


RESEARCH ARTICLE

Open Access



Structural landscape of the complete genomes of dengue virus serotypes and other viral hemorrhagic fevers

Riccardo Delli Ponti* and Marek Mutwil 

Abstract

Background: With more than 300 million potentially infected people every year, and with the expanded habitat of mosquitoes due to climate change, Dengue virus (DENV) cannot be considered anymore only a tropical disease. The RNA secondary structure is a functional characteristic of RNA viruses, and together with the accumulated high-throughput sequencing data could provide general insights towards understanding virus biology. Here, we profiled the RNA secondary structure of > 7000 complete viral genomes from 11 different species focusing on viral hemorrhagic fevers, including DENV serotypes, EBOV, and YFV.

Results: In our work we demonstrated that the secondary structure and presence of protein-binding domains in the genomes can be used as intrinsic signature to further classify the viruses. With our predictive approach, we achieved high prediction scores of the secondary structure (AUC up to 0.85 with experimental data), and computed consensus secondary structure profiles using hundreds of in silico models. We observed that viruses show different structural patterns, where e.g., DENV-2 and Ebola virus tend to be less structured than the other viruses. Furthermore, we observed virus-specific correlations between secondary structure and the number of interaction sites with human proteins, reaching a correlation of 0.89 in the case of Zika virus. We also identified that helicases-encoding regions are more structured in several flaviviruses, while the regions encoding for the contact proteins exhibit virus-specific clusters in terms of RNA structure and potential protein-RNA interactions. We also used structural data to study the geographical distribution of DENV, finding a significant difference between DENV-3 from Asia and South-America, where the structure is also driving the clustering more than sequence identity, which could imply different evolutionary routes of this subtype.

Conclusions: Our massive computational analysis provided novel results regarding the secondary structure and the interaction with human proteins, not only for DENV serotypes, but also for other flaviviruses and viral hemorrhagic fevers-associated viruses. We showed how the RNA secondary structure can be used to categorise viruses, and even to further classify them based on the interaction with proteins. We envision that these approaches can be used to further classify and characterise these complex viruses.

Keywords: Genome structure, Secondary structure, Virus

* Correspondence: riccardo.ponti@ntu.edu.sg; mutwil@ntu.edu.sg
School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore 637551, Singapore



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Background

Dengue virus (DENV) is a mosquito-borne virus that can potentially infect more than 300 million people a year in more than 120 countries [1, 2]. DENV infection can further evolve into a severe hemorrhagic fever (severe dengue), which could lead to shock and death. Due to climate change, the disease is now threatening an increasing number of countries, with cases reported in Europe [2]. The existence of four different serotypes (DENV-1, 2, 3, 4), with also a fifth recently reported [3], complicates the development of an effective vaccine [4]. The four serotypes show not only significant differences in sequence similarity [5] but also distinctive infection dynamics. For example, DENV-1 is the most widespread serotype, followed by DENV-2 [6, 7], which is also more often associated with severe cases [8]. However, the mechanisms behind DENV infections and the complete set of differences between the serotypes are still unclear.

In severe cases, DENV can manifest as Viral Haemorrhagic Fever (VHF). The definition of VHF is complex since the symptoms can be mild or rare, but mainly caused by single-stranded RNA viruses from different families, such as flavivirus and filovirus [9]. However, not all the flaviviruses are associated with VHF, such as in the case of Zika virus (ZIKV), or the hemorrhagic symptoms can be secondary, for example intracranial hemorrhage causing paralysis or coma for Japanese Encephalitis virus (JEV [10]);). In general, VHFs can be extremely dangerous in humans, as in the case of Ebola virus (EBOV). Other VHFs show not only similarities to severe dengue in terms of symptoms, but also in the transmission vector. For example, Yellow Fever virus (YFV) is also a mosquito-borne VHF, with higher mortality but a slower rate of evolutionary change compared to DENV [11]. The mild VHF Chikungunya virus (CHIKV) shares the same vector with DENV, mosquito *Aedes aegypti*, and the two viruses can even coexist in the same mosquito [12]. However, while clinical and experimental analysis are the gold standard when comparing viruses, we still rely on sequence similarity approaches to understand the similarities between the thousands of available viral genomes.

The secondary structure of RNA viruses is fundamental for many viral functions, from encapsidation to egression from the cell and host defence [13–15]. Specific structures in the UTRs were found to be functional, for example, in DENV, but also in HIV and coronaviruses [13, 16, 17]. Other structural regions, including the 3' UTR, were found conserved not only in DENV serotypes but also between DENV and ZIKV [18]. Moreover, the single-stranded RNA (ss-RNA) viruses preserve their structure (folding) even if their sequence mutates rapidly [19, 20]. Thus, the folding shows the potential to be used

to classify different viral species and subspecies. 'Selective 2' Hydroxyl Acylation analyzed by Primer Extension' (SHAPE) is a chemical-probing technique that uses different chemical agents (1 M7, NAI, NMIA) to bind single-stranded RNA regions in order to experimentally profile the RNA secondary structure. The technique was successfully applied to the complete genome of different viruses, including HIV-1, DENV, and recently SARS-CoV-2 [18, 21, 22].

However, while the RNA secondary structure is an informative element to characterise viruses, the secondary structure of only a few viral genomes has been experimentally characterized [18, 21]. Consequently, while thousands of viral genomes have been sequenced, we can only rely on in silico data to study their secondary structure. Furthermore, predicting the RNA secondary structure of entire viral genomes can be challenging, due to usually large sizes of > 10,000 nucleotides (nt), where most thermodynamic algorithms used to model the secondary structure drop in performance after 700 nt [23]. In our work, we computationally profiled the RNA secondary structure of > 7000 viral genomes (prioritising DENV serotypes and in general VHFs) using Computational Recognition of Secondary Structure (CROSS), a neural network trained on experimental genome-wide secondary structure profiling, including chemical-probing data, such as SHAPE, and enzyme-based, such as 'Parallel Analysis of RNA Structure' (PARS). The algorithm was successfully applied to predict HIV genome structure [24]. Furthermore, we mapped the secondary structure properties of the viruses on the world map, to study the genome interaction with proteins, and to further classify and understand the viruses.

Results

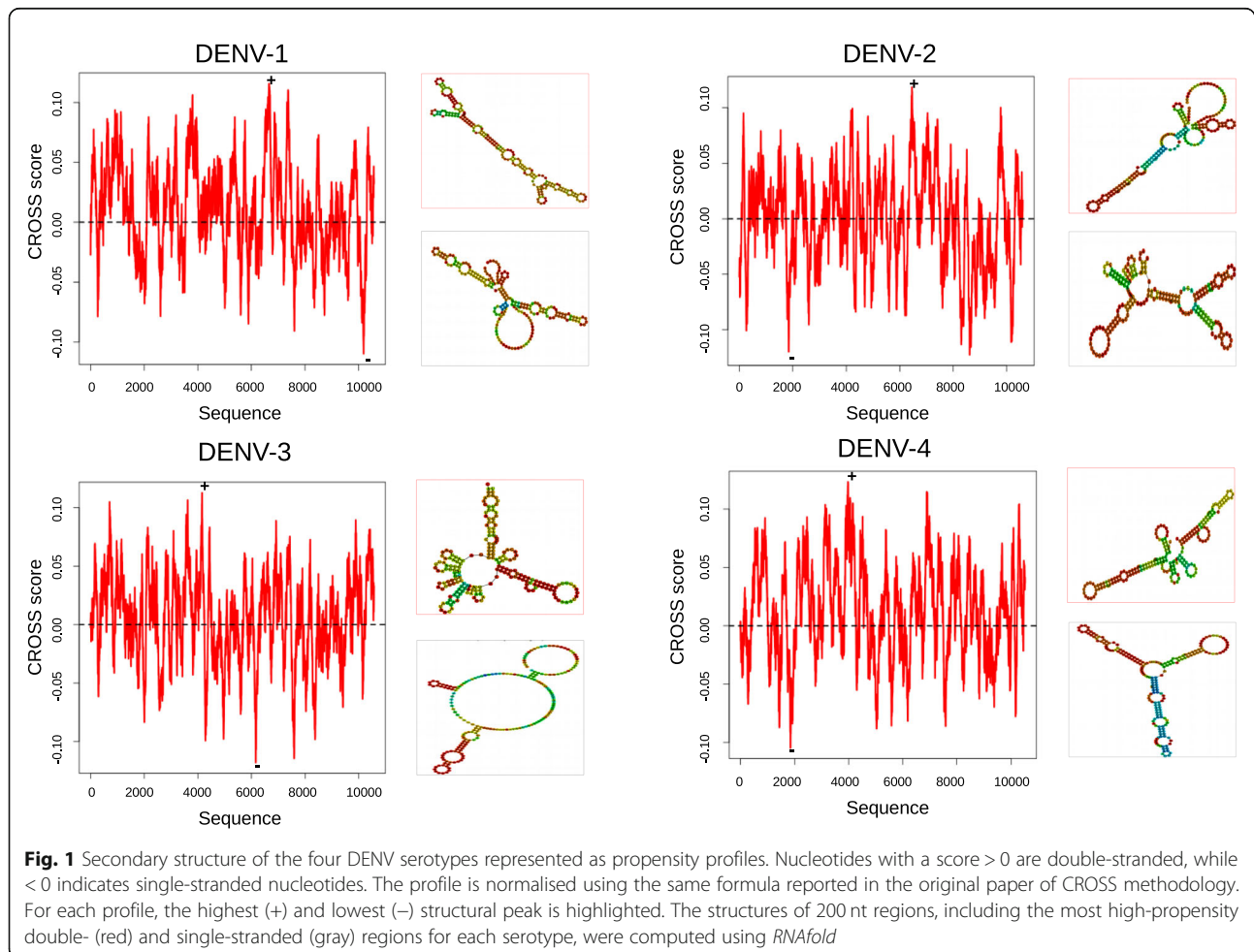
Structural properties of the DENV genomes

Here, we analysed the secondary structure profiles of the complete genomes of more than 7000 ss-RNA viruses (Table 1; Methods: Source of viral genomes). The structural profiles were generated using the CROSS algorithm, a fast and comprehensive alternative to profile the structural content (i.e., % of double-stranded nucleotides) of long and complex RNA molecules, such as viruses ([24]; see Methods: RNA secondary structure).

