Virus Evolution, 2020, 6(1): veaa028

doi: 10.1093/ve/veaa028
Research article

Intrinsically disordered regions are abundant in simplexvirus proteomes and display signatures of positive selection

Alessandra Mozzi,[1,*,†] Diego Forni,[1] Rachele Cagliani,[1] Mario Clerici,[2,3] Uberto Pozzoli,[1] and Manuela Sironi[1]

[1]Scientific Institute, IRCCS E. MEDEA, Bioinformatics, Bosisio Parini 23842, Italy, [2]Department of Physiopathology and Transplantation, University of Milan, Milan 20090, Italy and [3]Don C. Gnocchi Foundation ONLUS, IRCCS, Milan 20148, Italy

*Corresponding author: E-mail: alessandra.mozzi@lanostrafamiglia.it
†https://orcid.org/0000-0003-3911-1042

Abstract

Whereas the majority of herpesviruses co-speciated with their mammalian hosts, human herpes simplex virus 2 (HSV-2, genus *Simplexvirus*) most likely originated from the cross-species transmission of chimpanzee herpesvirus 1 to an ancestor of modern humans. We exploited the peculiar evolutionary history of HSV-2 to investigate the selective events that drove herpesvirus adaptation to a new host. We show that HSV-2 intrinsically disordered regions (IDRs)—that is, protein domains that do not adopt compact three-dimensional structures—are strongly enriched in positive selection signals. Analysis of viral proteomes indicated that a significantly higher portion of simplexvirus proteins is disordered compared with the proteins of other human herpesviruses. IDR abundance in simplexvirus proteomes was not a consequence of the base composition of their genomes (high G + C content). Conversely, protein function determines the IDR fraction, which is significantly higher in viral proteins that interact with human factors. We also found that the average extent of disorder in herpesvirus proteins tends to parallel that of their human interactors. These data suggest that viruses that interact with fast-evolving, disordered human proteins, in turn, evolve disordered viral interactors poised for innovation. We propose that the high IDR fraction present in simplexvirus proteomes contributes to their wider host range compared with other herpesviruses.

Key words: HSV-2; simplexviruses; positive selection; intrinsically disordered regions (IDRs), virus–host interactors.

1. Introduction

Herpesviruses (order *Herpesvirales*) constitute a diverse family of enveloped, double-stranded DNA viruses that infect a wide range of animals, usually resulting in life-long infections (McGeoch, Rixon, and Davison 2006). Herpesviruses have complex genomes, which characteristically contain regions of unique sequence flanked by direct or inverted repeats. Although herpesviruses express non-coding RNAs and miRNAs, protein-coding regions occupy the great majority of their genomes (McGeoch, Rixon, and Davison 2006).

Nine herpesviruses, all of them belonging to the *Herpesviridae* family, naturally infect humans: herpes simplex viruses 1 and 2 (HSV-1 and -2), varicella-zoster virus (VZV), Epstein–Barr virus (EBV), human cytomegalovirus (HCMV), human herpesviruses 6A and 6B (HHV-6A and -6B), human herpesvirus 7 (HHV-7), and human herpesvirus 8 (HHV-8, also known as Kaposi's sarcoma-associated herpesvirus). These viruses are among the most successful human pathogens in terms of global distribution, persistence in the host, and transmissibility. Most adults are infected with at least one herpesvirus and the seroprevalence of some of these viruses is as high as 90 per cent in

human populations (Balfour et al. 2013). Primary infection of immunocompetent human subjects generally results in a mild disease, although severe consequences may develop in specific groups of individuals. For instance, intra-uterine HCMV infection is the leading infectious cause of deafness and intellectual disability in children (Manicklal et al. 2013), whereas HSV-1, -2, and VZV are common etiologic agents of non-epidemic acute encephalitis in Western countries (Venkatesan 2013). When contracted at birth, human simplexviruses (HSV-1 and -2) can also cause neonatal invasive infection (Whitley 2004).

Viruses related to HHVs have been described in non-human primates (NHPs; Tischer and Osterrieder 2010). In these animals, seroprevalence can be very high and, similarly to what is observed in humans; infection can either cause a mild disease or can be asymptomatic (Tischer and Osterrieder 2010). In analogy to observations in other mammalian and non-mammalian hosts, herpesvirus infection is remarkably species-specific in primates, with most viruses naturally infecting a single host. Nonetheless, cases of cross-species transmission have been described (Eberle and Jones-Engel 2017). One of these refers to the macacine herpesvirus 1 (also known as B-virus), a simplexvirus indigenous in Asiatic macaques. In its natural host, B-virus infection is almost asymptomatic, but transmission to humans results in a severe, often fatal form of encephalomyelitis (Tischer and Osterrieder 2010; Eberle and Jones-Engel 2017). Likewise, infection of New World monkeys with HSV-1 is almost invariably fatal (Tischer and Osterrieder 2010).

Not only these examples highlight the potential zoonotic threat posed by herpesviruses but they also point to long-standing host–virus association. Indeed, the phylogenetic relationships among hepeviruses tend to mirror those among their hosts, indicating that viral lineages frequently arose through co-speciation with host lineages (McGeoch, Rixon, and Davison 2006). Among HHVs, though, an exception to this general tendency is represented by HSV-2. In fact, this virus most likely originated around 1.6 million years ago from the cross-species transmission of Chimpanzee herpes virus 1 (PanHV-3) to an ancestor of modern humans (Severini et al. 2013; Wertheim et al. 2014; Underdown, Kumar, and Houldcroft 2017).

Host-shift events are often accompanied by bursts of positive selection, as the virus adapts to infect and to be efficiently transmitted in a new species (Longdon et al. 2014; Sironi et al. 2015). We thus exploited the availability of the PanHV-3 genome, as well as of several HSV-2 sequences, to search for adaptive variants that arose or increased in frequency after the cross-species transmission event. Our results indicate that positive selection mainly occurred in HSV-2 protein regions that are intrinsically disordered (intrinsically disordered regions, IDRs)—that is, domains that do not adopt compact three-dimensional structures (Dyson and Wright 2005). We also show that, compared with other herpesviruses, IDRs are particularly abundant in the proteomes of primate-infecting simplexviruses and that the average extent of disorder in viral proteins tends to parallel that of the host interactors.

2. Methods

2.1 Sequences and alignments

HSV-2 genome sequences were retrieved from the NCBI (http://www.ncbi.nlm.nih.gov/, last accessed 20 January 2020) database. A list of accession IDs is reported in Supplementary Table S1.

Genome sequences were chosen to be representative of the genetic diversity of circulating HSV-2 strains. The pool of HSV-2 genomes included an equal number of strains sampled in different geographic areas. Only isolates that were directly sequenced with no (or very limited) *in vitro* passages were included. The genome sequence of the Chimpanzee alpha-1 herpesvirus (PanHV-3, NC_023677) was also retrieved from NCBI.

For each viral genome, we retrieved coding sequences of all annotated open reading frames (ORFs). For non-annotated genomes, ORFs were deduced by whole genomes alignments performed with Progressive MAUVE 2.3.1 (Darling et al. 2004; Darling, Mau, and Perna 2010), and orthology was inferred according to MAUVE attribution.

Fully or partially overlapping ORFs were merged (i.e. *UL26/UL26A*), completely removed (i.e. *UL15/UL16/UL17*), or partially analyzed (i.e. *UL13/UL14*, *US8/US8A*, and *US10/US11*), depending on whether they are translated in the same frame or not. Likewise, the *UL29*, *UL30*, and *UL39* genes were excluded from the analysis because most HSV-2 strains carry fragments deriving from recombination with HSV-1 (Burrel et al. 2017; Casto et al. 2020).

For each gene, alignments were generated using MAFFT (Katoh and Standley 2013), setting sequence type as codons; unreliably aligned codons were filtered using GUIDANCE2 (Sela et al. 2015) with a codon score of 0.90 (Privman, Penn, and Pupko 2012). The resulting alignments were manually inspected.

