RESEARCH ARTICLE

Prediction and analysis of human-herpes simplex virus type 1 protein-protein interactions by integrating multiple methods

Xianyi Lian[1], Xiaodi Yang[1], Jiqi Shao[2], Fujun Hou[3], Shiping Yang[4,*], Dongli Pan[3], Ziding Zhang[1,*]

[1] State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University, Beijing 100193, China
[2] National Demonstration Center for Experimental Biological Sciences Education, College of Biological Sciences, China Agricultural University, Beijing 100193, China
[3] Department of Medical Microbiology and Parasitology, and Department of Infectious Diseases of Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou 310058, China
[4] State Key Laboratory of Plant Physiology and Biochemistry, College of Biological Sciences, China Agricultural University, Beijing 100193, China
* Correspondence: zidingzhang@cau.edu.cn, shi_ping_yang@163.com

Background: Herpes simplex virus type 1 (HSV-1) is a ubiquitous infectious pathogen that widely affects human health. To decipher the complicated human-HSV-1 interactions, a comprehensive protein-protein interaction (PPI) network between human and HSV-1 is highly demanded.

Methods: To complement the experimental identification of human-HSV-1 PPIs, an integrative strategy to predict proteome-wide PPIs between human and HSV-1 was developed. For each human-HSV-1 protein pair, four popular PPI inference methods, including interolog mapping, the domain-domain interaction-based method, the domain-motif interaction-based method, and the machine learning-based method, were optimally implemented to generate four interaction probability scores, which were further integrated into a final probability score.

Results: As a result, a comprehensive high-confidence PPI network between human and HSV-1 was established, covering 10,432 interactions between 4,546 human proteins and 72 HSV-1 proteins. Functional and network analyses of the HSV-1 targeting proteins in the context of human interactome can recapitulate the known knowledge regarding the HSV-1 replication cycle, supporting the overall reliability of the predicted PPI network. Considering that HSV-1 infections are implicated in encephalitis and neurodegenerative diseases, we focused on exploring the biological significance of the brain-specific human-HSV-1 PPIs. In particular, the predicted interactions between HSV-1 proteins and Alzheimer's-disease-related proteins were intensively investigated.

Conclusion: The current work can provide testable hypotheses to assist in the mechanistic understanding of the human-HSV-1 relationship and the anti-HSV-1 pharmaceutical target discovery. To make the predicted PPI network and the datasets freely accessible to the scientific community, a user-friendly database browser was released at http://www.zzdlab.com/HintHSV/index.php.

Keywords: human-virus interaction; protein-protein interaction; prediction; herpes simplex virus type 1; Alzheimer's disease

Author summary: HSV-1 infects a large proportion of the global population with the ability to cause both fatal diseases and disturbing recurrent infections. However, the pathogenic mechanisms of HSV-1 infections and human immune responses have not been fully clarified. To decipher the complicated human-HSV-1 interactions, a comprehensive PPI network between human and HSV-1 is highly demanded. Here an integrative bioinformatics prediction method was employed to provide a relatively complete and reliable human-HSV-1 interactome. We hope our predictions can provide testable hypotheses to accelerate drug development for the treatment of diseases associated with HSV-1 infections.

INTRODUCTION

Herpes simplex virus type 1 (HSV-1) is a neurotropic, enveloped, and double-stranded linear DNA virus [1–4]. The genome of HSV-1 is roughly 152 kb, encoding more than 74 different genes [3]. As a widespread infectious virus, it can be transmitted from person to person through direct contact. Around 3.7 billion people under the age of 50 are estimated by the World Health Organization to be infected with HSV-1 worldwide [5]. Once entering the human body from the skin or mucosa, HSV-1 can enter sensory neurons and be transported through axons to the trigeminal ganglion where a latent infection is established.

When stimulated, the latent virus can be reactivated to cause symptomatic or asymptomatic recurrent infections, leading to common cold sores, blisters, and various serious diseases [2–4,6]. HSV-1 can also reach the central nervous system (CNS), occasionally leading to fatal neurological diseases, such as the herpes simplex encephalitis (HSE) [7,8]. Moreover, an increasing evidence points to a strong association between HSV-1 infection and the Alzheimer's disease (AD) [9]. There is no existing antiviral drug known that would eliminate an HSV-1 infection as the virus can undergo latent infection and thereby evade drug interactions. Therefore, more fundamental research efforts are required to decipher the complicated human-HSV-1 interactions to provide hints for developing novel prophylactic or therapeutic methods against viral infections.

Investigations on protein-protein interactions (PPIs) between the host and the pathogen can reveal key biological processes concerning the interaction as well as elucidate the underlying mechanisms of infectious diseases and thereby support the development of novel therapeutic strategies. As an important branch of host-pathogen PPI studies, human-virus PPI has always been a focus given the close relationship with human diseases. Current research efforts may focus on individual viral proteins at a time, such as glycoproteins involved in the HSV-1 entry into the host cell [10], ICP34.5 (neurovirulence factor) [11], ICP0 (viral E3 ubiquitin ligase) [12], ICP8 (single-stranded DNA-binding protein) [13], ICP4 (major viral transcription factor) [14] and so on. Therefore, it is still essential to decipher the interactome between human and HSV-1 proteins from a global perspective. Additional available data would enable a more robust PPI network to be built between human and HSV-1, which would make our understanding more comprehensive. In general, the experimental identification of PPIs, including the human-virus PPIs, is time-consuming, labor-intensive, and expensive. In this context, cost-effective computational prediction methods play an increasingly important role in supplementing the experimental identification of PPIs.

A plethora of host-pathogen PPI prediction methods including human-virus PPI were previously developed [15–18], mainly originating from intra-species PPI prediction methods [19–21]. In principle, traditional intra-species PPI prediction methods, such as the interolog mapping (IM) [22], the domain-domain interaction (DDI)-based method [22,23], and the domain-motif interaction (DMI)-based method [24], can be readily adapted to the prediction of human-virus PPIs. The IM can be used indirectly as a remedy for data scarcity by homolog knowledge transfer based on the assumption that the interacting protein pairs in one species are likely to be conserved in their cousins [25]. Interacting domain pairs are considered as the building blocks of PPI networks. Itzhaki's research [26] showed that interacting domain pairs potentially mediate human-herpesvirus interactions. The DMI-based method is slowly being revealed to be useful, given the extensive mimicry of host protein short linear motifs by viruses [27,28]. With the accumulation of experimentally verified human-virus PPI data, machine learning (ML)-based prediction methods were increasingly popular in the past decade, which made them worthy to be applied to the prediction of human-HSV-1 PPIs. Although none of the existing human-virus PPI prediction methods can achieve satisfactory performance, it is common knowledge that more powerful and robust predictive performance can be achieved by the integration of multiple prediction methods, which was implemented in a series of studies [21,29,30].

In this work, four PPI inference methods (*i.e.*, the IM, DDI, DMI, and ML-based method) were integrated for high-confidence PPI prediction between human and HSV-1 across the entire proteome. In addition to the ML-based method that can output predicted scores, the other three traditional PPI prediction methods were also refined, so that each prediction method could yield an interaction probability score for any query protein pair. The four predictive scores for the query protein pair were further integrated into a final score. PPIs with higher final scores (integration score > 0.5) were singled out for further analysis. In addition to the general functional and network topology analyses of HSV-1 targeting human proteins, the biological significance of the predicted human-HSV-1 interactome was further explored with a focus on brain tissue-specific PPIs. In particular, the potential mechanisms of the HSE and AD in the context of the human-HSV-1 interactome were investigated.

