is consistent with the findings of several other studies identifying genes involved in the "host-pathogen arms race" or other co-evolutionary processes and is seen throughout nature. Chitinase and other plant defense proteins show evidence of diversifying selection, and mutagenesis studies have confirmed the functional importance of the identified sites [32,38]. The *wsp* gene of the bacterium *Wolbachia*, which encodes an outer membrane protein, shows evidence of diversifying selection when in a parasitic relationship with arthropods, but not in a mutualistic relationship with nematodes [39]. In other viruses, all the major genes of HIV [40,41], and the capsid protein of Foot-and-Mouth Disease virus (FMDV) [33] experience diversifying selection. Importantly, the amino acids identified computationally in the FMDV capsid are known to be antigenic sites identified by monoclonal antibody escape mutants [33]. In host species, diversifying selection has been shown in the antigen recognition sites of the major histocompatibility (MHC) gene [42,43].

As has been demonstrated with the FMDV capsid protein [33], interaction between an antigen and the immune response can drive diversifying selection. Several studies have found evidence of diversifying selection in surface proteins on viruses, including HIV env, influenza hemagglutinin, and others [44-46]. Recent proteomics studies have identified the major and minor virion-associated proteins [16,17]. In this study, 39% of the genes identified are associated with the virion (Figure 3). Virion structural components and virion-associated enzymes may have greater exposure to antibody responses, leading to diversifying selection.

A few small whole genomes have been analyzed for diversifying selection, showing that diversifying selection can play strikingly different roles in the molecular evolution of organisms. A study of *Picornaviridae* shows evidence of diversifying selection in structural proteins but not in non-structural proteins [47]. Only a few codons in the astrovirus genome are under diversifying selection [48], while 9-38% of sites in human rhinoviruses experience diversifying selection [49]. The poxvirus genome is significantly larger than any other viral genome analyzed by this method, and the results indicate a more important role for diversifying selection in poxvirus genome evolution than in other viruses. This may be related to the large size of the poxvirus genome and the large number of accessory "non-essential" genes such as the host-response modifiers or the large number of encapsidated proteins. Over 1700 genes of the *Streptococcus* genome were analyzed and approximately 8% of the genes were found to be under diversifying selection [50]. Of those, 29% are related to virulence, and many others show tissue specific expression during invasive disease. A large fraction of poxvirus genes under diversifying selection are also known to be virulence factors, echoing the findings in *Streptococcus*.

In *Streptococcus*, several essential core function genes were identified, indicating that virulence is more complex than simply the presence of pathogen associated genes [50]. Similarly, analysis of *E. coli* genomes found 29 genes common to pathogenic and non-pathogenic strains which showed evidence of diversifying selection, and many of which are involved in functions like DNA metabolism and nutrient acquisition [51]. Several core poxvirus genes, not typically thought of as virulence factors, were also found to be under diversifying selection in the poxvirus genome. Some (but not all) of these are packaged in the virion and may be exposed to the host antibody response. However, this suggests that other poxvirus systems, in addition to manipulation of the host response, may be important in virulence. In attempting to explain virulence differences between strains of poxviruses, it may therefore be important to consider not only major genomic differences such as gene complement, but also diversifying selection in well conserved genes.

Conclusions

The identification of 79 genes in the orthopoxviruses that experience diversifying selection implicates this as a major mechanism of poxvirus evolution. Because many of these genes either interact with host defense mechanisms or may be targets of the immune response, interaction with the host may be the basis for adaptive

molecular evolution. Understanding the mechanisms of poxvirus evolution may shed light on important aspects of poxvirus biology such as adaptation of monkeypox virus to become more sustainable in human populations. Combined with experimental data, identification of sites under diversifying selection may be important in guiding future mutagenesis studies, as these residues may be important in host specific interactions such as protein-protein contacts or immune epitopes. Further, many genes with unknown or poorly defined functions were shown to have sites under diversifying selection. These data may provide a basis for investigating the function of these as host-interacting proteins or virulence factors, and may identify highly conserved functional residues or diverse host-specific residues that are of particular interest.