2.2 Analysis of selective patterns in HSV-2

Analyses were performed with gammaMap that uses intra-specific variation and inter-specific diversity to estimate, along coding regions, the distribution of selection coefficients (γ). In this framework, γ is defined as $2PN_es$, where P is the ploidy, N_e is effective population size, and s is the fitness advantage of any amino acid-replacing derived allele (Wilson et al. 2011).

For each gene, the corresponding coding sequence of PanHV-3 was used as the outgroup.

We assumed θ (neutral mutation rate per site), k (transitions/transversions ratio), and T (branch length) to vary within genes following log-normal distributions, whereas P (probability of adjacent codons to share the same selection coefficient) following a log-uniform distribution. For each gene, we set the neutral frequencies of non-STOP codons (1/61). For selection coefficients, we considered a uniform Dirichlet distribution with the same prior weight for each selection class. For each gene, we performed two runs with 100,000 iterations each and with a thinning interval of ten iterations. Runs were merged after checking for convergence and sites showing a cumulative probability higher than 0.75 of having $\gamma \geq 1$ (Quach et al. 2013) were defined as positively selected.

HSV-2 genes were classified as 'core' or 'non-core' following the characterization by Davison (2007).

The dN–dS parameter was calculated using the single-likelihood ancestor counting method (Kosakovsky Pond and Frost 2005).

2.3 Analysis of intrinsic disorder in herpesvirus and human proteins

IUPred2 (https://iupred2a.elte.hu/plot, last accessed 20 January 2020; Meszaros, Simon, and Dosztányi 2009; Mészáros, Erdos, and Dosztányi 2018; Dosztanyi 2018) was used to predict the fraction of disordered residues (score > 0.5) in the proteomes of

HSV-2, of viruses in the *Simplexvirus* genus, as well as of other HHVs spanning all the three *Herpesviridae* subfamilies (*Alphaherpesvirinae*, *Betaherpesvirinae*, and *Gammaherpesvirinae*, Supplementary Table S2). The whole proteome of each viral species was retrieved from NCBI database.

A long disorder prediction type was selected for all analyses. We defined as IDR, a protein region with at least thirty consecutive amino acids showing a IUPred2 score > 0.5 (Ward et al. 2004; Dosztanyi 2018).

The binomial test was performed by counting, for the sixty-seven analyzed ORFs, the number of disordered regions carrying at least one positively selected site (to avoid the non-independence among nearby sites), the total number of all regions (disordered or not) carrying a positively selected site, and the fraction of IDR length compared with the total length of the proteome.

Human–virus protein–protein interactions were obtained from the virus Mentha database (https://virusmentha.uniroma2.it/, last accessed 20 January 2020; Calderone, Castagnoli, and Cesareni 2013; Calderone, Licata, and Cesareni 2015). As few interactions were available for HSV-2, and because HSV-1 and -2 are closely related, we merged interactors of both viruses. Interactors of HHV-6B and VZV were not analyzed, as few data were available. IDRs for human proteins were calculated as for viral molecules.

HSV-1, -2, and VZV ortholog core genes were retrieved from Davison (2007).

All statistical tests were performed in the R environment (version 3.6.2, http://www.r-project.org, last accessed 7 January 2020).

3. Results

3.1 IDRs in HSV-2 proteins display abundant signals of positive selection

As mentioned earlier, HSV-2 is phylogenetically related to PanHV-3, as it is thought to have originated from the cross-species transmission of this latter virus to an intermediate host, possibly *Paranthropus boisei*, eventually reaching an ancestor of modern humans (Underdown, Kumar, and Houldcroft 2017). To investigate the selective events that accompanied the adaptation of HSV-2 to our species, we applied a population genetics–phylogenetics approach. Specifically, we used the gammaMap program (Wilson et al. 2011), that leverages intra-species variation and inter-species diversity to estimate the distribution of fitness effects (i.e. selection coefficients, γ, expressed as discrete categories from -500 to 100) along coding regions. In practical terms, γ values can be considered a measure of the fitness consequences of new non-synonymous mutations. Thus, we used the PanHV-3 sequence as an outgroup and we analyzed a representative phylogeny of fifty-three HSV-2 strains derived from clinical isolates sampled worldwide (Supplementary Table S1); among these, we included strains belonging to the recently identified African HSV-2 lineage (Burrel et al. 2017). For each HSV-2-coding gene annotated in the reference HG52 strain, orthologs were retrieved and aligned (see Section 2).

After discarding overlapping ORFs and genes (*UL29*, *UL30*, and *UL39*) deriving from recombination with HSV-1 (Burrel et al. 2017; Casto et al. 2020), the distribution of selection coefficients was estimated along sixty-seven ORFs (Fig. 1). These included both 'core' genes that are shared by all herpesviruses, and 'non-core' genes, which are specific for the *Simplexvirus* genus or for HSV-1/-2 only.

In general, most median values of γ were comprised between -10 and -1, both for core and for non-core genes, indicating that the majority of sites are subject to weak purifying selection. Only eight genes evolved under a stronger negative constraint (median $\gamma = -50$). Some of these latter code for proteins playing a key role in viral replication and spread (*UL31*, *UL34*, *UL40*, *UL41*, *UL4*, and *US9*) or for fundamental viral structural components (*UL18*).

To identify signals of positive selection, we estimated codon-wise posterior probabilities for each selection coefficient. Specifically, we defined a codon as positively selected if its cumulative posterior probability of $\gamma \geq 1$ was > 0.75. According to these criteria, we identified positively selected sites in twenty genes, eight cores (fraction selected= 22.2%), and twelve non-core (fraction selected = 38.7%; Fig. 1 and Supplementary Table S3).

Analysis of positively selected sites indicated that ~29 per cent of them map to regions with biased amino acid composition (e.g. proline-rich, alaninine-rich; Fig. 2). Because such regions are often disordered, we investigated whether positive selection is more likely to occur in IDRs. Thus, the presence of IDRs was predicted with IUPred2 (Mészáros, Simon, and Dosztányi 2009; Mészáros, Erdos, and Dosztányi 2018; Dosztanyi 2018) for all HSV-2 proteins we analyzed with gammaMap. We found that the large majority of selected sites fall within IDRs. In particular, we defined IDRs as regions with at least thirty consecutive residues with a IUPred2 score higher than 0.5 (Ward et al. 2004; Dosztanyi 2018). Statistical analysis indicated that IDRs are significantly more likely to harbor at least one positively selected site than expected based on their relative proportion in HSV-2 proteins (Binomial test, $P = 2.3 \times 10^{-07}$, see Section 2). This finding is in line with the fact that a higher fraction of non-core genes are positively selected, as their encoded proteins harbor more disordered residues compared with proteins encoded by core genes (Wilcoxon rank sum test, $P = 0.023$; Fig. 3A).

Previous analysis of IDRs, mostly from cellular organisms, indicated that these regions frequently display signatures of positive selection and that they are subject to weaker selective constraint than ordered regions (Brown, Johnson, and Daughdrill 2010). We thus compared the disorder score of HSV-2 protein residues with different selection coefficients (ranging from strongly deleterious or lethal to strongly beneficial). Results indicated a clear-cut difference in the distribution of disorder scores, with most constrained sites being preferentially located in ordered protein regions (Fig. 3B).

gammaMap models the allele frequency spectrum (conditioned on the ancestral allele) within a population and combines it with a model of the substitution process between species. The method thus intrinsically differs from approaches based on the comparison of the non-synonymous (dN) and synonymous (dS) substitution rates, which are best-suited to study inter-species diversity (i.e. long-term evolutionary processes; Kryazhimskiy and Plotkin 2008). As a comparison, for the twenty genes with at least one positively selected site detected by gammaMap, we calculated dN–dS. This metric was preferred over the conventional dN/dS ratio because it is not rendered to infinite for dS values equal to 0. As expected, virtually no correlation was observed between dN–dS and γ values (Kendall's rank correlation, tau = 0.048, $P = 4.4 \times 10^{-10}$; Supplementary Fig. S1).