RESULTS AND DISCUSSION

The landscape of predicted human-HSV-1 PPIs

In this work, an integrative computational framework was

applied to predict the interactions between 74 different proteins of HSV-1 strain KOS as well as 20,412 reviewed human proteins. Four methods (IM, DDI, DMI, and ML) were used in our computational framework to predict whether two proteins interact (Fig. 1). Briefly, IM is based on the experimentally validated interactions of multiple homologous protein pairs (*i.e.*, interologs) of the query human-HSV-1 protein pair; DDI/DMI relies on the detection of the known or possible domain-domain/ motif interactions in the query protein pair to infer the interaction probability; ML is trained from the known PPIs between human and HSV-1, the feature encoding schemes of which include the sequence features extracted from protein pairs and the network properties of human proteins in the corresponding human PPI network. Finally, the four interaction probability scores (Pr_{IM}, Pr_{DDI}, Pr_{DMI}, and Pr_{ML}) were combined into an integration score (Pr) representing the interaction probability of the human-HSV-1 protein pair. It is hard to precisely rank the performance of the four individual methods due to data limitation and bias of known human-HSV-1 PPIs. Thus, each method in the final integration was treated independently and assigned with the same weight. More methodological details are available in "Materials and methods".

The number of PPIs predicted by each method was calculated separately. As shown in Fig. 2A, the number of PPIs predicted by the DMI was the largest (41,828), followed by the DDI (13,579), the IM (7,805), and the ML (6,341). In general, the percentages of overlapping PPIs among different methods are low, implying that different methods are distinctive and complementary. Due to similarity in the methodologies, the DDI achieved relatively more consistent PPI prediction results compared to the IM and the DMI (the overlap rate in both cases accounted for about 10% of its total). After integrating the results of the four methods, the number of predicted PPIs with $Pr > 0$ was 65,673. Although a higher Pr should correspond to a higher reliability, it is still necessary to set a reasonable and convincing threshold for high-confidence predictions. The solution was sought from high-throughput human-virus PPI identification studies. Taking the number of experimentally validated human-HIV-1 PPIs as a reference, $100-200$ interactions with human proteins were identified for each HIV-1 protein in some high-throughput experimental studies [31]. Supplementary Fig. S1 showed the number of PPIs obtained under different confidence cutoffs. In general, a low threshold will result in too many predictions, which inevitably contain false positives. On the contrary, a high threshold will yield too few predictions, and many potential interactions will be
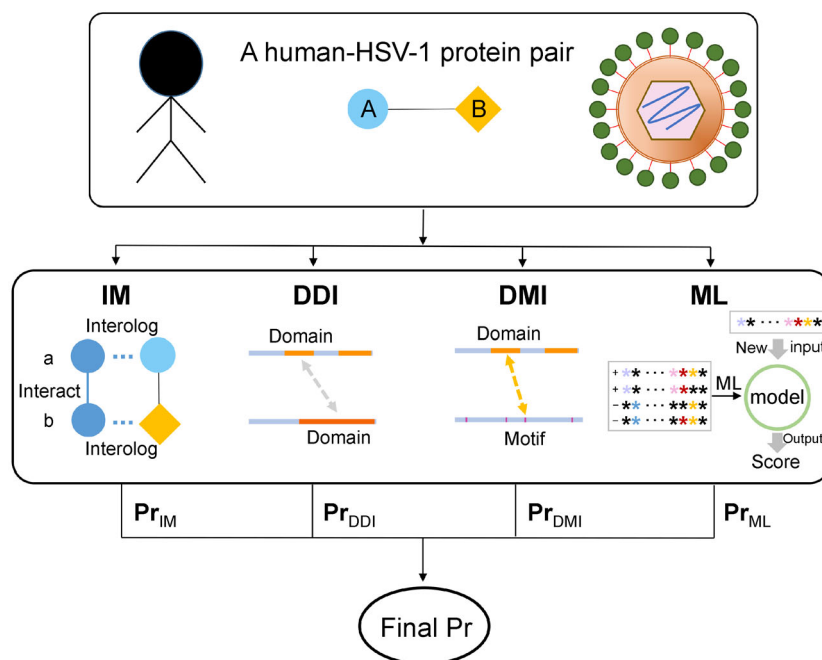

Figure 1. Workflow for the prediction of human-HSV-1 PPIs. The interaction probability for each human-HSV-1 protein pair by interolog mapping (IM), domain-domain interaction (DDI), domain-motif interaction (DMI), or machine learning (ML)-based methods was evaluated. The four interaction probability scores (Pr_{IM}, Pr_{DDI}, Pr_{DMI}, and Pr_{ML}) subsequently formed the final probability score (Pr).

ignored. Thus, the threshold of PPI predictions was empirically set to Pr > 0.5 and 10,432 PPIs were singled out as the most likely interacting protein pairs (Supplementary Fig. S1). On average, each viral protein interacts with 141 human proteins, which is a relatively reasonable number range in comparison to high-throughput PPI experimental identifications between human and HIV-1. Moreover, we found that 690 of 728 experimentally verified PPIs (collected from the HPIDB database and used in the ML method) overlaps with our 10,432 predicted results (Supplementary Data Set S1), 601 of which PPIs could be predicted by more than one method. Figure 2B showed that the IM method accounted for the largest proportion among these 10,432 high-confidence PPIs.

Functional and network analyses showing the reliability of predicted human-HSV-1 PPIs

The 10,432 high-confidence PPIs were further analyzed. First, the number of human proteins targeted by each HSV-1 protein was counted (Fig. 3). On average, one HSV-1 protein interacted with 145 human proteins and the top ten HSV-1 proteins contributed to 5,963 interactions (approximately 57%) in the predicted human-HSV-1 interactome. The HSV-1 protein UL22 was predicted to have the most interactions with human proteins, and the predicted interaction partners were significantly enriched in the category of membrane-bounded organelle components (hypergeometric test, corrected p-value = 3.37×10^{-51}). Previous studies suggested that UL22, also called as the envelope glycoprotein H (gH), complexed with glycoprotein L (gL, UL1) and interacted with glycoproteins B (gB, UL27) and D (gD, US6) to form a viral membrane fusion machine, thereby driving the fusion of the virus with the host membranes to allow the enter or spread of the virus between the host cells [32]. It is, therefore, reasonable to predict that this viral protein interacts with multiple human proteins especially membrane proteins. RL2, E3 ubiquitin ligase (ICP0), was predicted to interact with several human proteins that belong to the host cellular interferon-related proteins category (hypergeometric test, corrected p-value = 1.32×10^{-9}), which may indicate that the RL2 is a weapon of the HSV-1 to counteract the intrinsic- and interferon-based antiviral responses. Thus, the predicted viral targets play an important role in the viral infection process, indicating the reliability of our human-virus PPI prediction.

Viral proteins tend to target some important host (human) proteins, such as the "hub" (high-degree centrality) and "bottleneck" (high-betweenness centrality) nodes of the human PPI network, to hijack and utilize host cells for viral life cycles [33]. Therefore, the degree and betweenness centrality of target proteins (proteins in the human PPI network that are targeted by the HSV-1) and non-target proteins (proteins in the human PPI network that are not targeted by HSV-1) from the perspective of network biology were also calculated. It can be seen from Fig. 4 that, whether in degree or betweenness centrality, the values of target proteins were significantly higher than those of the non-target proteins (Wilcoxon rank-sum test, p-value < 2.2×10^{-16}), which is in accordance with previous observations inferred from human-pathogen PPI network analyses [34].