Methods

Sequence selection and alignment

A subset of viruses of the Chordopoxvirus subfamily was selected based on close phylogenetic relationships [4]. Fully sequenced genomes of the Orthopoxvirus, Leporipoxvirus, Yatapoxvirus, Capripoxvirus, Suipoxvirus and Mule Deer pox genera were used (Additional file 1, Table S1). Other Chordopoxvirus genera were excluded because their sequences were typically too divergent to generate reliable alignments and have very different % GC content compared to other poxviruses. Gene families were defined by the Viral Orthologous Clusters (VOCs) database [2]. CPXV-BR virus was used as a reference genome; all gene positions used are those in CPXV-BR, and only gene families with a member in CPXV-BR were used in the analysis. Gene families with duplicated genes (paralogs) and gene families that contained less than 6 genes were excluded.

All protein and nucleotide sequences were acquired from the VOCs database. Using perl scripts, identical nucleotide sequences and their corresponding amino acid sequences were removed and then parsed into their respective gene families. Codon based alignments of nucleic acid sequences of gene families were generated using PRANK [52]. PRANK has been shown to generate more reliable alignments resulting in fewer false positives than other alignment programs [53].

Tree construction

Maximum likelihood trees were constructed from the nucleotide alignments using the DNAml executable within the PHYLIP program. The trees were used as starting points for subsequent analysis using PAML4.

Detection of evolutionary pressures with PAML4

Each gene family was analyzed using the codeml program within the PAML4 package [54] to assess the selective forces acting on the genes. Codon based models were used since they have the greatest power to detect differing selective pressures acting on a protein [10,11]. To identify the presence of a codon site class under diversifying selection, two separate site model comparisons were used. In the first, the null model M1a, which allows for two site categories, $\omega = 0$ and $\omega = 1$, was compared to the alternative model M2a, that has an additional site class in which ω is allowed to exceed 1. In the second comparison, the null model (M7) that allows for sites with $0 < \omega < 1$, was compared to the alternative model (M8) that allows for an additional site class in which ω may exceed 1. Each gene family was analyzed on a 156 core cluster composed of 2.2GHz AMD Opteron cores. The computational time taken for each gene family varied by the number of sequences per gene family and the length of the sequences and ranged from several minutes to 31 days. Perl scripts were then used to compile results.

Differences in log likelihood values between null and alternative hypotheses were analyzed using the Likelihood Ratio Test (LRT) and the resulting values compared to appropriate chi squared values to determine significance, using a confidence threshold of $p < 0.05$ and $df = 1$.

Bayes empirical Bayes

For genes in which the alternative model showed a significantly better fit of the data, Bayes theorem [18] was applied to calculate the posterior probability that each site belongs to a particular ω class. This was performed through the codeml executable of the PAML4 package [54].

Additional material

Additional File 1: Viral genomes used in this analysis.

Additional File 2: Gene families excluded from analysis.

Additional File 3: Basic statistics of genes used in analysis.

Additional File 4: Genes under diversifying selection identified by model M8 only.

Additional File 5: Proportions and omega values of significant genes.

Additional File 6: Sites in significant genes (models M2a and M8) under diversifying selection determined by Bayes empirical Bayes analysis.

Additional File 7: Sites in significant genes (model M8 only) under diversifying selection determined by Bayes Empirical Bayes analysis.

Acknowledgements

The authors would like to acknowledge Greg Priest-Dorman, Phil Tully and Marc Smith (Vassar College) for assistance with the server and for writing perl scripts, and Brian Trapp for writing perl scripts. This work was supported by start-up funds from Vassar College.

Authors' contributions

DJE conceived of the study, participated in its design and coordination, carried out the analysis and drafted the manuscript. APH participated in the design of the study, carried out the analysis, and helped to draft the manuscript. All authors have read and approved the final manuscript.