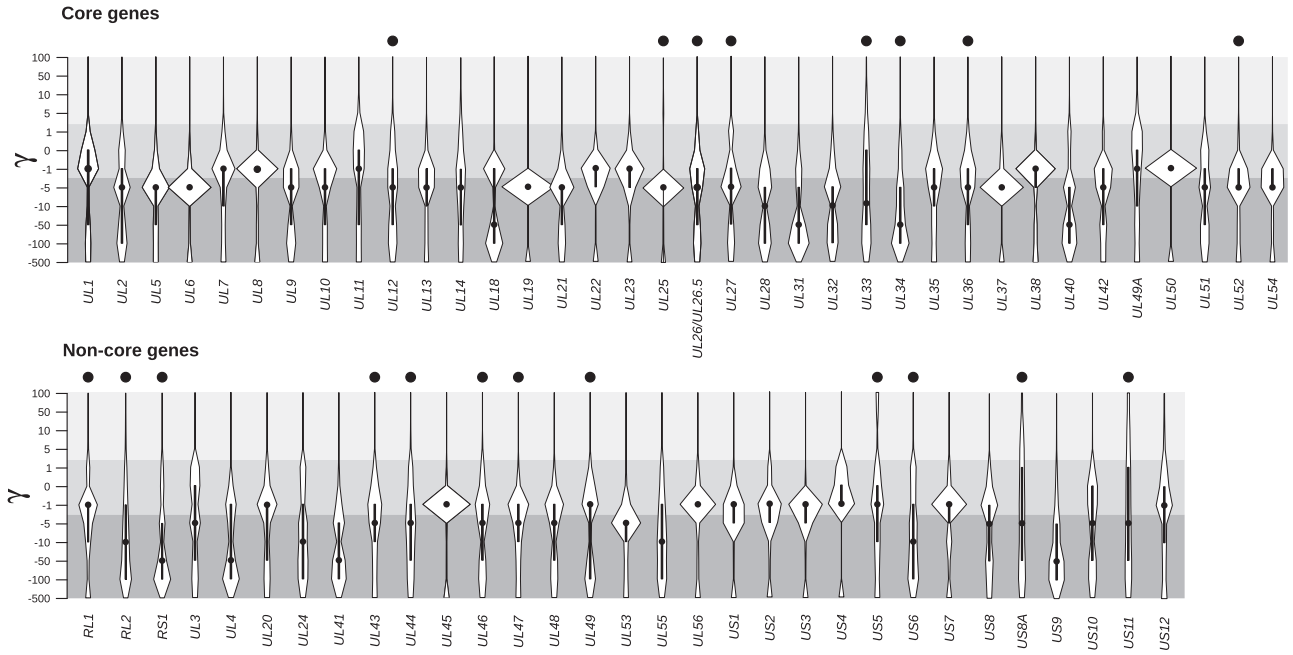

Figure 1. Population genetics–phylogenetics analysis of HSV-2 genes. Violin plots (median, black dot; interquartile range, black bar) of selection coefficients for HSV-2 core and non-core genes. The gray shading denotes different degrees of constraint based on selection coefficients ($\gamma \leq 5$, moderately or strongly deleterious, inviable; $-1 \leq \gamma \leq 1$, neutral or weakly deleterious/beneficial; $\gamma \geq 5$, moderately or strongly beneficial). Black dots above the plots indicated genes in which positively selected sites were detected by gammaMap (see Section 2).

3.2 Primate simplexvirus proteomes are particularly rich in IDRs

Given the results above, we investigate whether a similar proteome fraction is occupied by IDRs in HSV-2 and in other HHVs. We thus used IUPred2 to calculate the fraction of disordered residues in the proteomes of HSV-2 and of other human alphaherpesviruses (HSV-1 and VZV), betaherpesviruses (HHV-6B and HCMV), and gammaherpesviruses (EBV and HHV-8). Results indicated that a significantly higher portion of simplexvirus proteins is disordered compared with the proteins of other HHVs (Fig. 4A). Among these latter, HHV-6B and HHV-8 showed the lowest fraction of IDRs (Fig. 4A).

We next assessed whether a high fraction of disordered residues is a general feature of simplexviruses by analyzing the proteomes of viruses that infect NHPs and other mammals. We found that the IDR fraction of NHP simplexvirus proteins is similar to that of HSV-1 and -2. Conversely, simplexviruses infecting other mammals tended to have a lower fraction of their proteomes occupied by disordered residues, although the difference with the human viruses was not statistically significant (Fig. 4B).

Because both HSV-1/-2 and VZV belong to the *Alphaherpesvirinae* subfamily but display very different fractions of IDRs, we compared HSV-2 and VZV orthologous core genes in terms of disordered fraction in the encoded proteins. With the exception of a few proteins with similar fraction of IDRs, most HSV-2 proteins showed a higher fraction of disordered residues than VZV proteins, indicating that the results we obtained at the level of the whole proteome (Fig. 4C) are not due to a minority of outliers. Very similar results were obtained when HSV-1 and VZV orthologs were analyzed (Supplementary Fig. S2).

3.3 Interaction with Host Proteins Influences the Disordered Fraction of Simplexvirus Proteins

Previous studies showed that the level of protein disorder in viral proteomes is influenced by genome size and base composition (G + C content; Pushker et al. 2013). We noticed no association between IDR fraction and genome size for HHV or HSVs that infect NHPs and other mammals (Fig. 4A and B). Primate simplexviruses, however, had higher G + C content (both calculated over all codon positions and for the third position only) compared with all other HHVs. Simplexviruses infecting non-primate mammals had intermediate G + C levels. Other than this effect, we observed no covariation between G + C content and IDR fraction. For instance, the relatively G + C rich HCMV and EBV genomes encode proteins with a similar proportion of IDRs as the G + C poor HHV-6B and VZV genomes (Fig. 4A). To further address the role of base composition as a determinant of protein disorder in Simplexviruses, we correlated the IDR fraction of all HSV-1 and -2 proteins with the G + C content at the third codon position of their respective ORFs. No correlation was observed (Fig. 5), suggesting that the abundance of disordered regions in these viruses is not simply a consequence of their having G + C rich genomes.

Because IDRs are frequently involved in protein–protein interactions, we hypothesized that HHV proteins that interact with human factors have a higher fraction of disordered regions. Human–virus interactions were obtained from virus Mentha (https://virusmentha.uniroma2.it/; Calderone, Licata, and Cesareni 2015). As few interactions were available for HSV-2, and because HSV-1 and -2 are closely related, we assigned to both viruses the interactions reported for either. Conversely, HHV-6B and VZV were not analyzed, as very limited interaction data were available. Results indicated that the fraction of disordered residues is significantly higher in viral proteins that interact with human factors for HSV-1, -2, and HHV-8. This was not the case for EBV and HCMV (Fig. 6A).

3.4 Human interactors of simplexvirus proteins have a high fraction of disordered regions

Previous works have shown that human proteins that interact with viruses have a high fraction of IDRs, although differences

