1 1. Title page

2

3 **Title:** Consensus clustering of herpesvirus protein interaction networks
4 provides insights into their evolutionary relationship with bacteriophages

5 **Subtitle:** New insights on herpesviruses evolution from biological network
6 analysis

7 Anna Hernández Durán^{1,2}, Todd M. Greco³, Benjamin Vollmer², Ileana M.
8 Cristea³, Kay Grünewald^{2,4†}, Maya Topf^{1,†}

9

10 ¹ Institute of Structural and Molecular Biology, Birkbeck College, University of
11 London, Malet street, London, WC1E 7HX, UK;

12 ² Division of Structural Biology, Wellcome Centre for Human Genetics,
13 University of Oxford, Roosevelt Drive, Oxford, OX3 7BN, UK;

14 ³ Department of Molecular Biology, Princeton University, Lewis Thomas
15 Laboratory, Washington Road, Princeton, NJ 08544, USA;

16 ⁴ Department of Structural Cell Biology of Viruses, Centre for Structural
17 Systems Biology, Heinrich Pette Institute, Leibnitz Institute of Experimental
18 Virology, University of Hamburg, Notkestr. 85, Hamburg, D-22607, Germany

19

20 [†]To whom correspondence should be addressed: E-mails:
21 m.topf@cryst.bbk.ac.uk; Tel.: +44 (0)20 7631 6886; E-mail:
22 kay.gruenewald@cssb-hamburg.de, Tel. +49 (0)40 8998 87700.

23

24

25

26 **Abbreviations**

- 27 CAM: consensus agreement matrix
- 28 CLR: collagen-like repeat
- 29 DIP: Database of Interacting Proteins
- 30 dsDNA: double-stranded DNA
- 31 EBV: Epstein-Barr virus
- 32 EGFP: enhanced green fluorescent protein
- 33 EMDB: Electron Microscopy Data Bank
- 34 ePPI: experimentally detected PPI
- 35 GO: gene ontology
- 36 HCMV: human cytomegalovirus
- 37 HMM: Hidden Markov Model
- 38 HPI: hours post infection
- 39 HSV1: herpes simplex virus type 1
- 40 IP: immunoaffinity purification
- 41 LCC: largest connected component
- 42 MCP: major capsid protein
- 43 MS: mass spectrometry
- 44 PDB: Protein Data Bank
- 45 PPI: protein-protein interaction
- 46 oPPI: PPI in orthologous species
- 47 pPPI: computationally predicted PPI
- 48 TGN: trans-Golgi network
- 49 TMT: tandem mass tags
- 50 VZV: varicella-zoster virus

51

52 **2. Abstract**

53

54 Infections with human herpesviruses are ubiquitous and a public health
 55 concern worldwide. Current treatments reduce the morbidity of some
 56 manifested symptoms but neither remove the viral reservoir from the infected
 57 host nor protect from the recurrent symptom outbreaks that characterize
 58 herpetic infections. The difficulty in therapeutically tackling these viral systems
 59 stems in part from their remarkably large proteomes and the complex
 60 networks of physical and functional associations that they tailor. This study
 61 presents our efforts to unravel the complexity of the interactome of herpes
 62 simplex virus type 1 (HSV1), the prototypical herpesvirus species. Inspired by
 63 our previous work, we present an improved computational pipeline for the
 64 protein-protein interaction (PPI) network reconstruction in HSV1, combining
 65 both experimentally supported and bioinformatic predictions. Our newly-
 66 developed consensus clustering approach allows us to extend the analysis
 67 beyond binary physical interactions and reveals higher order functional
 68 associations including that of pUS10 with capsid proteins. In-depth
 69 bioinformatics sequence analysis unraveled structural features of this 34-36
 70 kDa protein reminiscent of those observed in some capsid-associated
 71 proteins in tailed bacteriophages, with which herpesviruses are thought to
 72 share a common ancestry. This suggests that pUS10 could represent an
 73 evolutionary vestige between these two viral lineages. Using immunoaffinity
 74 purification-mass spectrometry we found that pUS10 specifically co-isolated
 75 with the inner tegument protein pUL37, which binds cytosolic capsids,

76 contributing to initial tegumentation and eventual virion maturation. In
77 summary, this study unveils new insights at both the system and molecular
78 levels that can help better understand the complexity behind herpesvirus
79 infections.

80

81 **3. Introduction**

82 Herpesviruses infect a wide range of eukaryotic organisms, and are the
83 etiologic agent of severe diseases in livestock and humans. At present, there
84 are nine species of herpesviruses known to routinely infect human. They are
85 referred to as the human herpesviruses, and their infections are associated
86 with symptoms ranging from fever and cutaneous lesions, to encephalitis,
87 meningitis, and a number of cancerous malignancies [1]. A link between
88 herpetic infections and the neurodegenerative Alzheimer's disease was
89 recently confirmed, emphasising the socio-economic burden associated with
90 these viruses [2,3].

91

92 Herpesviruses are enveloped viruses that assemble into a morphologically
93 unique extracellular particle (i.e. virion), which is organised in concentric
94 structural layers [4]. At the innermost of the virion particle, an icosahedral
95 protein shell of ~ 120nm diameter - the capsid - encloses the double-stranded
96 DNA (dsDNA) genome of the virus. Surrounding the capsid, a protein matrix -
97 the tegument - occupies about two thirds of the virion particle volume. The
98 tegument contains proteins that are delivered into the host cytoplasm upon
99 infection and therefore are ready to initiate their function prior to the
100 transcription of any viral genes. The entire particle is coated by a lipid bilayer -

101 the envelope - that contains several proteins and glycoproteins that are crucial
 102 for host cell entry and cell-to-cell spread, as well as for modulation of the
 103 host's immune response. As for other enveloped viruses, entry into the host
 104 cell occurs via fusion of the envelope with plasma membrane or endocytic
 105 cellular membranes, which results in release of the virion content into the host
 106 cell cytosol [5]. Altogether, herpesviruses are composed of between ~70 to
 107 ~170 different protein species and achieve diameters between ~150 to
 108 ~250nm [6,7].

109

110 To understand the behavior of a system, the relationships among the
 111 components are as important as the components themselves. The need for a
 112 comprehensive description of the relationships among viral proteins, and with
 113 host factors, has prompted numerous protein-protein interaction (PPI) studies.
 114 In 2016 we published a compilation of PPI data in herpes simplex virus type 1
 115 (HSV1), the prototypical species of the *Herpesviridae* family based on
 116 integration of experimentally supported data from five public resources [8].
 117 Here, we present an improved computational pipeline for PPI network
 118 reconstruction, collating data from seven public repositories with
 119 computationally predicted PPIs inferred using conservative sequence
 120 similarity assessments. The reliability of all gathered PPIs is measured using
 121 a new scoring scheme. Moreover, we applied network analysis approaches to
 122 gain insights into the functional organization of the extracellular viral particle.
 123 We developed a clustering consensus protocol to study the community
 124 structure underlying the binary interactions among virion proteins. The
 125 resulting higher order relationships predicted previously unrecognized

functional viral protein associations, such as between the inner tegument protein pUL37 and pUS10. We further investigated this relationship using bioinformatic sequence analysis, and provided experimental support using immunoaffinity purification-mass spectrometry in primary human fibroblasts infected with HSV1.

4. Results

PPI network assembly

Binary Protein-Protein interaction (PPI) data were obtained from five molecular interaction repositories (BioGRID [9], the Database of Interacting Proteins (DIP) [10], IntAct [11], Mentha [12], and VirHostNet 2.0 [13]) and two structural databases (Protein Data Bank (PDB) [14] and the Electron Microscopy Data Bank (EMDB) [15]) (Fig 1A). We collected PPIs that have been detected for all nine human herpesvirus species, and three closely-related non-human herpesviruses (Fig 1B and S1 and S2 Tables). The latter were included as these species are frequently used as animal models of herpetic human infections (see Materials and Methods). The resulting joined non-redundant data set contained 2854 unique pieces of evidence, 2363 unique PPIs, and 758 unique protein sequences. From this data set, PPIs experimentally detected in HSV1 were directly added to the network (Fig 1C). PPIs experimentally detected in species other than HSV1 were used to predict new PPIs in HSV1 (Fig 1D). To predict PPIs, we used an orthology-based method referred to as *interologues mapping* [16]. This method predicts an interaction in HSV1, if both putative interactors have homologues known to

151 interact in other species (see Materials and Methods). Homology relationships
152 are inferred based on sequence-based Hidden Markov Model (HMM) profile
153 alignments [17] in combination with a conservative multi-criteria threshold to
154 filter out potential spurious matches from the homology search results (see
155 Materials and Methods).

156

157 **Fig 1. Network assembly framework.** (A and B) PPI data for a total of twelve
158 herpesvirus species (nine human and three non-human herpesviruses,
159 together covering members of all three subfamilies, i.e. the α -, β - and γ -
160 *Herpesvirinae*, S1 and S2 Tables) were collected from seven public resources
161 [9-15]. (C) PPIs experimentally detected in HSV1 were transferred to its
162 interactome (ePPIs). (D) PPIs detected in any of its orthologous herpesvirus
163 species (oPPIs) were used to predict PPIs in HSV1 (pPPIs). Predictions were
164 conducted based on a sequence-based interologues mapping [16] approach
165 (green box), and included the following steps: for each protein involved in a
166 binary oPPI, (1) sequence-based homologous sequences in the HSV1
167 proteome were searched for using HHblits [17]; (2) a conservative homology
168 threshold was applied to filter out potential spurious matches among the list of
169 candidates returned by HHblits. From the remaining matches, the best scoring
170 sequence was selected as the most reliable putative HSV1 homologue; (3) if
171 potential HSV1 homologous sequences were found for both proteins in the
172 initial oPPI, an interaction between the two HSV1 sequences was predicted.
173 (E) Predicted and experimentally supported PPIs were joined into a non-
174 redundant data set, and scored based on their supporting evidence (see
175 Materials and Methods).

176

177 Experimentally detected interactions in HSV1 and computational predictions
 178 were compiled into a PPI network, and the confidence of each interaction
 179 based on its cumulative evidence was assessed using a new scoring function
 180 (see Materials and Methods). The new function is a modification of the
 181 MIscore function [18], which was developed by the Proteomics Standards
 182 Initiative [19], and is compliant with standardised protocols for the assessment
 183 and representation of molecular interaction data [20]. Our function essentially
 184 adds a conservative penalty function term to the original MIscore function for
 185 interactions without experimental support. This term includes information on
 186 the prediction method (in this case sequence-based alignment) and the total
 187 number of homologous species from which the prediction was inferred. The
 188 function gives higher confidence scores to interactions that have been
 189 predicted from a larger number of species. For instance, if an interaction can
 190 be predicted in HSV1 based on interactions experimentally detected in three
 191 other species of herpesviruses (e.g. VZV, EBV and HCMV), the prediction will
 192 score higher than if it had been obtained from a single experimental
 193 observation (e.g. only in VZV).

194

195 The reconstructed network (Fig 2 and S3 Table) contained 369 PPIs, formed
 196 among 68 proteins, and supported by 643 unique pieces of evidence. Among
 197 the 369 PPIs, there were 159 experimentally supported interactions, and 250
 198 computationally predicted; 40 interactions were supported by both
 199 experimental and computational evidence. All these network data are

available through our new version of the HVint database [8]: <http://topf-group.ismb.lon.ac.uk/hvint/>.