Received: 9 July 2010 Accepted: 23 May 2011 Published: 23 May 2011

References

1. Moss B: **Poxviridae: The Viruses and Their Replication.** In *Fields Virology*.. 5 edition. Edited by: Knipe DM,Howley PM,Griffin DG, et al. Philadelphia: Lippincott Williams 2007:.
2. Upton C, Slack S, Hunter AL, Ehlers A, Roper RL: **Poxvirus orthologous clusters: toward defining the minimum essential poxvirus genome.** *J Virol* 2003, **77(13)**:7590-600.
3. Gubser C, Hué S, Kellam P, Smith GL: **Poxvirus genomes: A phylogenetic analysis.** *J Gen Virol* 2004, **85(1)**:105-117.
4. Lefkowitz EJ, Wang C, Upton C: **Poxviruses: past, present and future.** *Virus Res* 2006, **117(1)**:105-18.
5. Seet BT, Johnston JB, Brunetti CR, Barrett JW, Everett H, Cameron C, Sypula J, Nazarian SH, Lucas A, McFadden G: **Poxviruses and immune evasion.** *Annu Rev Immunol* 2003, **21**:377-423.
6. Perdiguero B, Esteban M: **The interferon system and vaccinia virus evasion mechanisms.** *Journal of Interferon and Cytokine Research* 2009, **29(9)**:581-598.
7. Taylor JM, Barry M: **Near death experiences: Poxvirus regulation of apoptotic death.** *Virology* 2006, **344(1)**:139-150.
8. Hughes AL, Friedman R: **Poxvirus genome evolution by gene gain and loss.** *Mol Phylogenet Evol* 2005, **35(1)**:186-95.
9. Nielsen R, Yang Z: **Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene.** *Genetics* 1998, **148(3)**:929-936.
10. Yang Z: **Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution.** *Mol Biol Evol* 1998, **15(5)**:568-573.
11. Yang Z, Nielsen R, Goldman N, Pedersen AK: **Codon-substitution models for heterogeneous selection pressure at amino acid sites.** *Genetics* 2000, **155(1)**:431-449.
12. Anisimova M, Bielawski JP, Yang Z: **Accuracy and power of Bayes prediction of amino acid sites under positive selection.** *Mol Biol Evol* 2002, **19(6)**:950-958.
13. **The Virus Bioinformatics Resource Center.** [http://www.biovirus.org].
14. Esteban DJ, Buller RML: **Ectromelia virus: The causative agent of mousepox.** *J Gen Virol* 2005, **86(10)**:2645-2659.
15. Gubser C, Smith GL: **The sequence of camelpox virus shows it is most closely related to variola virus, the cause of smallpox.** *J Gen Virol* 2002, **83(Pt 4)**:855-72.
16. Yoder JD, Chen TS, Gagnier CR, Vemulapalli S, Maier CS, Hruby DE: **Pox proteomics: Mass spectrometry analysis and identification of Vaccinia virion proteins.** *Virology Journal* 2006, **3**:10.
17. Chung C, Chen C, Ho M, Huang C, Liao C, Chang W: **Vaccinia virus proteome: Identification of proteins in vaccinia virus intracellular mature virion particles.** *J Virol* 2006, **80(5)**:2127-2140.
18. Yang Z, Wong WSW, Nielsen R: **Bayes empirical Bayes inference of amino acid sites under positive selection.** *Mol Biol Evol* 2005, **22(4)**:1107-1118.
19. McLysaght A, Baldi PF, Gaut BS: **Extensive gene gain associated with adaptive evolution of poxviruses.** *Proc Natl Acad Sci USA* 2003, **100(26)**:15655-15660.
20. Elde NC, Child SJ, Geballe AP, Malik HS: **Protein kinase R reveals an evolutionary model for defeating viral mimicry.** *Nature* 2009, **457(7228)**:485-489.
21. Hughes AL: **Origin and evolution of viral interleukin-10 and other DNA virus genes with vertebrate homologues.** *J Mol Evol* 2002, **54(1)**:90-101.
22. Born TL, Morrison LA, Esteban DJ, VandenBos T, Thebeau LG, Chen N, Spriggs MK, Sims JE, Buller RML: **A poxvirus protein that binds to and inactivates IL-18, and inhibits NK cell response.** *Journal of Immunology* 2000, **164(6)**:3246-3254.
23. Spriggs MK, Hruby DE, Maliszewski CR, Pickup DJ, Sims JE, Buller RML, VanSlyke J: **Vaccinia and cowpox viruses encode a novel secreted interleukin-1-binding protein.** *Cell* 1992, **71(1)**:145-152.
24. Sakala IG, Chaudhri G, Buller RM, Nuara AA, Bai H, Chen N, Karupiah G: **Poxvirus-encoded gamma interferon binding protein dampens the host immune response to infection.** *J Virol* 2007, **81(7)**:3346-3353.
25. Brandt TA, Jacobs BL: **Both carboxy- and amino-terminal domains of the vaccinia virus interferon resistance gene, E3L, are required for pathogenesis in a mouse model.** *J Virol* 2001, **75(2)**:850-856.