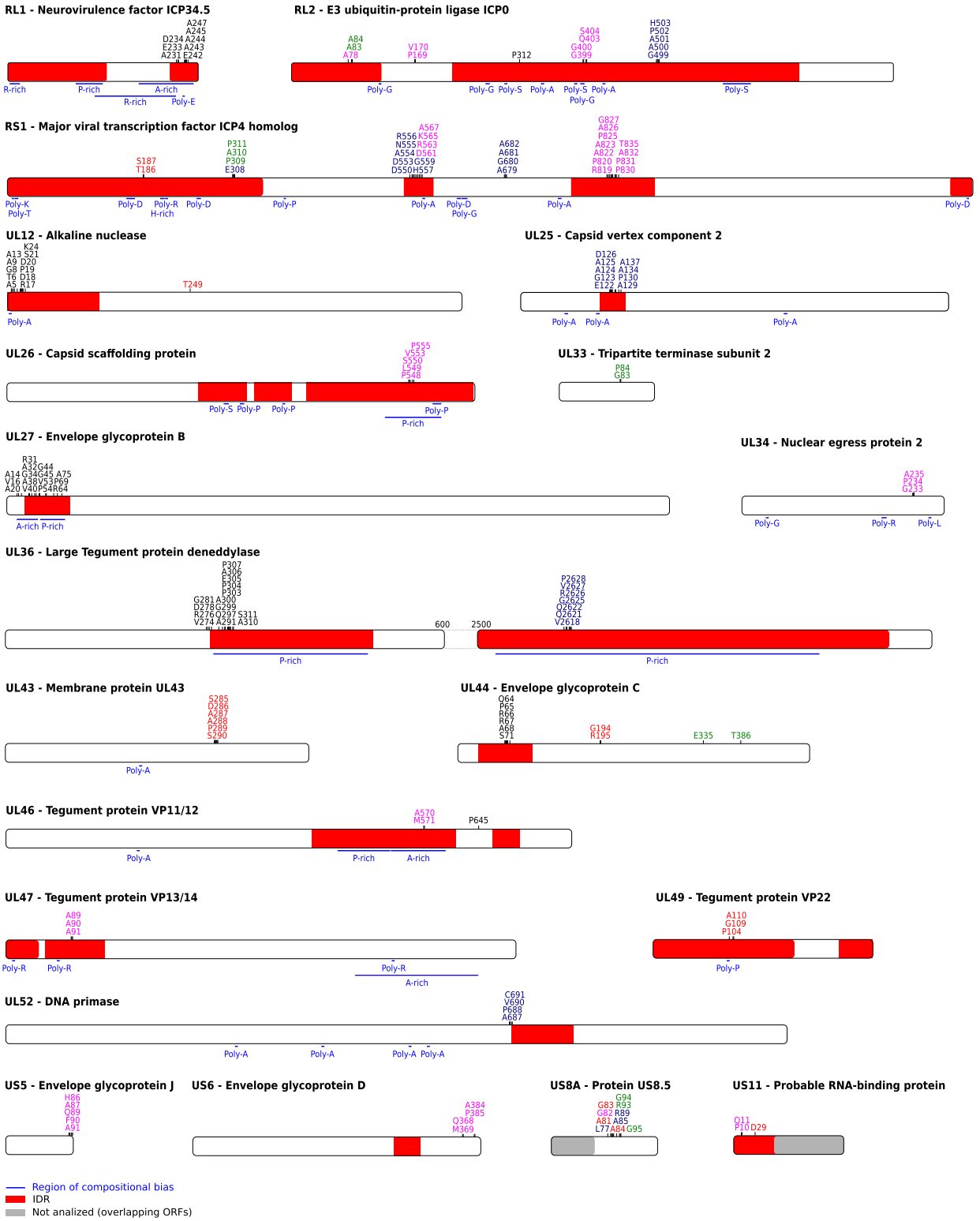

Figure 2. Distribution of positively selected sites. Sites that were detected by gammaMap as positively selected were mapped onto HSV-2 proteins together with the location of IDRs (red). Sites are color coded based on their highest posterior probability of γ (1, black; 5, blue; 10, green; 50, red; 100, magenta). Overlapping coding regions not considered in gammaMap analysis are in gray. The compositional biased region is also indicated along the protein sequence (blue lines). Positions refer to the reference HG52 strain (NC_001798).

were evident depending on the virus (Lou et al. 2016; Bosl et al. 2019). We wanted to investigate whether the different representation of IDRs in herpesvirus proteins is paralleled by different levels of disorder in the human interactors. We thus calculated IDRs for all herpesvirus-interacting human proteins, as derived from virus Mentha. Analysis indicated that indeed this is the

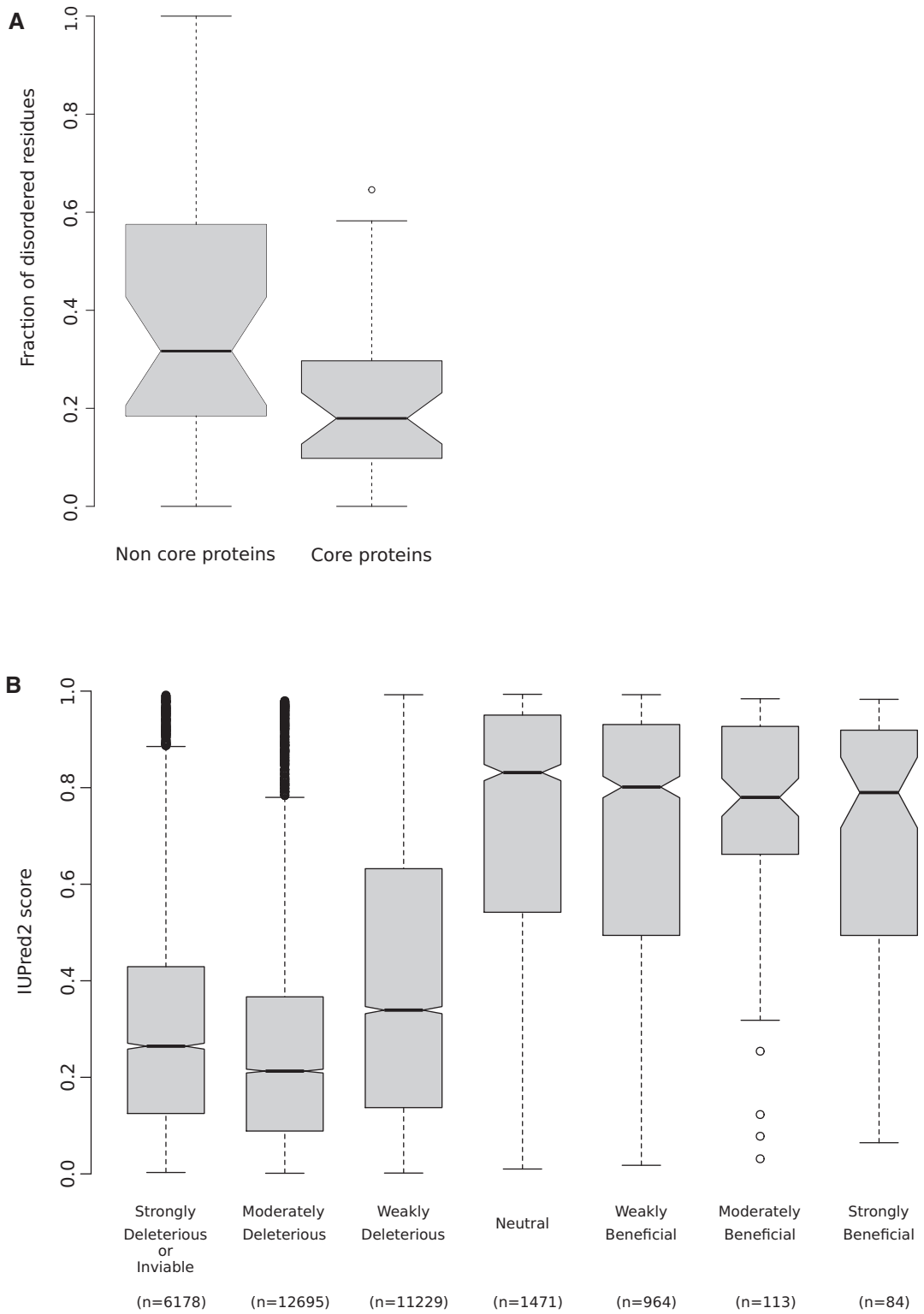

Figure 3. Intrinsically disordered residues in HSV-2 proteins. (A) The fraction of disordered residues of each HSV-2 non-core or core protein is plotted as boxplot. (B) IUPred2 score calculated for each residue of all HSV-2 protein is plotted as boxplot. Residues are grouped based on their highest posterior probability of γ, as calculated by gammaMap. Selection coefficients were defined as: strongly deleterious/inviable ($\gamma \leq -50$), moderately deleterious ($\gamma = -10$ or $\gamma = -5$), weakly deleterious ($\gamma = -1$), neutral ($\gamma = 0$), weakly beneficial ($\gamma = 1$), moderately beneficial ($\gamma = 5$ or $\gamma = 10$), and strongly beneficial ($\gamma = 50$ or $\gamma = 100$).