Fig 2. Reconstructed PPI network for HSV1. Nodes represent proteins and their size is proportional to the number of interactions associated with each protein in the network (degree). Nodes are colour-coded as follows: cyan for capsid and capsid-associated proteins, orange for tegument proteins, yellow for envelope glycoproteins, blue for non-glycosylated envelope proteins, respectively, and grey for proteins that are not present in the mature virion particle (i.e. typically only expressed during intra-cellular stages). Edge (or link) thickness reflects the confidence score for the interaction (the thicker, the higher the confidence). Edges are colour-coded to indicate the type of supporting evidence behind it, i.e. blue for experimentally supported interactions, red for computationally predicted, and green for interactions with both experimental and computational supporting evidence.

Selection of high confidence PPI predictions for experimental testing

We assessed the ability of our protocol for predicting meaningful interactions for experimental testing. We conducted a manual literature review in search of experimental support for our best predictions (5% top scoring interactions) among published literature that had not been included in the input dataset (i.e. it was not used to infer the predictions). This search returned experimental

225 support for 4 out of the total 11 (top 5% scoring) predicted PPIs examined
 226 [8,21,22] (Table 1).

227

228

Table 1. Top 5% computationally predicted PPIs.		
Interacting protein A	Interacting protein B	Score
UL32 (P10216)	UL32 (P10216)	0.347
UL32 (P10216)	UL40 (P10224)	0.297
UL5 (P10189)	UL52 (P10236)	0.287
UL25 (P10209)	UL25 (P10209)	0.266
UL15 (P04295)	UL50 (P10234)	0.263
UL40 (P10224)	UL52 (P10236)	0.26
UL39 (P08543)	UL50 (P10234)	0.257
UL23 (P03176)	UL33 (P10217)	0.254
UL21 (P10205)	UL33 (P10217)	0.254
UL5 (P10189)	UL16 (P10200)	0.249
UL12 (P04294)	UL16 (P10200)	0.236
Interacting proteins (A and B) are referred to with their open reading frame name and UniProtKB accession numbers. Shaded rows indicate predicted interactions with additional experimental support not included as input during network reconstruction.		

229

230

231 **Analysis of the community structure in the HSV1 network**

232

233 Using the HSV1 network data compiled at this point, we undertook the
 234 analysis of the community structure of the subnetwork formed by virion
 235 proteins. These are defined as components of the extracellular viral particle
 236 (i.e. capsid, tegument, and envelope proteins), which are currently well
 237 identified in this species [23-26]. We applied two further steps to remove noise
 238 (e.g. false positive PPIs) in the network to reduce potential sources of
 239 artefacts in the clustering results (Fig 3A, see Materials and Methods). First,
 240 we removed PPIs in the network taking place between capsid (and capsid-
 241 associated) and envelope proteins, which are likely unfeasible due to spatial
 242 constraints imposed by the virion tegument. Second, we extracted, from the
 243 network obtained in the previous step, the largest connected component (LCC
 244 – the largest group of nodes in which any pair of them can be connected by a
 245 path). The resulting network contained 40 nodes and 94 edges.

246

247 Next, 14 different clustering algorithms were applied to this LCC network (Fig
 248 3B and S4 Table). Those output partitions yielding positive modularity (11 in
 249 total) were integrated into a consensus agreement matrix (Fig 3C). Values in
 250 the consensus agreement matrix cells represented, for each pair of proteins in
 251 the network, the fraction of partitions that clustered the pair together. Using a
 252 threshold-based method guided by modularity maximization a consensus
 253 partition was inferred from the consensus agreement matrix (Fig 3D). This
 254 partition achieved a modularity of 0.43 (for thresholds in [0.55, 0.60]) (S1 Fig
 255 and S5 Table); values between 0.3 and 0.7 are considered indicative of a
 256 significant community structure [27,28].

257

Fig 3. Consensus clustering protocol. Starting from an input network, (A) a series of steps were applied to remove potential false positives. Here, these steps included selection of virion components only, and removal of PPI between capsid (and capsid-associated) proteins to envelope proteins. (B) N=14 clustering algorithms were applied to the network. (C) The output base partitions with positive modularity were integrated into a consensus agreement matrix. (D) A final consensus clustering partition was derived by filtering out the values in the consensus agreement matrix based on a modularity-maximisation threshold.

The consensus partition divided the network into four communities (Fig 4). Their biological consistency was then assessed using functional annotation data manually curated from Gene Ontology (GO) [29,30] annotations and published literature.

Fig 4. Community structure inferred for the HSV1 virion network. Depiction of nodes and edges follows the same criteria as described in Fig 2. Four separate communities are indicated by grey shaded areas.

Community A

Community A illustrates almost entirely the composition of the capsid (Fig 4). The only capsid protein missing is pUL6 (the portal protein), which appears disconnected from the LCC in the virion subnetwork. This was surprising as the localisation of pUL6 in the capsid, forming the capsid portal complex, is

283 well established [31]. However, the input data sets used to feed the network
284 assembly framework did not contain evidence for this interaction.

285

286 Community A also contains the protein pUS10. pUS10 is a poorly
287 characterised α -subfamily specific protein of 312 residues, which has been
288 found in the nuclear, peri-nuclear and cytoplasmic cellular compartments and
289 co-precipitating with capsids, yet direct interactions with capsid proteins have
290 not been reported. It exists in two phosphorylation states and is regarded as a
291 minor component of the tegument [32]. A consensus zinc finger was identified
292 in pUS10 homologues although the protein failed to bind nucleic acids
293 (common among zinc finger proteins) during experimental testing [32,33].
294 Finally, HSV1 pUS10 presents a high proline content at its N-terminus and a
295 four-residue long polyproline sequence located centrally. To assess whether
296 the clustering of pUS10 with capsid proteins had a functional significance or it
297 was an artefact of a low number of interactions for pUS10 in the input graph,
298 we sought further characterization of the protein through primary sequence
299 analysis (Fig 5 and S2 Fig, see Materials and Methods). An initial search for
300 potential sequence homologues did not identify any candidates beyond
301 pUS10 counterparts, yet it highlighted the presence of seven tandem
302 collagen-like repeats (CLRs) located toward the N-terminus of the protein
303 sequence. Interestingly, through a manual curation of the literature, we found
304 support for this prediction in a 30-year old study [34] (note that pUS10 was
305 annotated by its MW in this study, which could explain why this prediction was
306 not included in databases such as UniProtKB). This prominent feature
307 prompted compelling hypothesis on its functions and evolutionary history (see

Discussion). Secondary structure predictions indicated that the N- and C-termini of the protein are structurally distinct. Whilst the N-terminus was predicted to be disordered, the C-terminus was rich in α -helices. Additionally, our predictions revealed a potential single-pass transmembrane segment at the very C-terminus of the protein. This prediction overlaps with the potential zinc finger motif, which could explain why so far no functional evidence has been provided.

315

Fig 5. Sequence characterization of pUS10. Both previously described and newly identified features are indicated. Predicted disordered regions, α -helices and transmembrane helices are highlighted in blue, yellow and pink, respectively. Collagen-like repeats (CLRs) are framed in red boxes. Prolines are highlighted in red. The four-residues polyproline sequence is framed in a black box. The previously found consensus zinc finger sequence [33] is underscored.

323

The binary interaction that connects pUS10 with other components within its community is the interaction with pUL37. This interaction was predicted by our orthology-based method, and we aimed to further validate it experimentally (see below). pUL37 HSV1 is an essential inner tegument protein [35] that contains a binding site for HSV1 pUL36, through which it is recruited to cytosolic capsids before secondary envelopment [36,37] or to sites of secondary envelopment in the absence of capsids [38].

331

Experimental support of the pUS10-UL37 interaction by immunoaffinity purification – mass spectrometry

To provide evidence in support of the predicted pUS10-UL37 interaction, we isolated pUL37 from infected cells using immunoaffinity purification and analysed the co-isolated proteins by quantitative mass spectrometry (IP-MS). Specifically, protein complexes were isolated from human fibroblasts synchronously infected with HSV1 strains encoding either pUL37 tagged with enhanced green fluorescent protein (pUL37-EGFP) or EGFP alone as a control. Isolations were performed at two functionally distinct time points of infection, 8 and 20 hours post infection (hpi) (S6 Table). Eight hpi represents an early time point of pUL37 expression, which is prior to secondary envelopment. This is consistent with our observation that pUL37-EGFP appears diffusely localized in the cytoplasm by epifluorescence microscopy (Fig 6A). In contrast, at 20 hpi secondary envelopment and virus particle release is in progress, reflected by pUL37-EGFP fluorescence visualized as capsid-associated puncta within the cytoplasm and maturing virions (Fig 6A). The observed temporal kinetics and localization of pUL37 induction were consistent with the prior characterization of this pUL37-EGFP HSV1 strain [37].

After establishing the temporal expression and cellular distribution of pUL37, we performed IP-MS analyses of pUL37-EGFP and EGFP complexes isolated from cytoplasmic-enriched lysates (Fig 6B). The isolation of pUL37-EGFP from the input lysates was confirmed by western blot (Fig 6C). Next, the co-

isolated proteins were identified and relatively quantified using labelling with isobaric with Tandem Mass Tags (TMT). Specifically, proteins were digested with trypsin, and the resulting peptides were derivatised with distinct TMT labels. This approach offered multiplexing, as the labelled pUL37-EGFP and EGFP IPs performed in biological replicates (N = 2 replicates) were combined and analysed by tandem MS (Fig 6B). Therefore, both the specificity of the pUL37 interactions (via comparison to the control EGFP IP) and the relative abundance of the interactions at the different time points of infection (8 and 20 hpi) could be simultaneously assessed in this quantitative MS workflow. Nine viral proteins, including pUS10, that appear as interacting with pUL37 in our network were quantitatively enriched by ≥ 2 -fold in pUL37-EGFP vs EGFP control IPs in at least one-time point and in both replicate experiments (S6 Table). After normalizing the pUS10 abundance to the abundance of pUL37-EGFP bait calculated by TMT-MS, we observed that the relative amount of pUS10 co-isolated with pUL37 was increased at the later stage of infection. At this stage, pUL37-EGFP is localized to sites of secondary envelopment and associates with maturing virions, consistent with the formation of pUL37-EGFP puncta at 20 hpi (Fig 6A). These data suggest a potential role for the pUL37-US10 interaction in the cytoplasm during virion maturation.

