

12-1-2019

# Identification and Visualization of Functionally Important Domains and Residues in Herpes Simplex Virus Glycoprotein K(gK) Using a Combination of Phylogenetics and Protein Modeling

Paul J.F. Rider  
*School of Veterinary Medicine*

Lyndon M. Coghill  
*Louisiana State University*

Misagh Naderi  
*Louisiana State University*

Jeremy M. Brown  
*Louisiana State University*

Michal Brylinski  
*Louisiana State University*

*See next page for additional authors*

Follow this and additional works at: [https://digitalcommons.lsu.edu/animalsciences\\_pubs](https://digitalcommons.lsu.edu/animalsciences_pubs)

---

## Recommended Citation

Rider, P., Coghill, L., Naderi, M., Brown, J., Brylinski, M., & Kousoulas, K. (2019). Identification and Visualization of Functionally Important Domains and Residues in Herpes Simplex Virus Glycoprotein K(gK) Using a Combination of Phylogenetics and Protein Modeling. *Scientific Reports*, 9 (1)  
<https://doi.org/10.1038/s41598-019-50490-9>

This Article is brought to you for free and open access by the School of Animal Sciences at LSU Digital Commons. It has been accepted for inclusion in Faculty Publications by an authorized administrator of LSU Digital Commons. For more information, please contact [ir@lsu.edu](mailto:ir@lsu.edu).

---

## Authors

Paul J.F. Rider, Lyndon M. Coghill, Misagh Naderi, Jeremy M. Brown, Michal Brylinski, and Konstantin G. Kousoulas

**OPEN**

# Identification and Visualization of Functionally Important Domains and Residues in Herpes Simplex Virus Glycoprotein K(gK) Using a Combination of Phylogenetics and Protein Modeling

Paul J. F. Rider<sup>1</sup>, Lyndon M. Coghill<sup>2,3</sup>, Misagh Naderi<sup>2</sup>, Jeremy M. Brown<sup>2,4</sup>, Michal Brylinski<sup>2,3</sup> & Konstantin G. Kousoulas<sup>1</sup>

Alphaherpesviruses are a subfamily of herpesviruses that include the significant human pathogens herpes simplex viruses (HSV) and varicella zoster virus (VZV). Glycoprotein K (gK), conserved in all alphaherpesviruses, is a multi-membrane spanning virion glycoprotein essential for virus entry into neuronal axons, virion assembly, and pathogenesis. Despite these critical functions, little is known about which gK domains and residues are most important for maintaining these functions across all alphaherpesviruses. Herein, we employed phylogenetic and structural analyses including the use of a novel model for evolutionary rate variation across residues to predict conserved gK functional domains. We found marked heterogeneity in the evolutionary rate at the level of both individual residues and domains, presumably as a result of varying selective constraints. To clarify the potential role of conserved sequence features, we predicted the structures of several gK orthologs. Congruent with our phylogenetic analysis, slowly evolving residues were identified at potentially structurally significant positions across domains. We found that using a quantitative measure of amino acid rate variation combined with molecular modeling we were able to identify amino acids predicted to be critical for gK protein structure/function. This analysis yields targets for the design of anti-herpesvirus therapeutic strategies across all alphaherpesvirus species that would be absent from more traditional analyses of conservation.

Herpesviruses are a significant cause of morbidity and mortality in human and animal populations. To date there are greater than 100 species of viruses recognized in the order herpesvirales including 8 human herpesviruses<sup>1–3</sup>. Mammal, bird, and reptile herpesviruses constitute the family *Herpesviridae* which is divided into alpha, beta, or gamma, subfamilies based on their biological properties, and gene content<sup>1,3,4</sup>. There are approximately 40 genes shared among all subfamilies of herpesvirus, referred to as the core herpesvirus genes<sup>5,6</sup>. These genes encode structural components of the viral nucleocapsid, as well as proteins with functions related to aspects of the viral life cycle that are common to all members of the *Herpesviridae*, such as DNA replication. In addition to the core set of herpesvirus genes, there are genes specific to members of each subfamily. These genes are expected to be responsible for the biological properties of each subfamily that distinguish it from other subfamilies<sup>7</sup>.

<sup>1</sup>Division of Biotechnology and Molecular Medicine and Department of Pathobiological Sciences, School of Veterinary Medicine, Louisiana State University, Baton Rouge, LA, USA. <sup>2</sup>Department of Biological Sciences, Louisiana State University, Baton Rouge, LA, USA. <sup>3</sup>Center for Computation & Technology, Louisiana State University, Baton Rouge, LA, USA. <sup>4</sup>Museum of Natural Science, Louisiana State University, Baton Rouge, LA, USA. Paul J. F. Rider, Lyndon Coghill and Misagh Naderi contributed equally. Correspondence and requests for materials should be addressed to K.G.K. (email: [vtgusk@lsu.edu](mailto:vtgusk@lsu.edu))

Received: 17 January 2019  
 Accepted: 12 September 2019  
 Published online: 10 October 2019

Alphaherpesviruses are distinguished from the beta and gamma subfamilies by their ability to establish and maintain latency in the central and/or peripheral nervous systems of the host<sup>7</sup>. After the establishment of latency, these viruses will periodically reactivate and spread to sites of innervation causing clinical symptoms of herpesvirus infection. The ability of alphaherpesviruses to enter cells of the nervous system is central to the pathogenesis of infection by these viruses and to date remains poorly understood. One strategy for identifying genes involved in infection of the nervous system is to focus on those genes unique to alphaherpesviruses. While several of these alphaherpesvirus-specific genes have been implicated in neurovirulence, our understanding of how and which of them facilitate entry, establishment and maintenance of latent infection in cells of the nervous system is incomplete.

gK and UL20 are highly conserved among neurotropic alphaherpesviruses. gK and UL20 have been shown to interact and these interactions are important for virus entry, membrane fusion, cytoplasmic virion envelopment and egress<sup>8–19</sup>. Mechanistically, the gK/UL20 protein complex interacts with the viral fusogen gB and this interaction is necessary for gB-mediated membrane fusion during virus entry, as well as virus-induced cell-to-cell fusion and the formation of syncytia<sup>20,21</sup>. HSV-1 gK is expressed on virion envelopes and functions in virion entry<sup>16,22,23</sup>. Specifically, we have shown that a 39 amino acid N-terminal deletion of gK inhibits virus entry into neuronal axons *in vitro* and *in vivo*, but not other cell types (fibroblasts, epithelial cells)<sup>24–26</sup>.

HSV-1 is the cause of oral herpes and is the prototypical alphaherpesvirus<sup>7</sup>. gK and UL20 are both present in the virion envelope<sup>16,22,27</sup>. Previously, we have shown that gK has three extracellular and two intracellular domains separated by four transmembrane domains<sup>9</sup>. Mutations of gK result in viruses with domain-specific defects in entry, assembly, or egress<sup>17,19,28–31</sup>. These data suggest the domains of gK are functionally distinct, with some involved in entry and others in assembly and egress. At present, the mechanism by which gK functions to facilitate membrane fusion during virus entry and spread, as well as cytoplasmic virion envelopment, is not well understood and is the focus of ongoing work.

