 doi:10.20944/preprints202103.0039.v1

Article

Evaluation of conserved RNA secondary structures within and between geographic lineages of Zika virus

Kevin Nicolas Calderon [1], Johan Fabian Galindo[2] and Clara Isabel Bermudez-Santana[1,*]

[1] Departamento de Biología, Universidad Nacional de Colombia, Bogotá, Colombia.
[2] Departamento de Química, Universidad Nacional de Colombia, Bogotá, Colombia.
* Correspondence: cibermudezs@unal.edu.co; Tel.: +57 1 3165000 ext. 11305

Abstract: Zika virus (ZIKV), without a vaccine or no effective treatment approved as yet, have globally spread since the past century. The infection caused by ZIKV in humans has changed progressively from mild to subclinical in the last years, causing epidemics with greater infectivity, tropism towards new tissues, and other related symptoms as a product of various emergent ZIKV-host cell interactions. However, it is still unknown why or how the RNA genome structure impacts those interactions in differential evolutionary origin strains. Moreover, genomic comparison of ZIKV strains from the sequence-based phylogenetic analysis is well known, but differences from RNA structure comparisons are less known. Thus, in order to understand the RNA genome variability of lineages of various geographic distributions better, 412 complete genomes in a phylogenomic scanning were used for studying the conservation of structured RNAs. We found specific genomic regions, which highlight their patterns of conserved RNA structures at the level of inter-geographical comparisons. We have proposed these structures as candidates for further experimental validation to establish their potential role in vital functions of the viral cycle of ZIKV and their possible associations with the singularities of different outbreaks that occurred in specific geographic regions.

Keywords: Zika Virus, Phylogenomics, Viral Genomic Variability, Conserved RNA structures

1. Introduction

Zika virus (ZIKV) was first identified in Rhesus monkeys in the Zika forests of Uganda in 1947. Its spread from Africa reached out to Southeast Asia, limiting its associated symptoms to feverish symptoms, conjunctivitis, and joint pain during the 20th century [1]. At the beginning of the XXI century, outbreaks of the virus began in countries of the island complex of Oceania, where new symptoms were associated, such as Guillain-Barré Syndrome (GBS), an autoimmune disease in which the immune system attacks the nervous system of the affected person [2]. Since its dispersion in America in 2015, the virus infection started to be associated with congenital fetal microcephaly and neurological damage caused by the virus's vertical transfer between mother and fetus [3,4]. Declared as public health emergency by WHO in 2016 and due to the absence of a vaccine in preventive terms or specific treatment, ZIKV is currently one of the most significant concerns of health systems in tropical countries [5], where its transmission occurs mainly by vector mosquitoes of the *Aedes* genus [6].

ZIKV is a single-stranded positive-sense RNA arbovirus within the *Flavivirus* genus [7], the genus to which other known viruses of public health importance belong, such as Dengue (DENV), yellow fever (YFV), or West Nile Virus (WNV) [8]. The genome length is close to 10.8kb, and is composed of two non-translatable regions (UTR) located at the 5' and 3' ends, of 106 nt and 428 nt in length, respectively [9]; and a coding region (CDS) of 10.3kb. CDS codes for three structural proteins (Capsid (C), Membrane (M), and Enve-

 doi:10.20944/preprints202103.0039.v1

lope (E)), and for seven non-structural proteins (NS1, NS2A, NS2B, NS3, NS4A, NS4B, and NS5), necessary to complete its viral replicative cycle [8].

ZIKV genome folds in a secondary RNA structure shapes like other RNA viruses and RNA-type molecules. These structures perform pivotal functions during the viral cycle, i.e., such as regulating translation, promoting replication, and evading host cell antiviral responses [10,11,12]. Some well-known examples of these structures are in the 3' UTRs, which flaviviruses use to produce the non-coding RNAs (ncRNAs) and flaviviral subgenomic RNAs (sfRNAs). sfRNAs in particular have been associated with the antiviral response's evasion by negatively affecting the immune response mediated by interferon type 1 (IFN-I) of the host cell [13,14]. It has been reported that the structures of the 5' UTR region in ZIKV regulate the initial viral translation by promoting the placement of a CAP at the 5' end and mimicking the mRNA of the affected cells [15]. The viral replication in Flaviviruses is also initiated by changing their linear genome into a circular genome through the interaction of the RNA structures located at the two ends of the strand [16]. Additionally, many RNA viruses potentially encode precursor structures of microRNAs processed by canonical and non-canonical pathways in the host cell [17,18]. These microRNAs can affect the host cell's metabolism or regulate antiviral response from the genes' expression [19-22].

However, patterns of RNA secondary structures in viruses with RNA genomes, despite their importance, have been poorly studied [23]. We present an extensive survey to screen for ZIKV conserved genomic subregions restricted to specific continental geographical origins. We found that some structures are conserved by each geographic location or even by lineage classification. The fact that some patterns are region-specific suggests a potential selection by unknown pressure factors. We detected critical subregions in the ZIKV genome with our strategy and targeted future studies to establish their function on the viral cycle with the possible association in its infective capacity [12].

2. Materials and Methods

2.1. Genomic data source

An exhaustive search of complete Zika virus genomes was performed in NCBI and VipR databases [24,25] (search date: July 2020). A total of 1023 genomes were found, and redundant sequences were filtered using BLASTn [26], sequences that exceeded 0.05% of unassigned nucleotides "N" were removed using Bioperl v1.7.7 [27]. To work with equal lengths of sequence size, they were statistically analyzed with R v4.0.2 software, and the irregularly short sequences (lower outliers) were removed [28]. Therefore, we finished with a total of 412 complete ZIKV genome sequences to work in further steps.