Functional analysis of brain-specific human-HSV-1 PPIs

Among several diseases caused by HSV-1 infection, sporadic but often fatal HSE in the brain is of great concern. Therefore, additional focus was placed on PPIs in which the human proteins are specifically expressed in the brain tissue. 569 PPIs containing 283 brain-specific human proteins from the 10,432 high-confidence PPIs were selected. According to the Gene Ontology (GO)
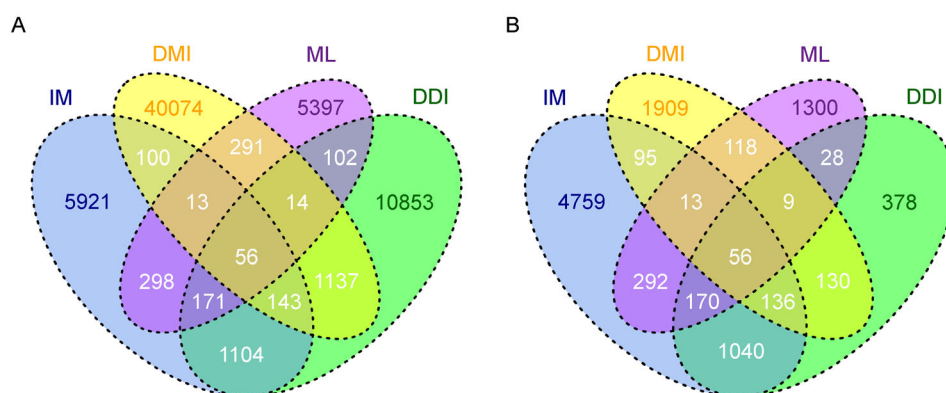

Figure 2. Overlaps of the predicted PPIs among the four individual methods. (A) All the predictions (Pr > 0). (B) High-confidence predictions (Pr > 0.5).
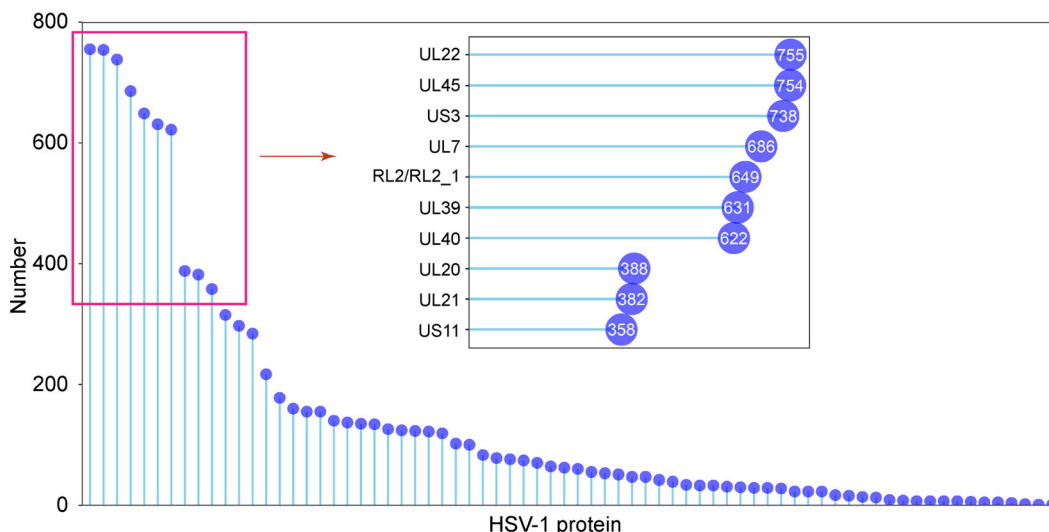
Figure 3. **The number of human proteins predicted to interact with HSV-1 proteins.**
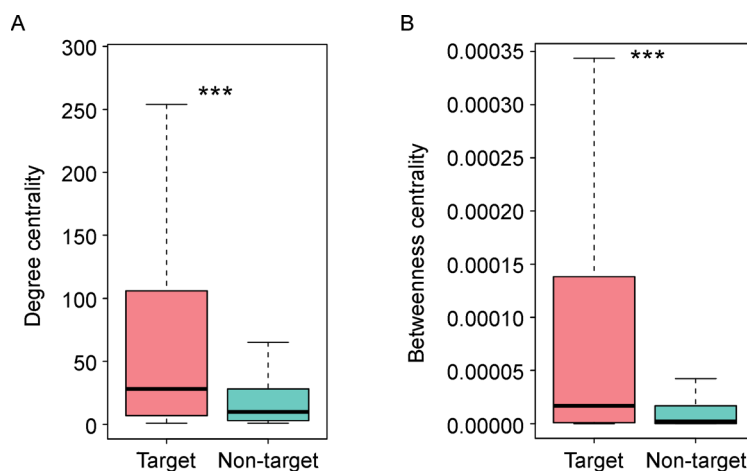

Figure 4. **Degree and betweenness centrality of human target proteins and non-target proteins.** ***, stands for statistically significant results (Wilcoxon test, *p*-value $< 2.2 \times 10^{-16}$).

enrichment analysis (Fig. 5), cell adhesion-related biological process (BP) terms, such as "cell adhesion", "biological adhesion" and "cell-cell adhesion", were found to be significantly enriched (Fig. 5A, corrected *p*-value $= 2.33 \times 10^{-7}$, 2.33×10^{-7} and 2.2×10^{-14}, respectively), which indicated the reliance of HSV-1 on the intricate events of attachment and fusion to enter cells, especially by utilizing its envelope proteins (envelope glycoproteins) to interact with cell adhesion molecules to mediate this process [35]. In our results, 55 cellular adhesion molecules were predicted to interact with HSV-1 proteins. In the cellular component (CC) category, human proteins were found to be significantly enriched in microtubule or microtubule cytoskeleton (Fig. 5B,

corrected *p*-value $= 1.14 \times 10^{-4}$ and 4.82×10^{-4}, respectively). Microtubules are major components of the cytoskeleton and are known to be involved in transport in all eukaryotic cells. Therefore, the above enriched GO terms are in accordance with previous knowledge about the transportation of viral capsids to and from the nucleus to complete the replication cycle after entering the host cell. This is particularly relevant to the processes associated with the establishment of latent infection and reactivation in neurons, during which the transport of capsids along microtubules in long axons is required.