26. Wasilenko ST, Stewart TL, Meyers AF, Barry M: **Vaccinia virus encodes a previously uncharacterized mitochondrial-associated inhibitor of apoptosis.** *Proc Natl Acad Sci USA* 2003, **100(24)**:14345-50.
27. Huang J, Huang Q, Zhou X, Shen MM, Yen A, Yu SX, Dong G, Qu K, Huang P, Anderson EM, Daniel-Issakani S, Buller RM, Payan DG, Lu HH: **The poxvirus p28 virulence factor is an E3 ubiquitin ligase.** *J Biol Chem* 2004, **279(52)**:54110-6.
28. Harte MT, Haga IR, Maloney G, Gray P, Reading PC, Bartlett NW, Smith GL, Bowie A, O'Neill LA: **The poxvirus protein A52R targets Toll-like receptor signaling complexes to suppress host defense.** *J Exp Med* 2003, **197(3)**:343-51.
29. Moss B: **Smallpox vaccines: Targets of protective immunity.** *Immunol Rev* 2011, **239(1)**:8-26.
30. Davies DH, Molina DM, Wrammert J, Miller J, Hirst S, Mu Y, Pablo J, Unal B, Nakajima-Sasaki R, Liang X, Crotty S, Karem KL, Damon IK, Ahmed R, Villarreal L, Felgner PL: **Proteome-wide analysis of the serological response to vaccinia and smallpox.** *Proteomics* 2007, **7(10)**:1678-1686.
31. Odom MR, Curtis Hendrickson R, Lefkowitz EJ: **Poxvirus protein evolution: Family wide assessment of possible horizontal gene transfer events.** *Virus Res* 2009, **144(1-2)**:233-249.
32. Bishop JG: **Directed mutagenesis confirms the functional importance of positively selected sites in polygalacturonase inhibitor protein.** *Mol Biol Evol* 2005, **22(7)**:1531-1534.
33. Haydon DT, Bastos AD, Knowles NJ, Samuel AR: **Evidence for positive selection in foot-and-mouth disease virus capsid genes from field isolates.** *Genetics* 2001, **157(1)**:7-15.
34. Krumm B, Meng X, Li Y, Xiang Y, Deng J: **Structural basis for antagonism of human interleukin 18 by poxvirus interleukin 18-binding protein.** *Proc Natl Acad Sci USA* 2008, **105(52)**:20711-20715.
35. Esteban DJ, Buller RML: **Identification of residues in an orthopoxvirus interleukin-18 binding protein involved in ligand binding and species specificity.** *Virology* 2004, **323(2)**:197-207.
36. Campbell S, Hazes B, Kvansakul M, Colman P, Barry M: **Vaccinia virus F1L interacts with Bak using highly divergent Bcl-2 homology domains and replaces the function of Mcl-1.** *J Biol Chem* 2010, **285(7)**:4695-4708.
37. Kvansakul M, Yang H, Fairlie WD, Czabotar PE, Fischer SF, Perugini MA, Huang DCS, Colman PM: **Vaccinia virus anti-apoptotic F1L is a novel Bcl-2-like domain-swapped dimer that binds a highly selective subset of BH3-containing death ligands.** *Cell Death Differ* 2008, **15(10)**:1564-1571.
38. Bishop JG, Dean AM, Mitchell-Olds T: **Rapid evolution in plant chitinases: Molecular targets of selection in plant-pathogen coevolution.** *Proc Natl Acad Sci USA* 2000, **97(10)**:5322-5327.
39. Jiggins FM, Hurst GDD, Yang Z: **Host-symbiont conflicts: Positive selection on an outer membrane protein of parasitic but not mutualistic Rickettsiaceae.** *Mol Biol Evol* 2002, **19(8)**:1341-1349.
40. Zanotto PMDA, Kallas EG, De Souza RF, Holmes EC: **Genealogical evidence for positive selection in the nef gene of HIV-1.** *Genetics* 1999, **153(3)**:1077-1089.
41. Yang W, Bielawski JP, Yang Z: **Widespread adaptive evolution in the human immunodeficiency virus type 1 genome.** *J Mol Evol* 2003, **57(2)**:212-221.
42. Hughes AL, Hughes MK, Howell CY, Nei M: **Natural selection at the class II major histocompatibility complex loci of mammals.** *Philosophical transactions of the Royal Society of London.Series B: Biological sciences* 1994, **346(1317)**:359-366.
43. Yang Z, Nielsent R: **Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages.** *Mol Biol Evol* 2002, **19(6)**:908-917.
44. Endo T, Ikeo K, Gojobori T: **Large-scale search for genes on which positive selection may operate.** *Mol Biol Evol* 1996, **13(5)**:685-690.
45. Bush RM, Fitch WM, Bender CA, Cox NJ: **Positive selection on the H3 hemagglutinin gene of human influenza virus A.** *Mol Biol Evol* 1999, **16(11)**:1457-1465.
46. Yamaguchi-Kabata Y, Gojobori T: **Reevaluation of amino acid variability of the human immunodeficiency virus type 1 gp120 envelope glycoprotein**