To analyse DENV secondary structure, we selected one strain for each serotype, focusing on strains that were widely used in previous publications [25]. In general, the four serotypes show significant differences in sequence, with around 65–70% sequence similarity [5]. Their secondary structure also shows notable differences (Fig. 1). For example, the 3' UTR of DENV-1 shows a peculiar structural valley, compared to the others. Interestingly, DENV-1 and DENV-2 share the highest structural peak around 6000 nucleotides, while DENV-3 and

Table 1 The information regarding the number of genomes available, the family, and the average nucleotide length of each family for all the viruses used in our analysis. Hemorrhagic fevers marked with "a" shows symptoms only rarely or mild, while the ones with "b" are also reported as hemorrhagic diseases by WHO

Genome	Symbol	Family	Hemorrhagic	Number of genomes	Avg Size
Dengue 1 virus	DENV-1	Flavivirus	Yes ^a (severe)	1634	10,500
Dengue 2 virus	DENV-2	Flavivirus	Yes ^a (severe)	1184	9750
Dengue 3 virus	DENV-3	Flavivirus	Yes ^a (severe)	772	10,590
Dengue 4 virus	DENV-4	Flavivirus	Yes ^a (severe)	176	9730
Zika virus	ZIKV	Flavivirus	No	258	10,120
Chikungunya virus	CHIKV	Togavirus	Yes ^a (mild)	522	10,500
Japanese encephalitis virus	JEV	Flavivirus	Yes ^a (rare cases)	279	10,500
Yellow fever virus	YFV	Flavivirus	Yes	124	8530
West Nile fever virus	WNV	Flavivirus	Yes	1528	10,390
Tick-borne encephalitis virus	TBV	Flavivirus	Yes ^b	121	9770
Ebola virus	EBOV	Filovirus	Yes ^b	530	18,200



DENV-4 also have the highest structural peak in common, but at position 4000.

We further expanded our analysis to cover the 4 different serotypes of DENV, comprising ~4000 genomes (Fig. 2; Table 1). The analysis revealed that DENV-2 and DENV-3 are less structured than DENV-1 and DENV-4.

To confirm our approach, we compared our predictions with SHAPE experiments performed on DENV genomes [18]. Using the Area Under the ROC curve (AUC) to distinguish ranked SHAPE reactivities, we obtained an AUC ranging from 0.75 to 0.85 for DENV-2 and DENV-1 (Supplementary Figure 1a, b). This further supports the power of our in silico approach, which can generate thousands of secondary structure profiles with high performances on experimental data.

Comparison of structural properties of the VHF genomes

Interestingly, DENV serotypes tend to be less structured than other flaviviruses, such as West Nile Fever virus (WNV), Yellow Fever virus (YFV), Tick-borne

Encephalitis virus (TBV), and Japanese Encephalitis virus (JEV; Fig. 2). Even if not properly a VHF, we also used as comparison Zika virus (ZIKV), due to the similarities with DENV not only in the vector (*Aedes aegypti*), but also in terms of secondary structure domains [18]. Interestingly, while TEV and ZIKV genomes are more structured (average double-stranded nucleotides >56%), WNF and JEV have a similar structural distribution, especially since they are also close in the species tree [26]. To further compare and classify the secondary structure of viral families outside of flavivirus, we also included >500 genomes of EBOV, one of the most severe VHF, and CHIKV, exhibiting only mild and rare hemorrhagic symptoms but showing similarities with DENV and ZIKV in terms of vector and spreading (Fig. 2, Table 1). The analysis revealed that the other viruses are significantly more structured than DENV (mean structural content for Flaviviruses and DENV serotypes is 0.55, 0.51, respectively; Kolmogorov-Smirnov < 2.2e-16), with the exception of EBOV, which is predicted as one of the less structured (mean structural content 0.50).

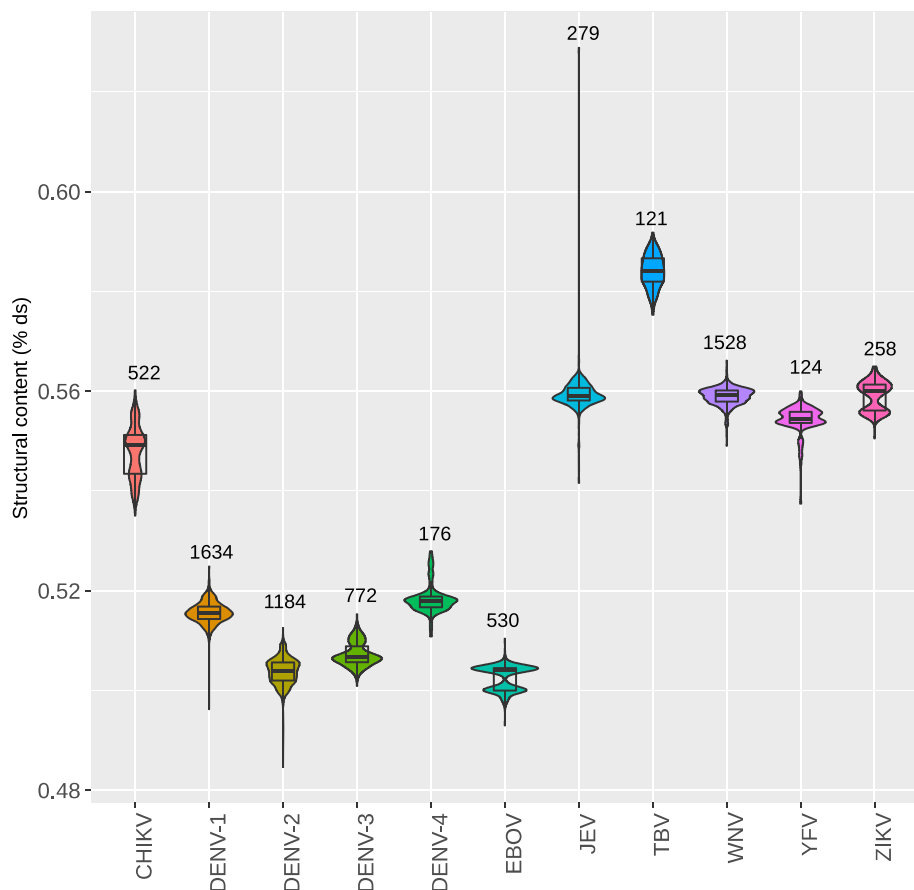


Fig. 2 Structural content (% double-stranded nucleotides) for all the genomes for the 11 species. The number above each violin plot indicates the number of genomes used in each species

Structural properties of the terminal regions including untranslated regions of VHF genomes

To further study the secondary structure content for the >7000 viral genomes, we also analysed the terminal regions including the 5' and 3' UTRs (first 1000 nt considered including the 5' UTR; last 1000 nt considered including the 3' UTR; Fig. 3a, b). Worth to specify that the terminal regions we are considering could have an overlap with coding genes, and this can go to less than 20% for TBV, up to 60% for ZIKV. DENV-3 is the only serotype with both terminal regions including UTRs more structured than the entire genome (5' UTR = 0.55 and 3' UTR = 0.53; Fig. 2), while DENV-1 has a more structured terminal region including the 5' UTR (structural content = 0.53). The results are also consistent when considering only the UTRs (from ~70 to ~700 nt depending on

the viral species; Supplementary Figure 2), highlighting a generally structured 3' UTR for the flaviviruses, as expected for the presence of complex structures [27]. This result is in line with the experimental Parallel Analysis of RNA Structure (PARS) data coming from human RNAs, where the UTRs were more structured than the CDS [28]. This suggests that some viruses tend to mimic the secondary structure of human mRNAs to be efficiently translated by the cellular machinery [29]. This is also further supported in DENV, where a complex structure at the 3' UTR was shown to mimic the absent polyA, to enhance translation [27]. Interestingly, EBOV has the least structured terminal regions including the UTRs (structural content 5' UTR = 0.46; 3' UTR = 0.41). In ZIKV, the terminal region including the 3' UTR is more structured than the 5' (Fig. 3c; 3' UTR = 0.56, 5' UTR = 0.50).