2.2. Genomic Alignments according to the geographical origin

All multiple alignments were performed by Clustal Omega v1.2.4 [29], setting two iterations per alignment and using the other parameters by default. The first alignment included all sequences (global context), and their extremes were trimmed as long as the gap content was greater than 33%, using UNIPRO-Ugene v33.0 software [30]. Subsequently, the sequences corresponding to different continental geographical origins (Africa, Asia, Oceania, and America) were extracted from this alignment. Once the group of sequences was set, they were de-aligned and re-aligned according to their continental groups.

2.3. Phylogenomics Analysis

The following R packages were used to evaluate genomic sequence relationships: APE, seqinr, and Phangorn [31-33]. A distance matrix was built using the dist.alignment function based on the square root of pairwise distances from multiple sequence alignments. Neighbor-Joining (NJ) tree was rooted in the yellow fever virus sequence (NC_002031.1), which also belongs to the *flavivirus* genus. The dendrogram branches' supports were made by 1000 subsamples (1000 bootstraps) of the distance matrix. A Maximum Likelihood (ML) analysis was also performed. To determine the base model for the ML tree, the modeltest function implemented in the phangorn package was used using the AIC index as the main criterion of the model selection, and the construction of the Maximum likelihood dendrogram was performed by PhyML v3.0.1[34]. Both types of dendrograms (NJ and ML) were plotted with R software.

Finally, to describe the variability and conservation of sequences, percentages of paired identity and the number of identical sites per nucleotide columns of the alignment were calculated using Geneious Prime v2020.1 [35]. From the phylogenetic relationships obtained from the first alignment (global context), subdivisions of clades that are contained within each large continent were proposed: Asia continental, Southeast Asia, Brazil Block, Colombia Block, Mexico Block, and Caribbean Block (Table A1). Thus, we obtained eleven groups of sequences that allow us to perform comparative analysis, at different scales, within and between geographic lineages.

2.4. Prediction of Conserved Secondary Structures

In order to analyze the geographic similarity of the RNA conserved structures, RNAz v2.1.1 software [36] was employed. An experimental design was carried out testing six different combinations of two factors: window size (150nt, 120nt, 100nt) and sliding window size (20nt and 40nt), in order to set the screening parameters and minimize the rate of false-positive prediction. For each parameter combination, 100 randomizations of each alignment were performed using the RNAz script, RandomAlign.pl, to determine false positives (FPrandom) and positive detections of the original alignment (PNative). The relationship between these two indices was established as a quantitative quality criterion (FPrandom / Pnative). The lower the numerical value of this relationship, the higher the specificity and the higher the sensitivity. Based on this, a two-way analysis of variance (ANOVA) was performed (homoscedasticity and normality assumptions were verified), and Tukey plots and Boxplots were made to guide a statistical decision regarding the selection of the sliding window size. Once the best combination of the sliding window was set, we followed the pipeline of the RNAz program, described in Gruber et al. (2010) [29], using as a filter: P> 0.9 and Z value <-2; as well as the "no-reference" and "both-strands" parameters. Finally, the genomic positions of conserved secondary structures were plotted using ggplot2 v3.3.3 [37]. The index produced by RNAz in HTML was used to graphically extract the most representative RNA secondary structures of each position, considering as selection criteria: the Z Value, SCI, SVM decision value, and MFE. These representative structures, specifically their consensus structure sequences, were compared using BLASTN against those stored in the RFAM repository for the Zika virus [38]

Finally, as a statistical support of the secondary RNA structures two different randomization alignment routes were performed: General (to each alignment) and specific (to each window). In the general randomization, 100 randomized alignments from each geographic or sub-geographic region were obtained (n = 53,100 windows per alignment). These were analyzed entirely in RNAz, keeping the size of the sliding window, Z value,

 doi:10.20944/preprints202103.0039.v1

and P-value parameters. The relative false positive detection rate (FP) and the percentage of specificity (% Specificity) were determined. The specific randomization (200 windows, n = 200) per each of the windows detected as structured in the original alignment was generated, and they were analyzed with the same RNAz parameters mentioned above, removing from the analysis those windows that had a rate of false positives greater than 0.05 (% FP> 0.05).

3. Results

The NJ approach shows the diverse relationships between the viral genomes and the continents (Figure 1A). A group of African genomes is observed in the cladogram's basal position, representing their ancestral status with respect to other continents. Likewise, Asia and America form two large uniform clades, with solid support in their basal branches (bootstrap> 0.9); and they agree in the order of ancestry: first Asia and then America. Additionally, the derived branches that do not have adequate support (bootstrap <0.7) coincide in having very short branches (<0.01), and their paired identity distances are minimal; thus, the pairs of sequences in these branches are very similar sequences. Finally, Oceania does not present a concise continental separation (bootstrap <0.7), and it is included within the American clade.

The circular design of the cladogram was changed to a traditional design (Figure 1B) to have a better detail of the continental sub-grouping. The distances between paired sequences are not taken into account in Figure 1B, but the groups of sequences and their support are better appreciated. We can see that Asia is divided into Continental Asia and Southeast Asia, with good branching support (Bootstrap> 0.9). In the same way, America has been split into at least four subgroups (or blocks). Colombia and Mexico blocks are well supported in their basal branching (Bootstrap> 0.9), and they are also composed of countries from northern South America and Central America, respectively (detailed list of the set of countries in Table A1). Most of the Caribbean Countries are in another block, which is well supported in a single group (Bootstrap> 0.9) and contains the island countries from the Caribbean and the coast of the United States. However, Puerto Rico is separated from this homogeneous group and is a particular case, even though it is at the same geographical position as other Caribbean countries. Finally, we can see a Brazilian block, which permeates all the other groups in the Americas in the cladogram, reflecting their condition as the original location of the outbreak in the Americas, since the spread was derived from there to the other American countries of America. This representation agrees and explains the little differentiation found in the branches with low support and short distance in Figure 1A.