Besides, one strategy usurped by the HSV-1 is to guide the entry pathway by the manipulation of various cell signaling cascades [36]. In the GO enrichment analysis
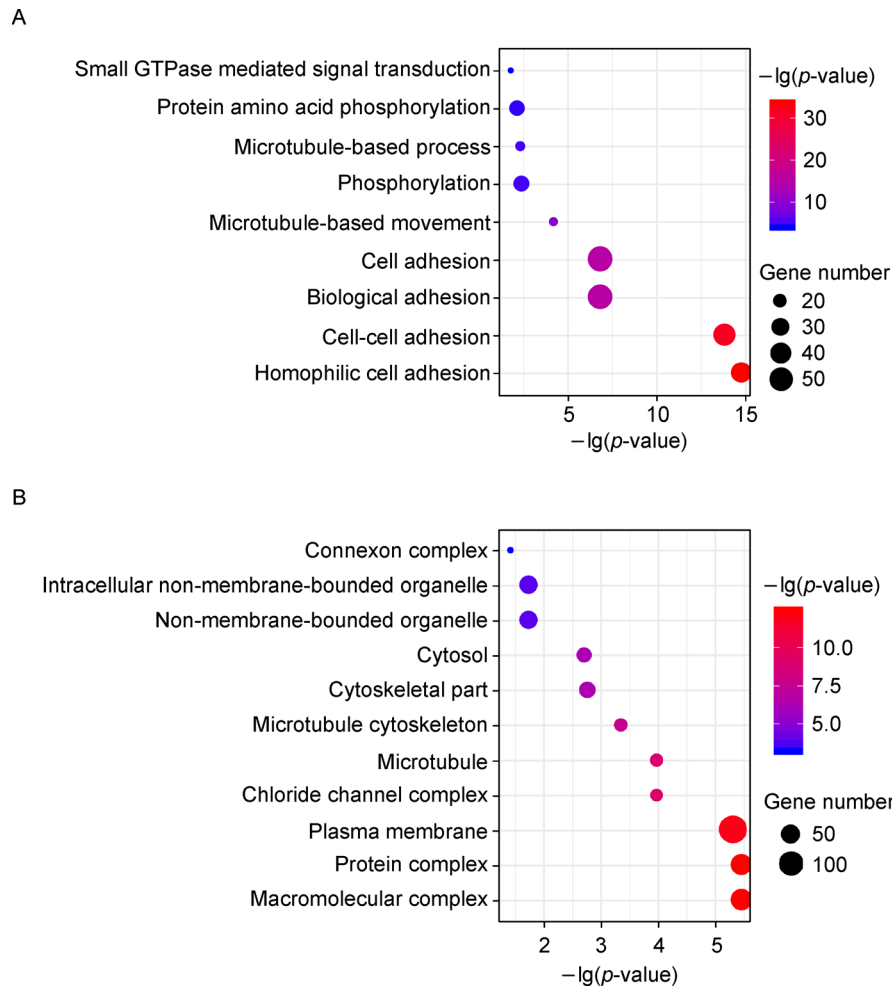
A

B

Figure 5. Enriched GO terms of the brain-specific human proteins predicted to interact with HSV-1 proteins in the biological processes (A) and cellular component (B) categories.

results of molecular function (MF) entries (Supplementary Fig. S2), the GO term of "calcium ion binding" was found to be significantly enriched. Ca^{2+} is one of the most prominent and common signal carriers and is known to modulate several steps during virus replication. The entry of HSV-1 is triggered by the interaction of the gH protein with cellular integrin, which eventually triggers Ca^{2+}-mediated signaling pathways within the cell to ensure effective nucleocapsid translocation into the cytoplasm [36]. Although the relationship between chloride channels and viral infections has so far received less attention, previous studies showed that chloride channels play an important role in the HSV-1 entry [37]. Here, the CC enrichment of the chloride channel complex and the MF enrichment of the chloride channel activity were also found to be significant, further supporting the association between the chloride channel and the HSV-1 entry.

Collectively, the GO enrichment results of the HSV-1-interacting human brain-specific proteins were consistent with known functions associated with the HSV-1 replication cycle, suggesting that the PPIs between HSV-1 and human disrupt the normal function of proteins in the brain cells, which may cause inflammation and damage leading to HSE. These data also support the overall reliability of the predicted PPIs. A vital subnetwork (Supplementary Fig. S3) of the human-HSV-1 interactome is expected to be formed by the 569 PPIs, which may enhance the mechanism-wise understanding of diseases related to HSV-1 infection (*e.g.*, HSE) as well as providing new hints to the discovery of novel therapeutic targets.

The association of the HSV-1 with the AD in the context of human-HSV-1 PPIs

Increasing evidence points to the association of HSV-1 brain infection with AD. HSV-1 is present in the latent
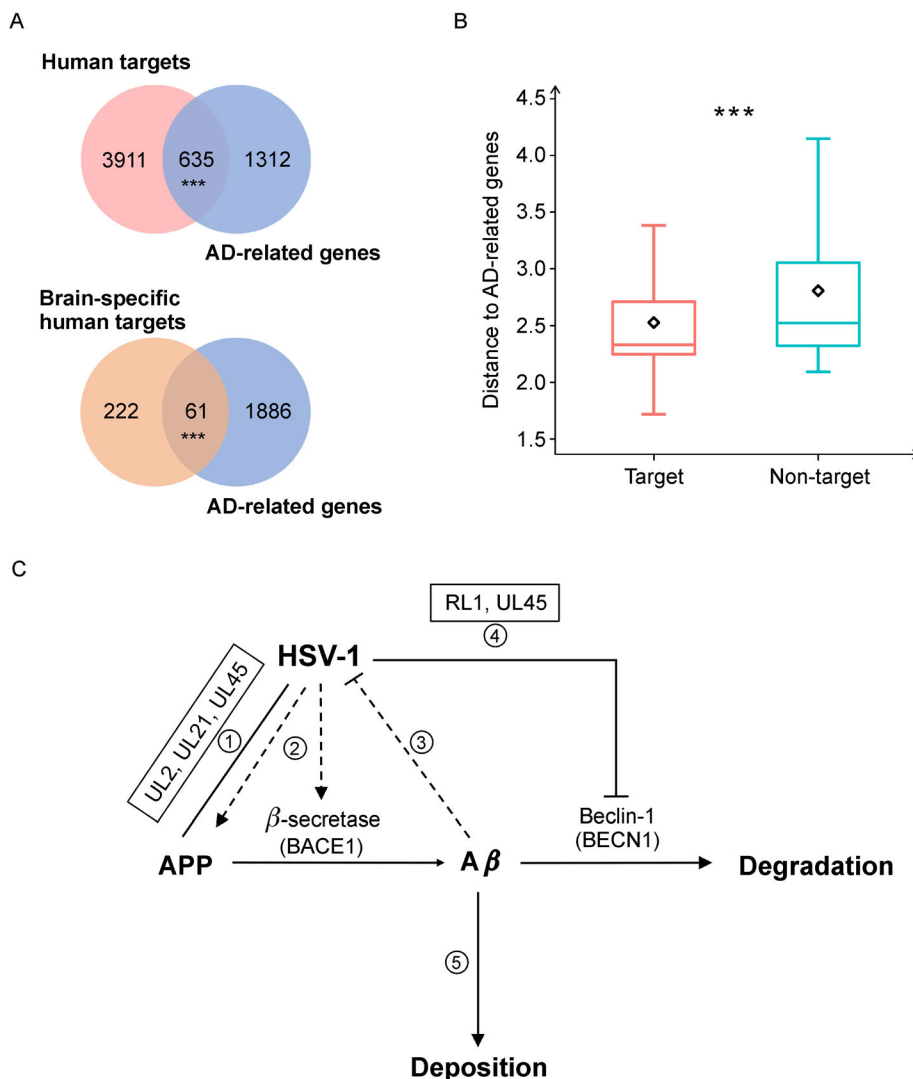
Figure 6. Association between human-HSV-1 PPIs and the AD. (A) Overlaps between human targets and AD-related genes. (B) Differences in network distance between target proteins and non-target proteins to AD-related genes in the human PPI network. *** denotes statistical significance (Wilcoxon test, p-value $< 2.2 \times 10^{-16}$). The mean value is represented by the small diamond box. (C) The possible relationship between APP, Aβ, and HSV-1. ① Interaction of HSV-1 proteins (UL2, UL21, and UL45) with the APP. ② APP phosphorylation and increased BACE1 activity induced by the HSV-1 infection, resulting in the conversion of APP to Aβ. ③ Aβ inhibits viral activity by encapsulating viral proteins. ④ HSV-1 proteins (RL1 and UL45) interact with Beclin-1, suppressing the degradation of Aβ via the inhibition of its autophagy-lysosome pathway. ⑤ To sum up, the deposition of Aβ is led by the imbalance between the Aβ production and degradation.

state in a high proportion of elderly brains. Intermittent reactivation from the latent state may cause local damage and inflammation, accumulation of which might eventually lead to AD [7].