case, as a significantly higher fraction of human proteins that interact with HSV-1 and -2 is occupied by IDRs compared with proteins that interact with EBV and HCMV products (Fig. 6B). Human interactors of HHV-8 proteins also displayed a consistent fraction of IDRs, although not as high as for HSV-1/-2 (Fig. 6B). These results were also confirmed when only

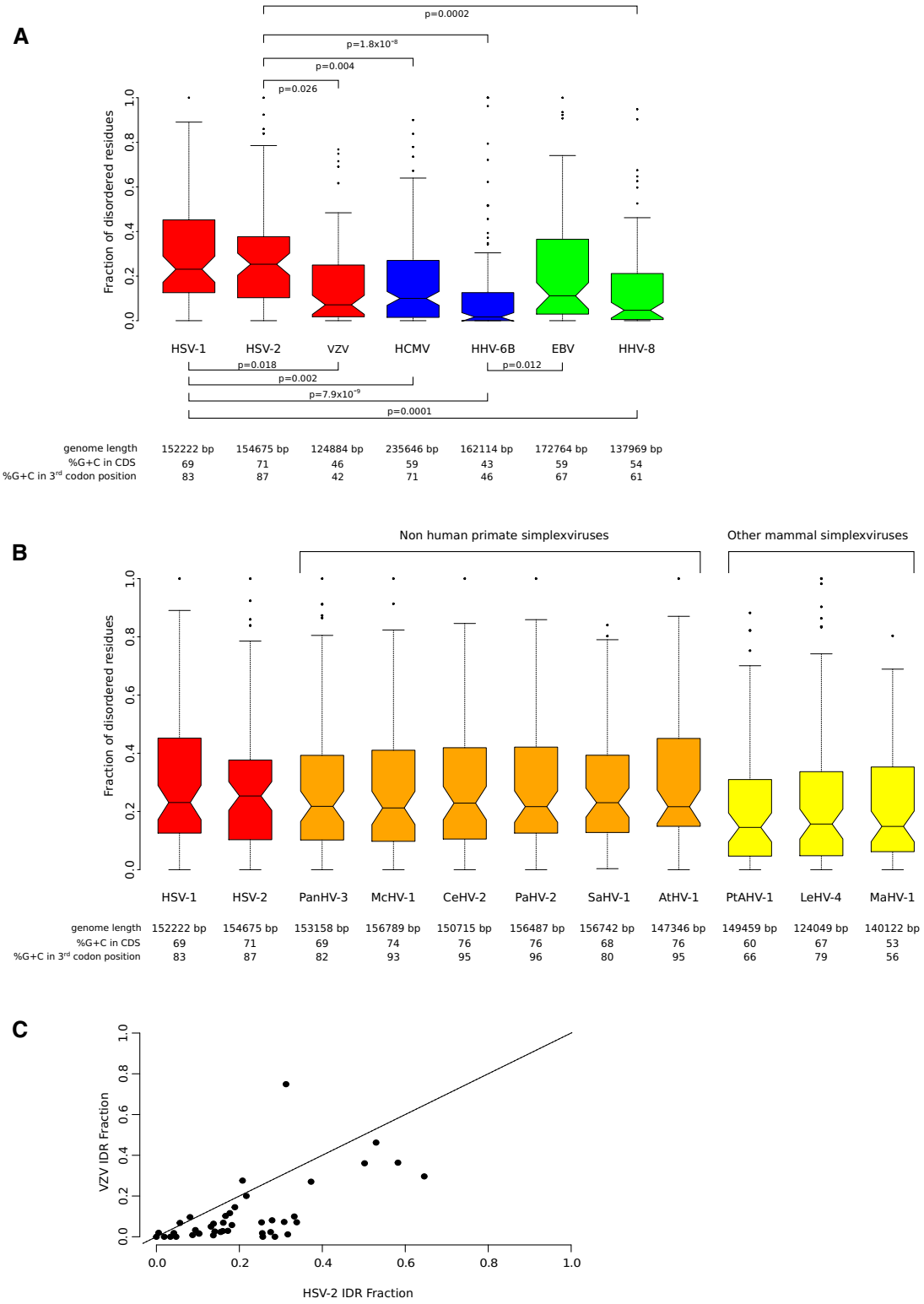

Figure 4. Intrinsically disordered residues in herpesviruses. (A) Boxplot representation of the fraction of disordered residues for HHV proteins. The genome length of each virus is also reported, along with the percentage of G + C content calculated for the whole-coding sequence or considering the third codon position only. Viruses are colored based on their subfamily: red, *Alphaherpesvirinae*; blue, *Betaherpesvirinae*; green, *Gammaherpesvirinae*. Statistically significant Nemenyi *post-hocs* after the Kruskal–Wallis test are reported. (B) Boxplot representation as in panel (A) for some representative species of the *Simplexvirus* genus. NHP simplexviruses are shown in orange, other mammal simplexviruses are shown in yellow (see Supplementary Table S2). (C) Scatter plot of the fraction of disordered residues for HSV-2 and VZV. Each dot represents an orthologous core protein.

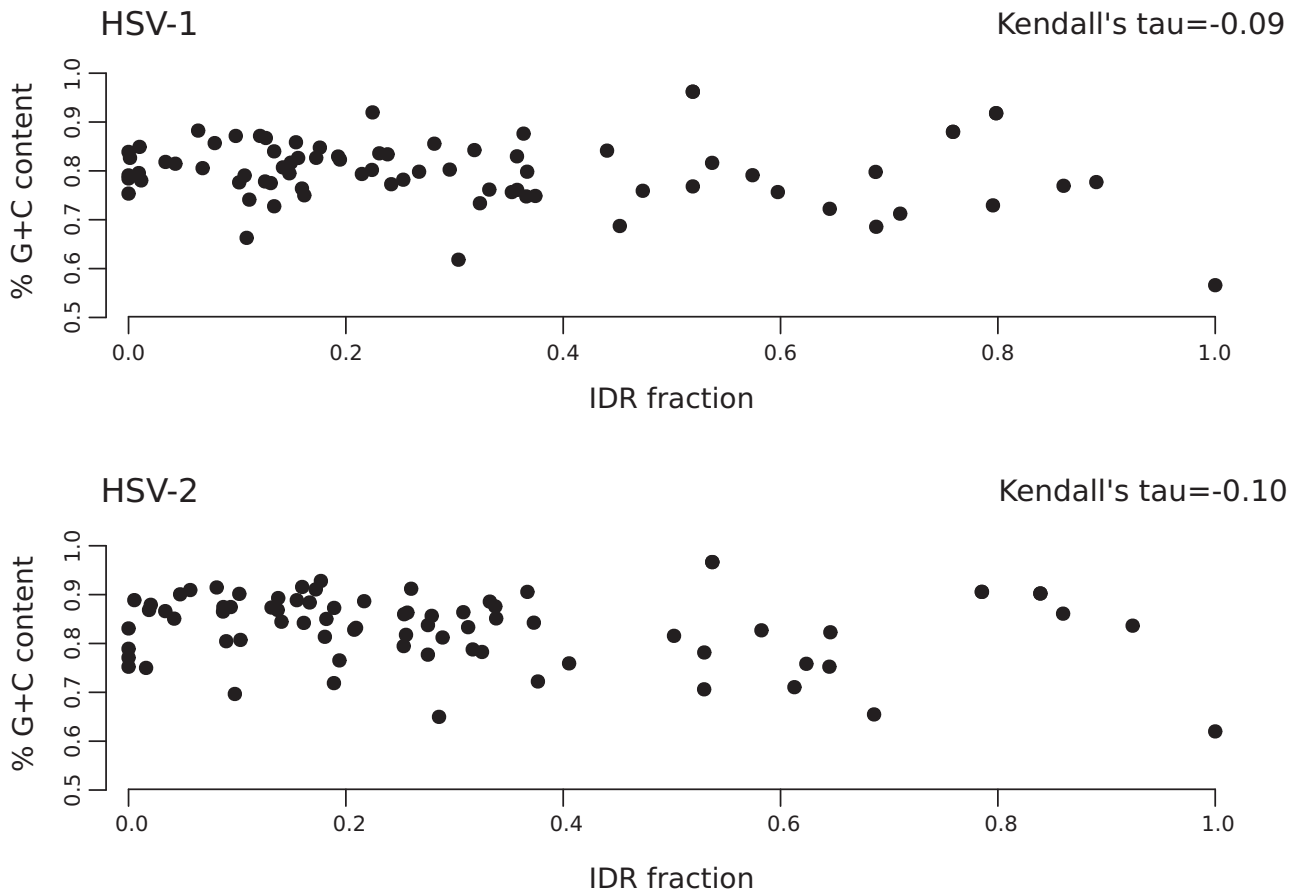

Figure 5. Correlation between G + C content and disordered residues. The percentage of G + C content calculated for the third codon position of each HSV-1 or -2 gene is plotted against the fraction of disordered residues for the protein encoded by the same gene.

interactors specific for each virus were analyzed (Supplementary Fig. S3).