 doi:10.20944/preprints202103.0039.v1

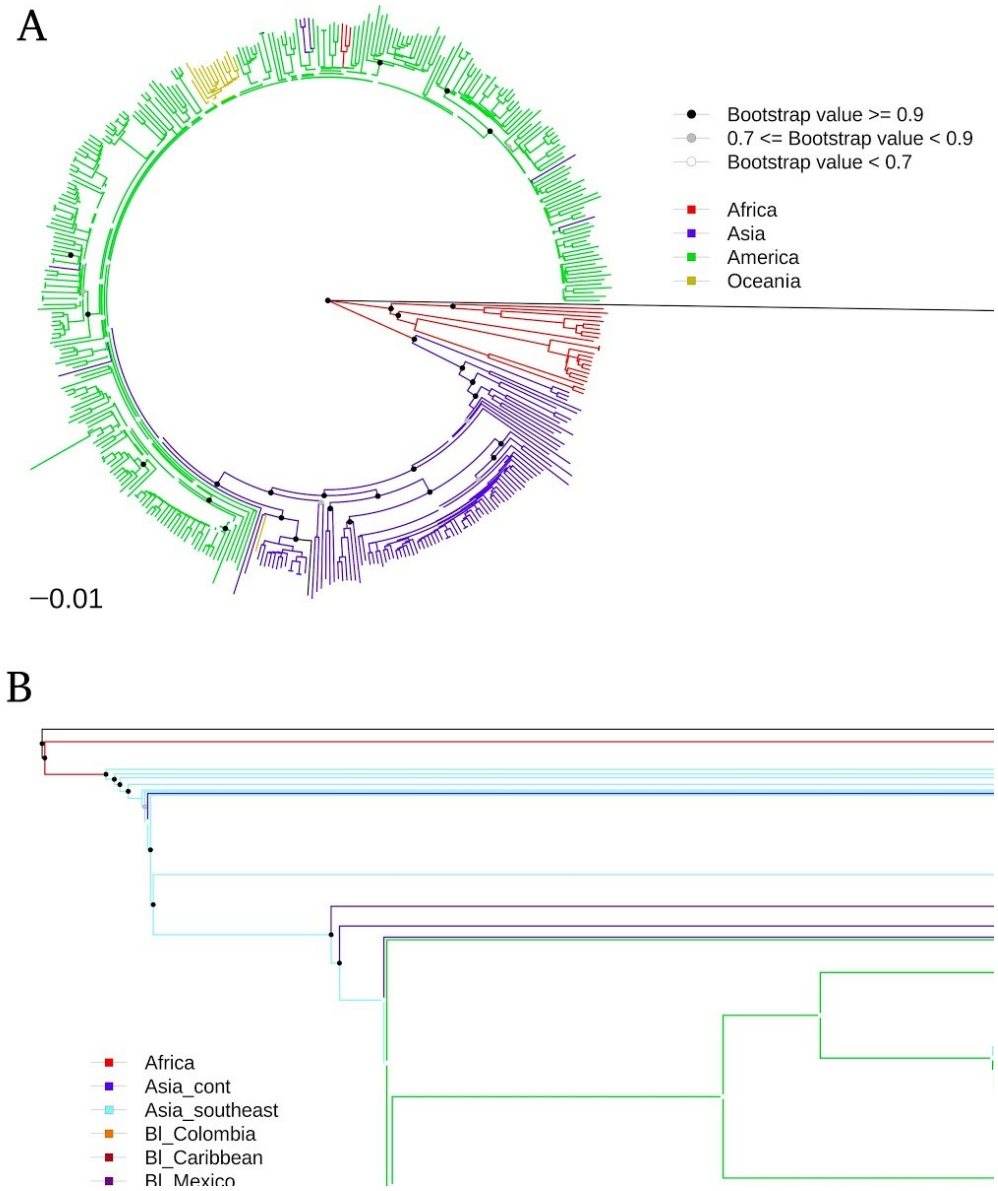

Figure 1. Phylogram of ZIKV produced by Neighbor-Joining approach. (A) Fan plot type, sequences according to their inter-geographic lineages (B) Phylogram classic-plot type, sequences according to their intra-geographic lineages.

Regarding ML dendrograms (Figure 2), these are consistent with the topology generated by NJ. Similar blocks to the NJ methodology are observed, i.e., Caribbean, Colombia, Mexico, Brazil, Southeast Asia, Continental Asia, and African country blocks (Figure 2B). Therefore, these groupings are independent of possible biases related to the dendrogram graphing methodology. The main difference between both methods lies in that ML separates the sequences from Oceania, and the resolution distance in the cladogram of continents (Figure 2A) is even smaller than NJ. Therefore, the GTR evolutionary model, despite having the best fit according to its AIC index, has an even lower resolution than the mathematical approach of distance based on pairwise sequence identities of Figure 1A.