and prediction of new discontinuous epitopes. *J Virol* 2000, **74(9)**:4335-4350.
47. Simmonds P: **Recombination and selection in the evolution of picornaviruses and other mammalian positive-stranded RNA viruses.** *J Virol* 2006, **80(22)**:11124-11140.
48. Van Hemert FJ, Lukashov VV, Berkhout B: **Different rates of (non-) synonymous mutations in astrovirus genes; correlation with gene function.** *Virology Journal* 2007, **4**.
49. Lewis-Rogers N, Bendall ML, Crandall KA: **Phylogenetic relationships and molecular adaptation dynamics of human rhinoviruses.** *Mol Biol Evol* 2009, **26(5)**:969-981.
50. Anisimova M, Bielawski J, Dunn K, Yang Z: **Phylogenomic analysis of natural selection pressure in Streptococcus genomes.** *BMC Evolutionary Biology* 2007, **7**:154.
51. Chen SL, Hung C-, Xu J, Reigstad CS, Magrini V, Sabo A, Blasiar D, Bieri T, Meyer RR, Ozersky P, Armstrong JR, Fulton RS, Latreille JP, Spieth J, Hooton TM, Mardis ER, Hultgren SJ, Gordon JI: **Identification of genes subject to positive selection in uropathogenic strains of Escherichia coli: A comparative genomics approach.** *Proc Natl Acad Sci USA* 2006, **103(15)**:5977-5982.
52. Löytynoja A, Goldman N: **Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis.** *Science* 2008, **320(5883)**:1632-1635.
53. Fletcher W, Yang Z: **The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection.** *Mol Biol Evol* 2010, **27(10)**:2257-2267.
54. Yang Z: **PAML 4: Phylogenetic analysis by maximum likelihood.** *Mol Biol Evol* 2007, **24(8)**:1586-1591.

doi:10.1186/1471-2164-12-261
Cite this article as: Esteban and Hutchinson: **Genes in the terminal regions of orthopoxvirus genomes experience adaptive molecular evolution.** *BMC Genomics* 2011 **12**:261.