376

Fig 6. Experimental validation of the pUL37-US10 interaction in HSV-1 infected primary human fibroblasts. (A) Visualization of pUL37-EGFP during HSV1 infection of human foreskin fibroblasts using live cell epifluorescence microscopy. Images show a representative field of infected cells at 8 and 20 hpi. Zoomed images show localization of pUL37-EGFP

(green) in the same cell at 8 and 20 hpi. Scale bar = 50 μ m. (B) IP-MS workflow. Human fibroblasts were synchronously infected (multiplicity of infection = 10) with either pUL37-EGFP or EGFP HSV1, with two replicates per condition. HSV1-UL37GFP was collected at 8 and 20 hpi and HSV1-GFP at 20 hpi (HSV1-GFP). pUL37-EGFP and its interactions were isolated from the cytoplasmic cell fraction by immunoaffinity purification using anti-GFP antibodies. Proteins were digested with trypsin, and the resulting peptides from each sample were labelled with unique TMT reagents and then combined prior to nanoliquid chromatography-tandem mass spectrometry analysis. (C) The recovery of pUL37-EGFP and EGFP in the immunoisolates was assessed by western blot with an anti-GFP antibody. 10% of each sample was analysed. (D) PPIs around pUL37 present in the reconstructed HSV1 network (Fig 2) and supported by immunoaffinity purification results. (E) The abundance of pUS10 interaction with pUL37 at 8 and 20 hpi (average $\pm$ range, N=2). The relative amount of pUS10 was calculated by TMT-MS quantification and normalized by the respective pUL37 TMT abundance in each IP.

Community B

In community B (Fig 4), proteins pUL41, pUL47, pUL48, and pUL49 are associated, directly or indirectly, and are all known to modulate gene expression [39-41]. Another major group of proteins, (pUL11, pUL14, pUL16, pUL21, and pUL6) are related to early virion morphogenesis, i.e. recruitment of capsids and virion components to the enveloping membranes during early

stages of tegumentation [42-45]. The tight relationships among the gene expression modulation proteins have been well characterized, yet the steps that guide their incorporation into the virion particle are not yet clearly defined. The results of our clustering analysis suggest that this process could be guided by the coordinated action of cluster B members involved in virion morphogenesis.

Community C

Community C is defined by proteins involved in late stages of secondary envelopment and virion release. This community is enriched in envelope glycoproteins (gM/gN, gK/UL20, pUL45, pUL43, gJ, and gG) that regulate membrane-associated events. These include internalization of proteins from the plasma membrane, membrane fusion rates, and modulation of immune responses. Tegument proteins in this community are also annotated with the trafficking of virion components at late stages of virion morphogenesis [46-52].

An interesting observation from community C comes from the presence of pUL55, another poorly characterised α -subfamily-specific protein [53]. Our primary sequence analysis of this 186-residue protein failed to find characteristic structural features but predicted pUL55 to be an α/β protein. The analysis of the hierarchical structure of community C on the other hand (S3 Fig, see Materials and Methods), did suggest a very stable clustering tendency with four other proteins in the cluster, namely gK/UL20, pUL45, and pUS2. The cluster assignment together with the particularly strong association

with the latter four proteins [54-57] suggest new functional scenarios in late envelopment and virion release events for the currently poorly-understood pUL55 protein.

Community D

Community D is the most functionally inconclusive community. We hypothesise that this is due to the lack of known binding partners (either due to missing data in the input data sets or to failed homology mappings) in the network. Specifically, the lack of proteins gI and pUS9, both associated to gE (gI physically and pUS9 functionally [58,59], as well as proteins gB and gD, which are the members of the fusion machinery (together with gH/gL [5]).

5. Discussion

Human herpesviruses are ubiquitous human pathogens with a severe socioeconomic impact worldwide [1]. Key to a successful infection is the delicate coordination of the complex network of virus-virus and virus-host protein interactions. Human herpesviruses encode distinctively large proteomes, which translate into large numbers of possible intra-viral PPIs. Several studies have attempted to address the description of these networks using a variety of, mostly, experimental approaches [60-63].

The aim of this study was to undertake a systematic and comprehensive analysis of PPI data in HSV1, the prototypical species of human

herpesviruses, to characterize the binary and higher order relationships among its encoded proteins. We addressed this using computational approaches to integrate and complement experimental binary PPI data, and analyse its modular structure.

461

A first step was to design a new computational framework for reconstructing the PPI network. The three pillars of this framework are: a) data integration; b) standardization; and c) reproducibility. Our method integrates experimentally supported and computationally predicted binary PPI data in a non-redundant manner using standardized molecular interaction data formats. Sequence-based orthology relationships are the basis of our PPI prediction approach. The search for homologues is conducted using the HHblits algorithm, the sensitivity and selectivity of which currently outperforms other available methods [17]. Next, our method applies more conservative criteria than previously [8] to select only the most confident candidate homologues among the results returned by HHblits. Additionally, all interactions are assessed under a common scoring scheme inspired by the standardized MIscore function [18], and which incorporates a new term to penalise the scores of computational predictions in a non-linear way. This term applies to interactions predicted based on protein conservation, and includes information on the prediction method (here, sequence-based interologues mapping), and the number of species from which the interaction is predicted. This aims to reflect that a broader cross-species conservation of protein interactions is indicative of a greater functional relevance of the conservation of the interaction. As a result of our updated, more conservative, protocol, we

observe that although the initial PPI dataset is larger than that used in our previous study (due to additional input databases) [8], the final PPI network is smaller. The reconstructed network is freely available to explore through the newly updated HVint database [8] interface: <http://topf-group.ismb.lon.ac.uk/hvint/>.

We tested our predictive strategy for the top 5% predicted interactions and we were able to find independent experimental support (i.e. not used to build the network) for 4 out of these 11 interactions. These included PPIs between pUL40 and pUL32, pUL40 and pUL52, pUL5 and pUL52, and the homodimer formed by pUL25 [8,21,22]. This highlights the power of our PPI prediction protocol in proposing new interactions for testing.

The community structure analysis performed on the virion subnetwork of HSV1 indicated the presence of four large communities, consistent with distinct stages of the virion formation process. This analysis suggests new functional relationships among virion components. The delineated communities illustrate events related to capsid formation, early tegumentation at peri- and juxtanuclear regions, and late tegumentation and virion release steps taking place in TGN-derived vesicles and the plasma membrane, respectively.

Community A drew our attention to protein pUS10, so far, a poorly characterized minor component of the virion specific to the *α -Herpesvirinae* subfamily. Our primary sequence analysis of pUS10 predicts the sequence to

be structurally divided into a disordered N-terminus, and an α -helical C-terminus likely to embed a single-pass transmembrane segment. Additionally, our analysis identified a seven-repeat length CLR, centrally in the protein sequence. CLR sequences are characterized by the GXY pattern, where X and Y are any amino acid, and adopt left-handed helical conformations that tend to tightly pack into trimeric right-handed helices [64,65]. CLRs have been identified in a range of organisms, from multicellular eukaryotes to unicellular bacteria and viruses, and they participate in a wide range of processes including e.g. adhesion, morphogenesis, and regulation of signaling cascades among others [64-67]. These features became especially interesting as they are highly reminiscent to those observed in another CLR-containing protein, gp12, from the bacteriophage SPP1 [68,69]. Importantly, SPP1 is one of the species of tailed dsDNA bacteriophages that have been shown to share a common evolutionary ancestry with herpesviruses [70]. Protein gp12 from SPP1 is a capsid-auxiliary protein that contains 8 CLRs centrally located in its sequence, and it is predicted to encode an α -helical C-terminus (Fig 7). The functional oligomeric state of gp12 is a trimer, assembled through the CLR region [69]. This was supported by circular dichroism analyses and confirmed by the cryo-electron microscopy reconstruction of the protein, which also indicated protein flexibility at one end [68,69]. Given the structural similarities with pUS10, it seems reasonable to assume that the functional form of pUS10 is also likely to be a trimer. Gp12 forms spikes that sit at the surface of the capsid, reversibly binding in a temperature-dependent manner, to the centre of the 60 capsid hexons. Binding of gp12 is thought to strengthen the adhesion properties of the virus [69]. Notably, in agreement with this, our

532 clustering analysis, also predicts close functional association of pUS10 capsid
533 proteins in HSV1.

534 **Fig 7. Structural and functional features of gp12 and pUS10.** (A)

535 Schematic comparison of the structural features of gp12 from bacteriophage
536 SPP1 and pUS10 of HSV1. Protein sequences are shown in grey, CLRs as
537 red boxes with the involved residues annotated, and predicted α -helical
538 regions are highlighted as yellow boxes. (B, C) Comparison of the known
539 biological function of gp12 (reversible binding to capsid hexons linked to
540 phage surface adhesion, (B)) and that hypothesised for pUS10 (C). The
541 interaction of pUS10 with HSV1 pUL37, predicted and experimentally
542 supported in this study, would take place in sites of secondary envelopment,
543 such as near Golgi- or trans-Golgi-network (TGN) derived membranes.
544 Accordingly, pUS10 might facilitate the dynamic interactions with membranes
545 in secondary envelopment of capsids through dynamic interactions with the
546 latter, in a similar fashion to gp12's interaction to adhesion surfaces, as well
547 as potentially with membranes by means of its predicted transmembrane
548 segment. Capsids (light blue) in (B) and (C) are drawn approximately to scale.

549 The identified evolutionary links between capsids of herpesviruses and tailed
550 dsDNA bacteriophages have so far included the conserved HK97-core fold
551 adopted by their major capsid proteins (MCPs), and their general assembly
552 pathways (transitioning from an immature procapsid to a mature, stable,
553 capsid) [68,71,72]. However, no evolutionary ancestry could be traced, based
554 on either sequence or structural features, amongst capsid-associated or
555 capsid auxiliary proteins of the two lineages. Here, we suggest that pUS10

could represent the first of such evolutionary vestiges. Clearly, SPP1 virions differ from those of herpesviruses in that they lack envelope, and, therefore, even in the presence of a common ancestor with gp12, one would expect pUS10 to have partially evolved new functionalities. In this context, the prediction of a transmembrane segment in the C-terminus of pUS10, which is not present in gp12, is interesting. Importantly, our network predicts an interaction between pUS10 and the inner tegument protein pUL37, which was further supported by our immunoaffinity experiments. The results of these experiments demonstrated a higher enrichment of pUS10 co-isolating with pUL37 later in infection (20 hpi). At these time points, pUL37 localises in secondary envelopment sites [37,38]. The functional role of an interaction between pUS10 and pUL37 could be that of facilitating the recruitment of the former to sites of secondary envelopment, to cooperatively function in subsequent capsid tegumentation events (Fig 7). An association of pUS10 to capsids via pUL37 would also explain why pUS10 has not yet been observed in the recent high-resolution structures of herpesviral capsids, [73-75]. On the basis of the structural similarities with gp12, it is reasonable to think that pUS10 could also exhibit reversible capsid/inner tegument-binding properties, similar to gp12, and establish dynamic interactions with capsid or capsid-associated proteins whilst, at the same time, interact with Golgi- or trans-Golgi-derived membranes, at the prospective capsids proximal pole of the forming virion, through its predicted transmembrane segment, and promote the final step in capsid trafficking during virion morphogenesis.

579

580 Together with pUS10, 9 out of the 14 proteins that our network data predict
581 interact with pUL37, were co-purified with the latter in our immunoaffinity
582 purification experiments (Fig 6D). Although these data cannot conclusively
583 answer which binary interactions are direct ones among the co-isolated
584 proteins, they add further support to the likely functional association of these
585 proteins, and therefore add to the confidence of the PPI involved in the
586 pUL37-centered subnetwork. Other interesting features of pUS10 will surely
587 help defining its functionalities. For instance, the polyproline sequence located
588 between two of the predicted C-terminus α -helices is also likely to be involved
589 in PPI [76].