 doi:10.20944/preprints202103.0039.v1

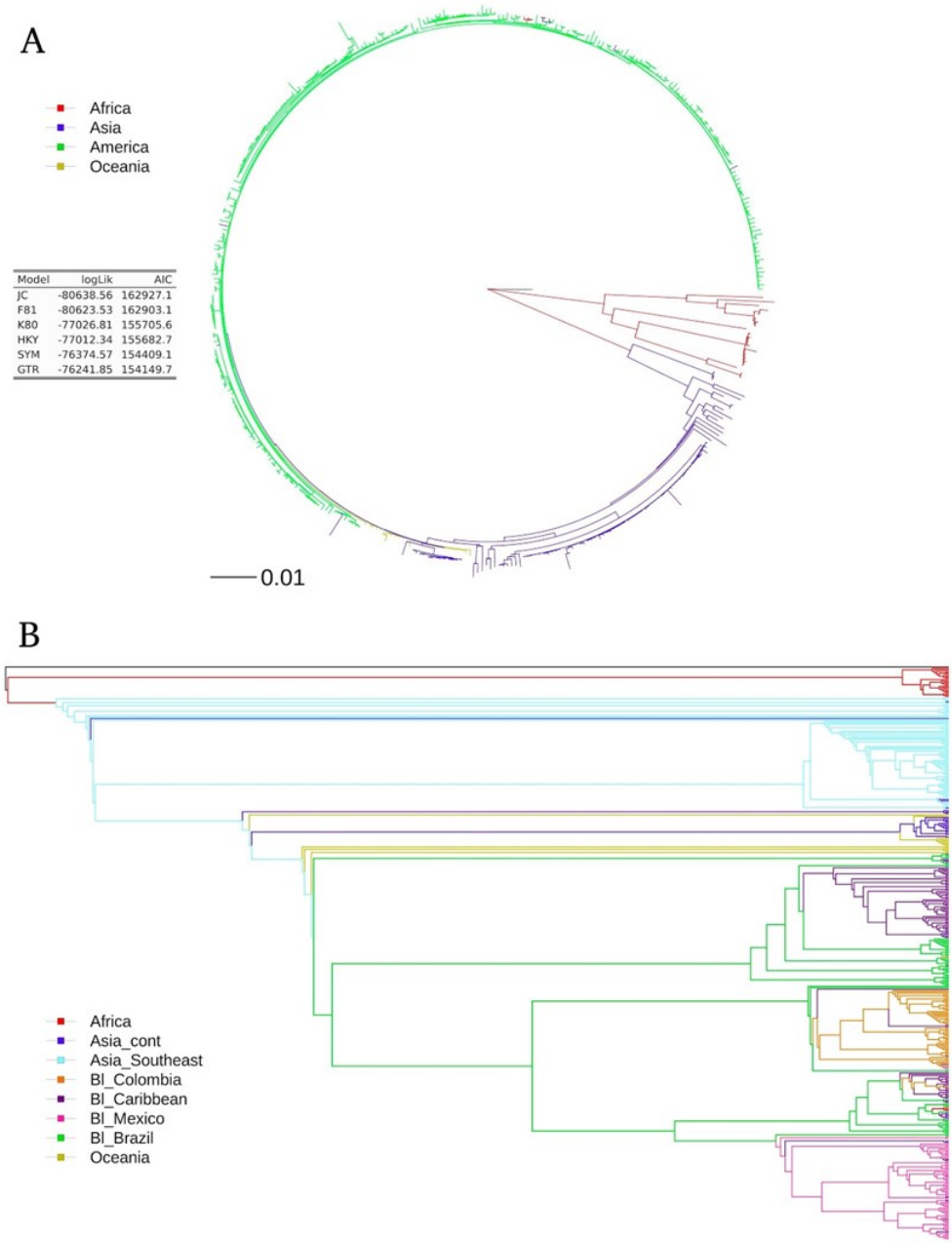

Figure 2. Phylogram of ZIKV produced by Maximum-Likelihood approach. (A) Fan plot type, sequences according to their inter-geographic lineages (B) Phylogram classic-plot type, sequences according to their intro-geographic lineages

In the descriptive analysis of sequence variability, summarized in Table 1, we appreciate a high sequence conservation level. The percentage of paired identity, taking all genomes as a set, is 98.29%, and in none of the proposed blocks was less than 99%. This result reflects the high degree of conservation between the sequences. Africa stands out as the region whose sequences show the least similarity among them (94.11%). Despite the high degree of similarity observed between sequences, the number and percentage of nucleotide columns, which are identical in each alignment, are always lower than their percentage of paired identity. Finally, the median of the sequences was 10,729 nt, which is close to the length of the ZIKV reference sequences (10.8kb).

		Length		Identical Sites		Mean pairwise identity	
Region	**Sequences(n)**	**Median**	**Range**	**N° Columns**	**%**	**%**	**SD**
Global	410	10729	10368-11119	7205	64,8	98,29	0,031
Africa	24	10782	10617-11119	8917	80,2	94,11	0,037
Asia	106	10762	10415-10808	8970	83	99,01	0,009
Oceania	14	10644	10585-11155	11021	98,8	99,86	0,001
America	266	10692	10368-10864	8973	82,6	99,59	0,001
Bl_Brazil	58	10752	10455-10864	10288	94,7	99,65	0,001
Bl_Colombia	53	10659	10385-10808	10375	96	99,8	0,002
Bl_Mexico	66	10696	10398-10807	10191	94,3	99,75	0,001
Bl_Caribbean	89	10727	10368-10808	9986	92,4	99,55	0,002

Table 1. Summary table of descriptive data of the sequences analyzed.

From the ANOVA analysis significant differences in the choice of window size were found (to select the initial parameters of RNAz: window and slide size, Figure S1), but not in terms of sliding size; the interaction between these two parameters is not significant (Figure S1 (C)). The Tukey test and the boxplots suggest that we can choose a window size of 150 or 100 nucleotides. The 150 nucleotide size allows the possibility of detecting RNA secondary structures in their complete form and appreciates hairpin-like structures, which are possible to be targeted by Dicer or any other cytoplasmic microprocessor. Therefore, 150 window and 20 sliding window size were the selected parameters.