4. Discussion

We exploited the peculiar evolutionary history of HSV-2 to investigate the relatively recent adaptive events that turned a chimpanzee herpesvirus into a successful human pathogen. To provide a genome-wide picture of the selective patterns acting on coding genes, we calculated the distribution of selection coefficients using gammaMap, which is relatively insensitive to the effects of demography and recombination (Wilson et al. 2011). As expected, we found that the majority of genes evolved under purifying selection. However, our major interest was to identify variants that increased in frequency as a result of positive selection after the cross-species transmission event. Such variants are expected to represent an adaptation to replication in human cells and to transmission in human populations. The most striking observation was the distribution of positively selected sites, as most of them occurred within IDRs. The clustering of selection signals, we observed in these regions, as well as in ordered portions, is expected to be minimally due to the approach we used, as the gammaMap sliding window model we applied (as recommended) only causes a slight increase in the probability of positive selection at nearby sites (Wilson et al. 2011). In any case, the significant enrichment of positive selection signals is independent of clustering, as we calculated the probability of IDRs and of non-disordered regions to harbor at least one positively selected site.

The general tendency of IDRs to be fast evolving was previously reported (Brown, Johnson, and Daughdrill 2010; Schlessinger et al. 2011; Toth-Petroczy and Tawfik 2011), and large-scale analyses in budding yeast and mammals indicated that significantly stronger positive selection is observed in intrinsically disordered compared with ordered regions (Nilsson, Grahn, and Wright 2011; Afanasyeva et al. 2018). Although to our knowledge, no systematic analysis was performed for viral proteomes, instances of adaptive evolution involving IDRs were described for some RNA viruses (Ortiz et al. 2013; Gitlin et al. 2014; Charon et al. 2018) and for polyomaviruses (Lauber et al. 2015). In particular, an analysis of the nodavirus polymerase indicated that the fast evolving, highly disordered C-terminus displays high functional robustness to amino acid replacements (Gitlin et al. 2014), whereas a study on the potyvirus genome-linked protein showed that amino acid changes that increase disorder expand the host range (Charon et al. 2018). Thus, disordered regions in viral proteins were suggested to afford evolutionary plasticity while preserving protein function (Gitlin et al. 2014; Charon et al. 2018). Our data strongly support this view and provide proteome-wide evidence that HSV-2 adaptation to the human/hominid host mainly occurred through changes within IDRs.

Previous studies indicated that, compared with cellular organisms, viral proteomes have a wider variation in IDR fraction (Pushker et al. 2013; Xue et al. 2014; Peng et al. 2015).

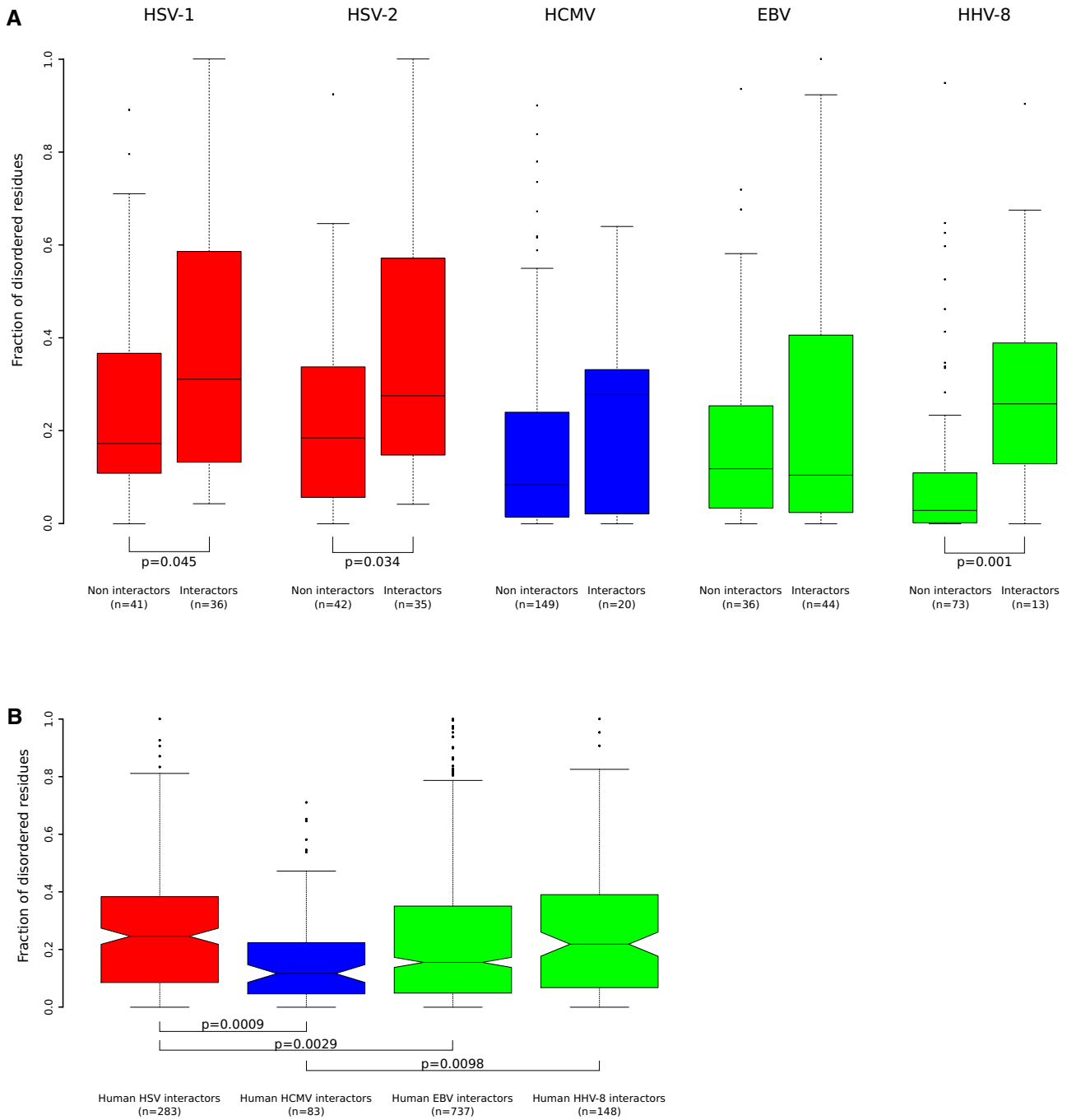

Figure 6. Fraction of disordered residues among interacting proteins. (A) Boxplots of the fraction of disordered residues for viral proteins having or not at least one known human protein interactor (see Section 2 for details). Statistically significant Wilcoxon rank sum tests are also reported. Colors are as in Fig. 4. (B) Fraction of disordered residues for human proteins that interact with herpesvirus proteins. Colors reflect panel (A). Statistically significant Nemenyi *post-hocs* after the Kruskal–Wallis test are reported.