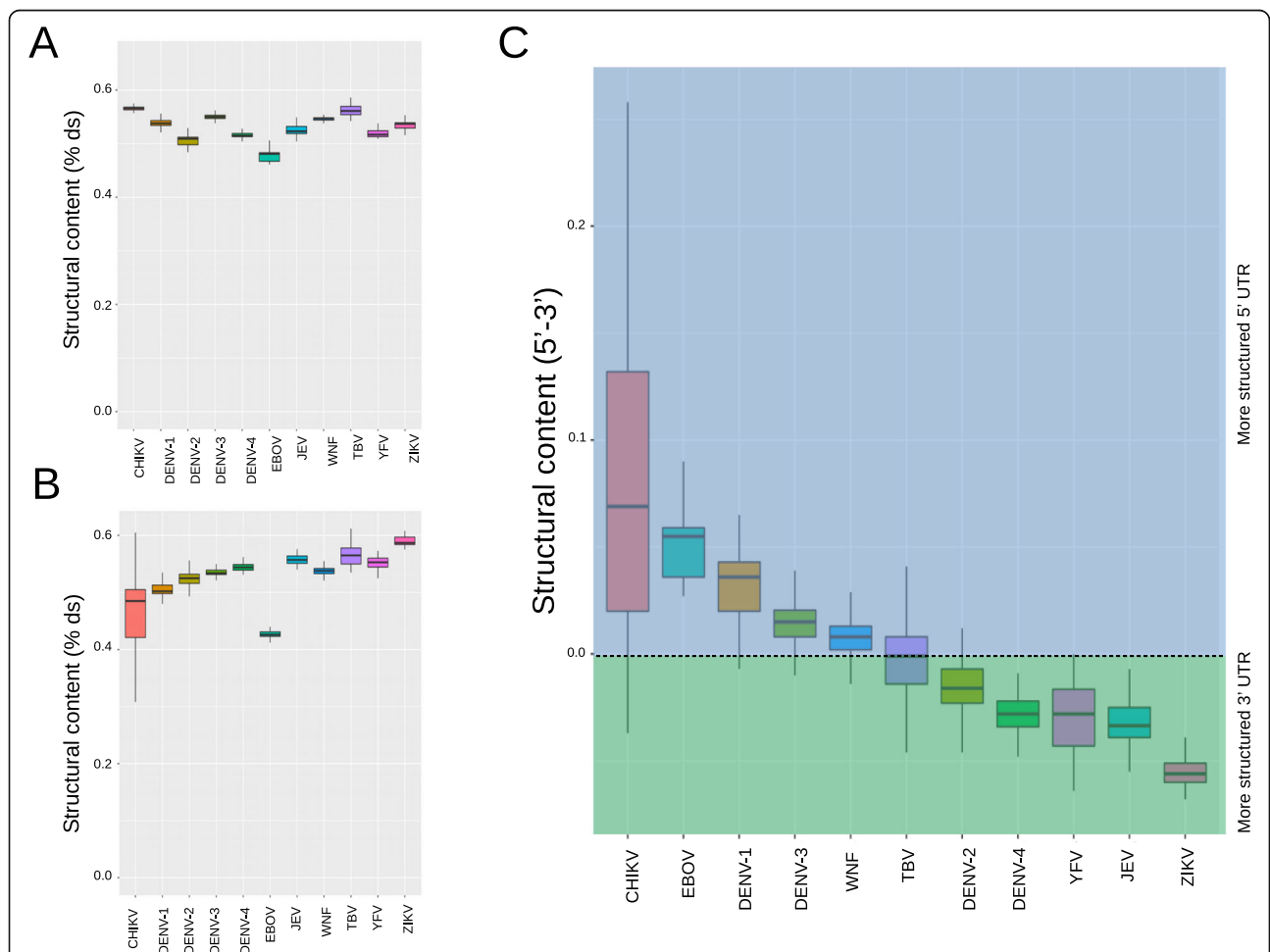


Fig. 3 Structural content of the terminal regions including the UTRs of the 11 viral species. **a** Structural content (% double-stranded nucleotides) for all the genomes for the 11 species for the terminal region including the 5' UTR. To have an equal comparison between the different species, we considered the 5' UTR included in the first 1000 nt. The name used for the viruses is reported in Table 1. **b** Structural content (% double-stranded nucleotides) for all the genomes for the 11 species for the terminal region including the 3' UTR. To have an equal comparison between the different species, we considered the 3' UTR included in the last 1000 nt. **c** The difference for each individual genome between the structural content of the terminal regions including the 5' and 3' UTR. Viruses with more structured terminal region including the 5' UTR are >0 (blue area), while <0 indicates more structured 3' UTRs (green area)

CHIKV shows not only the highest structural variability in the terminal region including the 3' UTR (standard deviation 3' UTR = 0.16, Fig. 3b), with a more structured region at 5' (3' UTR = 0.43, 5' UTR = 0.51; Fig. 3a). Finally, EBOV, DENV-1, and DENV-3 exhibit a more structured terminal region including the 5' UTR, especially when compared with DENV-2, DENV-4 and JEV, which tend to be more structured (Fig. 3c).

Structural content can be used to classify VHFs

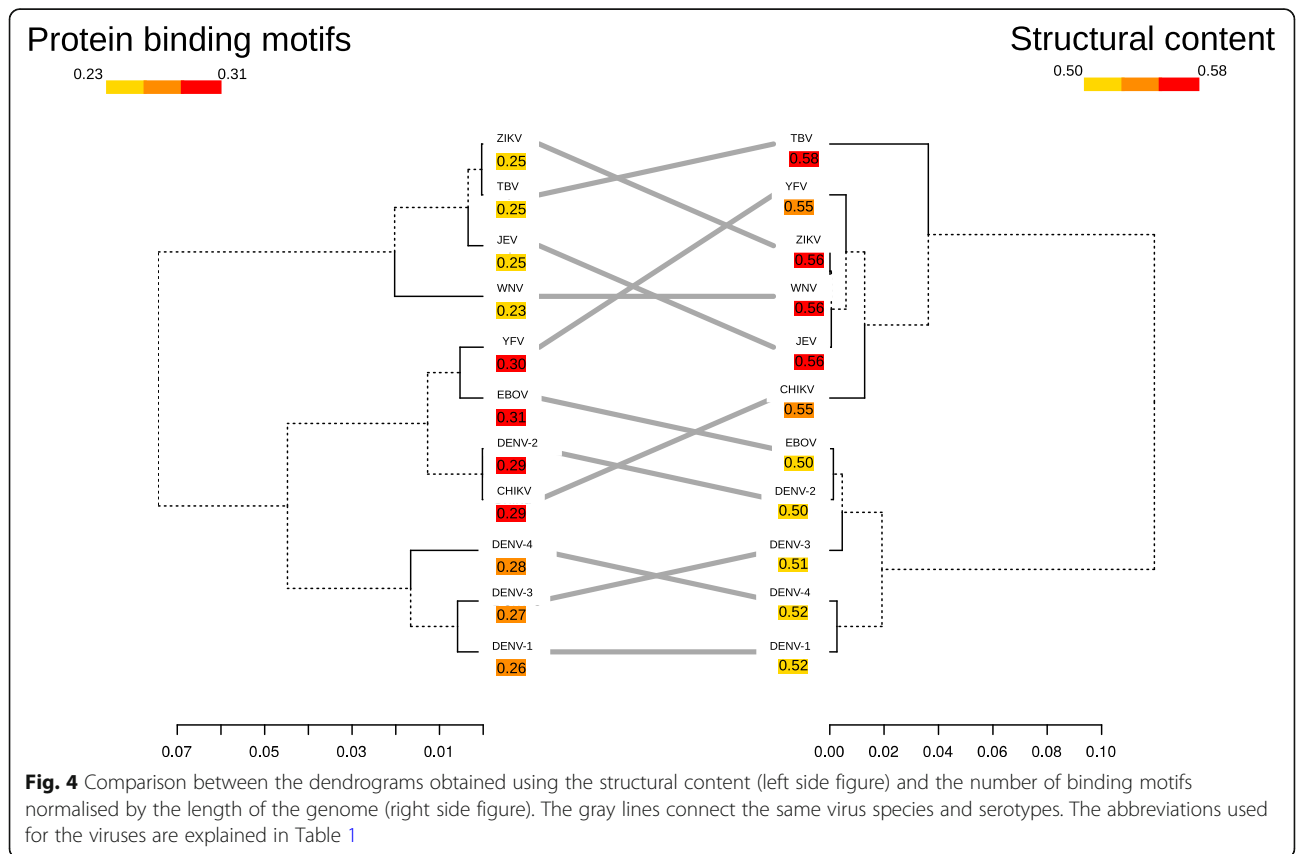
The overall similarities and differences in structure are an additional feature that could be employed to characterise the different viruses. For the next step, the structural content (mean of the % of double-stranded nucleotides for all the viral genomes) in a specific species was used to hierarchically group the 11 different viruses (Methods: Hierarchical clustering; Table 1). The resulting dendrogram clustered the DENV serotypes, showing that they are more structurally similar compared to other viruses (Fig. 4). The structural similarities of DENV serotypes, together with the similarity between WNV and JEV, are in agreement with the Phylogenetic Tree of Viral Hemorrhagic Fevers [26].

The structural content revealed interesting clustering of the viruses. For example, while having a different genome sequence, DENV also clusters together with EBOV,

since they share a less structured genome. According to its structural content, ZIKV is also part of the sub-cluster, together with WNV and JEV. The mosquito-transmitted YFV and CHIKV form a cluster, indicating that their structural content is similar. This is partially in agreement with the VHFs tree [26]. Interestingly, since TBV is more structured than any of these viruses, it forms an outlier. This is not a surprise, since it was previously shown that the secondary structure of mosquito- and tick-borne flaviviruses are more different, especially in the 3' UTR [30]. To conclude, these results indicate that the level of secondary structure inside a viral genome can be used as a metric to build a tree of similarities, which could be further employed to classify viruses.