To investigate whether the prediction results could provide supportive evidence for the association between the AD and HSV-1 infection, 1,947 AD-related human genes were compared with the 4,546 predicted HSV-1 target proteins (human proteins present in the 10,432 predicted PPIs), and 635 were found to be overlapping

(Fig. 6A, hypergeometric test, p-value $= 1.37 \times 10^{-12}$). Meanwhile, the overlap between AD-related genes and target proteins specifically expressed in brain tissue was calculated and found to be still significant (hypergeometric test, p-value $= 4.18 \times 10^{-10}$). The average network distance of AD-related genes to target proteins and non-target proteins in the human PPI network was also calculated with results showing that AD-related genes were closer to target proteins (Fig. 6B). The above network analyses may suggest the strong association of

many HSV-1 target proteins with the AD, and it can be hypothesized that the virus may also indirectly affect these AD-related genes by interacting with other proteins to enhance their ability to influence the AD risk and predisposition.

The amyloid precursor protein (APP) is a single-pass transmembrane protein that is widely expressed in tissues, especially at high levels in the brain neurons, and is subsequently metabolized rapidly [38]. Two pathways are known for the proteolysis of the APP (Fig. 6C), one of which includes its cleavage by α-secretase, generating the sAPPα fragment, and the other includes its cleavage by β-secretase (BACE1), producing neurotoxic amyloid β (Aβ) [38]. One of the commonly recognized hallmarks of the AD is the accumulation of the Aβ. First, HSV-1 uses its capsid proteins to physically interact with the APP, thereby hijacking the APP to transport newly generated virions in infected cells through a rapid anterograde transport mechanism [2]. Although such behavior changes the intracellular distribution of the APP and seems to prevent it from its conversion to Aβ partially, HSV-1 infection triggers an intra-CNS anti-microbial innate immune response to induce APP phosphorylation and activates the BACE1 activity, which jointly promotes the production of Aβ [39]. The Aβ would encapsulate the HSV-1 virions to facilitate their clearance by autophagy [40,41]. HSV-1 also employs virulence factors to counter-attack, inhibiting the autophagy-lysosome pathway of Aβ through interaction with the Beclin-1 [11]. The imbalance between the production and elimination of the Aβ caused by the HSV-1 infection accounts for excessive intracellular neurotoxic Aβ deposition within autophagosomes and endosomes, thus inducing neuronal apoptosis, which in turn can drive the degeneration of CNS tissue and the development of AD. Our predicted PPIs showed that three HSV-1 proteins (UL2, UL21, and UL45) interacted with the APP, two of which were in line with the experimental observation. Besides, the RL1 and UL45 were also predicted to play a virulence factor role in the interaction with the Beclin-1. In summary, the recapitulated interactions between the HSV-1, APP, and Aβ further argue for a mechanistic basis for the association between the HSV-1 infection and the risk of the AD (Fig. 6C).

Interactive web interface

The predicted 10,432 high-confidence PPIs were stored in a database to which an interactive web interface was provided (http://www.zzdlab.com/HintHSV/index.php) to facilitate user access. We have provided a search box for 72 HSV-1 proteins participating in these 10,432 PPIs, so any protein can be selected to view the corresponding interactions. For each HSV-1 protein, a table is provided to display all the prediction scores for each human target protein (including four individual prediction scores and one integrative score) and a subnetwork to show the PPIs, which are available for download. Human proteins can also be searched by the users to find possible PPIs with HSV-1. The 569 brain-specific PPIs, 690 known PPIs, and other datasets used in this work are also downloadable in the web interface.

Limitations of our work

The current work is inevitably subjected to the following limitations since the number of experimentally known human-HSV-1 PPIs is not sufficient. Firstly, some parameter settings were empirically selected since sufficient data for strict parameter optimization was not available. Secondly, the integration of different PPI inference methods was also hindered by the lack of data availability. In case of sufficient amount of known PPI data, some more powerful integration methods, such as the logistic regression can be tested. Thirdly, the reliability of the prediction results could not be directly assessed either. Even so, the prediction results are believed to become an important data resource, after the careful implementation of state-of-the-art PPI inference methods to provide useful PPI candidates for further experimental validation. Moreover, the new human-HSV-1 PPIs identified by experimental scientists in the future will continuously answer the overall reliability of the current predictions.

CONCLUSION

In this work, four popular PPI inference methods were used to predict the PPIs between human and HSV-1. To maximize the reliability of predictions, the interaction probability scores from the four methods were integrated into a final probability score and a stringent threshold (Pr > 0.5) was selected to single out high-confidence PPIs. The subsequent functional and network topology analyses also proved an overall reasonable reliability in methodology for the prediction strategy. To investigate the associations between the HSV-1 infection and neurodegenerative diseases (*e.g.*, the HSE and the AD), the focus was placed on brain-specific PPIs between human and HSV-1, and a subnetwork containing 569 inter-species PPIs was established. Functional analysis shows that human proteins involved in the entry, intracellular transport pathways, and various regulatory pathways, are utilized or hijacked by the HSV-1 through complicated inter-species PPIs. Collectively, the established human-HSV-1 PPI network provides a global landscape regarding the human-HSV-1 interactome, as well as new insights into the pathogenesis of the HSV-1 infection.

MATERIALS AND METHODS

Data sets

HSV-1 and human proteins

In this work, the focus was placed on the PPI prediction between the HSV-1 strain KOS and human. All the proteins of the HSV-1 strain KOS were downloaded from GenBank (https://www.ncbi.nlm.nih.gov/nuccore/952947517/). By merging two redundant proteins (Protein RL2 repeats with protein RL2_1; Protein RS1 repeats with protein RS1_1; the results are presented as RL2/RL2_1 and RS1/RS1_1, respectively), 74 HSV-1 proteins were obtained (Supplementary Data Set S2). 20,412 reviewed human proteins used for prediction were downloaded from the UniProt database [42] (Supplementary Data Set S3).

Brain-specific human genes

Brain-specific genes revealing elevated expression in the cerebral cortex were downloaded from the Human Protein Atlas (www.proteinatlas.org). By UniProt ID mapping, 1,442 brain-specific human proteins were obtained.

AD-related human genes

Gene-disease associations were downloaded from DisGeNET (http://www.disgenet.org/). The resulting 1,947 AD-related human genes were obtained by the UniProt ID mapping tool.

Human PPI network

The human interaction network was collected in our previous work [43], consisting of 345,064 PPIs and 18,473 proteins. It was used for network parameter analyses and network-based encoding in the development of the ML-based predictive model. The R package called the igraph [44] was used to calculate the network parameters of protein nodes in the network.

PPI prediction methods

To ensure that the predictions are robust and reliable, four prediction methods were used to infer the PPIs between 74 HSV-1 proteins and 20,412 human proteins. The four methods gave the probability scores $(0-1)$ of the interaction for $74 * 20,412$ protein pairs. Finally, the four scores were combined into one final score (Pr_{final})

according to the integration method used in the STRING database [30]. It was calculated in a naïve Bayesian manner under the assumption of the independence of various methods. The formulas to infer Pr_{final} are as follows:

$$Pr_i = \frac{Pr_i - p}{1 - p}, i = IM, DDI, DMI, ML, \qquad (1)$$

$$Pr_{total} = 1 - (1 - Pr_{IM}) * (1 - Pr_{DDI}) * (1 - Pr_{DMI})$$
$$* (1 - Pr_{ML}), \qquad (2)$$

$$Pr_{final} = Pr_{total} + p * (1 - Pr_{total}). \qquad (3)$$

Here p denotes a prior factor, which is set as 0.041 following the setting provided by STRING. Pr_{IM}, Pr_{DDI}, Pr_{DMI}, and Pr_{ML} stand for the interaction probability score for the IM, DDI, DMI, and ML method, respectively. Each method is briefly described in the following subsections.