Since domains of gK are functionally distinct, we hypothesized that they may exhibit variation in evolutionary rate. Such variation is likely driven by several, interacting factors. First, virion assembly, like DNA replication, may be expected to involve primarily conserved viral protein-protein interactions, while entry is partly the result of interactions between viral entry mediators with host receptor proteins. Viral protein domains on the surface of virions may evolve faster than internal virion proteins<sup>32</sup>. Additional selective pressure may be expected to occur across a virus family interacting with divergent host species, since they may be subject to diversifying selection imposed by co-evolution with hosts<sup>33</sup>. Extracellular domains (1, 3, and 5) are more likely to be involved in entry as well as be exposed to host antibody binding therefore these domains may be expected to evolve more quickly than intracellular domains (2 and 4). Second, some extramembrane domains may exhibit more structural restraints, while others could be less ordered. Elastic network models allow us to estimate the residue fluctuations, or mobility. We might expect that ordered domains with less fluctuation (e.g., domains 1 and 2; see below) should evolve more slowly than those with less well-defined structures and high fluctuation (e.g., domain 4). Lastly, transmembrane domains need to exhibit a stable core that maintains the integrity of a helical bundle, while residues on the sides of the helical bundle need to remain hydrophobic due to their interactions with membrane phospholipids. Transmembrane domains may also interact with other viral proteins (e.g., UL20)<sup>13,34,35</sup>. Given these constraints, they are generally expected to evolve slowly.

The various factors outlined above may result in contrasting predictions about a domain's expected evolutionary rate. For instance, an intracellular domain of gK is likely to be involved in virion assembly. Given this role, the domain should have experienced a relatively stable set of interactions during the diversification of alphaherpesviruses, and would be expected to evolve slowly in order to maintain the fidelity of these interactions. However, structural predictions may suggest that the same domain also has a less well-defined structure than other parts of gK, requiring less precision in folding to maintain function. In this case, the domain's primary sequence may evolve relatively quickly. The contrasts between these predictions allow us to elucidate the relative strength of the predictions.

Overall, our integrated phylogenetic and structural analyses provide new insights in the location and potential functions of gK protein domains. This approach can be extended to the prediction of functional domains and residues of other proteins and glycoproteins, including domains that can be potential targets for antiviral therapy for multiple evolutionary related pathogens.

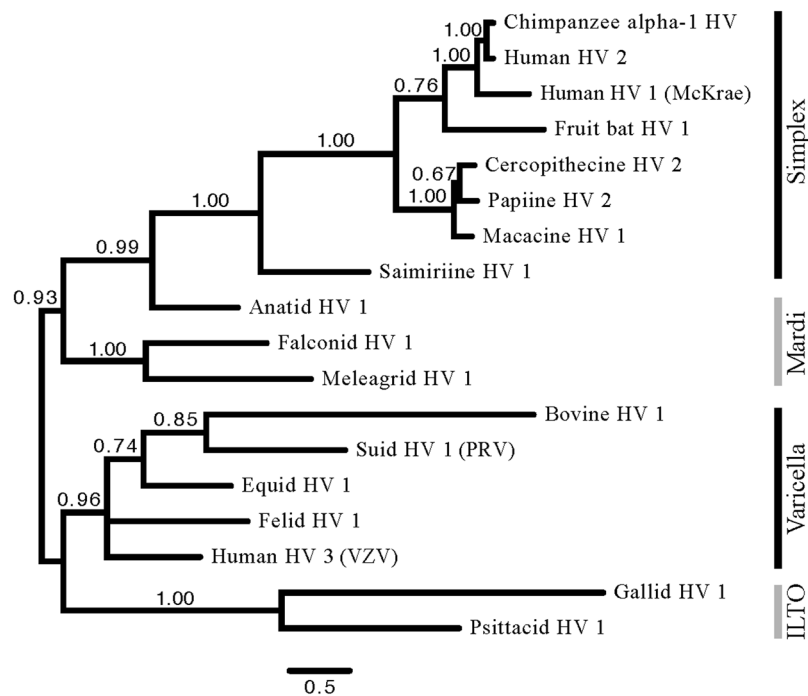
## Results

**Phylogenetic analysis and the relative evolutionary rate of individual gK domains.** We inferred the phylogenetic relationships among 18 species of alphaherpesviruses using a set of genes orthologous to HSV-1 gK (Table 1, Fig. 1). These analyses employed a model of rate variation with a fixed partitioning scheme, including a separate rate multiplier for each gK domain. The topology of the inferred tree is similar to previously published relationships based on whole genome sequences or gene content comparisons<sup>3,36,37</sup>.

To test the hypothesis that domains vary substantially in evolutionary rate based on function, we inferred these rates for each of the 5 extramembrane and four transmembrane domains (Fig. 2A) across the 18 gK orthologs listed in Table 1. To determine the locations of the transmembrane domains for the HSV-1 gK orthologs, sequences were submitted to tmfoldweb<sup>9</sup>. With one exception, all gKs were predicted by tmfoldweb to possess 4 transmembrane domains and five extramembrane domains with a similar structure to HSV-1 gK. The only exception was PRV gK, which was predicted to have 6 extramembrane domains and 5 transmembrane domains. This alternative prediction was due to the separation of domain 2 into two separate domains (one cytoplasmic and one extracellular). Since all 17 other gK sequences were predicted to possess the same topology as HSV-1, we combined the two domains predicted for PRV domain 2, yielding five extramembrane domains across all orthologs.

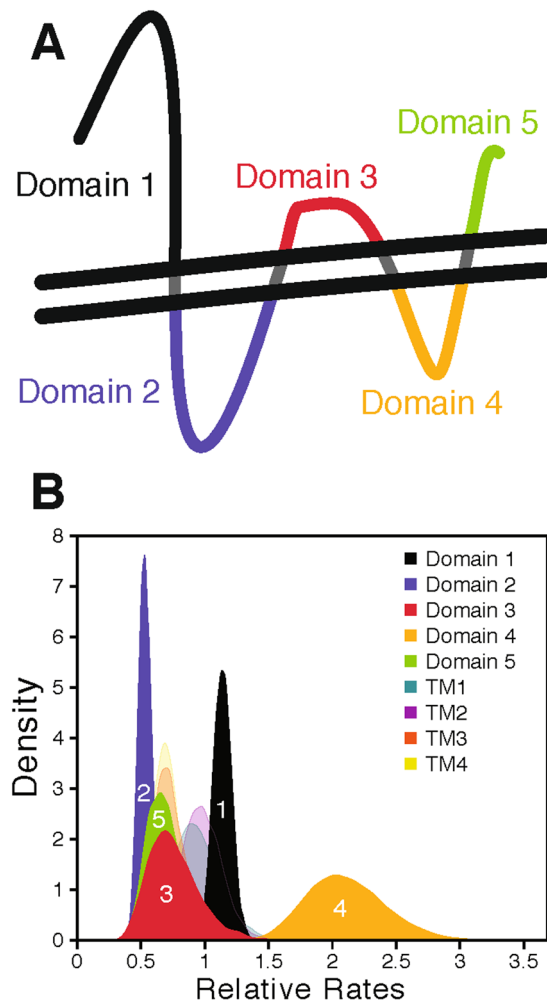
gK Sequence	Accession Number	% Identity with HSV1
Anatid herpesvirus 1	YP_003084366.1	49.79
<b>Bovine herpesvirus 1</b>	<b>NP_045305.1</b>	<b>49.58</b>
<b>Cercopithecine herpesvirus 2</b>	<b>YP_164497.1</b>	<b>71.75</b>
<b>Chimpanzee alpha-1 herpesvirus</b>	<b>YP_009011041.1</b>	<b>89.12</b>
<b>Equid herpesvirus 1</b>	<b>YP_053051.1</b>	<b>50.20</b>
Falconid herpesvirus 1*	YP_009046554.1	48.47
<b>Felid herpesvirus 1</b>	<b>YP_003331525.1</b>	<b>53.34</b>
<b>Fruit bat alphaherpesvirus 1</b>	<b>YP_009042116.1</b>	<b>76.56</b>
Gallid herpesvirus 1	YP_182337.1	32.42
<b>Human herpesvirus 2</b>	<b>YP_009137206.1</b>	<b>88.49</b>
Human herpesvirus 3 (VZV)	NP_040128.1	46.65
Macacine herpesvirus 1	AIA09548.1	72.17
Meleagrid herpesvirus 1	NP_073348.1	45.81
Papiine herpesvirus 2	YP_443901.1	70.92
Psittacid herpesvirus 1	NP_944381.1	42.05
Saimiriine herpesvirus 1	YP_003933787.1	56.27
Suid herpesvirus 1 (PRV)	YP_068321.1	39.53
<b>Human herpesvirus 1 (McKrae)</b>	<b>AFP86418</b>	<b>100.00</b>
Average		57.85

**Table 1.** Alphaherpesvirus gK orthologs analyzed in this study. Species in bold were used for genome-wide phylogenetic analyses. \*Presently not recognized as a formal virus species.