Such variation only weakly depends on genome size, whereas a stronger effect was described for base composition. This is partially expected, as residues enriched in disordered regions are mainly encoded by G and C rich codons (Pushker et al. 2013; Basile et al. 2019). Clearly, this observation opens the question as to whether the fraction of IDRs is simply a consequence of base composition or if it is an adaptive feature driven by specific viral/host characteristics. Results herein document a high variability of disordered regions within the *Herpesviridae* family. We confirm that genome size is not a major determinant of the extent of protein disorder in herpesviruses and we show that base composition is also unlikely to explain the high fraction of simplexvirus proteomes occupied by IDRs. Although simplexviruses have G+C rich genomes, we found no correlation between protein disorder and codon composition at the third codon position. Instead, we show that non-core genes, most of which are involved in determining cellular tropism, immune evasion, and transactivation (Davison, Dargan, and Stow 2002), encode proteins with a higher fraction of IDRs than core genes. Consistently, viral proteins that interact with host components

are more disordered than those without known human interactors. Notably, this effect was also observed for HHV-8 proteins: although the virus encodes proteins with a low IDR fraction, viral proteins that interact with the host have an extent of disorder comparable with that of HSV-1 and -2. These observations clearly point to the fact that viral protein function, rather than genome base composition, determines the fraction IDRs. This conclusion is strengthened by the observation that human proteins that interact with HSV-1/-2 also have a higher fraction of disorder than those interacting with EBV or HCMV, with HHV-8 interactors having intermediate levels of disorder. Thus, although human proteins that interact with viruses are, in general, more disordered than those with no viral binding partners (Lou et al. 2016; Bosl et al. 2019), differences exist depending on the virus and on the viral proteome. Indeed, a low level of disorder was previously reported for EBV interactors (Bosl et al. 2019). Conversely, the IDR fraction of human proteins that interact with other HHVs had never been investigated before. Interestingly, though, Lou and coworkers recently identified the DNA repair protein Nbs1 as an interactor of the HSV-1 protein ICP0 (encoded by *RL2*; Lou et al. 2016). Specifically, human Nbs1 binds the viral protein through a disordered region, and differences in this same region across primate species account for the species-specific effect of Nbs1 on viral replication. The authors thus suggested that, because of the fast evolution of IDRs, genetic arms-races between hosts and viruses may commonly involve disordered regions (Lou et al. 2016), as these are likely to form interaction surfaces and thus are expected to experience the strongest selective pressure (Sironi et al. 2015). Our data indirectly confirm this prediction for simplexviruses, as we found that HSV-2 adaptation to its human host mostly involved disordered regions, which are, in turn, more abundant among viral proteins that interact with host components. We also found that the average extent of disorder in viral proteins tends to parallel that of the host interactors, which is consistent with the genetic conflict hypothesis, as IDRs also evolve faster in humans and mammals (Afanasyeva et al. 2018). Thus, these data suggest that viruses that interact with fast-evolving, disordered human proteins in turn evolve disordered viral interactors poised for innovation.

Of course, a major open question concerns whether the marked differences in disordered fraction of herpesvirus proteomes reflect specific viral features. An interesting possibility is that the fraction of disorder in the viral proteome contributes to determine viral host range in terms of animal species and/or cell type. In fact, although most natural herpesvirus infections are species-specific, *in vitro* experiments and accidental cross-species transmission events indicate that alphaherpesviruses have a broader host range than betaherpesviruses and gammaherpesvirsus (Spear and Longnecker 2003; Azab et al. 2018). In the case of HCMV and EBV, post-cell entry events play a major role in limiting the infection to our species, indicating that these viruses cannot efficiently hijack the cellular machinery of non-human cells to promote their replication (Fioretti et al. 1973; Lafemina and Hayward 1988; Ellsmore, Reid, and Stow 2003; Jurak and Brune 2006; Schumacher et al. 2010; Muhe and Wang 2015). Among alphaherpesvirsus, VZV, with a low fraction of IDRs, is an exception, showing high human-specificity and the ability to infect few cell types (Zerboni et al. 2014). Conversely, HSV-1 and -2 can infect a variety of cell types from different mammals (Karasneh and Shukla 2011). Indeed, *in vivo* models of HSV-1/-2 pathogenesis were developed in rodents and NHPs (Kollias et al. 2015), and the accidental transmission to Old World as well as New World monkeys was documented (Azab et al. 2018). Thus, the high fraction of disordered protein regions in simplexvirus proteomes may provide flexibility in terms of cellular binding partners, possibly affording a wider host range.

Data availability

All sequences used in this article are publicly accessible through the NCBI database. The GenBank Accession numbers of all sequences used in this article are available in Supplementary Material.

Supplementary data

Supplementary data are available at *Virus Evolution* online.

Funding

This work was supported by the Italian Ministry of Health ('Ricerca Corrente 2019–20' to M.S. and 'Ricerca Corrente 2018–20' to D.F.).

Conflict of interest: None declared.

References

Afanasyeva, A. et al. (2018) 'Human Long Intrinsically Disordered Protein Regions Are Frequent Targets of Positive Selection', *Genome Research*, 28: 975–82.

Azab, W. et al. (2018) 'How Host Specific Are Herpesviruses? Lessons from Herpesviruses Infecting Wild and Endangered Mammals', *Annual Review of Virology*, 5: 53–68.

Balfour, H. H. et al. (2013) 'Age-Specific Prevalence of Epstein-Barr Virus Infection among Individuals Aged 6-19 Years in the United States and Factors Affecting Its Acquisition', *The Journal of Infectious Diseases*, 208: 1286–93.

Basile, W. et al. (2019) 'Why Do Eukaryotic Proteins Contain More Intrinsically Disordered Regions?', *PLoS Computational Biology*, 15: e1007186.

Bosl, K. et al. (2019) 'Common Nodes of Virus-Host Interaction Revealed through an Integrated Network Analysis', *Frontiers in Immunology*, 10: 2186.

Brown, C. J., Johnson, A. K., and Daughdrill, G. W. (2010) 'Comparing Models of Evolution for Ordered and Disordered Proteins', *Molecular Biology and Evolution*, 27: 609–21.

Burrel, S. et al. (2017) 'Ancient Recombination Events between Human Herpes Simplex Viruses', *Molecular Biology and Evolution*, 34: 1713–21.

Calderone, A., Castagnoli, L., and Cesareni, G. (2013) 'Mentha: A Resource for Browsing Integrated Protein-Interaction Networks', *Nature Methods*, 10: 690–1.

——, Licata, L., and —— (2015) 'VirusMentha: A New Resource for Virus-Host Protein Interactions', *Nucleic Acids Research*, 43: D588–92.

Casto, A. M. et al. (2020) 'Large, Stable, Contemporary Interspecies Recombination Events in Circulating Human Herpes Simplex Viruses', *The Journal of Infectious Diseases*, 221: 1271–9.

Charon, J. et al. (2018) 'First Experimental Assessment of Protein Intrinsic Disorder Involvement in an RNA Virus Natural Adaptive Process', *Molecular Biology and Evolution*, 35: 38–49.

Darling, A. C. et al. (2004) 'Mauve: Multiple Alignment of Conserved Genomic Sequence with Rearrangements', *Genome Research*, 14: 1394–403.

Darling, A. E., Mau, B., and Perna, N. T. (2010) 'progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement', *PLoS One*, 5: e11147.

Davison, A. J. (2007) 'Comparative Analysis of the Genomes', in A. Arvin. (eds.), *Human Herpesviruses: Biology, Therapy, and Immunoprophylaxis*, pp. 10–26. Cambridge: Cambridge University Press.

——, Dargan, D. J., and Stow, N. D. (2002) 'Fundamental and Accessory Systems in Herpesviruses', *Antiviral Research*, 56: 1–11.

Dosztanyi, Z. (2018) 'Prediction of Protein Disorder Based on IUPred', *Protein Science*, 27: 331–40.