In evaluating conserved RNA secondary structured regions in the geographical inter-lineages comparison (Figure 3), the globally conserved structured regions are only those of the initial and final parts of the sequences (5 'and 3', respectively). These accomplish crucial functions throughout the flavivirus genus, and their presence reflects a positive control in the detection methodology of structured areas for the viral genome (Figure 3). Additionally, African sequences present a unique pattern of structured regions, containing lineage-specific structures at positions 2.8kb and 9.6kb (Figure 3B). On the other hand, Asia, Oceania, and America share four structured regions at places 1.1kb, 4.5kb, 7.1kb, and 8.5kb (Figure 3C – 3E). Similarly, Oceania and America share a structured region at the 3.1kb position (Figure 3D, 3E). Finally, Oceania has a particular structured area at position 5.2kb (Figure 3D). This graph allows us to appreciate three types of patterns: 1. There are conserved structured regions in all sequences; 2. There are unique conserved structured regions based on the particular geographic lineages; 3. Patterns in the geographic lineage groups are formed because they share certain conserved structured regions.

 doi:10.20944/preprints202103.0039.v1

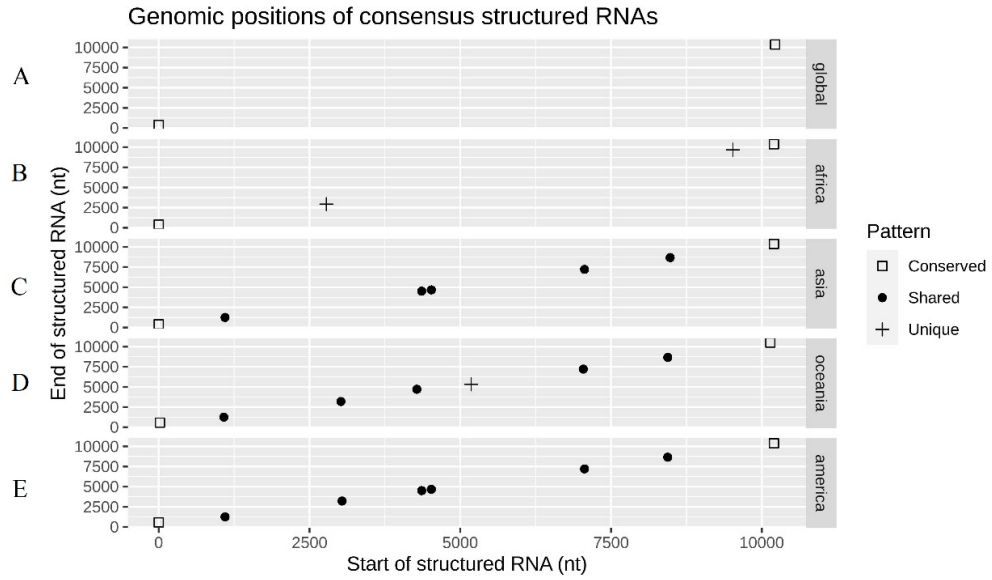

Figure 3. Genomic position and patterns of the structured regions of RNA found in inter-geographic lineages of the Zika virus. (A) Global. (B) Africa. (C) Asia. (D) Oceania. (E) America.

In the evaluation within geographical lineages of the conserved secondary RNA regions, the Caribbean block is the only one presenting two structured zones, which differs from all the other American subregions at positions 5.2kb and 9.2kb (Figure 4A). It is worth clarifying that the double points generated in the position close to 4kb in the Americas continent and in the Brazil Block subregion, are the consequence of a discontinuity in the sliding window of the RNAz pipeline, and they do not represent a different structured region. Additionally, the structured position of Southeast Asia at position 5.7kb is the only one different from other sub-regions of the Asian continent (Figure 4B). Finally, the same case of the double points, previously mentioned, occurs in the position of 4.5 kb between Asia and Southeast Asia. In general, there is little variation in terms of the presence-absence of structured regions at intra-geographical level regions.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 March 2021 doi:10.20944/preprints202103.0039.v1