**Figure 1.** Phylogenetic analysis of alphaherpesvirus glycoprotein K. Phylogeny inferred from Bayesian phylogenetic analysis of gK orthologs from 18 alphaherpesviruses, arbitrarily rooted. Branches are labeled with posterior probabilities. Branch lengths are given as the expected number of substitutions.

We found that domain 2 had the lowest inferred rate of evolution, while domain 4 had a substantially faster rate than all others (2–4 fold faster; Table 2, Fig. 2B), despite both being intracellular. The extracellular domains (1, 3, and 5) had intermediate inferred rates, with domain 1 evolving faster than domains 3 and 5. The trans-membrane domains also had intermediate rates, roughly spanning the same range as the extracellular domains. Unsurprisingly, domains with more residues (e.g., 1 and 2) produced more precise estimates of rate (Fig. 2B).



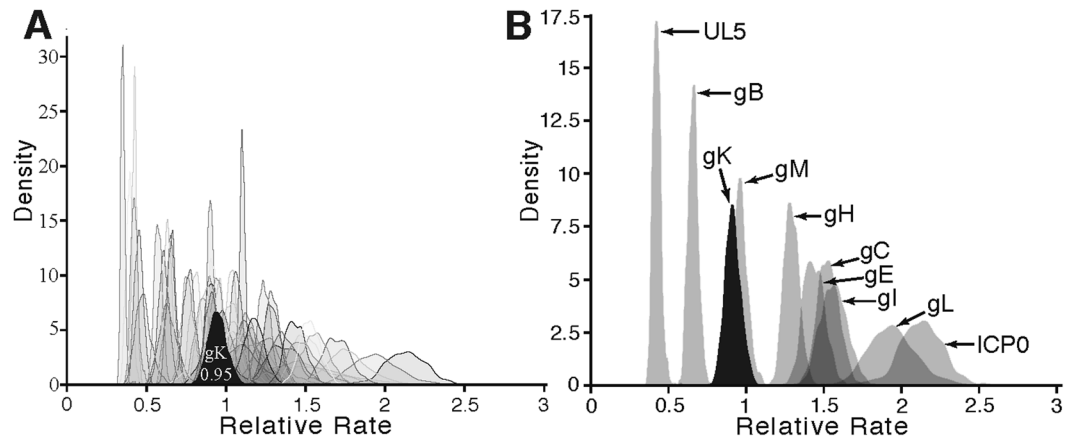
**Figure 2.** Relative rate of change for individual domains of alphaherpesvirus gK. (A) A cartoon representation of gK's topology with individual domains labeled. The two parallel horizontal lines denote cell membrane. (B) Posterior distributions of relative substitution rates for each domain in gK. "TM" indicates a transmembrane domain, which are numbered in the same direction as extramembrane domains. Colors corresponding to extramembrane domains illustrated in A.

	Domain 1	Domain 2	Domain 3	Domain 4	Domain 5	TM1	TM2	TM3	TM4
mean	1.1877	0.3377	0.6368	2.3745	0.4787	0.6938	1.0055	0.8736	0.7968
stdev	0.0953	0.0529	0.2429	0.3909	0.1509	0.1747	0.2047	0.2091	0.1661

**Table 2.** Relative rates of change for gK domains. A value of 1 is the average rate across all gK residues.

**Analysis of the relative evolutionary rate of glycoprotein K (gK).** We created a concatenated alignment of 59 orthologs present in all alphaherpes viruses and employed a partitioned model with independent rate multipliers for each gene, allowing us to compare the rate of evolution of gK relative to the rest of the genome (Fig. 3A). To improve mixing and convergence of these Markov chain Monte Carlo (MCMC) analyses, we fixed the tree topology using the majority-rule consensus from the gK results. The estimated rate multiplier for gK was 0.95, indicating that it is evolving at a rate roughly average across the 59 orthologs. Other orthologs varied substantially in their estimated rates, with multipliers ranging from roughly 0.4 to greater than 2 (Fig. 3a). This range roughly parallels the range of estimated rates across domains within gK.

Focusing only on the shared glycoproteins, gK has a rate that is second lowest after glycoprotein B, the major mediator of herpesvirus fusion (Fig. 3B). This relationship between the evolutionary rates for glycoproteins was previously demonstrated in an analysis restricted to HSV-1 and HSV-2<sup>38</sup> and holds true for comparisons among all alphaherpesvirus species in our analysis. For comparison, we included ICP0 and UL5 in these comparisons, although they are not glycoproteins. ICP0 and UL5 are known to have high and low evolutionary rates, respectively. In agreement with previous results<sup>39</sup>, ICP0 had a higher evolutionary rate than all glycoproteins (Fig. 3B) and UL5 exhibited the lowest rate of change. ICP0 is ascribed multiple, disparate functions including E3 ubiquitin



**Figure 3.** Rate of change for alphaherpesvirus gK relative to conserved alphaherpesvirus ORFs. **(A)** The posterior distribution of gK's relative substitution rate compared to those from 58 other alphaherpesvirus genes. **(B)** Posterior distributions of relative substitution rates for a subset of genes found in **(A)**, focusing on glycoproteins. UL5 (slowest) and ICP0 (fastest) are also included as references.

ligase and transactivation activities<sup>7</sup>. UL5 is involved in DNA replication, a function that is expected to exhibit a high degree of conservation regardless of host variation<sup>7</sup>.

**Analysis of the relative evolutionary rate of individual gK amino acids.** To illustrate sequence features of gK, orthologs from simplex virus genera (HSV-1, HSV-2) and varicellovirus genera (VZV and PRV) were aligned using clustal omega<sup>40,41</sup> (Fig. 4). Two important features can be seen in this alignment. First, all four of these gK sequences possess two putative N-linked glycosylation sites separated by 9–10 amino acids in domain 1. These sites occur at positions 49 and 60 in the alignment (Fig. 4) for the simplex viruses and positions 60 and 70 for the varicelloviruses. Second, domain 2, which has the lowest rate of change for all gK domains (Fig. 2), possesses multiple conserved tyrosine residues suggesting this domain may be involved in protein-protein interactions that facilitate assembly.

To develop a comprehensive and quantitative understanding of variation in evolutionary rate across individual residues, including identification of those that may be under strong constraints, we employed a novel finite mixture model (FMM) for among-site rate variation. While site rate estimates did vary somewhat as the number of mixture components changed (Figs S1–3), the rates were quite consistent across analyses using either 4 or 6 components (Fig. S3). Individual sites exhibited more than 20-fold differences in inferred rate, with a strong mode at intermediate rates (Fig. S4; tree length  $\approx 12$ ). The overall distribution of site rates was roughly symmetric, but included smaller modes at slower (tree length  $\approx 6$ ) and faster (tree length  $\approx 17$ ) rates.