590

591 Although further investigations on the roles of pUS10 during viral infection are
592 required, our analysis has revealed insightful new data on the structural
593 features and potential evolutionary history of this intriguing and, so far
594 understudied protein. Combined, our analyses have allowed to pose
595 functional hypotheses that will further our understanding of the intricate
596 tegumentation process of herpesvirus virions.

597

598 A second community sheds light to early tegumentation stages, specifically to
599 the recruitment of a group of transcription and translation modulators, playing
600 key roles at early stages of the lytic life cycle, into the virion particle. pUL48,
601 also called VP16 or α -TIF, is a transcription activator that migrates with
602 capsids to the host nucleus upon infection, and activates transcription of lytic
603 genes [77]. The roles of pUL47 and pUL49 are more diverse, however both of
604 them, together with pUL48 participate in regulating the activity of pUL41, also

605 known as the virus host shut-off (VHS) protein, which inhibits protein
606 synthesis [78]. Additionally, binary interactions among the four proteins seem
607 to be required for their incorporation to the virion [39,40,79,80]. Yet, the stage
608 at which this happens and the events that lead the process are unclear. Our
609 results suggest that the recruitment of this module into the particle could be
610 guided by the coordinated action of proteins pUL16, pUL11, pUL21, pUL14
611 and pUL56, potentially in peri- or juxtanuclear regions.

612

613 Two other communities are enriched in late stages of the tegumentation
614 process and virion release. We observed a large number of proteins
615 annotated with trafficking of TGN-derived vesicles and membrane-regulatory
616 events. Particularly, our analysis suggests a strong clustering tendency
617 between gK/UL20, pUL45, pUS2, and the poorly characterized α -
618 *Herpesvirinae*-specific protein pUL55 [53]. Primary sequence analysis
619 indicated pUL55 is a globular α/β protein with no predicted transmembrane
620 regions. However, all four proteins gK, pUL20, pUL45, and pUS2, are
621 membrane or membrane-associated proteins during intracellular stages of the
622 virus life cycle, and gK, pUL20, and pUL45 are also envelope-associated
623 proteins [54,81,82]. This finding supports the idea that pUL55 is a component
624 of the outer most layer of the tegument, and its recruitment might be
625 dependent on gK/UL20, pUL45, pUS2.

626

627 The outcomes of our protocol for predicting both binary and higher-order
628 protein associations, bring novel insights and direction to future experimental
629 testing. Further, we are currently working on the implementation of the

introduced network assembly and analysis frameworks on other species of human herpesviruses. Widening this analysis to other members of this important group of human pathogens will also underscore both conserved and species-specific features of their interactome organization, and help to explain the observed phenotypes and evolution of their pathogenic strategies.

Materials and Methods

Data collection, curation, and integration

Binary PPI data from molecular interaction repositories were downloaded in the standardised MITAB 2.5 format [20] (Fig 1). Binary PPIs from PDB and EMDb entries were manually extracted. Only entries with assigned PubMed identifier were considered. In the case of EMDb entries, only those with resolutions $\leq 5\text{\AA}$ at the time of this work were taken into account. When fitted atomic models were available, binary interactions between two proteins were extracted if the interface between them was larger than $500\text{\AA}^2$ [83-85]. When fitted atomic models were not available, binary interactions were assigned as described in the primary citation. The resulting set of interactions were next manually annotated following the MITAB 2.5 format.

The data were collected for the species in S1 and S2 Tables. Only binary PPIs for which both interacting proteins were annotated with a taxonomic identifier corresponding to these species (all strains considered) were selected. Taxonomic identifiers were obtained from the National Center for

655 Biotechnology Information (NCBI) [86] Taxonomy database
656 (<https://www.ncbi.nlm.nih.gov/taxonomy>).

657

658 Protein identifiers were mapped, where possible, to UniProtKB [7] accession
659 numbers. PPIs for which such mapping was not possible for one or both of the
660 interacting proteins, were not further considered in our pipeline. The seven
661 input PPI data sets were merged into a single non-redundant collection.

662

663

664 **Computational prediction of PPIs**

665

666 PPI predictions were obtained using an interologues mapping approach [16].
667 This method (Fig 1D) predicts an interaction in species X if two proteins,
668 known to interact in species Y, are conserved in both species X and Y. Here,
669 protein conservation was assessed using sequence-based orthology
670 predictions computed with the iterative Hidden Markov Model (HMM) profile
671 comparison algorithm implemented by HHblits [17]. For each query sequence,
672 the best match found in HSV1 satisfying all of the following conditions was
673 considered a reliable putative homologue: $\geq 20\%$ sequence identity, $\geq 30\%$
674 sequence similarity, $\geq 50\%$ query HMM profile coverage, and $\geq 95\%$ probability
675 of being a true positive. The resulting set of candidate homology relationships
676 were used to infer interologues in each target interactomes.

677

678 **Integration of validated and predicted PPIs**

679

Experimentally validated and computationally predicted interactions were merged into a single interactome data set. Strain redundancy was removed by mapping all protein sequences to reference strain accession numbers (HSV1 strain 17) using UniRef90 clusters.

PPI scoring function

The scoring scheme integrated in our framework is inspired by the standardised MIscore function [18]. Under this scheme, PPIs that had experimental support in the target species (with or without additional support from computational predictions) were scored using the MIscore function. This was done through the MImerge service [18]. PPIs that did not have experimental support were scored with a new scoring function, defined as in Equation 1.

$$Score = MIscore \times (Penalty\ function) \text{ (Equation 1)}$$

The new scoring function consists of first scoring an interaction using the MIscore [18], and next applying a penalty function to the returned score. This scaling factor (Equation 2) takes as reference the structure of the terms used in the MIscore function, but it redefines the meaning of their parameters to incorporate information on the number of species and prediction method used.

$$Scaling\ factor = \log_{(2a+1)}(a + 1) \text{ (Equation 2)}$$

705 where

$$a = scv \times n$$

706

707 where *scv* is the score associated to the PPI prediction method (in this case
708 interology mapping), and *n* is the number of different species from which the
709 interaction was predicted.

710 The value of the penalty function increases asymptotically from ~0.5 to 1 with
711 the number of species from which an interaction is predicted. Because we
712 considered 10 orthologues to each target species, the values of the scaling
713 factor in this study fall in the [0.5, 0.6] range. After applying the penalty
714 function the values are normalised within [0, 1].

715

716 **Consensus clustering framework**

717

718 Prior to clustering, the signal-to-noise ratio in the network data was increased
719 as follows (Fig 3A). First, the subnetwork formed by proteins found in the
720 mature virion particle of HSV1 (i.e. HSV1 virion subnetwork) was extracted
721 from the full network. Next, self-interactions were removed as they do not add
722 topological information about the network. Interactions between capsid and
723 envelope proteins were also eliminated as they are precluded by the presence
724 of the tegument in between the two former layers. Finally, the LCC was
725 obtained and used as input to the consensus clustering framework.

726

727 A total of 14 different clustering algorithms, representing a broad spectrum of
 728 clustering techniques, were then applied to the LCC (Fig 3B), i.e. the K-means
 729 algorithm [87], agglomerative hierarchical clustering [88], K-
 730 means/hierarchical clustering [89], Fuzzy C-means [90], Model-based
 731 clustering [91,92], Markov Cluster algorithm (MCL) algorithm [93], Density-
 732 based clustering [94], Edge betweenness [95,96], Louvain method [97], and
 733 Leading eigenvector [98], Fast greedy [28], Walktrap [99], and InfoMap [100]
 734 algorithms. Where needed, partition selection and parameter optimisation
 735 were conducted based on modularity [95] maximisation. The clustering results
 736 of those partitions yielding positive modularity were integrated into a
 737 consensus agreement matrix (CAM) (Fig 3C and S4 Table). The condition of
 738 positive modularity discarded partitions where each and every node was
 739 singled out as a cluster by itself, the entire network was treated as a super-
 740 cluster, or the graph under such partition had less local structure than random
 741 models. The columns and rows of the consensus agreement matrix
 742 represented the network proteins; the matrix values indicated the fraction of
 743 accepted base partitions classifying two proteins in the same cluster.

744

745 A final consensus partition was derived from the CAM by applying a filtering
 746 threshold (Fig 3D and S5 Table). For thresholds between 0 and 1, by
 747 increases of 0.5, we iteratively calculated a consensus partition in the
 748 following way. CAM cells with values equal to or below the threshold were
 749 ignored; the remaining cells defined a partition for which the modularity was
 750 calculated. At the end of this iterative process, the partition associated to the
 751 highest modularity was chosen as consensus.

752

753

754 **Sequence analysis**

755

756 The canonical sequence of pUS10 and pUL55 in the HSV1 reference
 757 proteome were obtained from UniProtKB [7]. For each sequence, the
 758 following analysis was conducted. ScanProsite [101] was used to scan the
 759 sequences for sequence motifs. Potential sequence homologues within the
 760 entire UniProtKB database [7], were searched for using HHblits [17]. Next, a
 761 consensus prediction of secondary structure elements was inferred from the
 762 results of four different secondary structure prediction tools, i.e. SPIDER²
 763 [102], PSIPRED [103], JPred4 [104], and PSSpred [105]. Probabilities
 764 associated to the returned predictions were not integrated in the consensus
 765 analysis. Similarly, consensus predictions for transmembrane segments were
 766 derived from five different algorithms, i.e. Dense Alignment Surface (DAS)
 767 [106], Phobius [107], PHDhtml [108], TMPred [109], and MEMSAT-SVM [110].
 768 From TMPred predictions only significant regions (defined as regions with
 769 score above 500) and core residues were taken into consideration. From
 770 Phobius, only residues with probability of belonging to a transmembrane
 771 region above 0.1 were considered. Finally, a consensus prediction for
 772 disordered regions was built from the results of two algorithms, i.e.
 773 DISOPRED [111] and MetaDisorder [112].

774

775 **Infection of human fibroblasts with HSV1 and live cell imaging**

We used the HSV1(17⁺)Lox-UL37GFP strain, a generous gift from B. Sodeik and previously characterized in [37], here denoted HSV1-UL37EGFP, and as a control, the HSV1(17⁺)Lox-P_{MCMV}GFP strain, denoted HSV1-EGFP, which expresses EGFP alone inserted between the pUL55 and pUL56 ORFs, under the control of the murine cytomegalovirus promoter [113]. Viruses were propagated, isolated, and titered in Vero cells (ATCC CCL81) grown in DMEM containing 10% FBS and 1% penicillin/streptomycin (P/S), as previously described [8]. Primary human foreskin fibroblast cells were infected with HSV-1 strains at 10 plaque forming units/cell using a cold-synchronized protocol [114]. The progression of infection was visualized by live-cell imaging on a Nikon Ti-Eclipse epifluorescence inverted microscope from 2 hpi to 24 hpi. Images were viewed and analysed by ImageJ [115].

Immunoaffinity Purification Quantitative Mass Spectrometry

For IP-MS experiments, cells were infected as above with HSV1-UL37GFP or control HSV1-GFP, in duplicate. Infected cells were collected at 8 and 20 hpi (HSV1-UL37GFP) or 20 hpi (HSV1-GFP) in ice-cold PBS and pelleted by centrifugation ($\sim 1 \times 10^7$ per time point per replicate). Cell pellets were washed in ice-cold PBS and lysed hypotonically. Cytosolic lysates were adjusted to 20 mM HEPES-KOH, pH 7.4, containing 0.11 M potassium acetate, 2 mM MgCl₂, 0.1% Tween 20, 1 μ M ZnCl₂, 1 μ M CaCl₂, 250 mM NaCl, and 0.5% NP-40, mixed by Polytron homogenization, and centrifuged at 8,000 x g for 10 min at 4°C. The supernatant was recovered and subjected to immunoaffinity

800 purification using magnetic beads conjugated with in-house generated rabbit
801 anti-GFP antibodies, as previously described [114,116].