The IM method

The IM method is a widely used PPI inference method. The core idea of IM is to infer unknown PPIs from known homologous PPIs (termed as interologs) in other organisms. Previous IM applications often used the PPI templates from one or several model species to infer unknown PPIs. To maximize the IM method, we extended the species source range of template PPIs to cover most of the experimentally identified PPIs, including both of intra-species and inter-species PPIs. Here, 571,359 template PPIs with relatively complete information were collected from seven public databases, including Bio-GRID [45], DIP [46], HPIDB [47], IntAct [48], PATRIC [49], InnateDB [50] and VirHostNet [51]. We employed the strategy of HIPPIE [52] to evaluate the quality of each PPI template. For each PPI template, a quality score (S_{temp}) ranging from 0 to 1 was assigned by accounting for three conditions (*i.e.,* the experimental methods for the PPI determination, the literature reporting the PPI, and the species included in the PPI). The six parameter values in the formula are as set in HIPPIE. To identify the interologs for a query protein pair between human and HSV-1, BLAST searching was conducted to identify their homologs, and the criteria for two proteins to be considered homologous are as follows: E-value $\leqslant 10^{-5}$, sequence identity $\geqslant 30\%$, and alignment coverage of query protein $\geqslant 40\%$. In case n homologous pairs were identified for the query pair, the IM-based interaction probability (Pr_{IM}) can be defined as:

$$\mathrm{Pr}_{\mathrm{IM}} = 1 - \prod_{i=1}^{n}(1-s_i),$$

$$s_i = \begin{cases} 0, & \text{if protein pair } i \text{ not in PPI templates} \\ S_{\mathrm{temp}}, & \text{if protein pair } i \text{ in PPI templates} \end{cases} . \quad (4)$$

The DDI-based method

Considering that the interaction between two proteins may be mediated through evolutionarily-conserved, interacting domain pairs existing in the proteins, the DDI method was developed for PPI prediction. The list of known DDIs can be downloaded from the 3did database [53]. To construct as large DDI library as possible, the expectation-maximization (EM)-based algorithm proposed by Liu *et al.* [54] was also employed to mine domain pairs that are frequently used in known PPIs. Here, the domain definition was based on the Pfam database [55], and hmmscan [56] was employed to search for protein domains (E-value $\leqslant 10^{-5}$). Among the known PPIs collected in this study, 918,116 PPIs conformed to the requirement that the corresponding two protein partners should contain Pfam domains. The probability of DDIs contained in these PPIs was evaluated using the EM algorithm. Because some domains frequently occurred in proteins that may not participate in PPIs, to avoid the introduction of potential noise, domains that occurred in such a highly frequent manner were not taken into account in the subsequent implementation of the EM algorithm. Finally, a comprehensive DDI library was compiled by combining the known DDIs in 3did and the inferred DDIs through the EM algorithm. With the principle that DDIs collected from 3did should be more reliable, the confidence score (S_{DDI}) for each DDI in the library was assigned based on the following formula:

$$S_{\mathrm{DDI}} = \frac{1}{2} * (S_{\mathrm{DDI\text{-}EM}} + S_{\mathrm{DDI\text{-}known}}), \quad (5)$$

where $S_{\mathrm{DDI\text{-}known}}$ takes 1 or 0 respectively to represent whether the DDI is known to be from 3did or not, and $S_{\mathrm{DDI\text{-}EM}}$ is the score of the DDI from the EM algorithm, ranging from 0 to 1. The probability of interaction ($\mathrm{Pr}_{\mathrm{DDI}}$) between one HSV-1 protein and one human protein was inferred from the n domain pairs they contain, which is defined as:

$$\mathrm{Pr}_{\mathrm{DDI}} = 1 - \prod_{i=1}^{n}(1-s_i),$$

$$s_i = \begin{cases} 0, & \text{if domain pair } i \text{ not in DDI library} \\ S_{\mathrm{DDI}}, & \text{if domain pair } i \text{ in DDI library} \end{cases} . \quad (6)$$

The DMI-based method

DMI is also considered to be an important way to mediate human-virus PPIs. Like the DDI method, the DMI method can also infer PPIs. The DMI library is also a combination of known DMIs and the inferred DMIs with the assistance of the EM algorithm. Known DMIs was also be downloaded from 3did. Here, domain assignment is the same as in case of the DDI method. The motif of each protein was identified only from those motif patterns that were contained in known DMIs. Moreover, like the filtering strategy used in the DDI method, the evaluated DMIs containing the highly frequently occurred domains or motifs were removed before their scoring was undertaken with the EM algorithm. Finally, the confidence score (S_{DDI}) for each DDI in the library was defined using the following equation:

$$S_{\mathrm{DMI}} = \frac{1}{2} * (S_{\mathrm{DMI\text{-}EM}} + S_{\mathrm{DMI\text{-}known}}), \quad (7)$$

where $S_{\mathrm{DMI\text{-}known}}$ takes 1 or 0 respectively to represent whether the DMI is known to be from 3did or not, and $S_{\mathrm{DMI\text{-}EM}}$ is the score of the DMI from the EM algorithm. The interaction probability ($\mathrm{Pr}_{\mathrm{DMI}}$) of a human-HSV-1 protein pair containing n domain-motif pairs was further inferred from the following formula:

$$\mathrm{Pr}_{\mathrm{DMI}} = 1 - \prod_{i=1}^{n}(1-s_i),$$

$$s_i = \begin{cases} 0, & \text{if domain-motif pair } i \text{ not in DMI library} \\ S_{\mathrm{DMI}}, & \text{if domain-motif pair } i \text{ in DMI library} \end{cases} . \quad (8)$$

The ML-based method

During the development of ML prediction models, both positive and negative samples are required. Positive and negative samples for human-virus PPI predictions are known to be highly skewed in the real application. The ratio of positive and negative samples used in the training of ML-based PPI prediction models remains an open issue. Instead of using balanced or extremely unbalanced training sample ratios, a relatively imbalanced ratio is often adopted. Based on the above considerations, the ratio of positive to negative samples was empirically set to 1:10. Therefore a training dataset containing 728 positive samples (*i.e.*, known human-HSV-1 PPIs) and 7,280 negative samples (*i.e.*, human-HSV-1 non-PPIs) was compiled to develop an ML-based predictor. The positive samples were collected from HPIDB 3.0 (the download date is December 2018), in which HSV-1 proteins from

different strains (not just the strain KOS) were taken into account, while the negative samples were randomly selected from human-HSV-1 protein pairs with unidentified interaction relationships. Moreover, two encoding schemes were employed to transform protein pairs into feature vectors, including a sequence-based encoding scheme called the CKSAAP as well as a network property-based encoding scheme called the NetTP. The CKSAAP calculated the composition of k-space amino acid pairs for protein pairs. The NetTP encoding scheme considered that human proteins targeted by viral proteins have different network properties from those that are not targeted. Six network topology parameters were used to infer the NetTP encoding, including the degree centrality, betweenness centrality, closeness centrality, eigenvector centrality, PageRank centrality, as well as eccentricity. More details about these two encoding schemes are available in our previous publication [43]. Subsequently, the predictive models of the two encoding methods were both trained by the random forest method, and they were subsequently integrated into a stronger predictive model through logistic regression. The performance of the two individual models as well as the integrative model was evaluated through a 5-fold cross-validation (Supplementary Fig. S4). In general, the integrative model could outperform each ML model. For each query protein pair, the final prediction model generated a prediction score (S_{ML}) ranging from 0 to 1. Note that the F1 value was chosen to comprehensively evaluate the performance of the model, which is the harmonic mean of precision and recall of the model. When the F1 reaches the maximum under a certain threshold, the precision and recall of the model would achieve an optimal balance. The definitions of *precision, recall,* and *F*1 are as follows:

$$Precision = \frac{TP}{TP + FP}, \tag{9}$$

$$Recall = \frac{TP}{TP + FN}, \tag{10}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

$$= \frac{2 \times TP}{2 \times TP + FP + FN}, \tag{11}$$

where *TP, TN, FP,* and *FN* denote the numbers of true positives, true negatives, false positives, and false negatives, respectively. We calculated the *F*1 values of the model in the 5-fold cross-validation according to different thresholds and took the threshold value of 0.363 corresponding to the maximum value of *F*1 as the final criterion to determine whether the query pair had interaction or not. Furthermore, the prediction score was converted into the ML-based interaction probability score (Pr_{ML}):

$$Pr_{ML} = \begin{cases} S_{ML}, & S_{ML} \geq threshold \\ 0, & S_{ML} < threshold \end{cases}. \tag{12}$$

ID mapping

The online UniProt ID mapping tool (https://www.uniprot.org/uploadlists/) was used to convert other IDs (*e.g.*, human or viral gene IDs) into UniProt IDs.

GO enrichment analysis

The BiNGO plugin [57] in Cytoscape [58] was used for the GO enrichment analysis. The enrichment analysis of the UL22-targeted human proteins was conducted against the background of 20,412 reviewed human proteins, and the GO category of the CC was selected. To explore why the HSV-1 targets these 283 human proteins that are specifically expressed in brain tissues, a GO enrichment analysis of the three categories (BP, CC, and MF) was conducted by taking the 1,442 brain-specific human proteins as the background (reference set). Statistical significance was inferred from the hypergeometric test and enriched terms were selected with a significance level of 0.05 after the Benjamini and Hochberg False Discovery Rate correction.

COMPLIANCE WITH ETHICS GUIDELINES

The authors Xianyi Lian, Xiaodi Yang, Jiqi Shao, Fujun Hou, Shiping Yang, Dongli Pan and Ziding Zhang declare that they have no competing interests.

The article does not contain any human or animal subjects performed by any of the authors.

REFERENCES

1. Conrady, C. D., Drevets, D. A. and Carr, D. J. J. (2010) Herpes simplex type I (HSV-1) infection of the nervous system: is an immune response a good thing? J. Neuroimmunol., 220, 1–9

2. Piacentini, R., De Chiara, G., Li Puma, D. D., Ripoli, C., Marcocci, M. E., Garaci, E., Palamara, A. T. and Grassi, C. (2014) HSV-1 and Alzheimer's disease: more than a hypothesis. Front. Pharmacol., 5, 97

3. Watanabe, D. (2010) Medical application of herpes simplex virus. J. Dermatol. Sci., 57, 75–82

4. De Chiara, G., Piacentini, R., Fabiani, M., Mastrodonato, A., Marcocci, M. E., Limongi, D., Napoletani, G., Protto, V., Coluccio, P., Celestino, I., *et al.* (2019) Recurrent herpes simplex virus-1 infection induces hallmarks of neurodegeneration and cognitive deficits in mice. PLoS Pathog., 15, e1007617

5. Looker, K. J., Magaret, A. S., May, M. T., Turner, K. M. E., Vickerman, P., Gottlieb, S. L. and Newman, L. M. (2015) Global and regional estimates of prevalent and incident herpes simplex virus type 1 infections in 2012. PLoS One, 10, e0140765

6. Ashford, P., Hernandez, A., Greco, T. M., Buch, A., Sodeik, B., Cristea, I. M., Grünewald, K., Shepherd, A. and Topf, M. (2016) HVint: a strategy for identifying novel protein-protein interactions in herpes simplex virus type 1. Mol. Cell. Proteomics, 15, 2939–2953

7. Itzhaki, R. F. (2018) Corroboration of a major role for herpes simplex virus type 1 in Alzheimer's disease. Front. Aging Neurosci., 10, 324

8. Steiner, I. (2011) Herpes simplex virus encephalitis: new infection or reactivation? Curr. Opin. Neurol., 24, 268–274

9. Itzhaki, R. F., Lin, W. R., Shang, D., Wilcock, G. K., Faragher, B. and Jamieson, G. A. (1997) Herpes simplex virus type 1 in brain and risk of Alzheimer's disease. Lancet, 349, 241–244

10. Agelidis, A. M. and Shukla, D. (2015) Cell entry mechanisms of HSV: what we have learned in recent years. Future Virol., 10, 1145–1154

11. Orvedahl, A., Alexander, D., Tallóczy, Z., Sun, Q., Wei, Y., Zhang, W., Burns, D., Leib, D. A. and Levine, B. (2007) HSV-1 ICP34.5 confers neurovirulence by targeting the Beclin 1 autophagy protein. Cell Host Microbe, 1, 23–35

12. Smith, M. C., Boutell, C. and Davido, D. J. (2011) HSV-1 ICP0: paving the way for viral replication. Future Virol., 6, 421–429

13. Bryant, K. F., Yan, Z., Dreyfus, D. H. and Knipe, D. M. (2012) Identification of a divalent metal cation binding site in herpes simplex virus 1 (HSV-1) ICP8 required for HSV replication. J. Virol., 86, 6825–6834

14. Dremel, S. E. and DeLuca, N. A. (2019) Herpes simplex viral nucleoprotein creates a competitive transcriptional environment facilitating robust viral transcription and host shut off. eLife, 8, e51109

15. Yang, S., Fu, C., Lian, X., Dong, X. and Zhang, Z. (2019) Understanding human-virus protein-protein interactions using a human protein complex-based analysis framework. mSystems, 4, e00303–e00318

16. Nourani, E., Khunjush, F. and Durmuş, S. (2016) Computational prediction of virus-human protein-protein interactions using embedding kernelized heterogeneous data. Mol. Biosyst., 12, 1976–1986

17. Eid, F.-E., ElHefnawi, M. and Heath, L. S. (2016) DeNovo: virus-host sequence-based protein-protein interaction prediction. Bioinformatics, 32, 1144–1150

18. Qi, Y., Tastan, O., Carbonell, J. G., Klein-Seetharaman, J. and Weston, J. (2010) Semi-supervised multi-task learning for predicting interactions between HIV-1 and human proteins. Bioinformatics, 26, i645–i652

19. Zhou, Y., Zhou, Y. S., He, F., Song, J. and Zhang, Z. (2012) Can simple codon pair usage predict protein-protein interaction? Mol. Biosyst., 8, 1396–1404

20. Guo, Y., Yu, L., Wen, Z. and Li, M. (2008) Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. Nucleic Acids Res., 36, 3025–3030

21. Kotlyar, M., Pastrello, C., Pivetta, F., Lo Sardo, A., Cumbaa, C., Li, H., Naranian, T., Niu, Y., Ding, Z., Vafaee, F., *et al.* (2015) *In silico* prediction of physical protein interactions and characterization of interactome orphans. Nat. Methods, 12, 79–84

22. Li, Z.-G., He, F., Zhang, Z. and Peng, Y.-L. (2012) Prediction of protein-protein interactions between *Ralstonia solanacearum* and *Arabidopsis thaliana.* Amino Acids, 42, 2363–2371