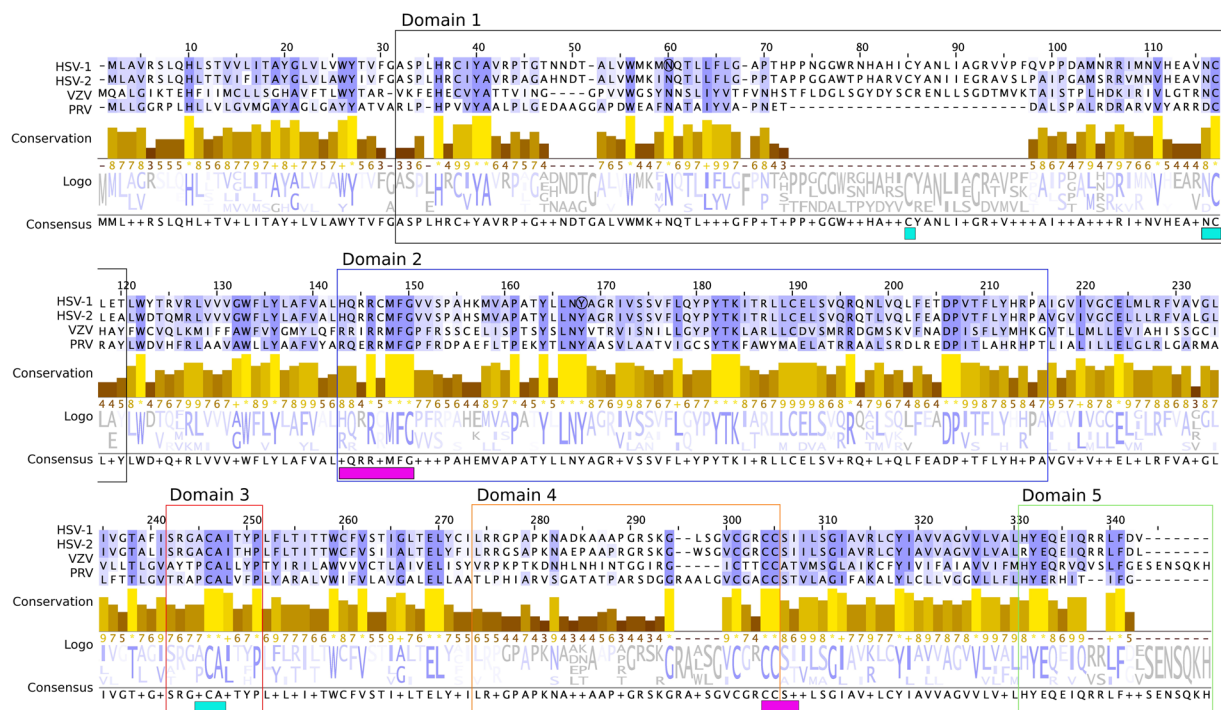
**Structural prediction of alphaherpesvirus gKs.** Structure based analyses can reveal protein features not evident from amino acid sequence alone<sup>6</sup>. These analyses are particularly valuable for comparison of gKs from the simplex and varicella genera of alphaherpesviruses that possess little amino acid sequence identity, but which are expected to have functional conservation. Additionally, gK has extensive hydrophobic regions that preclude the use of traditional methods to study structure, such as crystallization and circular dichroism. To determine the degree of structural conservation between representative varicello and simplex herpesvirus, we predicted the structures of HSV-2, VZV and PRV and compared these to our previously published HSV-1<sup>42</sup> predicted gK structure (Fig. 5). We found that all four alphaherpesvirus gKs were structurally congruent for domains that possess a discernable structure (domains 1 and 2). We have shown that mutations in domain 1 of gK render the virus incapable of entry into cells via neuronal axons<sup>26,42</sup>. Interestingly, the domain with the highest rate of change (domain 4) is also predicted to be the least ordered (Fig. 5).

To visualize conserved amino acids on the structure of HSV-1 gK, we generated a heat map with each amino acid colored according to its relative rate of change (Fig. 6). Overall, one can clearly see that spatially clustered sites often have similar rates, despite the fact that the model of sequence evolution we employed assumes that each site evolves independently. Figure 6 also shows the conservation of amino acids at the predicted interface between domains 1 and 3, as well as domains 2 and 4. Interestingly, the conserved, slow evolving residues of domain 2 form an almost mirror image across two predicted alpha helices that extend into the cytoplasm. Additionally, we found that the position of the YTK sequence of domain 2 is conserved at the tip of the first alpha helix in all alphaherpesvirus examined. We previously showed that this tyrosine is functionally important, as a virus with a Y183S mutation was severely attenuated *in vitro*<sup>17</sup>.

## Discussion

Alphaherpesviruses are significant human and animal pathogens, creating a substantial burden of morbidity and mortality across many species. They have also been subject to diverse evolutionary pressures, given their expansive host range. Future therapeutic efforts to prevent or control alphaherpesvirus infection will be most effective if they exploit viral characteristics that are evolutionarily conserved across the entire group and critical to their





**Figure 4.** Multiple sequence alignment of selected herpes simplex and varicella gKs. Amino acid sequences for gK are from HSV-1, HSV-2, VZV, and PRV. Residues are color coded according to Jalview BLOSUM62 color scheme. The consensus sequence represents the modal residue at each column and “+” signifies where the modal residue is ambiguous. Conservation scores are shown as histograms and indicate the degree of preservation of physicochemical properties for each residue. The highest score of 11 (depicted in the lightest shades and indicated with “+” below the histogram) shows fully conserved residues, while values close to 11 indicate some change in amino acid identity while maintaining physicochemical properties. Extramembrane domains are boxed and colored according to graphical representation of domains shown in Fig. 3. Magenta (Domains 2 and 4) and cyan (Domains 1 and 3) bars at the bottom represent amino acid clusters that are brought together in the folded gK structure featured in Fig. 6, circled by the same colors.

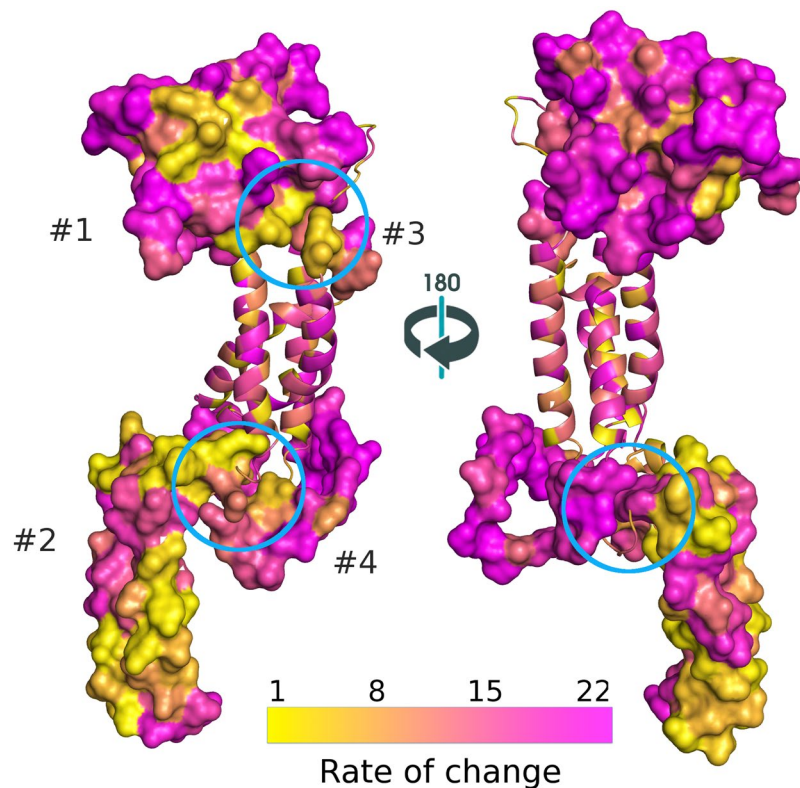
function. In this study, we employed a combination of phylogenetic and structural analyses to understand how the viral glycoprotein gK, which is required for virus entry into neuronal axons, has evolved across this diverse group of viruses and to predict which of its domains and residues are most important for maintaining critical functions.