802 Immunoisolated proteins were processed by a Filter-Aided Sample
803 Preparation method using Amicon ultrafiltration devices (Millipore, 30 kDa
804 MWCO) as described [117], except 0.1 M Tris-HCl, pH 7.9 was replaced with
805 0.1 M triethylammonium bicarbonate (TEAB). Following overnight trypsin
806 digestion and clean-up, peptides (4 µl) were analyzed by nanoliquid
807 chromatography-tandem mass spectrometry on a Dionex Ultimate 3000
808 RSLC coupled directly to a LTQ Orbitrap Velos ETD configured with a
809 Nanospray ion source (Thermofisher Scientific).

810 The Proteome Discoverer software (ver. 2.2) was used for post-acquisition
811 mass recalibration of precursor and fragment ions masses, MS/MS spectrum
812 extraction, peptide spectrum matching and validation, calculation of TMT
813 reporter ion intensities, and assembly of quantified into protein groups. Protein
814 groups and TMT protein abundances for herpesvirus proteins with a minimum
815 of 2 unique quantified peptides were exported to Excel. IP protein enrichment
816 ratios for each time point and replicate were calculated as the TMT
817 abundance ratio of pUL37GFP / GFP. Proteins with IP enrichment ratios of ≥
818 2-fold in at least one time point in both replicates were considered specific
819 associations. The TMT abundance ratio for proteins in the 20 vs 8 hpi
820 pUL37GFP IPs were calculated after normalization by the pUL37 TMT
821 abundance. Further details on data collection and analysis can be found in S1
822 Text.

823

824 **Acknowledgments**

825 We are grateful for funding from Human Frontiers Science Program
826 (RGY0079/2009-C) to K.G., M.T., and I.M.C., from the NIH (GM114141) and
827 a Mallinckrodt Scholar Award to I.M.C., from the Wellcome Trust
828 (209250/Z/17/Z) to M.T. and K.G., and (107806/Z/15/Z) and a WT Core
829 Award (203141/Z/16/Z) to K.G., and from the MRC (MR/M019292/1) to M.T.
830 and K.G.

831 **References**

832

- 833 1. Arvin A, Campadelli-Fiume G, Mocarski E, Moore PS, Roizman B,
834 Whitley R, et al. Human Herpesviruses: Biology, Therapy, and
835 Immunoprophylaxis. Cambridge: Cambridge University Press;
836 2007.
- 837 2. Eimer WA, Vijaya Kumar DK, Navalpur Shanmugam NK,
838 Rodriguez AS, Mitchell T, Washicosky KJ, et al. Alzheimer's
839 Disease-Associated β -Amyloid Is Rapidly Seeded by Herpesviridae
840 to Protect against Brain Infection. *Neuron*. 2018;99: 56–63.e3.
841 doi:10.1016/j.neuron.2018.06.030
- 842 3. Readhead B, Haure-Mirande J-V, Funk CC, Richards MA, Shannon
843 P, Haroutunian V, et al. Multiscale Analysis of Independent
844 Alzheimer's Cohorts Finds Disruption of Molecular, Genetic, and
845 Clinical Networks by Human Herpesvirus. *Neuron*. Cell Press;
846 2018;99: 64–82.e7. doi:10.1016/j.neuron.2018.05.023

- 847 4. Grünewald K, Desai P, Winkler DC, Heymann JB, Belnap DM,
848 Baumeister W, et al. Three-Dimensional Structure of Herpes
849 Simplex Virus from Cryo-Electron Tomography. *Science. American*
850 Association for the Advancement of Science; 2003;302: 1396–
851 1398. doi:10.1126/science.1090284

- 852 5. Spear PG, Longnecker R. Herpesvirus entry: an update. *J Virol.*
853 American Society for Microbiology (ASM); 2003;77: 10179–10185.
854 doi:10.1128/JVI.77.19.10179-10185.2003

- 855 6. Arvin A, Campadelli-Fiume G, Mocarski E, Moore PS, Roizman B,
856 Whitley R, et al. Comparative virion structures of human
857 herpesviruses. Cambridge: Cambridge University Press; 2007.

- 858 7. Consortium TU. UniProt: the universal protein knowledgebase.
859 *Nucleic Acids Res.* 2017;45: D158–D169. doi:10.1093/nar/gkw1099

- 860 8. Ashford P, Hernandez A, Greco TM, Buch A, Sodeik B, Cristea IM,
861 et al. HVint: A Strategy for Identifying Novel Protein-Protein
862 Interactions in Herpes Simplex Virus Type 1. *Mol Cell Proteomics.*
863 American Society for Biochemistry and Molecular Biology; 2016;15:
864 2939–2953. doi:10.1074/mcp.M116.058552

- 865 9. Stark C, Breitkreutz B-J, Reguly T, Boucher L, Breitkreutz A, Tyers
866 M. BioGRID: a general repository for interaction datasets. *Nucleic*
867 *Acids Res.* 2006;34: D535–9. doi:10.1093/nar/gkj109

- 868 10. Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM,
869 Eisenberg D. DIP: the database of interacting proteins. *Nucleic*

- 870 Acids Res. 2000;28: 289–291.
- 871 11. Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-
872 Carter F, et al. The MIntAct project--IntAct as a common curation
873 platform for 11 molecular interaction databases. Nucleic Acids Res.
874 2014;42: D358–63. doi:10.1093/nar/gkt1115
- 875 12. Calderone A, Castagnoli L, Cesareni G. mentha: a resource for
876 browsing integrated protein-interaction networks. Nat Methods.
877 Nature Publishing Group; 2013;10: 690–691.
878 doi:10.1038/nmeth.2561
- 879 13. Guirimand T, Delmotte S, Navratil V. VirHostNet 2.0: surfing on the
880 web of virus/host molecular interactions data. Nucleic Acids Res.
881 2015;43: D583–7. doi:10.1093/nar/gku1121
- 882 14. Berman H, Henrick K, Nakamura H. Announcing the worldwide
883 Protein Data Bank. Nat Struct Mol Biol. Nature Publishing Group;
884 2003;10: 980–980. doi:10.1038/nsb1203-980
- 885 15. Lawson CL, Baker ML, Best C, Bi C, Dougherty M, Feng P, et al.
886 EMDataBank.org: unified data resource for CryoEM. Nucleic Acids
887 Res. 2011;39: D456–64. doi:10.1093/nar/gkq880
- 888 16. Yu H. Annotation Transfer Between Genomes: Protein-Protein
889 Interologs and Protein-DNA Regulogs. Genome Research. Cold
890 Spring Harbor Lab; 2004;14: 1107–1118. doi:10.1101/gr.1774904
- 891 17. Remmert M, Biegert A, Hauser A, Söding J. HHblits: lightning-fast

- 892 iterative protein sequence searching by HMM-HMM alignment. Nat
893 Methods. Nature Publishing Group; 2011;9: 173–175.
894 doi:10.1038/nmeth.1818
- 895 18. Villaveces JM, Jiménez RC, Porras P, Del-Toro N, Duesbury M,
896 Dumousseau M, et al. Merging and scoring molecular interactions
897 utilising existing community standards: tools, use-cases and a case
898 study. Database (Oxford). 2015;2015: bau131–bau131.
899 doi:10.1093/database/bau131
- 900 19. Orchard S, Hermjakob H, Apweiler R. The proteomics standards
901 initiative. Proteomics. Wiley-Blackwell; 2003;3: 1374–1376.
902 doi:10.1002/pmic.200300496
- 903 20. Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L,
904 Ceol A, et al. The HUPO PSI's molecular interaction format--a
905 community standard for the representation of protein interaction
906 data. Nat Biotechnol. Nature Publishing Group; 2004;22: 177–183.
907 doi:10.1038/nbt926
- 908 21. Biswas N, Weller SK. The UL5 and UL52 subunits of the herpes
909 simplex virus type 1 helicase-primase subcomplex exhibit a
910 complex interdependence for DNA binding. J Biol Chem. American
911 Society for Biochemistry and Molecular Biology; 2001;276: 17610–
912 17619. doi:10.1074/jbc.M010107200
- 913 22. Dai X, Zhou ZH. Structure of the herpes simplex virus 1 capsid with
914 associated tegument protein complexes. Science. American

- 915 Association for the Advancement of Science; 2018;360: eaao7298.
916 doi:10.1126/science.aao7298
- 917 23. Bechtel JT, Winant RC, Ganem D. Host and Viral Proteins in the
918 Virion of Kaposi's Sarcoma-Associated Herpesvirus. J Virol.
919 American Society for Microbiology Journals; 2005;79: 4952–4964.
920 doi:10.1128/JVI.79.8.4952-4964.2005
- 921 24. Loret S, Guay G, Lippé R. Comprehensive characterization of
922 extracellular herpes simplex virus type 1 virions. J Virol. American
923 Society for Microbiology Journals; 2008;82: 8605–8618.
924 doi:10.1128/JVI.00904-08
- 925 25. Human cytomegalovirus virion proteins. Human Immunology.
926 Elsevier; 2004;65: 395–402. doi:10.1016/j.humimm.2004.02.008
- 927 26. Johannsen E, Luftig M, Chase MR, Weicksel S, Cahir-McFarland
928 E, Illanes D, et al. Proteins of purified Epstein-Barr virus. PNAS.
929 National Academy of Sciences; 2004;101: 16286–16291.
930 doi:10.1073/pnas.0407320101
- 931 27. Newman MEJ, Girvan M. Finding and evaluating community
932 structure in networks. Phys Rev E Stat Nonlin Soft Matter Phys.
933 American Physical Society; 2004;69: 026113.
934 doi:10.1103/PhysRevE.69.026113
- 935 28. Clauset A, Newman MEJ, Moore C. Finding community structure in
936 very large networks. Phys Rev E Stat Nonlin Soft Matter Phys.
937 American Physical Society; 2004;70: 066111.