Interaction between viral genomes and human host proteins can be used to classify VHFs

During translation and replication, ss-RNA viruses are naked RNA molecules inside human host cells. Previous studies already showed that genomes of the DENV interact with multiple human proteins during the infection and that the protein binding can enhance or inhibit the virulence [31]. Furthermore, RNA binding proteins tend to exhibit an altered activity during viral infection, in some cases due to the presence of highly abundant viral RNA,



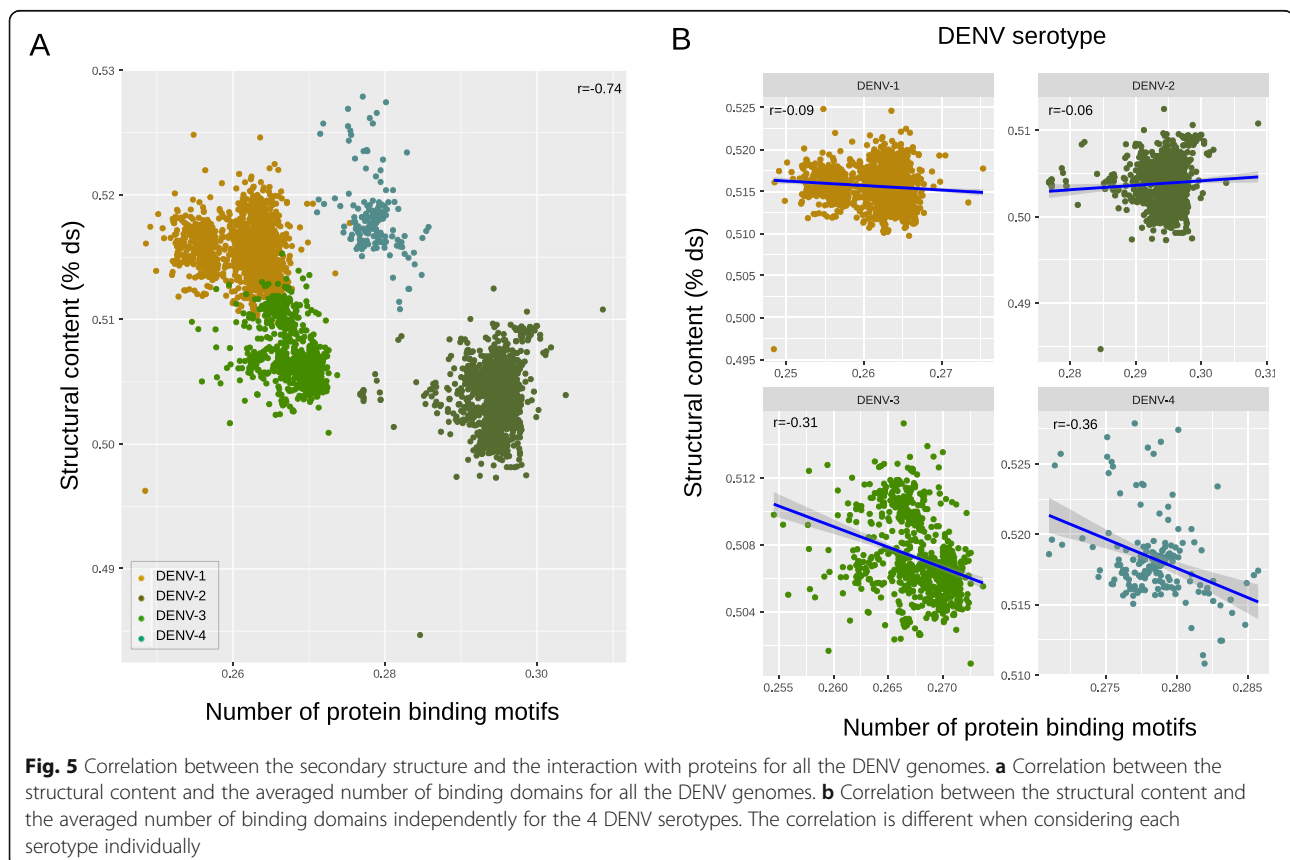
which can compete for the interaction with cellular RNA [32]. To study the relationship between human proteins and the viral RNA structures, we selected binding motifs from RNA Bind-n-Seq (RBNS) data of 78 human RNA-binding proteins [33], and searched the complete viral genomes for these motifs (Methods: Protein-RNA interactions). We observed that the 4 DENV serotypes have a different presence of protein binding domains, with DENV-2 showing the highest number of motifs, followed by DENV-4 (Supplementary Figure 3).

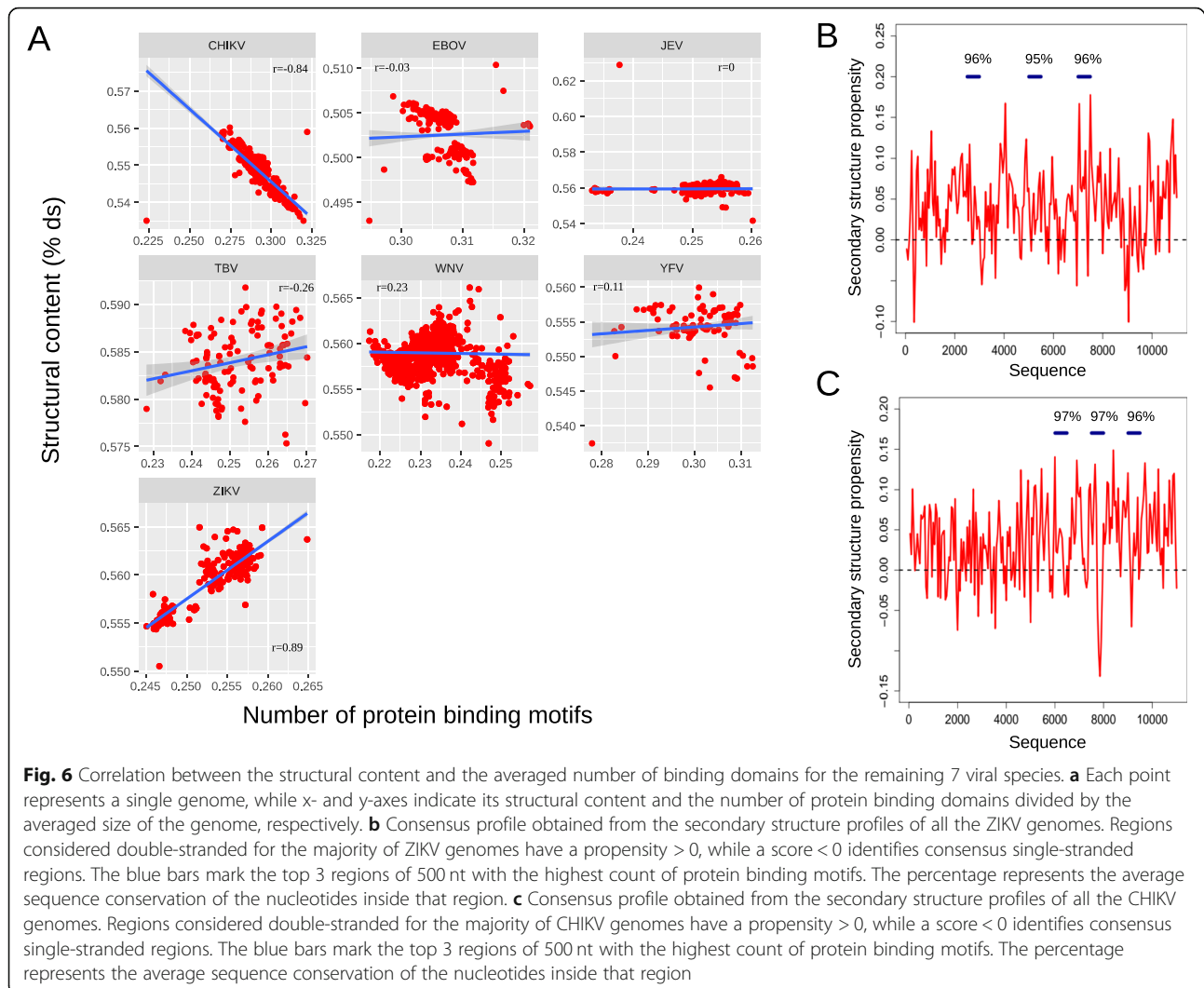
Similarly to the structural content analysis above, we used the number of protein binding domains to classify the viruses. To further understand how the connection between structure and interaction with proteins can classify viruses, we compared the resulting trees (Fig. 4). Interestingly, the DENV cluster is almost perfectly maintained, except that, for the number of protein interactions, DENV-2 is more similar to CHIKV than EBOV, which in turn is more related to YFV. Furthermore, clustering of WNV, JEV and ZIKV is partially maintained when using structure and interaction with proteins. To conclude, by analyzing thousands of different viral genomes, we identified specific clusters both in terms of secondary structure content and the potential number of interactions with proteins.

Relationship between the structural content and number of protein interaction motifs

Since both the structural content and the number of binding motifs can be used to classify the viruses, we hypothesized that there is a correlation between these two features. We found an overall high anti-correlation ($r = -0.74$; p -value $< 2.2e-16$) between the number of protein-binding motifs and the structural content in DENV, meaning that less structured DENV genomes tend to bind more proteins (Fig. 5a). Interestingly, the different DENV serotypes cluster together according to their structure and the interaction with proteins (Fig. 5a). Also, the serotypes show a different trend when independently analysed, with DENV-3 and DENV-4 exhibiting the highest influence of the structure on the number of possible interacting proteins (r is -0.31 and -0.36 ; p -value $< 2.2e-16$ and $7.177e-07$, respectively; Fig. 5b).