23. Schleker S., Garcia-Garcia, J., Klein-Seetharaman, J. and Oliva, B. (2012) Prediction and comparison of *Salmonella*-human and *Salmonella-Arabidopsis* interactomes. Chem. Biodivers., 9, 991–1018

24. Evans, P., Dampier, W., Ungar, L. and Tozeren, A. (2009) Prediction of HIV-1 virus-host protein interactions using virus and host sequence motifs. BMC Med. Genomics, 2, 27

25. Lee, S. A., Chan, C. H., Tsai, C. H., Lai, J. M., Wang, F. S., Kao, C. Y. and Huang, C. Y. (2008) Ortholog-based protein-protein interaction prediction and its application to inter-species interactions. BMC Bioinformatics, 9, S11

26. Itzhaki, Z. (2011) Domain-domain interactions underlying herpes-virus-human protein-protein interaction networks. PLoS One, 6, e21724

27. Franzosa, E. A. and Xia, Y. (2011) Structural principles within the human-virus protein-protein interaction network. Proc. Natl. Acad. Sci. USA, 108, 10538–10543

28. Hagai, T., Azia, A., Babu, M. M. and Andino, R. (2014) Use of host-like peptide motifs in viral proteins is a prevalent strategy in host-virus interactions. Cell Reports, 7, 1729–1739

29. Yang, S., Li, H., He, H., Zhou, Y. and Zhang, Z. (2019) Critical assessment and performance improvement of plant-pathogen protein-protein interaction prediction methods. Brief. Bioinform., 20, 274–287

30. von Mering, C., Jensen, L. J., Snel, B., Hooper, S. D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M. A. and Bork, P. (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res., 33, D433–D437

31. Bandyopadhyay, S., Ray, S., Mukhopadhyay, A. and Maulik, U. (2015) A review of *in silico* approaches for analysis and prediction of HIV-1-human protein-protein interactions. Brief. Bioinform., 16, 830–851

32. Szpara, M. L., Gatherer, D., Ochoa, A., Greenbaum, B., Dolan, A., Bowden, R. J., Enquist, L. W., Legendre, M. and Davison, A. J. (2014) Evolution and diversity in human herpes simplex virus genomes. J. Virol., 88, 1209–1227

33. Meyniel-Schicklin, L., de Chassey, B., André, P. and Lotteau, V.

(2012) Viruses and interactomes in translation. Mol. Cell. Proteomics, 11, M111.014738

34. Dyer, M. D., Murali, T. M. and Sobral, B. W. (2008) The landscape of human proteins interacting with viruses and other pathogens. PLoS Pathog., 4, e32

35. Zhang, N., Yan, J., Lu, G., Guo, Z., Fan, Z., Wang, J., Shi, Y., Qi, J. and Gao, G. F. (2011) Binding of herpes simplex virus glycoprotein D to nectin-1 exploits host cell adhesion. Nat. Commun., 2, 577

36. Azab, W., Gramatica, A., Herrmann, A. and Osterrieder, N. (2015) Binding of alphaherpesvirus glycoprotein H to surface α4β1-integrins activates calcium-signaling pathways and induces phosphatidylserine exposure on the plasma membrane. MBio, 6, e01552–15

37. Zheng, K., Chen, M., Xiang, Y., Ma, K., Jin, F., Wang, X., Wang, X., Wang, S. and Wang, Y. (2014) Inhibition of herpes simplex virus type 1 entry by chloride channel inhibitors tamoxifen and NPPB. Biochem. Biophys. Res. Commun., 446, 990–996.

38. O'Brien, R. J. and Wong, P. C. (2011) Amyloid precursor protein processing and Alzheimer's disease. Annu. Rev. Neurosci., 34, 185–204

39. Mayer-Proschel, M., Hogestyn, J. M. and Mock, D. J. (2018) Contributions of neurotropic human herpesviruses herpes simplex virus 1 and human herpesvirus 6 to neurodegenerative disease pathology. Neural Regen. Res., 13, 211–221

40. Readhead, B., Haure-Mirande, J.-V., Funk, C. C., Richards, M. A., Shannon, P., Haroutunian, V., Sano, M., Liang, W. S., Beckmann, N. D., Price, N. D., et al. (2018) Multiscale analysis of independent Alzheimer's cohorts finds disruption of molecular, genetic, and clinical networks by human herpesvirus. Neuron, 99, 64–82.e7

41. Bourgade, K., Garneau, H., Giroux, G., Le Page, A. Y., Bocti, C., Dupuis, G., Frost, E. H. and Fülöp, T. Jr. (2015) b-Amyloid peptides display protective activity against the human Alzheimer's disease-associated herpes simplex virus-1. Biogerontology, 16, 85–98

42. The UniProt Consortium. (2019) UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res., 47, D506–D515

43. Lian, X., Yang, S., Li, H., Fu, C. and Zhang, Z. (2019) Machine-learning-based predictor of human-bacteria protein-protein interactions by incorporating comprehensive host-network properties. J. Proteome Res., 18, 2195–2205

44. Csárdi, G. and Nepusz, T. (2006) The igraph software package for complex network research. InterJournal Complex Syst., 1695, 1–9

45. Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A. and Tyers, M. (2006) BioGRID: a general repository for interaction datasets. Nucleic Acids Res., 34, D535–D539

46. Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E.

M. and Eisenberg, D. (2000) DIP: the database of interacting proteins. Nucleic Acids Res., 28, 289–291

47. Ammari, M. G., Gresham, C. R., McCarthy, F. M. and Nanduri, B. (2016) HPIDB 2.0: a curated database for host-pathogen interactions. Database (Oxford), baw103

48. Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., et al. (2004) IntAct: an open source molecular interaction database. Nucleic Acids Res., 32, D452–D455

49. Wattam, A. R., Davis, J. J., Assaf, R., Boisvert, S., Brettin, T., Bun, C., Conrad, N., Dietrich, E. M., Disz, T., Gabbard, J. L., et al. (2017) Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center. Nucleic Acids Res., 45, D535–D542

50. Breuer, K., Foroushani, A. K., Laird, M. R., Chen, C., Sribnaia, A., Lo, R., Winsor, G. L., Hancock, R. E. W., Brinkman, F. S. L. and Lynn, D. J. (2013) InnateDB: systems biology of innate immunity and beyond – recent updates and continuing curation. Nucleic Acids Res., 41, D1228–D1233

51. Guirimand, T., Delmotte, S. and Navratil, V. (2015) VirHostNet 2.0: surfing on the web of virus/host molecular interactions data. Nucleic Acids Res., 43, D583–D587

52. Schaefer, M. H., Fontaine, J. F., Vinayagam, A., Porras, P., Wanker, E. E. and Andrade-Navarro, M. A. (2012) HIPPIE: Integrating protein interaction networks with experiment based quality scores. PLoS One, 7, e31826

53. Stein, A., Céol, A. and Aloy, P. (2011) 3did: identification and classification of domain-based interactions of known three-dimensional structure. Nucleic Acids Res., 39, D718–D723

54. Liu, X., Huang, Y., Liang, J., Zhang, S., Li, Y., Wang, J., Shen, Y., Xu, Z. and Zhao, Y. (2014) Computational prediction of protein interactions related to the invasion of erythrocytes by malarial parasites. BMC Bioinformatics, 15, 393

55. El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., et al. (2019) The Pfam protein families database in 2019. Nucleic Acids Res., 47, D427–D432

56. Eddy, S. R. (1998) Profile hidden Markov models. Bioinformatics, 14, 755–763

57. Maere, S., Heymans, K. and Kuiper, M. (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics, 21, 3448–3449

58. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res., 13, 2498–2504