- 938 doi:10.1103/PhysRevE.70.066111
- 939 29. research GOCNA, 2016. Expansion of the Gene Ontology
- 940 knowledgebase and resources. academicoupcom
- 941 .
- 942 30. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM,
- 943 et al. Gene ontology: tool for the unification of biology. The Gene
- 944 Ontology Consortium. Nat Genet. Nature Publishing Group;
- 945 2000;25: 25–29. doi:10.1038/75556
- 946 31. Kornfeind EM, Visalli RJ. Human herpesvirus portal proteins:
- 947 Structure, function, and antiviral prospects. Reviews in Medical
- 948 Virology. Wiley-Blackwell; 2018;28: e1972. doi:10.1002/rmv.1972
- 949 32. Yamada H, Daikoku T, Yamashita Y, Jiang YM, Tsurumi T,
- 950 Nishiyama Y. The product of the US10 gene of herpes simplex
- 951 virus type 1 is a capsid/tegument-associated phosphoprotein which
- 952 copurifies with the nuclear matrix. J Gen Virol. Microbiology
- 953 Society; 1997;78 (Pt 11): 2923–2931. doi:10.1099/0022-1317-78-
- 954 11-2923
- 955 33. Roger Holden V, Yalamanchili RR, Harty RN, O'Callaghan DJ.
- 956 Identification and characterization of an equine herpesvirus 1 late
- 957 gene encoding a potential zinc finger. Virology. Academic Press;
- 958 1992;188: 704–713. doi:10.1016/0042-6822(92)90525-T
- 959 34. Charalambous BM, Keen JN, McPherson MJ. Collagen-like

- 960 sequences stabilize homotrimers of a bacterial hydrolase. EMBO J.
961 European Molecular Biology Organization; 1988;7: 2903–2909.
- 962 35. Desai P, Sexton GL, McCaffery JM, Person S. A null mutation in
963 the gene encoding the herpes simplex virus type 1 UL37
964 polypeptide abrogates virus maturation. J Virol. American Society
965 for Microbiology; 2001;75: 10259–10271.
966 doi:10.1128/JVI.75.21.10259-10271.2001
- 967 36. Padeloup D, Beilstein F, Roberts APE, McElwee M, McNab D,
968 Rixon FJ. Inner tegument protein pUL37 of herpes simplex virus
969 type 1 is involved in directing capsids to the trans-Golgi network for
970 envelopment. Journal of General Virology. Microbiology Society;
971 2010;91: 2145–2151. doi:10.1099/vir.0.022053-0
- 972 37. Sandbaumhüter M, Döhner K, Schipke J, Binz A, Pohlmann A,
973 Sodeik B, et al. Cytosolic herpes simplex virus capsids not only
974 require binding inner tegument protein pUL36 but also pUL37 for
975 active transport prior to secondary envelopment. Cell Microbiol.
976 Wiley/Blackwell (10.1111); 2013;15: 248–269.
977 doi:10.1111/cmi.12075
- 978 38. Desai P, Sexton GL, Huang E, Person S. Localization of herpes
979 simplex virus type 1 UL37 in the Golgi complex requires UL36 but
980 not capsid structures. J Virol. American Society for Microbiology
981 Journals; 2008;82: 11354–11361. doi:10.1128/JVI.00956-08
- 982 39. Zhang Y, Sirko DA, McKnight JL. Role of herpes simplex virus type

- 983 1 UL46 and UL47 in alpha TIF-mediated transcriptional induction:
984 characterization of three viral deletion mutants. J Virol. American
985 Society for Microbiology; 1991;65: 829–841.
- 986 40. Smibert CA, Popova B, Xiao P, Capone JP, Smiley JR. Herpes
987 simplex virus VP16 forms a complex with the virion host shutoff
988 protein vhs. J Virol. American Society for Microbiology (ASM);
989 1994;68: 2339–2346.
- 990 41. Elliott G, Hafezi W, Whiteley A, Bernard E. Deletion of the herpes
991 simplex virus VP22-encoding gene (UL49) alters the expression,
992 localization, and virion incorporation of ICP0. J Virol. 2005;79:
993 9735–9745. doi:10.1128/JVI.79.15.9735-9745.2005
- 994 42. Klupp BG, Böttcher S, Granzow H, Kopp M, Mettenleiter TC.
995 Complex formation between the UL16 and UL21 tegument proteins
996 of pseudorabies virus. J Virol. American Society for Microbiology
997 Journals; 2005;79: 1510–1522. doi:10.1128/JVI.79.3.1510-
998 1522.2005
- 999 43. Yeh P-C, Meckes DG, Wills JW. Analysis of the interaction
1000 between the UL11 and UL16 tegument proteins of herpes simplex
1001 virus. J Virol. American Society for Microbiology Journals; 2008;82:
1002 10693–10700. doi:10.1128/JVI.01230-08
- 1003 44. Harper AL, Meckes DG, Marsh JA, Ward MD, Yeh P-C, Baird NL,
1004 et al. Interaction domains of the UL16 and UL21 tegument proteins
1005 of herpes simplex virus. J Virol. American Society for Microbiology;

- 1006 2010;84: 2963–2971. doi:10.1128/JVI.02015-09
- 1007 45. Chadha P, Han J, Starkey JL, Wills JW. Regulated interaction of
1008 tegument proteins UL16 and UL11 from herpes simplex virus. J
1009 Virol. American Society for Microbiology Journals; 2012;86: 11886–
1010 11898. doi:10.1128/JVI.01879-12
- 1011 46. Tunbäck P, Liljeqvist JA, Löwhagen GB, Bergström T. Glycoprotein
1012 G of herpes simplex virus type 1: identification of type-specific
1013 epitopes by human antibodies. J Gen Virol. Microbiology Society;
1014 2000;81: 1033–1040. doi:10.1099/0022-1317-81-4-1033
- 1015 47. Klupp BG, Altenschmidt J, Granzow H, Fuchs W, Mettenleiter TC.
1016 Identification and characterization of the pseudorabies virus UL43
1017 protein. Virology. 2005;334: 224–233.
1018 doi:10.1016/j.virol.2005.01.032
- 1019 48. Aubert M, Krantz EM, Jerome KR. Herpes simplex virus genes
1020 Us3, Us5, and Us12 differentially regulate cytotoxic T lymphocyte-
1021 induced cytotoxicity. Viral Immunol. Mary Ann Liebert, Inc. 2
1022 Madison Avenue Larchmont, NY 10538 USA; 2006;19: 391–408.
1023 doi:10.1089/vim.2006.19.391
- 1024 49. Aubert M, Chen Z, Lang R, Dang CH, Fowler C, Sloan DD, et al.
1025 The antiapoptotic herpes simplex virus glycoprotein J localizes to
1026 multiple cellular organelles and induces reactive oxygen species
1027 formation. J Virol. American Society for Microbiology Journals;
1028 2008;82: 617–629. doi:10.1128/JVI.01341-07

- 1029 50. Dollery SJ, Lane KD, Delboy MG, Roller DG, Nicola AV. Role of the
1030 UL45 protein in herpes simplex virus entry via low pH-dependent
1031 endocytosis and its relationship to the conformation and function of
1032 glycoprotein B. *Virus Res.* 2010;149: 115–118.
1033 doi:10.1016/j.virusres.2010.01.004
- 1034 51. Kasmi EI, Lippé R. Herpes Simplex Virus 1 gN Partners with gM
1035 To Modulate the Viral Fusion Machinery. Longnecker RM, editor. *J*
1036 *Virol.* American Society for Microbiology; 2015;89: 2313–2323.
1037 doi:10.1128/JVI.03041-14
- 1038 52. Lau S-YK, Crump CM. HSV-1 gM and the gK/pUL20 complex are
1039 important for the localization of gD and gH/L to viral assembly sites.
1040 *Viruses.* Multidisciplinary Digital Publishing Institute; 2015;7: 915–
1041 938. doi:10.3390/v7030915
- 1042 53. Yamada H, Jiang YM, Oshima S, Daikoku T, Yamashita Y, Tsurumi
1043 T, et al. Characterization of the UL55 gene product of herpes
1044 simplex virus type 2. *Journal of General Virology.* Microbiology
1045 Society; 1998;79: 1989–1995. doi:10.1099/0022-1317-79-8-1989
- 1046 54. Kang M-H, Roy BB, Finnen RL, Le Sage V, Johnston SM, Zhang H,
1047 et al. The Us2 gene product of herpes simplex virus 2 is a
1048 membrane-associated ubiquitin-interacting protein. *J Virol.*
1049 2013;87: 9590–9603. doi:10.1128/JVI.00994-13
- 1050 55. Roller RJ, Haugo AC, Yang K, Baines JD. The herpes simplex virus
1051 1 UL51 gene product has cell type-specific functions in cell-to-cell

1052 spread. J Virol. 2014;88: 4058–4068. doi:10.1128/JVI.03707-13

1053 56. Albecka A, Owen DJ, Ivanova L, Brun J, Liman R, Davies L, et al.
1054 Dual Function of the pUL7-pUL51 Tegument Protein Complex in
1055 Herpes Simplex Virus 1 Infection. Sandri-Goldin RM, editor. J Virol.
1056 American Society for Microbiology Journals; 2017;91: 448.
1057 doi:10.1128/JVI.02196-16

1058 57. Lu X, Huang C, Zhang Y, Lin Y, Wang X, Li Q, et al. The Us2 Gene
1059 Product of Herpes Simplex Virus 2 modulates NF-κB activation by
1060 targeting TAK1. Sci Rep. Nature Publishing Group; 2017;7: 8396.
1061 doi:10.1038/s41598-017-08856-4

1062 58. Dingwell KS, Johnson DC. The herpes simplex virus gE-gI complex
1063 facilitates cell-to-cell spread and binds to components of cell
1064 junctions. J Virol. American Society for Microbiology; 1998;72:
1065 8933–8942.

1066 59. DuRaine G, Wisner TW, Howard P, Williams M, Johnson DC.
1067 Herpes Simplex Virus gE/gI and US9 Promote both Envelopment
1068 and Sorting of Virus Particles in the Cytoplasm of Neurons, Two
1069 Processes That Precede Anterograde Transport in Axons.
1070 Longnecker RM, editor. J Virol. American Society for Microbiology
1071 Journals; 2017;91: 153. doi:10.1128/JVI.00050-17

1072 60. Fossum E, Friedel CC, Rajagopala SV, Titz B, Baiker A, Schmidt T,
1073 et al. Evolutionarily conserved herpesviral protein interaction
1074 networks. Sun R, editor. PLoS Pathog. Public Library of Science;

- 1075 2009;5: e1000570. doi:10.1371/journal.ppat.1000570
- 1076 61. Vittone V, Diefenbach E, Triffett D, Douglas MW, Cunningham AL,
1077 Diefenbach RJ. Determination of interactions between tegument
1078 proteins of herpes simplex virus type 1. J Virol. American Society
1079 for Microbiology Journals; 2005;79: 9566–9571.
1080 doi:10.1128/JVI.79.15.9566-9571.2005
- 1081 62. Lee JH, Vittone V, Diefenbach E, Cunningham AL, Diefenbach RJ.
1082 Identification of structural protein–protein interactions of herpes
1083 simplex virus type 1. Virology. Academic Press; 2008;378: 347–
1084 354. doi:10.1016/j.virol.2008.05.035
- 1085 63. Calderwood MA, Venkatesan K, Xing L, Chase MR, Vazquez A,
1086 Holthaus AM, et al. Epstein-Barr virus and virus human protein
1087 interaction maps. PNAS. National Academy of Sciences; 2007;104:
1088 7606–7611. doi:10.1073/pnas.0702332104
- 1089 64. Brodsky B, Ramshaw JA. The collagen triple-helix structure. Matrix
1090 Biol. 1997;15: 545–554.
- 1091 65. Yu Z, An B, Ramshaw JAM, Brodsky B. Bacterial collagen-like
1092 proteins that form triple-helical structures. Journal of Structural
1093 Biology. 2014;186: 451–461. doi:10.1016/j.jsb.2014.01.003
- 1094 66. Rasmussen M, Jacobsson M, Björck L. Genome-based
1095 Identification and Analysis of Collagen-related Structural Motifs in
1096 Bacterial and Viral Proteins. J Biol Chem. American Society for
1097 Biochemistry and Molecular Biology; 2003;278: 32313–32316.