We found that while glycoprotein K is highly conserved across all alphaherpesviruses, and it is known to perform critical roles in neuronal infection<sup>25,26,42</sup>, its rate of sequence evolution is not remarkable. Specifically, gK evolves at a rate roughly average across the herpes viral genomes examined. In contrast to its overall sequence evolution the predicted structure of gK for specific domains was predicted to be more highly conserved. The disconnect between relative rate of sequence evolution and structural conservation for gK may result from the need to adapt to diverse hosts while maintaining structural similarity. In support of this hypothesis, our combined analysis suggests that, while gK lacks correlation between structure and sequence evolution across its entirety, the correlation between protein domains with low rates of sequence evolution possess greater structural conservation.

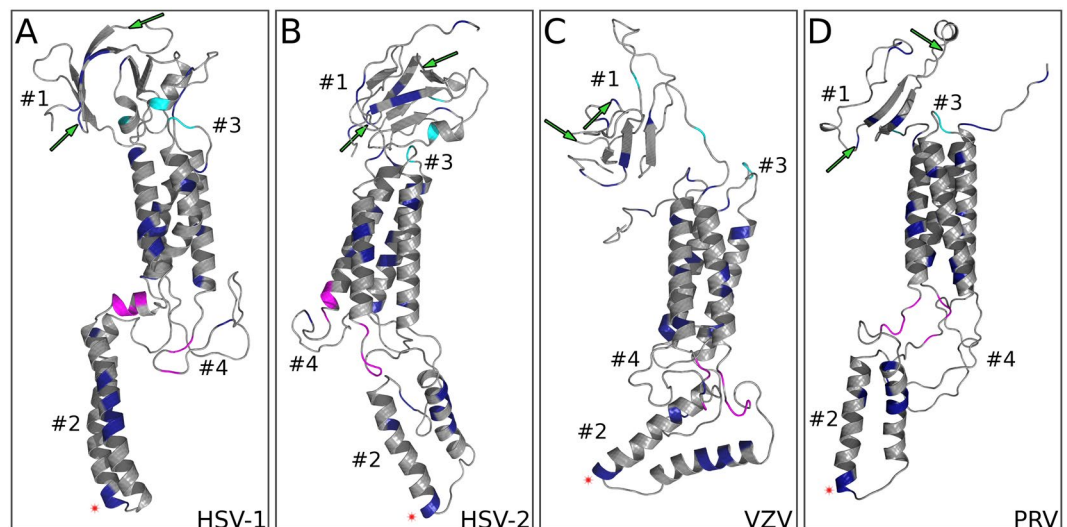
Extramembrane domains of gKs vary more in evolutionary rate than transmembrane domains, but their rate is not well predicted by their position in the protein nor their presumed function. Instead, the residue fluctuation (estimated Root Mean Square Factor (RMSF)) and the degree of ordering in their structures (Figs S5, 6) appears to be a more important predictor of a domain's average rate of evolution. Transmembrane domains (lighter colors in Fig. 3B) evolved at relatively similar rates, consistent with the expectation that they must maintain general structural features (e.g., stable helices with hydrophobic exteriors), but have flexibility in the precise identities of many amino acids. Interestingly, cytoplasmic domains (2 and 4) did not evolve consistently faster or slower than extracellular/luminal domains (1, 3, and 5). Instead domain 2 had the slowest relative rate, while domain 4 had the highest rate. This ranking suggests that the degree of ordering in a domain's structure is a more important factor in determining its rate of evolution than its probable function. The slow average rate for domain 2 also suggested that it harbored many sites constrained by selection, and that alteration of these residues may disrupt protein function more than similar changes in other domains.

Site-by-site variation in evolutionary rate was inferred to be 10-fold greater than variation in average rate across domains and genes (Figs 2, 3, 6, and S4). This result is not surprising, since averages are by definition intermediate, but it reinforces the need to examine site-specific properties in order to predict functional or evolutionary significance. Critical functions of proteins are likely to be controlled by a subset of residues, whose properties may easily be overlooked when averaged across surrounding sites. In turn, the importance of inferring





**Figure 6.** Site-specific relative rates of change mapped onto the predicted structure of HSV-1 gK. Relative rates of change for each amino acid in HSV-1 gK are mapped onto the predicted structure. Transmembrane and extramembrane domains are shown in ribbon and surface representations, respectively. Domains 1 and 3 proximal residues are marked with a cyan circle and proximal residues between domains 2 and 4 are marked with a magenta circle, corresponding to the alignment (Fig. 4). Purple and gold color scheme indicates high and low relative rate of change, respectively.



**Figure 5.** Structural predictions of selected simplex and varicella gKs. Cartoon representations of simplex (HSV-1, HSV-2), and Varicella (VZV, PRV) gK protein models are shown in (A)–(D). Residues conserved across all four sequences (Fig. 4) are shown in color. Magenta and Cyan colors correspond to the alignment (Fig. 4). Predicted N-glycosylation sites are marked by green arrows. Asterisk indicates YTK sequence in domain 2.

site-specific properties highlights the need to expand the available set of sequence evolution models. Here, we applied a novel finite mixture model (FMM) to infer site-specific rates. This model formulation has several advantages, including minimal assumptions about the distribution of rates across sites and a parameter space that

allows for efficient estimation of the posterior distribution. While our FMM is less flexible than the Dirichlet process prior model (DPP) model proposed by Huelsenbeck and Suchard (2007), our limited tests suggest that it is easier to apply and achieved stable rate estimates with a small number of categories (Fig. S3).

Despite being structurally naive, the model of sequence evolution that we applied resulted in site-specific rate estimates that are spatially clustered (Fig. 6) and often make sense in the context of a predicted functional role. For instance, despite the overall high rate of evolution inferred for domain 4, residues that are predicted to interact with domain 2 have very low site-specific rates (magenta circle in Fig. 6). The same is true for residues that interact between domains 1 and 3 (cyan circle in Fig. 6). While site-specific rate estimates can often be very informative, they also exhibit considerable uncertainty. By placing these estimates in the context of a predicted protein structure, we gain greater confidence that sites with low predicted rates are functionally important and worth exploring experimentally. Further, the rate estimates, by being structurally naïve support features of our structural models such as the potential interaction between domains.

A general region of potential functional importance was domain 2 of gK, given its overall slow rate. In structural predictions for HSV-2, VZV, and PRV, domain 2 exhibits a highly conserved double alpha helix that extends into the cytoplasm. This structure is consistent with its presumed role in virus assembly, a process conserved across alphaherpesviruses. Of specific interest are a number of conserved tyrosines within gK domain 2. Our molecular model and phylogenetic analyses indicate that these tyrosines may play important roles in how gK regulates gB-mediated membrane fusion. Similarly, cytoplasmic virion maturation is a coordinated process that is both spatially and temporally regulated<sup>7,43</sup>. These tyrosines may be involved in protein-protein interactions and/or the appropriate localization of gK to facilitate cytoplasmic virion envelopment. In this regard, we recently showed that the HSV-1 UL37 tegument protein interacts with gK via specific tyrosine residues located within conserved UL37 regions<sup>35</sup>.