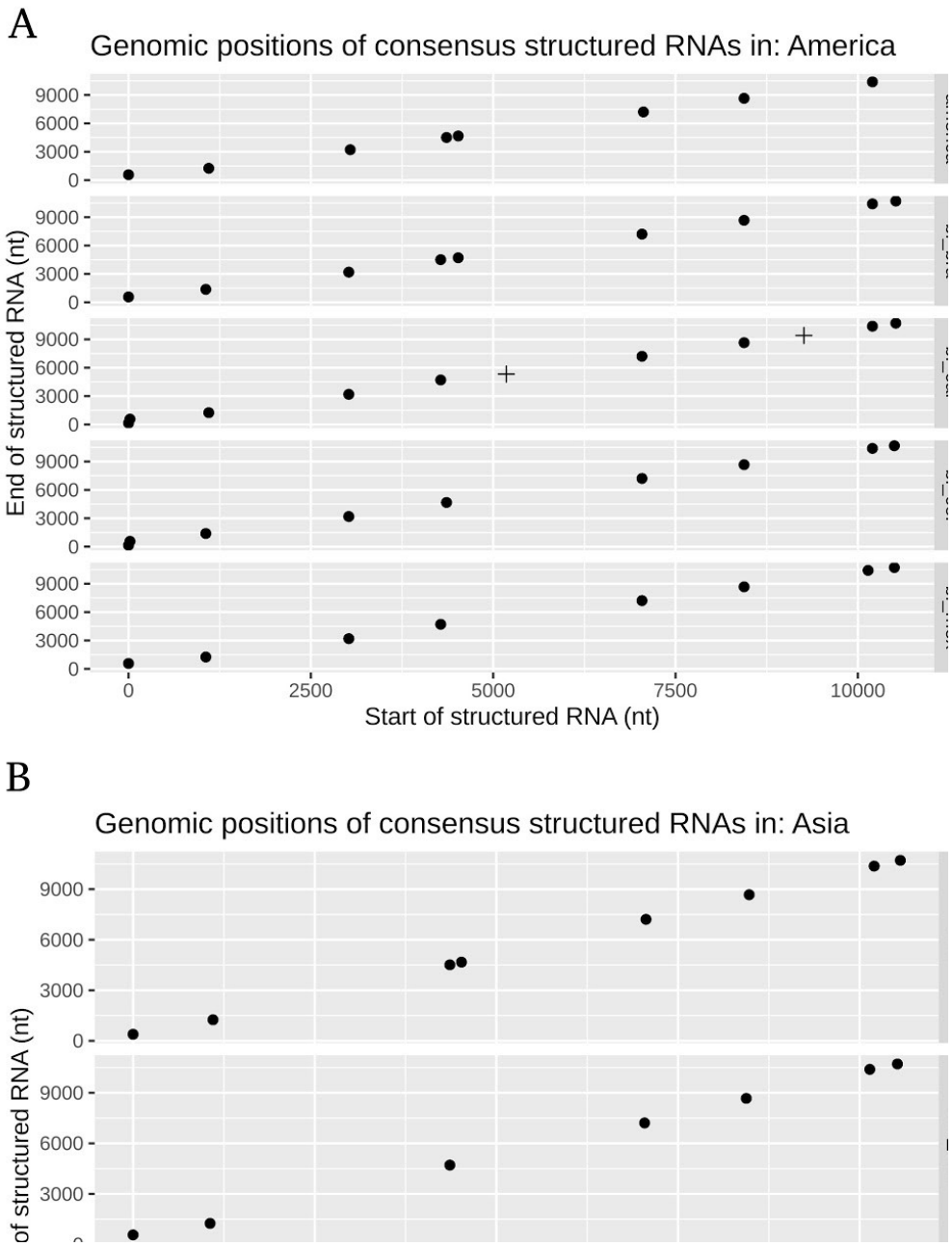

Figure 4. Genomic position and patterns of the structured regions of RNA found in the intra-geographic lineages of the Zika virus. (A) America and its respective sub-regions. (B) Asia and its respective sub-regions.

In the statistical validation by complete alignments, we obtained a false positive detection rate lower than 5% (FP <0.05), and a specificity index higher than 95% in all cases; therefore the filters used to run the pipeline of RNAz (Z value <-2, P> 0.9) were effective in selecting information of the detected structures and the results were statistically significant (Table 2). In the other approach of statistical validation, for each of the structured windows obtained from RNAz, cited in Tables S1 and S2, a total of 30 windows (FP > 0.05) of the analysis were removed. Window number 32 of the Bl_Colombia caught our attention, which was the only one that represented the removal of an entire structured locus.

 doi:10.20944/preprints202103.0039.v1

Region	% FP	Specificity (%)	N° Windows (n)
Global	0,003	99,65	53100
Africa	0,005	99,49	53100
Asia	0,019	98,12	53100
America	0,023	97,61	53100
Oceania	0,034	96,6	53100
Asi_SE	0,021	97,87	53100
Asi_cont	0,03	96,96	53100
Bl_Bra	0,032	96,83	53100
Bl_Car	0,032	96,85	53100
Bl_Col	0,034	96,6	53100
Bl_Mex	0,033	96,7	53100

Table 2. Statistical evaluation for complete alignments of geographic lineages. (%FP = False positives rate).

The results of the most representative secondary RNA structures at the inter-geographical region level are found in Figure 5, highlighting the one found in the envelope region (Figure 5B) since it has experimental validation [23]. The representative structures of the intra-geographical regions are in Figure S2.