1098 doi:10.1074/jbc.M304709200

1099 67. Mienaltowski MJ, Birk DE. Structure, physiology, and biochemistry
1100 of collagens. *Adv Exp Med Biol.* Dordrecht: Springer Netherlands;
1101 2014;802: 5–29. doi:10.1007/978-94-007-7893-1_2

1102 68. White HE, Sherman MB, Brasilès S, Jacquet E, Seavers P,
1103 Tavares P, et al. Capsid structure and its stability at the late stages
1104 of bacteriophage SPP1 assembly. *J Virol. American Society for*
1105 *Microbiology Journals*; 2012;86: 6768–6777.
1106 doi:10.1128/JVI.00412-12

1107 69. Zairi M, Stiege AC, Nhiri N, Jacquet E, Tavares P. The collagen-
1108 like protein gp12 is a temperature-dependent reversible binder of
1109 SPP1 viral capsids. *J Biol Chem. American Society for*
1110 *Biochemistry and Molecular Biology*; 2014;289: 27169–27181.
1111 doi:10.1074/jbc.M114.590877

1112 70. Baker ML, Jiang W, Rixon FJ, Chiu W. Common ancestry of
1113 herpesviruses and tailed DNA bacteriophages. *J Virol. American*
1114 *Society for Microbiology Journals*; 2005;79: 14967–14970.
1115 doi:10.1128/JVI.79.23.14967-14970.2005

1116 71. Gertsman I, Gan L, Guttman M, Lee K, Speir JA, Duda RL, et al.
1117 An unexpected twist in viral capsid maturation. *Nature.* 2009;458:
1118 646–650. doi:10.1038/nature07686

1119 72. Parent KN, Khayat R, Tu LH, Suhanovsky MM, Cortines JR,
1120 Teschke CM, et al. P22 coat protein structures reveal a novel

- 1121 mechanism for capsid maturation: stability without auxiliary proteins
1122 or chemical crosslinks. *Structure*. 2010;18: 390–401.
1123 doi:10.1016/j.str.2009.12.014
- 1124 73. Dai X, Zhou ZH. Structure of the herpes simplex virus 1 capsid with
1125 associated tegument protein complexes. *Science*. American
1126 Association for the Advancement of Science; 2018;360: eaao7298.
1127 doi:10.1126/science.aao7298
- 1128 74. Yu X, Jih J, Jiang J, Zhou ZH. Atomic structure of the human
1129 cytomegalovirus capsid with its securing tegument layer of pp150.
1130 *Science*. American Association for the Advancement of Science;
1131 2017;356: eaam6892. doi:10.1126/science.aam6892
- 1132 75. Dai X, Gong D, Lim H, Jih J, Wu T-T, Sun R, et al. Structure and
1133 mutagenesis reveal essential capsid protein interactions for KSHV
1134 replication. *Nature*. Nature Publishing Group; 2018;553: 521–525.
1135 doi:10.1038/nature25438
- 1136 76. Berisio R, Vitagliano L. Polyproline and triple helix motifs in host-
1137 pathogen recognition. *Curr Protein Pept Sci*. Bentham Science
1138 Publishers; 2012;13: 855–865. doi:10.2174/138920312804871157
- 1139 77. Wysocka J, Herr W. The herpes simplex virus VP16-induced
1140 complex: the makings of a regulatory switch. *Trends Biochem Sci*.
1141 2003;28: 294–304. doi:10.1016/S0968-0004(03)00088-4
- 1142 78. Smiley JR. Herpes simplex virus virion host shutoff protein: immune
1143 evasion mediated by a viral RNase? *J Virol*. American Society for

- 1144 Microbiology (ASM); 2004;78: 1063–1068.
- 1145 doi:10.1128/JVI.78.3.1063-1068.2004
- 1146 79. Elliott G, Mouzakis G, O'Hare P. VP16 interacts via its activation
- 1147 domain with VP22, a tegument protein of herpes simplex virus, and
- 1148 is relocated to a novel macromolecular assembly in coexpressing
- 1149 cells. J Virol. American Society for Microbiology (ASM); 1995;69:
- 1150 7932–7941.
- 1151 80. Taddeo B, Sciortino MT, Zhang W, Roizman B. Interaction of
- 1152 herpes simplex virus RNase with VP16 and VP22 is required for
- 1153 the accumulation of the protein but not for accumulation of mRNA.
- 1154 PNAS. National Academy of Sciences; 2007;104: 12163–12168.
- 1155 doi:10.1073/pnas.0705245104
- 1156 81. Dollery SJ, Lane KD, Delboy MG, Roller DG, Nicola AV. Role of the
- 1157 UL45 protein in herpes simplex virus entry via low pH-dependent
- 1158 endocytosis and its relationship to the conformation and function of
- 1159 glycoprotein B. Virus Res. Elsevier; 2010;149: 115–118.
- 1160 doi:10.1016/j.virusres.2010.01.004
- 1161 82. Lau S-YK, Crump CM. HSV-1 gM and the gK/pUL20 complex are
- 1162 important for the localization of gD and gH/L to viral assembly sites.
- 1163 Viruses. Multidisciplinary Digital Publishing Institute; 2015;7: 915–
- 1164 938. doi:10.3390/v7030915
- 1165 83. Jones S, Thornton JM. Principles of protein-protein interactions.
- 1166 PNAS. National Academy of Sciences; 1996;93: 13–20.

- 1167 84. Bogan AA, Thorn KS. Anatomy of hot spots in protein interfaces. J
1168 Mol Biol. 1998;280: 1–9. doi:10.1006/jmbi.1998.1843
- 1169 85. Day ES, Cote SM, Whitty A. Binding efficiency of protein-protein
1170 complexes. Biochemistry. American Chemical Society; 2012;51:
1171 9124–9136. doi:10.1021/bi301039t
- 1172 86. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K,
1173 Chetvernin V, et al. Database resources of the National Center for
1174 Biotechnology Information. Nucleic Acids Res. 2006;34: D173–80.
1175 doi:10.1093/nar/gkj158
- 1176 87. MacQueen J. Some methods for classification and analysis of
1177 multivariate observations. Berkeley, Calif.: University of California
1178 Press; 1967. pp. 281–297.
- 1179 88. Franklin J. The elements of statistical learning: data mining,
1180 inference and prediction. The Mathematical Intelligencer. Springer-
1181 Verlag; 2008;27: 83–85. doi:10.1007/BF02985802
- 1182 89. Bernard Chen, Harrison R, Yi Pan, Phang C Tai. Novel Hybrid
1183 Hierarchical-K-means Clustering Method (H-K-means) for
1184 Microarray Analysis. IEEE; pp. 105–108.
1185 doi:10.1109/CSBW.2005.98
- 1186 90. Liu H-C, Wu D-B, Yih J-M, Liu S-W. Fuzzy c-Mean Algorithm
1187 Based on Complete Mahalanobis Distances and Separable
1188 Criterion. IEEE; 2008. pp. 87–91. doi:10.1109/FSKD.2008.34

- 1189 91. Banfield JD, Raftery AE. Model-Based Gaussian and Non-
1190 Gaussian Clustering. *Biometrics*. 1993;49: 803.
1191 doi:10.2307/2532201
- 1192 92. Fraley C, Raftery AE. MCLUST: Software for Model-Based Cluster
1193 Analysis. *Journal of Classification*. Springer-Verlag; 2014;16: 297–
1194 306. doi:10.1007/s003579900058
- 1195 93. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for
1196 large-scale detection of protein families. *Nucleic Acids Res. Oxford*
1197 *University Press*; 2002;30: 1575–1584.
- 1198 94. Ester M, Kriegel H-P, Sander JOR, Xu X. A Density-based
1199 Algorithm for Discovering Clusters a Density-based Algorithm for
1200 Discovering Clusters in Large Spatial Databases with Noise. *AAAI*
1201 *Press*; 1996. pp. 226–231.
- 1202 95. Newman MEJ, Girvan M. Finding and evaluating community
1203 structure in networks. *Phys Rev E Stat Nonlin Soft Matter Phys.*
1204 *American Physical Society*; 2004;69: 268.
1205 doi:10.1103/PhysRevE.69.026113
- 1206 96. Girvan M, Newman MEJ. Community structure in social and
1207 biological networks. *PNAS. National Academy of Sciences*;
1208 2002;99: 7821–7826. doi:10.1073/pnas.122653799
- 1209 97. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding
1210 of communities in large networks. *J Stat Mech. IOP Publishing*;
1211 2008;2008: P10008. doi:10.1088/1742-5468/2008/10/P10008

- 1212 98. Newman MEJ. Finding community structure in networks using the
1213 eigenvectors of matrices. Phys Rev E Stat Nonlin Soft Matter Phys.
1214 American Physical Society; 2006;74: 036104.
1215 doi:10.1103/PhysRevE.74.036104
- 1216 99. Pons P, Latapy M. Computing Communities in Large Networks
1217 Using Random Walks. Computer and Information Sciences - ISCIS
1218 2005. Berlin, Heidelberg: Springer Berlin Heidelberg; 2005. pp.
1219 284–293. doi:10.1007/11569596_31
- 1220 100. Rosvall M, Bergstrom CT. Maps of random walks on complex
1221 networks reveal community structure. Proc Natl Acad Sci USA.
1222 National Acad Sciences; 2008;105: 1118–1123.
1223 doi:10.1073/pnas.0706851105
- 1224 101. de Castro E, Sigrist CJA, Gattiker A, Bulliard V, Langendijk-
1225 Genevaux PS, Gasteiger E, et al. ScanProsite: detection of
1226 PROSITE signature matches and ProRule-associated functional
1227 and structural residues in proteins. Nucleic Acids Res. 2006;34:
1228 W362–W365. doi:10.1093/nar/gkl124
- 1229 102. Yang Y, Heffernan R, Paliwal K, Lyons J, Dehzangi A, Sharma A,
1230 et al. SPIDER2: A Package to Predict Secondary Structure,
1231 Accessible Surface Area, and Main-Chain Torsional Angles by
1232 Deep Neural Networks. Methods Mol Biol. New York, NY: Springer
1233 New York; 2017;1484: 55–63. doi:10.1007/978-1-4939-6406-2_6
- 1234 103. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure

1235 prediction server. *Bioinformatics*. 2000;16: 404–405.

1236 104. Drozdetskiy A, Cole C, Procter J, Barton GJ. JPred4: a protein
1237 secondary structure prediction server. *Nucleic Acids Res*. 2015;43:
1238 W389–W394. doi:10.1093/nar/gkv332

1239 105. Yan R, Xu D, Yang J, Walker S, Zhang Y. A comparative
1240 assessment and analysis of 20 representative sequence alignment
1241 methods for protein structure prediction. *Sci Rep*. The Author(s) SN
1242 ; 2013;3: 93. doi:10.1038/srep02619

1243 106. Cserző M, Wallin E, Simon I, Heijne von G, Elofsson A. Prediction
1244 of transmembrane alpha-helices in prokaryotic membrane proteins:
1245 the dense alignment surface method. *Protein Eng*. 1997;10: 673–
1246 676.