Next, we compared the secondary structure and protein binding motifs of the other VHFs. The general picture is quite complex, with some viruses showing opposite trends between structure and interaction with proteins, providing a characteristic signature to further classify viruses into three categories. First, similarly to DENV, the mild hemorrhagic fever CHIKV shows a high anticorrelation ($r = -0.84$; p -value $< 2.2e-16$, Fig. 6a),





identifying a category of less structured but highly interactive viruses. DENV-3 and DENV-4 act similarly to CHIKV, having less structured genomes but highly interactive with proteins (Fig. 5b). Second, TBV and ZIKV show a positive correlation (Pearson correlation of 0.23 and 0.89; p-value 0.01 and < 2.2e-16 respectively; Fig. 6), identifying a category of high-structured viruses that are also highly interactive with host-proteins. Third, a class composed of JEV, WNV, and EBOV show almost no correlation between the structure and the possible interaction with proteins ($r \sim 0$). These results indicate a complex relationship between the structural content and protein-binding motifs, which can classify the viruses in 3 different categories: highly-structured and highly-interactive, poorly-structured and highly-interactive, or without a strong relationship between the overall structural content and interaction with proteins.

To further understand the different behaviour of ZIKV and CHIKV in terms of secondary structure and

interaction with proteins, we generated a consensus-profile between the hundreds of genomes available (Methods; Fig. 6b, c). We also highlighted for each species the 3 regions in windows of 500 nt which have the highest-presence of protein binding motifs (Fig. 6b, c; blue bars). Interestingly, this analysis validates our overall observation, where the most contacted regions in ZIKV are very structured in the consensus profile obtained from all ZIKV genomes (Fig. 6b). Moreover, the most structured region from ZIKV consensus profile and one of the most highly-contacted by proteins is the one encoding for the nonstructural protein 3 (NS3), a helicase essential for viral replication [34]. We speculate that this region needs a very specific structure in order to be highly regulated by proteins. Conversely, the most contacted regions for CHIKV consensus secondary structure profile fall into highly unstructured regions (Fig. 6c). The least structured region of CHIKV consensus secondary structure profile, and one of the most

regulated by proteins, encodes for the structural protein E3 [35]. Regardless of the structural propensity, the most interactive regions are also highly conserved in sequence both in ZIKV (~ 96% conservation) and slightly more in CHIKV (~ 97%) (Fig. 6b, c; Methods: Sequence similarity).

To extend and further validate this result, we checked the interactions between the most structured ZIKV and CHIKV viruses and > 1000 human RBPs. After selecting the 10 most and least structural genomes of ZIKV and CHIKV, we used the catRAPID algorithm [36] to predict $> 4 \times 10^6$ protein interactions between the genomes and human proteins (Methods: Protein-RNA interactions). Interestingly, the highly-structured ZIKV has stronger and more frequent interactions with proteins, reaching $\times 10$ more strong interactions with proteins, compared to CHIKV (Supplementary Table 1). This result supports our hypothesis that ZIKV interacts with human proteins mainly using double-stranded regions when compared to CHIKV.

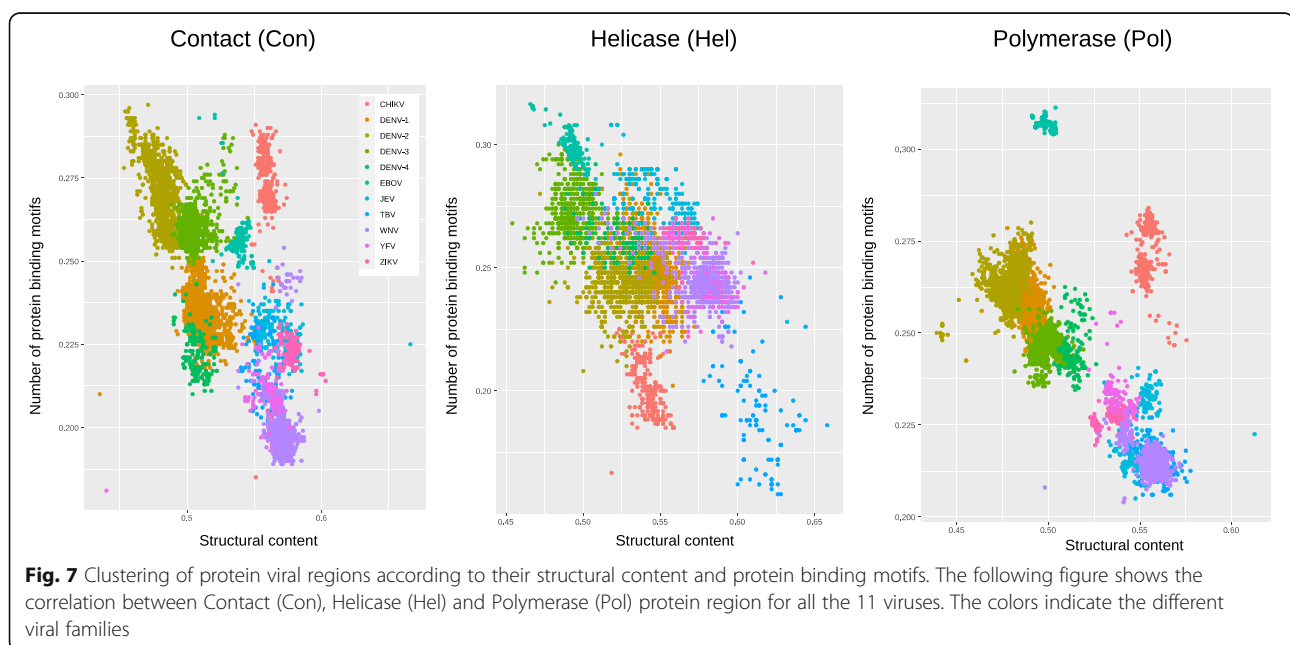
Structural and protein interaction analysis of functionally related protein regions

We selected using NCBI annotations genomic regions encoding for proteins with a related function between the different viral families [37–39]. We selected the regions encoding for polymerases (Pol), helicases (Hel), and the structural protein hypothetically responsible for entering the human cells, which we called contact (Con), for the flaviviruses, filoviruses (EBOV), and togaviruses (CHIKV). The average structural content in the aforementioned regions highlight an interesting further classification (Supplementary Figure 4a). For example, the

region associated with helicases (Hel) is more structured especially in flaviviruses, including DENV serotypes (DENV-1, DENV-2, DENV-4, WNV, TBV, YFV; Kolmogorov-Smirnov $< 2.2e-16$). Interestingly, the helicase region was also the most contacted region for ZIKV (Fig. 6b). Conversely, EBOV shows a significantly more structured Contact region (Con; Kolmogorov-Smirnov $< 2.2e-16$), associated to the GP protein, which plays a critical role in the host-cell entry process. To conclude, the structural content can further categorise viruses, as different genetic regions show a characteristic structural signature.

We further studied the amount of binding motifs and the structural content inside the regions corresponding to polymerase (Pol), helicase (Hel), and contact (Con) for each virus (Supplementary Figure 4b). Interestingly, the regions show an even further pattern that can be used for viral classification. The Hel region is highly-interactive with proteins in several flaviviruses (DENV-3, DENV-4, JEV, YFV, ZIKV; Kolmogorov-Smirnov $< 2.2e-16$). CHIKV is again an outlier, with a less interactive Hel region compared with Pol and Con.

An even stronger pattern emerges when clustering the viruses according to both interaction with proteins and secondary structure content (Fig. 7). The Pol region shows more aggregated clusters, especially for JEV, YFV, TBV, and WNV, while DENV serotypes, CHIKV and EBOV are the only one with a more defined cluster. This is in agreement with previous references, explaining how the polymerase tends to be more conserved in different viruses [38], following in our case a similar pattern of structure and regulation. While we did not observe clear clusters for the Hel region, probably due to the different



protein families involved in this activity [37], the Con region shows more defined clusters for almost every virus.