 doi:10.20944/preprints202103.0039.v1

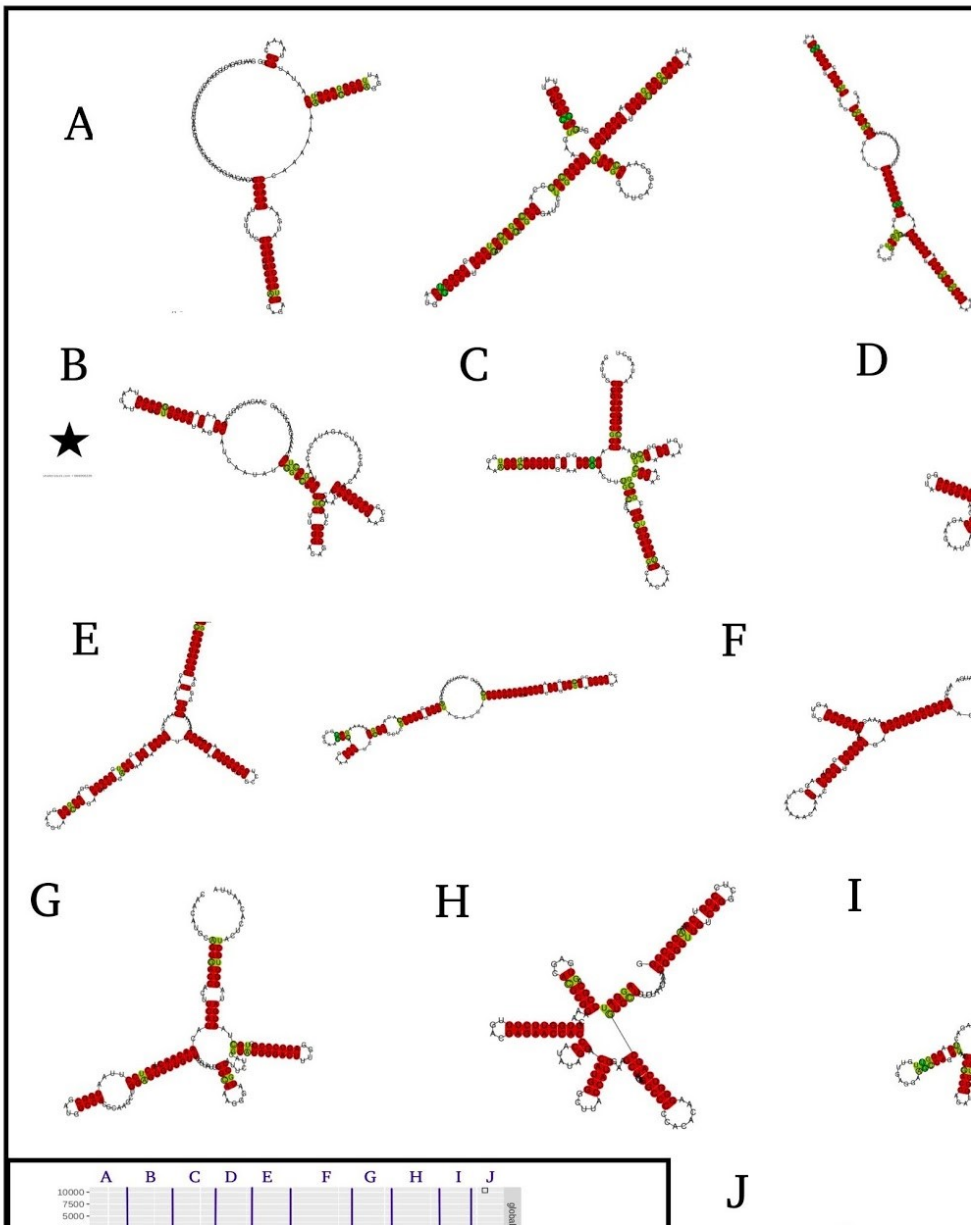

Figure 5. RNA most representative structures of the genomic regions at the inter-geographic lineages level. (A-J) Each structure, or group of structures, represents the genomic position indicated in the lower left box. The star indicates the structure related to the envelope coding region; its confirmation was made using a BLASTx search with parameters of 100% identity, 98% Query cover, and E value 6e-27 to the "envelope protein" ANC90422.1 of the Zika virus.

On the other hand, RNA structures overlapped with RNAs models from RFAM are shown in Table 3. However, the Flavi_CRE structure was not detected because its presence occurs at the end of the 3 'UTR genomic region (~ 10697 nt); therefore, the alignment quality clipping process, mentioned in the methodology, could have limited its structural detection.

 doi:10.20944/preprints202103.0039.v1

Annotated structure	window	p.ident	length	mismatch	gap.open	q.start	q.end	s.start	s.end	e.value	Bit.score
Flavivirus DB	16	100	29	0	0	122	150	1	29	3,29E-11	49,6
Flavi_SLA	1	100	57	0	0	1	57	17	73	4,48E-25	95,7
Flavi_CRE	no hits	/	/	/	/	/	/	/	/	/	/

Table 3. Match of the structured windows found in the analysis with the only structures reported in the RFAM for the Zika virus.

4. Discussion

The chronological order found in both cladograms agrees with the historical records of Zika virus outbreaks. The sequences from Africa are basal, as a viral origin; followed by Asia and last America [6]. According to ML analysis, the location of the sequences from Oceania, as a sister group to America, agrees with it being the place of viral origin introduced to America [39]. Otherwise, the inclusion of Oceania in the American clade, by the NJ method with low support, could suggest a difficulty in the tree resolution due to the little genomic differentiation between the sequences of both continents. Moreover, our dendrograms agree with the one generated by Metzky *et al*. (2017), in which Brazil is located as the geographical origin of the viral breakout in America [5]. In the basal part of the American clade, the short branches with low support agree with the sequences from Brazil as the origin of the outbreak. These sequences are probably dispersed throughout the continent since they were still very similar to each other. This pattern is associated with a rapidly spreading viral outbreak by the introduction of a new virus to a population without a history of immune memory to the same virus [5].

Additionally, the grouped sequences in the clades of both types of cladograms reflect the establishment of individual viral genotypes in geographically delimited regions, regardless of the methodology used. We always found the Colombia block, the Caribbean block, the Mexico block, Asia Southeast, Asia Continental, and Africa with good support. These groups agree with other dendrograms reported in the literature with a set of sequences previously reported [40,5].

The high degree of similarity seen between the Zika viral sequences can be contextualized with similar features found in other flaviviruses. For instance, the similarity within Dengue serotypes, where its most variable fragments, the 3 'UTR region, reaches 97% of paired identity [41]. Therefore, finding 98% global identity and 99% at the regional level are not an unusually high value and, indeed, it suggests selective purifying pressures on the Zika virus genome. This is feasible because, in the CDS region, its entire length encodes proteins, which are essential for evading the host's immune responses and completing their viral replication. Thus, the accumulation of drastic changes in its genome can affect the viral viability [42]. Other authors have suggested a purifying selection in viruses that handle a complete viral cycle inside humans. The majority of viral genomes used in phylogenomics studies come from clinical samples making it more difficult to uncover the virus´s true diversity. If viral genomes were sampled directly from ZIKV circulating in wild mosquitoes vectors, a greater diversity is to be expected [41]. Indeed, higher variability in the WNV virus found in vector insects has already been reported, whose vector, *Culex spp*, has a greater vector-specific viral diversity in contrast to the variety found in the host vertebrate [43].

On the other hand, a greater variation, comparing identical columns of the alignments, allows more significant variability in the secondary RNA structures, overcoming the coded sequences' variation. Consequently, a greater range in the variation of functional RNA structures can be observed, increasing the range of functions performed by the proteins encoded by the virus [11]. However, to confirm this hypothesis, it would be

necessary to analyze how many of the found mutations are synonymous, compensatory to structures, or not synonymous, inconsistent with the structures.