1247 107. Käll L, Krogh A, Sonnhammer ELL. A combined transmembrane
1248 topology and signal peptide prediction method. *J Mol Biol*.
1249 2004;338: 1027–1036. doi:10.1016/j.jmb.2004.03.016

1250 108. Rost B, Fariselli P, Casadio R. Topology prediction for helical
1251 transmembrane proteins at 86% accuracy-Topology prediction at
1252 86% accuracy. *Protein Science*. Wiley-Blackwell; 1996;5: 1704–
1253 1718. doi:10.1002/pro.5560050824

1254 109. Ikeda M, Arai M, Okuno T, Shimizu T. TMPDB: a database of
1255 experimentally-characterized transmembrane topologies. *Nucleic
1256 Acids Res*. Oxford University Press; 2003;31: 406–409.

- 1257 110. Nugent T, Jones DT. Transmembrane protein topology prediction
1258 using support vector machines. *BMC Bioinformatics*. BioMed
1259 Central; 2009;10: 159. doi:10.1186/1471-2105-10-159
- 1260 111. Ward JJ, McGuffin LJ, Bryson K, Buxton BF, Jones DT. The
1261 DISOPRED server for the prediction of protein disorder.
1262 *Bioinformatics*. 2004;20: 2138–2139.
1263 doi:10.1093/bioinformatics/bth195
- 1264 112. Kozlowski LP, Bujnicki JM. MetaDisorder: a meta-server for the
1265 prediction of intrinsic disorder in proteins. *BMC Bioinformatics*.
1266 2012;13: 111. doi:10.1186/1471-2105-13-111
- 1267 113. Snijder B, Sacher R, Rämö P, Liberali P, Mench K, Wolfrum N, et
1268 al. Single-cell analysis of population context advances RNAi
1269 screening at multiple levels. *Mol Syst Biol*. EMBO Press; 2012;8:
1270 579. doi:10.1038/msb.2012.9
- 1271 114. Lin AE, Greco TM, Döhner K, Sodeik B, Cristea IM. A proteomic
1272 perspective of inbuilt viral protein regulation: pUL46 tegument
1273 protein is targeted for degradation by ICP0 during herpes simplex
1274 virus type 1 infection. *Mol Cell Proteomics*. 2013;12: 3237–3252.
1275 doi:10.1074/mcp.M113.030866
- 1276 115. Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25
1277 years of image analysis. *Nat Methods*. Nature Publishing Group;
1278 2012;9: 671–675. doi:10.1038/nmeth.2089
- 1279 116. Cristea IM, Williams R, Chait BT, Rout MP. Fluorescent proteins as

1280 proteomic probes. Mol Cell Proteomics. American Society for
 1281 Biochemistry and Molecular Biology; 2005;4: 1933–1941.
 1282 doi:10.1074/mcp.M500227-MCP200

1283 117. Wiśniewski JR, Zougman A, Nagaraj N, Mann M. Universal sample
 1284 preparation method for proteome analysis. Nat Methods. Nature
 1285 Publishing Group; 2009;6: 359–362. doi:10.1038/nmeth.1322

1286

1287

1288 **Supporting Information**

1289 **S1 Text. Details on Immunoaffinity Purification Quantitative Mass** 1290 **Spectrometry experiments.**

1291

1292 **S1 Fig. Modularity values of the consensus partition as a function of the**
1293 **filtering threshold.** Modularity values achieved by each of the partitions
1294 derived from the consensus agreement matrix, by filtering values below the
1295 specified thresholds (ranging from 0 to 1 in increments of 0.5). The dashed
1296 line is only shown to emphasise the trend of the modularity values. The
1297 yellow-shaded area of the plot highlights the threshold range for which the
1298 highest modularity was achieved.

1299 **S2 Fig. Primary sequence analysis of US10.** Previously reported and newly
1300 identified features are indicated. Predictions from each software tool are
1301 shown. Predicted disordered regions, α -helices and transmembrane helices
1302 are indicated in blue, yellow and pink, respectively. The identified CLRs are
1303 shown in red boxes. Individual prolines are highlighted in red. The four-
1304 residues polyproline sequence is indicated with a black box. The previously
1305 identified consensus zinc finger sequence [33] is underscored. The final
1306 assignment of the secondary structure elements was based on the consensus
1307 of individual methods (prediction confidence scores were not taken into
1308 account).

1309 **S3 Fig. Primary sequence analysis of UL55.** Predictions from each
1310 software tool are shown. Predicted disordered regions, α -helices and β -
1311 strands are indicated in blue, yellow and green, respectively. The final

1312 assignment of the secondary structure elements was based on the consensus
1313 of individual methods (prediction confidence scores were not taken into
1314 account).

1315 **S1 Table. Herpesvirus species for which PPI data were collected as**
1316 **input for the PPI network assembly framework.**

1317 **S2 Table. Taxonomic identifiers associated to species in S1 Table and**
1318 **used to extract PPIs from input resources.**

1319 **S3 Table. Protein-protein interaction network reconstructed for HSV1.**
1320 For each interaction, the interacting proteins, detection methods, associated
1321 PubMed IDs, types of interaction, confidence score, and whether the
1322 interaction was computationally predicted and/or experimentally supported, is
1323 indicated.

1324 **S4 Table. Parameters associated to the partitions returned by the 14**
1325 **clustering algorithms.**

1326 **S5 Table. Modularity values associated to the partitions derived from the**
1327 **consensus agreement matrix, at different thresholds.**

1328 **S6 Table. Proteins co-purifying with pUL37 by immunoaffinity**
1329 **purification mass-spectrometry.**

Figure 2.

bioRxiv preprint doi: <https://doi.org/10.1101/473751>; this version posted November 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(A)

Noise reduction

- ▶ Virion components only
- ▶ Remove PPIs between capsid and envelope proteins
- ▶ Extract LCC

(B)

- ▶ Modularity maximisation-based optimisation

Modularity > 0

(C)

Modularity
maximisation-based
threshold

(D)

>sp | P06486 | US10_HHV11

MIKRRGNVEIRVYYESVRTLSRSHLK P SDRQQS P GHRVF P GS P GFRDH P ENLG N P EYRE	60
L P ET P GYRVT P GIHD N P GL P GS P GL P GS P GL P GS P GPH A PPANHVRLAGLYS P GKY A P LA	120
S P D P FS P QHGA Y ARARVGIHTAVRV P P TGS P TH T HLRQD P GDE P TSDDSGLY P LDARALA	180
HLV M L P AD H RAFFRTVV E VS R MC A ANVRD P P P P ATG A MLGRHARLVHTQWL R ANQETS P L	240
W P WRTAA I N F IT T MA P RVQ T HR H M H <u>DLLMACAFWCCLTHASTCSYAGLYS</u> THCLHL F GAF	300
GCGD P ALT P PLC	

(A)

(B)

(C)

(D)

(A)

(B) SPP1

(C) HSV1

- (A)
- Molecular interaction databases**
- ▶ IntAct
 - ▶ DIP
 - ▶ VirHostNet2.0
 - ▶ BioGRID
 - ▶ Mentha
- Structural repositories**
- ▶ PDB
 - ▶ EMDB

(B)

Taxons
(9 human herpesvirus species
+ PRV, MuHV1, MuHV4)

Input

oPPI
PPIs *detected* in
orthologous species

HSV1 PPI
PPIs *detected*
in HSV1

(C)

pPPI
PPIs *predicted* in
target species

(D)

pPPI
↓
Score
↑
ePPI

Target PPI
network

Data collection

Curation and prediction of PPIs

Data integration

MIKRRGNVEIRVYYESVRTLRSRSHLKPSDROQSPGHRVFPGPSGFRDHPENLGNPEYRE

Figure 1: Performance of protein disorder prediction methods. The figure shows a horizontal bar chart comparing 11 methods: MEMSAT-SVM, TMpred, Phobius, PHDhtm, DAS, SPIDER², PSIPRED, JPred 4, PSSpred, DISOPRED3, and MetaDisorder. The x-axis represents the number of proteins, with a scale from 0 to 100. The y-axis lists the methods. Each method's performance is shown as a horizontal bar with a color-coded segment indicating the number of proteins correctly predicted as disordered. The colors are: MEMSAT-SVM (blue), TMpred (blue), Phobius (blue), PHDhtm (blue), DAS (blue), SPIDER² (green), PSIPRED (green), JPred 4 (green), PSSpred (blue), DISOPRED3 (blue), and MetaDisorder (blue). The bars for SPIDER², PSIPRED, and JPred 4 are significantly shorter than the others, indicating lower performance.

LPETPGYRVTPGIHDNPGLPGSPGLPGSPGLPGSPGPHAPPANHVRLAGLYSPGKYAPLA

Method	Accuracy
MEMSAT-SVM	0.78
TMpred	0.76
Phobius	0.75
PHDhtm	0.74
DAS	0.73
SPIDER ²	0.72
PSIPRED	0.71
JPred 4	0.70
PSSpred	0.69
DISOPRED3	0.68
MetaDisorder	0.67

SPDPFSPOHGAYARARVGIHTAVRVPP TGSPTHTHLROD PGDEPTSDDSGLYPLDARALA

Method	Score (approximate)
MEMSAT-SVM	0.05
TMpred	0.05
Phobius	0.05
PHDhtm	0.05
DAS	0.05
SPIDER ²	0.95
PSIPRED	0.90
JPred 4	0.05
PSSpred	0.05
DISOPRED3	0.95
MetaDisorder	0.90

HLVML**P**ADHRAFFRTVVEVSRMCAANVRD**PPPP**ATGAMLGRHARLVHTOWLRANOETS**P**L

Figure 1: Comparison of protein disorder prediction methods. The figure shows a sequence of 100 residues with predicted disorder regions from various methods. MEMSAT-SVM, TMpred, Phobius, and PHDhtm show no disorder. DAS, SPIDER², PSIPRED, JPred 4, PSSpred, DISOPRED3, and MetaDisorder show several disorder regions. SPIDER² and PSIPRED show the most extensive disorder regions, covering most of the sequence. DAS shows disorder in the first 10 residues, residues 20-30, and residues 40-50. JPred 4 shows disorder in residues 60-70. PSSpred shows disorder in residues 80-90. DISOPRED3 and MetaDisorder show disorder in residues 90-100.

WPWRTAAINFITTMAPRVOTHRHMDLLMACAFWCCLTHASTCSYAGLYSTHCLHLEGAFL

Figure 1: Comparison of protein disorder prediction methods. The figure shows a sequence of 1000 residues with various prediction methods overlaid. MEMSAT-SVM, TMpred, Phobius, PHDhtm, and DAS are shown in pink, while SPIDER², PSIPRED, JPred 4, PSSpred, DISOPRED3, and MetaDisorder are shown in yellow. The sequence is divided into three regions: 1-300, 300-600, and 600-1000. The methods are compared against a reference sequence of 1000 residues.

GCGDPALTPPLC

-----	MEMSAT-SVM
-----	TMpred
-----	Phobius
-----	PHDhtm
T-----	DAS
-----E	SPIDER ²
-----	PSIPRED
-----	JPred 4
-----	PSSpred
-----	DISOPRED3
-----	MetaDisorder

H α helix
E β strand/sheet
D Disorder
T Transmembrane
 Collagen-like repeat
 Polyproline
 Zinc-finger consensus