While less conserved overall, other domains also contained individual residues predicted to be of functional importance. Domain 3 possesses remarkable conservation of a central cysteine residue across all alphaherpesvirus examined suggesting that this domain plays an important role in infectious virus production. Similarly, there are 3 conserved cysteines in the extracellular domain 1, which may be involved in conserved disulfide bonding between domains 1 and 2. We previously showed that mutation of the conserved C304S-307S in domain 4 resulted in a virus with attenuated growth<sup>17</sup>. Therefore, while domain 4 had the highest rate of evolution, it may still possess a conserved function. Conversely, the slow rate of evolution for domain 5 is surprising, since we previously showed that deleting this entire domain has little effect on viral replication *in vitro*<sup>17</sup>. Although deletion of domain 5 does not adversely affect virus replication *in vitro* domain 5 may be important for gK function *in vivo*. Transmembrane and near-transmembrane regions may also have conserved features to facilitate proper localization or protein-protein interactions. For instance, in our template based models of alphaherpesvirus gKs, residues near the ends of transmembrane helices are often conserved (Fig. 4), possibly due to interactions with different phospholipid composition on the two sides of the two membrane leaflets. Additionally, every helix in the transmembrane bundle exhibits a side with more stable residues towards the inside of the bundle and a more mutable side towards the membrane (Fig. 6), although we have not formally analyzed this pattern. The same general theme can be seen in Fig. 4, where more conserved residues are located in the transmembrane domains at almost every four amino acids.

In this work we focused on the identification of conserved regions in alphaherpesvirus gK. However, functionally important regions of divergence are likely to be involved in host-range specificity including virus entry, spread and overall *in vivo* pathogenicity. We have shown in previous work that the number of amino acids between N-linked glycosylation sites within gK domain 1 are correlated with host range<sup>23</sup>. For example distances of 11 amino acids are found in primate and non-human primate herpesviruses whereas distances of 8 amino acids are found in most avian herpesviruses that were examined. Additional work from our laboratory has demonstrated that mutations in gK domain 1 modify receptor specificity of HSV-1<sup>44</sup>. Finally, while simplex viruses infect and replicate in the epithelium before entering enervating neurons via their axonal termini other viruses have differing biology of infection. For example VZV is hematogenously disseminated after transfer to T-cells. After dissemination VZV infects neurons via the axonal termini or cell body<sup>45</sup>. In addition VZV lacks the receptor binding member of the receptor complex, gD, and is highly cell associated with the generation of very few free virus particles<sup>46</sup>. Using data generated in this paper, a future focus of research will be to test the function of both conserved and divergent residues of gK across multiple families of alphaherpesvirus species.

Our approach of combining evolutionary and structural analyses to predict functional sites is readily applicable to the study of other proteins and will likely complement standard approaches to understanding their function. While we have focused mostly on similarities across alphaherpesvirus gK orthologs, our analyses also identify areas of rapid evolution that may contribute to differences in infection and pathogenesis across species. Domain 1, in particular, is known to control cellular receptor utilization by the virus and has likely been the target of diverse selection pressures. The same approach that we have employed here could be used to investigate host-specific adaptation in the future. Overall, these analyses contribute to a greater understanding of both the evolution and structure of an important viral pathogenesis determinant and represent a general approach for the rational design of drugs to target multiple human and animal alphaherpesvirus pathogens.

## Materials and Methods

**Phylogenetic and evolutionary rate analysis.** Eight previously sequenced HSV strains were selected from Genbank to be included in genome-wide phylogenetic analyses (bolded species in Table 1). These sequences were chosen to represent all major groups of alphaherpesviruses, while also focusing on those found in humans (HSV-1, HSV-2, and VZV). In order to conduct phylogenetic analyses, homologous regions of the genome must be identified and aligned. We conservatively interpreted available annotations to identify homologous regions, resulting in 59 homologous gene products that were extracted from each genome and compiled into gene-by-gene

(amino acids) datasets. Each gene-by-gene dataset was aligned with MAFFT version 7 using the E-INS-i algorithm for the highest level of accuracy<sup>47</sup>. In order to select the best fitting model of amino acid substitution, we used the Bayesian Information Criterion (BIC) as implemented in protest version 3.2<sup>48</sup>. All available 120 amino acid substitution models were tested, and in each case the Jones92 model with invariant sites and a gamma distribution was preferred by BIC. All individual gene datasets were then concatenated, resulting in an alignment of approximately 41,543 amino acids from each of 59 genes (a total of ~2.5 million amino acids).

Bayesian analyses with fixed partitioning schemes were performed using MrBayes version 3.2.5<sup>49</sup>. Fixed partitioning requires making a priori decisions about which sites we expect to evolve with similar dynamics or at a similar rate. Similar sites are then assigned to the same model of sequence evolution. Fixed partitioning analyses included both those that estimated parameters independently across 59 genes (to compare the relative rate of gK to other genes), and those that estimated domain-specific parameters for gK. All substitution model parameters, with the exceptions of topology and branch lengths, were unlinked across data subsets, including relative rates of evolution for each gene or partition. Each analysis included 4 independent runs of 5 million generations with 4 Metropolis-coupled chains per run. Analyses were monitored for convergence by examining trace plots of scalar values in Tracer version 1.6<sup>50</sup>. Trace plots show the values of parameters sampled across MCMC generations. Similarity in these values through time, and across runs, improves confidence that we are accurately estimating the posterior distribution. Topological convergence was assessed using the average standard deviation of split frequencies (ASDSF). All analyses were conducted both with and without missing data (gaps). In all cases, recovered topologies and support values did not exhibit strong conflicts, so we only report results from complete datasets. A majority-rule consensus tree was constructed in MrBayes and plotted using FigTree version 1.4.2 for the gK analyses (Fig. 1)<sup>51</sup>. The relative rates of evolution of different genes or domains were compared using inferred rate multipliers.

For domain specific analyses, eighteen viral amino acid sequences for the envelope glycoprotein K gene (gK) were taken from Genbank (Table 1). These sequences were then aligned by hand, using structure as a guide, which resulted in an alignment with 478 positions. HSV-1 gK is 338 amino acids long and when aligned with other gKs the gaps result in an alignment that is overall 478 positions. This alignment was partitioned into individual domains, also based on structure (Figs 4–6, and Supplementary Material). Model selection was performed as described above.

To estimate the relative rates of evolution for each residue in gK, we employed a novel partitioning model, where the assignment of sites to rate categories is not fixed. This model is closely related to the nonparametric method of Huelsenbeck and Suchard, who considered the number of rate categories to be a random variable distributed according to a Dirichlet process prior model (DPP)<sup>52</sup>. We initially explored the use of the DPP, but found that mixing and convergence were difficult to achieve for our data, resulting in site-specific estimates that were not reliable. Instead, we employed a finite mixture model (FMM), where the number of rate categories ( $k$ ) is fixed *a priori*, although the assignment of sites to categories is random. Following the notation of Huelsenbeck and Suchard<sup>52</sup> the assignment of sites to categories can be given as a vector of mappings  $\sigma = (\sigma_1, \dots, \sigma_n)$ , where  $n$  is the number of sites in the alignment and  $\sigma_i \in \{1, \dots, k\}$ . We explored two versions of this FMM: one where the prior probability of each site's assignment to each category is fixed and equal,  $P(\sigma_i = 1, \dots, k) = 1/k$ , and another where a Dirichlet distribution is employed as a hyperprior on the probabilities of assignment,  $P(\sigma_i = 1, \dots, k) \sim \text{Dir}(\alpha)$ . In the latter form,  $\alpha = (\alpha_1, \dots, \alpha_k)$ , where  $\alpha_i$  is the concentration parameter for category  $i$ . All of our analyses employed a flat Dirichlet with  $\alpha = (1, \dots, 1)$ . For both versions of the FMM, we explored the use of 2, 4, and 6 discrete rate categories, each with an independent tree length. Replicate analyses were compared to ensure consistency of site-specific rate estimates. The number of categories was increased until rate estimates stabilized as more categories were added (Figs S1–3). Our primary goals were to estimate the rankings and relative magnitudes of site-specific rates, and we found that these estimates were relatively stable between the 4- and 6-category analyses. These analyses were performed with RevBayes v1.0.1<sup>53</sup> and the associated Rev scripts are available as supplementary material ([https://github.com/jembrown/Evolution2017/blob/master/Example\\_FMM.rb](https://github.com/jembrown/Evolution2017/blob/master/Example_FMM.rb)).