Dyson, H. J., and Wright, P. E. (2005) 'Intrinsically Unstructured Proteins and Their Functions', *Nature Reviews Molecular Cell Biology*, 6: 197–208.

Eberle, R., and Jones-Engel, L. (2017) 'Understanding Primate Herpesviruses', *Journal of Emerging Diseases and Virology*, 3: 10.16966/2473-1846.127.

Ellsmore, V., Reid, G. G., and Stow, N. D. (2003) 'Detection of Human Cytomegalovirus DNA Replication in Non-Permissive Vero and 293 Cells', *Journal of General Virology*, 84: 639–45.

Fioretti, A. et al. (1973) 'Nonproductive Infection of guinea Pig Cells with Human Cytomegalovirus', *Journal of Virology*, 11: 998–1003.

Gitlin, L. et al. (2014) 'Rapid Evolution of Virus Sequences in Intrinsically Disordered Protein Regions', *PLoS Pathogens*, 10: e1004529.

Jurak, I., and Brune, W. (2006) 'Induction of Apoptosis Limits Cytomegalovirus Cross-Species Infection', *The EMBO Journal*, 25: 2634–42.

Karasneh, G. A., and Shukla, D. (2011) 'Herpes Simplex Virus Infects Most Cell Types In Vitro: Clues to Its Success', *Virology Journal*, 8: 481.

Katoh, K., and Standley, D. M. (2013) 'MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability', *Molecular Biology and Evolution*, 30: 772–80.

Kollias, C. M. et al. (2015) 'Animal Models of Herpes Simplex Virus Immunity and Pathogenesis', *Journal of Neurovirology*, 21: 8–23.

Kosakovsky Pond, S. L., and Frost, S. D. (2005) 'Not So Different after All: A Comparison of Methods for Detecting Amino Acid Sites under Selection', *Molecular Biology and Evolution*, 22: 1208–22.

Kryazhimskiy, S., and Plotkin, J. B. (2008) 'The Population Genetics of dN/dS', *PLoS Genetics*, 4: e1000304.

Lafemina, R. L., and Hayward, G. S. (1988) 'Differences in Cell-Type-Specific Blocks to Immediate Early Gene Expression and DNA Replication of Human, Simian and Murine Cytomegalovirus', *Journal of General Virology*, 69: 355–74.

Lauber, C. et al. (2015) 'Interspecific Adaptation by Binary Choice at de Novo Polyomavirus T Antigen Site through Accelerated Codon-Constrained Val-Ala Toggling within an Intrinsically Disordered Region', *Nucleic Acids Research*, 43: 4800–13.

Longdon, B. et al. (2014) 'The Evolution and Genetics of Virus Host Shifts', *PLoS Pathogens*, 10: e1004395.

Lou, D. I. et al. (2016) 'An Intrinsically Disordered Region of the DNA Repair Protein Nbs1 is a Species-Specific Barrier to Herpes Simplex Virus 1 in Primates', *Cell Host & Microbe*, 20: 178–88.

Manicklal, S. et al. (2013) 'The "Silent" Global Burden of Congenital Cytomegalovirus', *Clinical Microbiology Reviews*, 26: 86–102.

McGeoch, D. J., Rixon, F. J., and Davison, A. J. (2006) 'Topics in Herpesvirus Genomics and Evolution', *Virus Research*, 117: 90–104.

Mészáros, B., Erdos, G., and Dosztányi, Z. (2018) 'IUPred2A: Context-Dependent Prediction of Protein Disorder as a Function of Redox State and Protein Binding', *Nucleic Acids Research*, 46: W329–37.

——, Simon, I., and —— (2009) 'Prediction of Protein Binding Regions in Disordered Proteins', *PLoS Computational Biology*, 5: e1000376.

Muhe, J., and Wang, F. (2015) 'Host Range Restriction of Epstein-Barr Virus and Related Lymphocryptoviruses', *Journal of Virology*, 89: 9133–6.

Nilsson, J., Grahn, M., and Wright, A. P. (2011) 'Proteome-Wide Evidence for Enhanced Positive Darwinian Selection within Intrinsically Disordered Regions in Proteins', *Genome Biology*, 12: R65.

Ortiz, J. F. et al. (2013) 'Siltberg-Liberles J. Rapid Evolutionary Dynamics of Structural Disorder as a Potential Driving Force for Biological Divergence in Flaviviruses', *Genome Biology and Evolution*, 5: 504–13.

Peng, Z. et al. (2015) 'Exceptionally Abundant Exceptions: Comprehensive Characterization of Intrinsic Disorder in All Domains of Life', *Cellular and Molecular Life Sciences*, 72: 137–51.

Privman, E., Penn, O., and Pupko, T. (2012) 'Improving the Performance of Positive Selection Inference by Filtering Unreliable Alignment Regions', *Molecular Biology and Evolution*, 29: 1–5.

Pushker, R. et al. (2013) 'Marked Variability in the Extent of Protein Disorder Within and Between Viral Families', *PLoS One*, 8: e60724.

Quach, H. et al. (2013) 'Different Selective Pressures Shape the Evolution of Toll-like Receptors in Human and African Great Ape Populations', *Human Molecular Genetics*, 22: 4829–40.

Schlessinger, A. et al. (2011) 'Protein Disorder–A Breakthrough Invention of Evolution?', *Current Opinion in Structural Biology*, 21: 412–8.

Schumacher, U. et al. (2010) 'Mutations in the M112/M113-Coding Region Facilitate Murine Cytomegalovirus Replication in Human Cells', *Journal of Virology*, 84: 7994–8006.

Sela, I. et al. (2015) 'GUIDANCE2: Accurate Detection of Unreliable Alignment Regions Accounting for the Uncertainty of Multiple Parameters', *Nucleic Acids Research*, 43: W7–14.

Severini, A. et al. (2013) 'Genome Sequence of a Chimpanzee Herpesvirus and Its Relation to Other Primate Alphaherpesviruses', *Archives of Virology*, 158: 1825–8.

Sironi, M. et al. (2015) 'Evolutionary Insights into Host-Pathogen Interactions from Mammalian Sequence Data', *Nature Reviews Genetics*, 16: 224–36.

Spear, P. G., and Longnecker, R. (2003) 'Herpesvirus Entry: An Update', *Journal of Virology*, 77: 10179–85.

Tischer, B. K., and Osterrieder, N. (2010) 'Herpesviruses—A Zoonotic Threat?', *Veterinary Microbiology*, 140: 266–70.

Toth-Petroczy, A., and Tawfik, D. S. (2011) 'Slow Protein Evolutionary Rates Are Dictated by Surface-Core Association', *Proceedings of the National Academy of Sciences*, 108: 11151–6.

Underdown, S. J., Kumar, K., and Houldcroft, C. (2017) 'Network Analysis of the Hominin Origin of Herpes Simplex Virus 2 from Fossil Data', *Virus Evolution*, 3: vex026.

Venkatesan, A. (2013) 'Advances in Infectious Encephalitis: Etiologies, Outcomes, and Potential Links with Anti-NMDAR Encephalitis', *Current Infectious Disease Reports*, 15: 594–9.

Ward, J. J. et al. (2004) 'Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life', *Journal of Molecular Biology*, 337: 635–45.

Wertheim, J. O. et al. (2014) 'Evolutionary Origins of Human Herpes Simplex Viruses 1 and 2', *Molecular Biology and Evolution*, 31: 2356–64.

Whitley, R. (2004) 'Neonatal Herpes Simplex Virus Infection', *Current Opinion in Infectious Diseases*, 17: 243–6.

Wilson, D. J. et al. (2011) 'A Population Genetics-Phylogenetics Approach to Inferring Natural Selection in Coding Sequences', *PLoS Genetics*, 7: e1002395.

Xue, B. et al. (2014) 'Structural Disorder in Viral Proteins', *Chemical Reviews*, 114: 6880–911.

Zerboni, L. et al. (2014) 'Molecular Mechanisms of Varicella Zoster Virus Pathogenesis', *Nature Reviews Microbiology*, 12: 197–210.