Geographical distribution of DENV serotypes

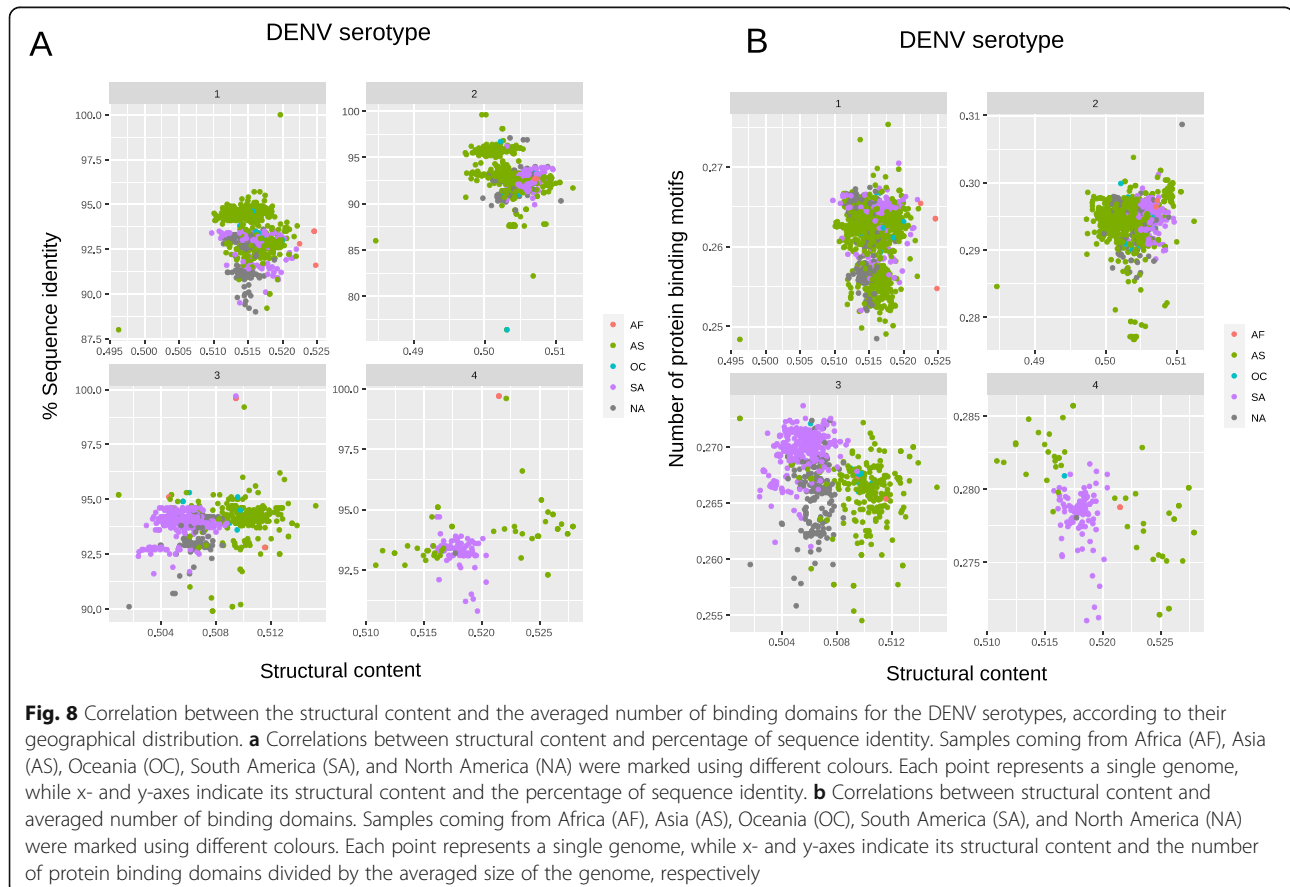
We also studied the geographical connection between DENV serotypes and the secondary structure content. We found that the African DENV-1 is predicted to be more structured than the other serotypes, even when compared with the Asian strains (Supplementary Figure 5 Kolmogorov-Smirnov = 0.001), while on the contrary, Asian DENV-3 is more structured than the African strains (Kolmogorov-Smirnov = 0.07; Supplementary Figure 5). The structural content was also compared with pairwise sequence identity (Methods: Sequence Similarity) in order to identify the driving signal (Fig. 8a). Interestingly, DENV-3 is the one showing a more neat clustering of the Asian (AS) and South American (SA) strains. However, it is interesting to notice how the structural content is the feature driving the clustering, with the Asian and South American strain mainly lying on a sequence identity of $\sim 93.5\%$.

Moreover, the correlation between binding motifs and secondary structure at the geographical level portrays a quite complex scenario (Fig. 8b). The clearest clusters are again identifiable for DENV-3, where the South-

American strains are not only less structured, but also highly interacting with proteins. DENV-3 from South America is also more distant from the Asian strains (Euclidean distance centroids $\times 1000 = 5$), compared, for example, with DENV-1 (Euclidean distance centroids $\times 1000 = 1$). Conversely, when comparing the protein binding motifs and sequence identity, the AS-SA cluster is disrupted, showing how the secondary structure information was essential for the identification of this cluster (Supplementary Figure 6). The analysis supports the hypothesis that secondary structure can be used to classify viruses and to discover possible differences identifiable or driven at geographic level.

Discussion

The genomes of viral hemorrhagic fevers show different levels of secondary structure, especially structured in the UTR regions [16, 18]. This secondary structure is thought to be needed for different viral mechanisms, such as packaging and egression [13–15]. However, a comprehensive secondary structural landscape for their genomes was lacking. In our work, we computationally profiled and analysed the secondary structure profiles of more than 7000 complete viral genomes, including almost 4000 DENV samples, and 3500 other viral



hemorrhagic fever-causing viruses. By studying the structural profiles, we observed that DENV is predicted to be less structured compared to viruses such as ZIKV, YFV, and WNV (Fig. 2). Conversely, DENV serotypes still tend to retain structured terminal regions including the UTRs, probably to be efficiently translated by the cellular machinery, similarly to human mRNAs [28, 40]. Single-stranded regions could be necessary to confer flexibility to the viral genomes, since flaviviruses need a high level of structural plasticity to undergo conformational changes during their life cycle, including circularization [41].

We also identified a correlation between the secondary structure and the number of protein binding domains, implying that the secondary structure is employed to regulate potential binding with proteins, as observed for human RNAs [42]. For viruses, the situation is more complex, as we observed a significantly positive (TBV and ZIKV) and negative (CHIKV, DENV-3, and DENV-4) relationship between secondary structure and the potential interaction with proteins. For example, ZIKV (positive RNA genome, positive correlation between secondary structure and interaction with proteins) and CHIKV (negative genome, negative correlation) belong to different viral families (Flavivirus and Togavirus respectively), and have a completely different capsid. Also while ZIKV shows a sequence similarity of 56% with YFV and JEV, it shows only 1.3% sequence identity with CHIKV [43]. Further investigation of additional viral families, especially togaviruses, are needed to elucidate the mechanism behind these opposing patterns.

Region encoding for viral helicases tend to be more structured in flaviviruses, while for other families, such as EBOV, contact protein regions show higher structural content. Moreover, region encoding for polymerases tend to be similarly structured across viruses and are potentially more bound by host-proteins, while the regions encoding for contact proteins often show species-specific clusters in terms of protein interactions and structure.

We also analysed the secondary structure at a geographical level, showing that DENV-3 strains from South-America and Asia have different patterns in their structure and the potential interaction with proteins, especially when compared with DENV-1 and DENV-2 (Fig. 8), and that the structure is a crucial elements for the identification of this cluster. Interestingly, this is in line with DENV-3 being the youngest serotype and the only one with a proposed origin not in Asia but in America [44]. This could explain the niched behaviour of DENV-3 in terms of structure and protein interactions, as well as supporting a possible independent origin of Asia and American DENV-3. Interestingly, DENV-4 shows a similar trend, but there are too few samples available to explain its evolution.

Conclusions

In our study we employed secondary structural content and the presence of protein-binding domains to build similarity trees between VHF in order to further characterise the viruses. The secondary structure and interaction with proteins can be used to cluster the viruses in agreement with previous phylogenetic trees, such as DENV serotypes, and JEV with WNV. Conversely, some relationships are surprising, as for example, DENV-2 is closer to EBOV when the secondary structure is used to establish similarity, but not when using the interaction with proteins. This result suggests how different measures, especially the secondary structure content, could be used to further classify and characterise different classes of viruses.

The inclusion of additional viral families or species, especially togaviruses and filoviruses, could further improve the analysis, providing even more data to explain some of the characteristic trends that we identified. Future experimental evidence, especially additional SHAPE profiles or cross-linking studies for the RBPs, will also help to extend and validate part of our analysis, as well as providing useful data to the scientific community.

Our massive computational analysis provided novel results regarding the secondary structure and the interaction with human proteins, not only for DENV serotypes but also for other viral hemorrhagic fevers. We envision that these approaches can be used by the scientific community to classify further and characterise these complex viruses.

Methods

Source of viral genomes

The viral genomes were downloaded from NCBI, selecting for each specific virus the fasta sequences containing the keyword *complete genome*. NCBI data was also used to extract the geographical information of DENV viruses as well as the serotypes. Fragmented or incomplete genomes were also removed. We also selected only complete genomes with only standard-nucleotides, filtering out sequences containing unknown (“N”) or degenerate (for example “R” or “Y”) nucleotides (Supplementary File 1).

RNA secondary structure

The secondary structure profiles were computed using the CROSS (Computational Recognition of Secondary Structure) algorithm. CROSS is a neural network-based machine learning approach trained on experimental data (SHAPE, PARS, NMR/X-Ray, and icSHAPE), able to quickly profile large and complex molecules such as viral genomes without length restrictions. CROSS was already used to profile the complete HIV genome, also showing an AUC of 0.75 with experimental ‘Selective 2’ Hydroxyl Acylation analyzed by Primer Extension’ (SHAPE) data

[24] and recently an AUC of 0.73 on SHAPE data for SARS-CoV-2 [45]. In the original manuscript, CROSS was also tested on crystallographic structures (AUC 0.72, PPV 0.74) and on DMS data for murine *XIST* (AUC 0.75) [24]. The algorithm was also updated and trained on structural in vivo data both with and without RNA methylation, and tested on the complete murine *XIST* [46]. For a comprehensive analysis, we used the *Global Score* model, considering nucleotides with a score > 0 as double-stranded, and < 0 as single-stranded. For the purpose of computing the structural content of a complete genome (i.e., % double-stranded nucleotides), the total number of nucleotides with a score > 0 was averaged for the total length of each genome. For the plots showing the complete secondary structure of DENV regions, we used the MFE structure computed using RNA-fold [47].