**Multiple sequence alignment.** Multiple sequence alignment (MSA) between four different strains of alphaherpesviruses HSV-1, HSV-2, PRV, and VZV (UniProtKB accession number P68333, G9I276, Q85230 and Q4JQX0). The MSA was constructed with Clustal Omega sequence profile alignment<sup>54</sup> by Jalview software<sup>55</sup> and visualized using the Jalview Blosum62 color scheme; in this color scheme gaps are colored white. If a residue matches the consensus sequence residue at that position it is colored dark blue. If it does not match the consensus residue but the 2 residues have a positive Blosum62 score<sup>56</sup> it is colored light blue. Conservation and sequence logo analysis alignment is generated and visualized by Jalview.

**Computational protein structure modeling.** Template-based modeling of gK proteins from four alphaherpesvirus, HSV-1, HSV-2, PRV, and VZV (UniProtKB accession number P68333, G9I276, Q85230 and Q4JQX0), were achieved by dividing the protein sequences into domains, modeling each domain separately using Modeller<sup>57</sup>. The extramembrane domains were separated by the transmembrane domains. The locations of the transmembrane domains were determined using tmsfoldweb server. Secondary protein structures were predicted using PSIPRED and JPRED. Secondary structure elements were enforced in the modeling process. In order to model the transmembrane domain into one alpha helix bundle rather than separate alpha helices, the best template structure was found using tmsfoldweb server. The X-ray crystal structures of E. coli YEDX (PDB code: 2igl, chain: D), MOUSE CLAUDIN-19 (PDB code: 3 × 29), Epstein-Barr virus gH and gL complex (PDB code: 3PHF, chain: Y) were used as templates for the N-terminal, the transmembrane and the domain 2 for HSV-1, HSV-2, and VZV. In case of modeling the N-terminal domain for PRV, Pichia Pastoris Lysyl Oxidase (PDB code: 1W7C, chain: A) was used as the template. All of these templates showed high scores with low E-values.

Individual domains were assembled according to the current understanding of each domain; the N-term and TM domains were specifically positioned relative to each other by superimposing separate domain structure models onto MOUSE CLAUDIN-19 protein crystal structure (PDB code: 3 × 29). Full-chain protein structures were remodeled using Modeller to refine loops and gaps in between domains.

**In silico alanine scanning.** Amino acid residues in the gK homology model are mutated into alanines one at a time by AlaScan feature in the protein design package FoldX Suite downloaded in 2018 (<http://foldxsuite.org.eu/>); the difference in the free Gibbs energy change ( $\Delta\Delta G$ ) before and after performing each mutation is calculated by FoldX Suite. The numbers for the specific mutations are reported as deviations from the average; a large positive  $\Delta\Delta G$  indicates the importance of an amino acid residue for structure stability.

## Data Availability

All of the data are fully accessible at the following url: <https://doi.org/10.5061/dryad.8g0d43c>.

## References

1. Davison, A. J. Evolution of the herpesviruses. *Veterinary microbiology* **86**, 69–88 (2002).
2. Ablashi, D. *et al.* Classification of HHV-6A and HHV-6B as distinct viruses. *Archives of virology* **159**, 863–870 (2014).
3. Pellet, P. E. *et al.* In *Virus taxonomy: Ninth report of the International Committee on Taxonomy of Viruses*. (eds Adams, M. J., King, A. M. Q., Carstens, E. B. & Lefkowitz, E. J.) p. 81–92 (Elsevier, Oxford, 2011).
4. Davison, A. J. *et al.* The order Herpesvirales. *Archives of virology* **154**, 171–177 (2009).
5. Alba, M. M., Das, R., Orenge, C. A. & Kellam, P. Genomewide function conservation and phylogeny in the Herpesviridae. *Genome research* **11**, 43–54 (2001).
6. Mocarski, E. S. In *Herpesviruses: Biology, Therapy, and Immunophylaxis*. (ed. Campadelli-Fiume G *et al.*) Ch. 4, (Cambridge University Press, 2007).
7. Roizman, B., Knipe, D. & Whitley, R. J. In *Fields Virology* (ed. Peter Howley David, M. Knipe) 2501–2602 (Lippincott Williams & Wilkins, 2006).
8. Dietz, P. *et al.* Pseudorabies virus glycoprotein K requires the UL20 gene product for processing. *J Virol* **74**, 5083–5090 (2000).
9. Foster, T. P., Alvarez, X. & Kousoulas, K. G. Plasma membrane topology of syncytial domains of herpes simplex virus type 1 glycoprotein K (gK): the UL20 protein enables cell surface localization of gK but not gK-mediated cell-to-cell fusion. *J Virol* **77**, 499–510 (2003).
10. Foster, T. P., Melancon, J. M., Baines, J. D. & Kousoulas, K. G. The Herpes Simplex Virus Type 1 UL20 Protein Modulates Membrane Fusion Events during Cytoplasmic Virion Morphogenesis and Virus-Induced Cell Fusion. *J Virol* **78**, 5347–5357 (2004).
11. Fuchs, W., Klupp, B. G., Granzow, H. & Mettenleiter, T. C. The UL20 gene product of pseudorabies virus functions in virus egress. *J Virol* **71**, 5639–5646 (1997).
12. Melancon, J. M., Fulmer, P. A. & Kousoulas, K. G. The herpes simplex virus UL20 protein functions in glycoprotein K (gK) intracellular transport and virus-induced cell fusion are independent of UL20 functions in cytoplasmic virion envelopment. *Virology* **4**, 120 (2007).
13. Foster, T. P., Chouljenko, V. N. & Kousoulas, K. G. Functional and physical interactions of the herpes simplex virus type 1 UL20 membrane protein with glycoprotein K. *J Virol* **82**, 6310–6323 (2008).
14. Foster, T. P., Melancon, J. M. & Kousoulas, K. G. An alpha-helical domain within the carboxyl terminus of herpes simplex virus type 1 (HSV-1) glycoprotein B (gB) is associated with cell fusion and resistance to heparin inhibition of cell fusion. *Virology* **287**, 18–29 (2001).
15. Foster, T. P., Melancon, J. M., Olivier, T. L. & Kousoulas, K. G. Herpes simplex virus type-1 (HSV-1) glycoprotein K (gK) and the UL20 protein are interdependent for intracellular trafficking and trans-Golgi network localization. *J Virol* (2004).
16. Jambunathan, N. *et al.* Site-specific proteolytic cleavage of the amino terminus of herpes simplex virus glycoprotein K on virion particles inhibits virus entry. *J Virol* **85**, 12910–12918 (2011).
17. Foster, T. P. & Kousoulas, K. G. Genetic analysis of the role of herpes simplex virus type 1 glycoprotein K in infectious virus production and egress. *J Virol* **73**, 8457–8468 (1999).
18. Hutchinson, L., Roop-Beauchamp, C. & Johnson, D. C. Herpes simplex virus glycoprotein K is known to influence fusion of infected cells, yet is not on the cell surface. *J Virol* **69**, 4556–4563 (1995).
19. Jayachandra, S., Baghian, A. & Kousoulas, K. G. Herpes simplex virus type 1 glycoprotein K is not essential for infectious virus production in actively replicating cells but is required for efficient envelopment and translocation of infectious virions from the cytoplasm to the extracellular space. *J Virol* **71**, 5012–5024 (1997).
20. Chouljenko, V. N., Iyer, A. V., Chowdhury, S., Chouljenko, D. V. & Kousoulas, K. G. The amino terminus of herpes simplex virus type 1 glycoprotein K (gK) modulates gB-mediated virus-induced cell fusion and virion egress. *J Virol* **83**, 12301–12313 (2009).
21. Chouljenko, V. N., Iyer, A. V., Chowdhury, S., Kim, J. & Kousoulas, K. G. The herpes simplex virus type 1 UL20 protein and the amino terminus of glycoprotein K (gK) physically interact with gB. *J Virol* **84**, 8596–8606 (2010).
22. Foster, T. P., Rybachuk, G. V. & Kousoulas, K. G. Glycoprotein K specified by herpes simplex virus type 1 is expressed on virions as a Golgi complex-dependent glycosylated species and functions in virion entry. *J Virol* **75**, 12431–12438 (2001).
23. Rider, P. J. F. *et al.* Cysteines and N-Glycosylation Sites Conserved among All Alphaherpesviruses Regulate Membrane Fusion in Herpes Simplex Virus 1 Infection. *J Virol* **91** (2017).
24. David, A. T., Baghian, A., Foster, T. P., Chouljenko, V. N. & Kousoulas, K. G. The herpes simplex virus type 1 (HSV-1) glycoprotein K(gK) is essential for viral corneal spread and neuroinvasiveness. *Curr Eye Res* **33**, 455–467 (2008).
25. David, A. T. *et al.* A herpes simplex virus 1 (McKrae) mutant lacking the glycoprotein K gene is unable to infect via neuronal axons and egress from neuronal cell bodies. *mBio* **3**, e00144–00112 (2012).
26. Saied, A. A., Chouljenko, V. N., Subramanian, R. & Kousoulas, K. G. A replication competent HSV-1(McKrae) with a mutation in the amino-terminus of glycoprotein K (gK) is unable to infect mouse trigeminal ganglia after cornea infection. *Curr Eye Res* **39**, 596–603 (2014).
27. Klupp, B. G., Baumeister, J., Dietz, P., Granzow, H. & Mettenleiter, T. C. Pseudorabies virus glycoprotein gK is a virion structural component involved in virus release but is not required for entry. *J Virol* **72**, 1949–1958 (1998).
28. Hutchinson, L. & Johnson, D. C. Herpes simplex virus glycoprotein K promotes egress of virus particles. *J Virol* **69**, 5401–5413 (1995).
29. Neubauer, A. & Osterrieder, N. Equine herpesvirus type 1 (EHV-1) glycoprotein K is required for efficient cell-to-cell spread and virus egress. *Virology* **329**, 18–32 (2004).
30. Mo, C., Suen, J., Sommer, M. & Arvin, A. Characterization of Varicella-Zoster virus glycoprotein K (open reading frame 5) and its role in virus growth. *J Virol* **73**, 4197–4207 (1999).
31. Musarrat, F., Jambunathan, N., Rider, P. J. F., Chouljenko, V. N. & Kousoulas, K. G. The Amino Terminus of Herpes Simplex Virus 1 Glycoprotein K (gK) Is Required for gB Binding to Akt, Release of Intracellular Calcium, and Fusion of the Viral Envelope with Plasma Membranes. *J Virol* **92** (2018).



32. Plotkin, J. B. & Dushoff, J. Codon bias and frequency-dependent selection on the hemagglutinin epitopes of influenza A virus. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 7152–7157 (2003).
33. Alfano, N. *et al.* Episodic Diversifying Selection Shaped the Genomes of Gibbon Ape Leukemia Virus and Related Gammaretroviruses. *J Virol* **90**, 1757–1772 (2016).
34. Haque, M., Stanfield, B. & Kousoulas, K. G. Bovine herpesvirus type-1 glycoprotein K (gK) interacts with UL20 and is required for infectious virus production. *Virology* **499**, 156–164 (2016).
35. Jambunathan, N. *et al.* Herpes simplex virus 1 protein UL37 interacts with viral glycoprotein gK and membrane protein UL20 and functions in cytoplasmic virion envelopment. *J Virol* **88**, 5927–5935 (2014).
36. Montague, M. G. & Hutchison, C. A. 3rd Gene content phylogeny of herpesviruses. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 5334–5339 (2000).
37. Gao, L. & Qi, J. Whole genome molecular phylogeny of large dsDNA viruses using composition vector method. *BMC evolutionary biology* **7**, 41 (2007).
38. Lamers, S. L. *et al.* Global Diversity within and between Human Herpesvirus 1 and 2 Glycoproteins. *J Virol* **89**, 8206–8218 (2015).
39. Szpara, M. L. *et al.* Evolution and diversity in human herpes simplex virus genomes. *J Virol* **88**, 1209–1227 (2014).
40. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular systems biology* **7**, 539 (2011).
41. Goujon, M. *et al.* A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic acids research* **38**, W695–699 (2010).
42. Jambunathan, N. *et al.* Deletion of a Predicted beta-Sheet Domain within the Amino Terminus of Herpes Simplex Virus Glycoprotein K Conserved among Alphaherpesviruses Prevents Virus Entry into Neuronal Axons. *J Virol* **90**, 2230–2239 (2015).
43. Owen, D. J., Crump, C. M. & Graham, S. C. Tegument Assembly and Secondary Envelopment of Alphaherpesviruses. *Viruses* **7**, 5084–5114 (2015).
44. Chowdhury, S., Chouljenko, V. N., Naderi, M. & Kousoulas, K. G. The amino terminus of herpes simplex virus 1 glycoprotein K is required for virion entry via the paired immunoglobulin-like type-2 receptor alpha. *J Virol* **87**, 3305–3313 (2013).
45. Zerboni, L., Sen, N., Oliver, S. L. & Arvin, A. M. Molecular mechanisms of varicella zoster virus pathogenesis. *Nature reviews. Microbiology* **12**, 197–210 (2014).
46. Oliver, S. L., Yang, E. & Arvin, A. M. Varicella-Zoster Virus Glycoproteins: Entry, Replication, and Pathogenesis. *Current clinical microbiology reports* **3**, 204–215 (2016).
47. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution* **30**, 772–780 (2013).
48. Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27**, 1164–1165 (2011).
49. Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574 (2003).
50. Suchard, M. A. *et al.* Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol* **4**, vey016 (2018).
51. Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular biology and evolution* **29**, 1969–1973 (2012).
52. Huelsenbeck, J. P. & Suchard, M. A. A nonparametric method for accommodating and testing across-site rate variation. *Syst Biol* **56**, 975–987 (2007).
53. Hohna, S. *et al.* RevBayes: Bayesian Phylogenetic Inference Using Graphical Models and an Interactive Model-Specification Language. *Syst Biol* **65**, 726–736 (2016).
54. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research* **22**, 4673–4680 (1994).
55. Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M. & Barton, G. J. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).
56. Eddy, S. R. Where did the BLOSUM62 alignment score matrix come from? *Nature biotechnology* **22**, 1035–1036, <https://doi.org/10.1038/nbt0804-1035> (2004).
57. Fiser, A. & Sali, A. Modeller: generation and refinement of homology-based protein structure models. *Methods in enzymology* **374**, 461–491 (2003).

## Acknowledgements

We would like to thank Sebastian Höhna for help designing the finite mixture model in RevBayes. We used high-performance computing resources provided by Louisiana State University for many of the analyses in this study.

## Author Contributions

P.R., L.C., M.N., J.M.B., M.B. and K.G.K. designed the experiments. L.C. and J.M.B. performed the phylogenetic experiments. M.N. and M.B. performed the molecular modeling experiments. P.R. wrote the manuscript with the assistance of J.M.B. and K.G.K. All authors reviewed the manuscript.

## Additional Information

**Supplementary information** accompanies this paper at <https://doi.org/10.1038/s41598-019-50490-9>.

**Competing Interests:** The authors declare no competing interests.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019