Protein-RNA interactions

To analyse the protein binding motifs in the viral genomes, we selected 5-mer motifs from the Table S3 of Dominguez et al. [33]. All of the 520 possible redundant motifs (270 non-redundant) were selected for further analysis. We scanned for the motifs on the complete genomes of the different strains, selecting only perfect matches. The number of motifs normalized by the average genome length of the different species of viruses was used to define a score for the number of potential interactions with proteins, according to the following formula:

$$\frac{m}{\text{avg}(n)}$$

where m is the number of exact motifs found in a genome, and $\text{avg}(n)$ is the average length of the genome of a specific species.

We also computed high-throughput predictions against the human proteome using the catRAPID Omics algorithm [36], which estimates the binding propensity between proteins and RNA by combining secondary structure, hydrogen bonding and van der Waals contributions. The algorithms computed more than 2 millions interactions between viral genomes and human proteins. The Discriminative Power (DP) was used to progressively filter for strong interactions. The Discriminative Power (DP) ranges from 0 to 1, where DP values above 0.5 indicate that the interaction is likely to take place.

Hierarchical clustering

The structural content and the averaged number of binding domains were employed to build dendrograms. To this end, we computed the Euclidean distance between the values associated with each virus using

statistical software R. We then used the *hclust* function based on a *ward.D2* module to build the dendrograms according to the hierarchical clustering.

Secondary structure consensus profile

We used hundreds of profiles generated using CROSS to build secondary structure consensus profiles for ZIKV and CHIKV. To build the consensus profiles we selected non-overlapping windows of 50 nucleotides across all the genomes of each species, and we averaged CROSS propensity scores for each window in all the genomes available. To avoid problems due to the different lengths of the genomes, we limited the sliding window till the average length of the genomes of that specific species (Table 1). Regions with a score > 0 are double-stranded regions in agreement in all the genomes of a species, while < 0 for single-stranded consensus regions.

Sequence similarity

To build pairwise sequence identities between the complete viral genomes we used the command line version of EMBOSS needle [48]. For a fast calculation we used a reference sequence for each dataset (same as used in Fig. 1), to align with all the sequences inside that specific dataset (for example all DENV-1 genomes). The algorithm was launched using standard parameters. The percentage of sequence identity was then used to identify the similarities in terms of primary structure (i.e. sequence) between the viruses. To extract conserved regions, we used a novel version of MAFFT [49] developed in April 2020, specifically built to perform multiple-alignments of huge viral genomes, such as the one of SARS-CoV-2. For each position in the alignment, we selected the nucleotide most present in that position, and we assigned the nucleotide and the associated percentage to a consensus profile. If a gap is identified as the most conserved, that position will have a value of 0.

Abbreviations

CROSS: Computational Recognition of Secondary Structure; SHAPE: Selective 2' Hydroxyl Acylation analyzed by Primer Extension; PARS: Parallel Analysis of RNA Structure; VHF: Viral Haemorrhagic Fevers; DENV: Dengue virus; ZIKV: Zika virus; CHIKV: Chikungunya virus; WNV: West Nile Fever virus; JEV: Japanese Encephalitis virus; TBV: Tick-borne Encephalitis virus; EBOV: Ebola virus; YFV: Yellow Fever virus; RBP: RNA binding protein; UTR: Untranslated region; CDS: Coding sequence; AUC: Area under the ROC curve; ss-RNA: Single-stranded RNA; MFE: Minimum free energy

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12864-021-07638-7>.

Additional file 1: Supplementary Figure 1. ROC curves of our predictions obtained using CROSS and experimental SHAPE data on DENV-2 (A) and DENV-1 (B). SHAPE data were ranked according to their reactivity, and the 5, 10 and 25% top/bottom nucleotides were selected. The AUC increases from 0.75 (25% top/bottom ranked SHAPE data; i.e.

half of the dataset) to 0.85 (5% top/bottom ranked SHAPE data). Uncharacterised SHAPE reactivities < 0 were removed from the ranking. **Supplementary Figure 2.** Boxplot showing the structural content, as made for Fig. 3, but specifically selecting the 5' and 3' UTR. **Supplementary Figure 3.** Violin plot showing the interaction with proteins for each DENV serotype, computed as the presence of RNA binding motifs on their genome, averaged for the mean of the length of each serotype. **Supplementary Figure 4.** Boxplots showing for each virus how the regions coding for helicases (Hel), polymerases (Pol), and contact protein (Con) are different in terms of (A) structural content and (B) number of binding motifs. **Supplementary Figure 5.** Barplot showing for each DENV serotype the differences in structural content (% double-stranded nucleotides) in different geographical samples coming from Africa (AF), Asia (AS), Oceania (OC), South America (SA) and North America (NA). **Supplementary Figure 6.** Correlations between percentage of sequence identity and averaged number of binding domains. Samples coming from Africa (AF), Asia (AS), Oceania (OC), South America (SA), and North America (NA) were marked using different colours. Each point represents a single genome, while y- and x-axes indicate the percentage of sequence identity and the number of protein binding domains divided by the averaged size of the genome, respectively. **Supplementary Table 1.** Number of predicted interactions of all the human proteome and the 10 most structured ZIKV and CHIKV genomes. An increasing threshold on the Discriminative Power (DP) of catRAPID algorithm was used to iteratively select stronger interactions.

Additional file 2.

Acknowledgments

The authors thank the other members of Mutwil's group for useful comments.

Authors' contributions

RDP conceived the study. MM and RDP designed the study. RDP performed the analysis. RDP and MM wrote the manuscript. The author(s) read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

All the data we used were available in NCBI (<https://www.ncbi.nlm.nih.gov/>) and downloaded as specified in the Methods on January/February 2020. Additional information and identifiers available in the Supplementary Files.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 4 June 2020 Accepted: 21 April 2021

Published online: 17 May 2021

References

- Bhatt S, Gething PW, Brady OJ, Messina JP, Farlow AW, Moyes CL, et al. The global distribution and burden of dengue. *Nature*. 2013;496(7446):504–7. <https://doi.org/10.1038/nature12060>.
- Brady OJ, Gething PW, Bhatt S, Messina JP, Brownstein JS, Hoen AG, et al. Refining the global spatial limits of dengue virus transmission by evidence-based consensus. *PLoS Negl Trop Dis*. 2012;6(8):e1760. <https://doi.org/10.1371/journal.pntd.0001760>.
- Normile D. Tropical medicine. Surprising new dengue virus throws a spanner in disease control efforts. *Science*. 2013;342(6157):415. <https://doi.org/10.1126/science.1242615>.
- The Lancet Infectious Diseases. The dengue vaccine dilemma. *Lancet Infect Dis*. 2018;18:123. [https://doi.org/10.1016/S1473-3099\(18\)30023-9](https://doi.org/10.1016/S1473-3099(18)30023-9).
- Anoop M, Mathew AJ, Jayakumar B, Issac A, Nair S, Abraham R, et al. Complete genome sequencing and evolutionary analysis of dengue virus serotype 1 isolates from an outbreak in Kerala, South India. *Virus Genes*. 2012;45(1):1–13. <https://doi.org/10.1007/s11262-012-0756-3>.
- Messina JP, Brady OJ, Scott TW, Zou C, Pigott DM, Duda KA, et al. Global spread of dengue virus types: mapping the 70 year history. *Trends Microbiol*. 2014;22(3):138–46. <https://doi.org/10.1016/j.tim.2013.12.011>.
- Bäck AT, Lundkvist A. Dengue viruses - an overview. *Infect Ecol Epidemiol*. 2013;3(1). <https://doi.org/10.3402/iee.v3i1.19839>.
- Fried JR, Gibbons RV, Kalayanarooj S, Thomas SJ, Srikiatkachorn A, Yoon I-H, et al. Serotype-specific differences in the risk of dengue hemorrhagic fever: an analysis of data collected in Bangkok, Thailand from 1994 to 2006. *PLoS Negl Trop Dis*. 2010;4(3):e617. <https://doi.org/10.1371/journal.pntd.0000617>.
- Cobo F. Viruses causing hemorrhagic fever. Safety laboratory procedures. *Open Virol J*. 2016;10(1):1–9. <https://doi.org/10.2174/1874357901610010001>.
- Feng Q, Chen Q, Bi X, Yu S, Wang J, Sun X, et al. Severe Japanese encephalitis with multiple intracranial hemorrhages: a case report. *Medicine (Baltimore)*. 2019;98:e17453. <https://doi.org/10.1097/MD.00000000000017453>.
- Sall AA, Faye O, Diallo M, Firth C, Kitchen A, Holmes EC. Yellow fever virus exhibits slower evolutionary dynamics than dengue virus. *J Virol*. 2010;84(2):765–72. <https://doi.org/10.1128/JVI.01738-09>.
- Leroy EM, Nkoghe D, Ollomo B, Nze-Nkoghe C, Becquart P, Grand G, et al. Concurrent chikungunya and dengue virus infections during simultaneous outbreaks, Gabon, 2007. *Emerging Infect Dis*. 2009;15(4):591–3. <https://doi.org/10.3201/eid1504.080664>.
- Lu K, Heng X, Summers MF. Structural determinants and mechanism of HIV-1 genome packaging. *J Mol Biol*. 2011;410(4):609–33. <https://doi.org/10.1016/j.jmb.2011.04.029>.
- Stewart H, Bingham RJ, White SJ, Dykeman EC, Zothner C, Tuplin AK, et al. Identification of novel RNA secondary structures within the hepatitis C virus genome reveals a cooperative involvement in genome packaging. *Sci Rep*. 2016;6(1):22952. <https://doi.org/10.1038/srep22952>.
- Witteveldt J, Blundell R, Maarleveld JJ, McFadden N, Evans DJ, Simmonds P. The influence of viral RNA secondary structure on interactions with innate host cell defences. *Nucleic Acids Res*. 2014;42(5):3314–29. <https://doi.org/10.1093/nar/gkt1291>.
- de Borja L, Villordo SM, Marsico FL, Carballada JM, Filomatori CV, Gebhard LG, et al. RNA structure duplication in the dengue virus 3' UTR: redundancy or host specificity? *MBio*. 2019;10(1). <https://doi.org/10.1128/mBio.02506-18>.
- Yang D, Leibowitz JL. The structure and functions of coronavirus genomic 3' and 5' ends. *Virus Res*. 2015;206:120–33. <https://doi.org/10.1016/j.virusres.2015.02.025>.
- Huber RG, Lim XN, Ng WC, Sim AYL, Poh HX, Shen Y, et al. Structure mapping of dengue and Zika viruses reveals functional long-range interactions. *Nat Commun*. 2019;10(1):1408. <https://doi.org/10.1038/s41467-019-09391-8>.
- Chursov A, Frishman D, Shneider A. Conservation of mRNA secondary structures may filter out mutations in *Escherichia coli* evolution. *Nucleic Acids Res*. 2013;41(16):7854–60. <https://doi.org/10.1093/nar/gkt507>.
- Rizvi TA, Panganiban AT. Simian immunodeficiency virus RNA is efficiently encapsidated by human immunodeficiency virus type 1 particles. *J Virol*. 1993;67(5):2681–8. <https://doi.org/10.1128/JVI.67.5.2681-2688.1993>.
- Watts JM, Dang KK, Gorelick RJ, Leonard CW, Bess JW, Swanstrom R, et al. Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature*. 2009;460(7256):711–6. <https://doi.org/10.1038/nature08237>.
- Manfredonia I, Nithin C, Ponce-Salvatierra A, Ghosh P, Wirecki TK, Marinus T, et al. Genome-wide mapping of therapeutically-relevant SARS-CoV-2 RNA structures. *BioRxiv*. 2020. <https://doi.org/10.1101/2020.06.15.151647>.
- Hajiaghayy M, Condon A, Hoos HH. Analysis of energy-based algorithms for RNA secondary structure prediction. *BMC Bioinformatics*. 2012;13(1):22. <https://doi.org/10.1186/1471-2105-13-22>.
- Delli Ponti R, Marti S, Armaos A, Tartaglia GG. A high-throughput approach to profile RNA structure. *Nucleic Acids Res*. 2017;45(5):e35. <https://doi.org/10.1093/nar/gkw1094>.
- Chin-Inmanu K, Suttitheptumrong A, Sangsrakru D, Mairiang D, Tangphatsornruang S, Malasit P, et al. Complete genome sequences of four serotypes of dengue virus prototype continuously maintained in the laboratory. *Microbiol Resour Announc*. 2019;8(19). <https://doi.org/10.1128/MRA.00199-19>.

26. Paul B, Tham WL. Controlling dengue: effectiveness of biological control and vaccine in reducing the prevalence of dengue infection in endemic areas. *Health*. 2016;08(01):64–74. <https://doi.org/10.4236/health.2016.81008>.
27. Clyde K, Kyle JL, Harris E. Recent advances in deciphering viral and host determinants of dengue virus replication and pathogenesis. *J Virol*. 2006; 80(23):11418–31. <https://doi.org/10.1128/JVI.01257-06>.
28. Mortimer SA, Kidwell MA, Doudna JA. Insights into RNA structure and function from genome-wide studies. *Nat Rev Genet*. 2014;15(7):469–79. <https://doi.org/10.1038/nrg3681>.
29. Reid DW, Campos RK, Child JR, Zheng T, Chan KWK, Bradrick SS, et al. Dengue virus selectively annexes endoplasmic reticulum-associated translation machinery as a strategy for co-opting host cell protein synthesis. *J Virol*. 2018;92(7). <https://doi.org/10.1128/JVI.01766-17>.
30. Proutski V, Gould EA, Holmes EC. Secondary structure of the 3' untranslated region of flaviviruses: similarities and differences. *Nucleic Acids Res*. 1997; 25(6):1194–202. <https://doi.org/10.1093/nar/25.6.1194>.
31. Viktorovskaya OV, Greco TM, Cristea IM, Thompson SR. Identification of RNA binding proteins associated with dengue virus RNA in infected cells reveals temporally distinct host factor requirements. *PLoS Negl Trop Dis*. 2016;10(8): e0004921. <https://doi.org/10.1371/journal.pntd.0004921>.
32. Garcia-Moreno M, Noerenberg M, Ni S, Järvelin AI, González-Almela E, Lenz CE, et al. System-wide profiling of RNA-binding proteins uncovers key regulators of virus infection. *Mol Cell*. 2019;74:196–211.e11. <https://doi.org/10.1016/j.molcel.2019.01.017>.
33. Dominguez D, Freese P, Alexis MS, Su A, Hochman M, Palden T, et al. Sequence, structure, and context preferences of human RNA binding proteins. *Mol Cell*. 2018;70:854–867.e9. <https://doi.org/10.1016/j.molcel.2018.05.001>.
34. Xu S, Ci Y, Wang L, Yang Y, Zhang L, Xu C, et al. Zika virus NS3 is a canonical RNA helicase stimulated by NS5 RNA polymerase. *Nucleic Acids Res*. 2019;47(16):8693–707. <https://doi.org/10.1093/nar/gkz650>.
35. Yap ML, Klose T, Urakami A, Hasan SS, Akahata W, Rossmann MG. Structural studies of Chikungunya virus maturation. *Proc Natl Acad Sci U S A*. 2017; 114(52):13703–7. <https://doi.org/10.1073/pnas.1713166114>.
36. Agostini F, Zanzoni A, Klus P, Marchese D, Cirillo D, Tartaglia GG. catRAPID omics: a web server for large-scale prediction of protein-RNA interactions. *Bioinformatics*. 2013;29(22):2928–30. <https://doi.org/10.1093/bioinformatics/btt495>.
37. Frick DN, Lam AMI. Understanding helicases as a means of virus control. *Curr Pharm Des*. 2006;12(11):1315–38. <https://doi.org/10.2174/138161206776361147>.
38. Venkataraman S, Prasad BVLS, Selvarajan R. RNA dependent RNA polymerases: insights from structure. *Funct Evol Viruses*. 2018;10(2). <https://doi.org/10.3390/v10020076>.
39. Villanueva RA, Rouillé Y, Dubuisson J. Interactions between virus proteins and host cell membranes during the viral life cycle. *Int Rev Cytol*. 2005;245: 171–244. [https://doi.org/10.1016/S0074-7696\(05\)45006-8](https://doi.org/10.1016/S0074-7696(05)45006-8).
40. Kramer MC, Gregory BD. Does RNA secondary structure drive translation or vice versa? *Nat Struct Mol Biol*. 2018;25(8):641–3. <https://doi.org/10.1038/s41594-018-0100-2>.
41. Mazeaud C, Freppel W, Chatel-Chaix L. The multiples fates of the flavivirus RNA genome during pathogenesis. *Front Genet*. 2018;9:595. <https://doi.org/10.3389/fgene.2018.00595>.
42. Cid-Samper F, Gelabert-Baldrich M, Lang B, Lorenzo-Gotor N, Ponti RD, Severijnen L-AWFM, et al. An integrative study of protein-RNA condensates identifies scaffolding RNAs and reveals players in fragile X-associated tremor/ataxia syndrome. *Cell Rep*. 2018;25:3422–3434.e7. <https://doi.org/10.1016/j.celrep.2018.11.076>.
43. Chang H-H, Huber RG, Bond PJ, Grad YH, Camerini D, Maurer-Stroh S, et al. Systematic analysis of protein identity between Zika virus and other arthropod-borne viruses. *Bull World Health Organ*. 2017;95(7):517–525. <https://doi.org/10.2471/BLT.16.182105>.
44. Costa RL, Voloch CM, Schrago CG. Comparative evolutionary epidemiology of dengue virus serotypes. *Infect Genet Evol*. 2012;12(2):309–14. <https://doi.org/10.1016/j.meegid.2011.12.011>.
45. Vandelli A, Monti M, Milanetti E, Armaos A, Rupert J, Zacco E, et al. Structural analysis of SARS-CoV-2 genome and predictions of the human interactome. *Nucleic Acids Res*. 2020;48(20):11270–83. <https://doi.org/10.1093/nar/gkaa864>.
46. Delli Ponti R, Armaos A, Vandelli A, Tartaglia GG. CROSSalve: a web server for predicting the in vivo structure of RNA molecules. *Bioinformatics*. 2020; 36:940–1. <https://doi.org/10.1093/bioinformatics/btz666>.
47. Lorenz R, Luntzer D, Hofacker IL, Stadler PF, Wolfinger MT. SHAPE directed RNA folding. *Bioinformatics*. 2016;32:145–7. <https://doi.org/10.1093/bioinformatics/btv523>.
48. Rice P, Longden I, Bleasby A. EMBOS: the european molecular biology open software suite. *Trends Genet*. 2000;16(6):276–7. [https://doi.org/10.1016/S0168-9525\(00\)02024-2](https://doi.org/10.1016/S0168-9525(00)02024-2).
49. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 2013; 30(4):772–80. <https://doi.org/10.1093/molbev/mst010>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