Regarding the conserved structured regions of RNA in all the Zika virus sequences, at the 5 'and 3' UTR ends, several of their essential functions have been reported in *the flavivirus* genus, especially in structures that were also found in the RFAM database [44,45]. Thus, the SLA structure found in the 5 'region is the structure recognized by RNA polymerase (NS5), fundamental in viral replication [36]. Additionally, this structure promotes the addition of 5' CAP during viral RNA synthesis, which is necessary for viral translation by recruiting the eukaryotic eIF4E binding factor and the subsequent recruitment of 48s and 60s ribosomal units [46,47]. Likewise, the DB structure of the 3 'UTR region has been related to the formation of sfRNAs structures and, indeed, a 30 nucleotides deletion in Dengue virus DB1 has generated an attenuated version of the virus, because it turns particularly susceptible to type 1 interferon, thus highlighting the importance of this structure in the viral cycle. Something similar may happen in the Zika virus [48]. The differences in structured regions between Africa and the rest of the world are possibly related to biotic particularities of the continent; for instance, the transmission in Africa is performed by another vector species: *Aedes Africanus*. It is acknowledged that secondary RNA structures are vital factors in flaviviruses for viral replication in their respective disease-transmitting insects [42]. For example, when DENV is cultured in human cells, structures sfRNA1 and 2 are mainly produced, while culturing the same serotype in mosquito cells produces more types of sfRNAs: 1, 2, 3, and 4. Interestingly, when human cells' viral culture went through five reproductive cycles in mosquito cells, the sfRNA3 and 4 were observed again. Therefore, sfRNA structures 3 and 4 can be performing some function in viral replication inside of insect vectors, while sfRNA structures 1 and 2 are required regardless of the type of cultured cells [49].

Another example is a WNV mutant which lacks the formation of the sfRNA1 structure and it does not survive in the intestine of the insect *Culex spp*, but it survives in its salivary glands; therefore, it is a key structure to complete the viral cycle, from the ingestion of blood to transmission by mosquito's saliva [14]. This is an interesting aspect because it has been shown that the Zika virus can replicate in the intestine and salivary glands of *Culex spp* insects [50]. The virus may be adapting to new vectors, and its outbreaks may have new scopes linked to the distribution of this other type of vector, all mediated by changes and adaptations in their particular RNA structures.

With respect to the structural region of RNA shared between Asia, Oceania, and America in the position close to 1.1kb, it is found to be intriguing because it has been experimentally validated, in vivo, and its importance in the function of the Zika virus has been reported [23]. The authors found an intramolecular interaction of the RNA structures present in the 5'UTR region (2 nt-43 nt) and the structures of the envelope coding region (E) (1089 nt-1134 nt), which occurs only in post-epidemic viral strains of Asia, Oceania, and America; but not in Africa. They evaluated four mutants, damaging their RNA structures of the coding position corresponding to the envelope, without altering the encoded protein, and obtained a reduction in viral infectivity. This infectivity was partially restored by reintroducing compensatory mutations to re-form the structure initially found [23]. This structure coincides with the structured region found in the present work, at 1.1 kb, corresponding to the envelope coding region (Figure 5B). Therefore, this pattern analysis of an RNA structure, which has been associated with viral infectivity, allows the possibility that other structured regions found in this work may also have critical functions in the viral cycle of ZIKV (with unique or shared patterns, Figure 3).

Finally, the RNA structures found in common between America and Oceania might contribute to the genomic particularity present in the outbreaks, where the tissue damage

and viral infectivity were superior in contrast to Asian and African lineages in experimental studies with mice [51]. Additionally, it is remarkable that fragmenting the alignments in sliding windows generates border effects, where in-silico structures may be incomplete in their prediction. However, the structures reported here show a strong signal of being a structured region of the genome and facilitates a later evaluation of the real 3D structure. An example of the relation of the RNA 3D structure and its function can be observed in pre-microRNAs. [52]

5. Conclusions

Performing this comparative analysis between Zika virus genomes and their RNA conserved secondary structures allowed the selection of regions and certain specific structures, which stand out for their patterns in genomic comparison at the inter-geographical lineage level. In this way, these patterns of structural conservation guided the selection of potential functionally relevant structures in the viral cycle of ZIKV. Further experimental analysis for associating new functions must be performed. In the future, these structures may have the potential to be targeted to negatively affect viral replication [12].

Supplementary Materials: Figure S1: Results of the two-way ANOVA, comparing the window size factor (100, 120 and 150), and the sliding size factor (20.40), Figure S2: RNA most representative structures of the genomic regions at intra-geographic lineages level., Table S1: Statistical evaluation in the randomized alignments performed by each window detected as positive for structural RNA at inter-geographic lineages level, Table S2: Statistical evaluation in the randomized alignments performed by each window detected as positive for structural RNA at intra-geographic lineages level.

Author Contributions: Conceptualization, C.I.B.; methodology, J.F.G. and C.I.B.; software, J.F.G. and C.I.B.; validation, K.N.C., J.F.G. and C.I.B.; formal analysis, K.N.C., J.F.G. and C.I.B.; investigation, K.N.C., J.F.G. and C.I.B.; resources, C.I.B.; data curation, K.N.C.; writing—original draft preparation, K.N.C.; writing—review and editing, J.F.G. and C.I.B.; visualization, K.N.C. and C.I.B.; supervision, J.F.G. and C.I.B.; project administration, C.I.B.; funding acquisition, J.F.G. and C.I.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: This work was carried out in the equipment provided by the theoretical RNomics group at the Universidad Nacional de Colombia. J.F.G. and C.I.B. thank to DIEB-UNAL for financial support. This work and the computational analysis were partially supported by the equipment donation from the German Academic Exchange Service-DAAD to the Faculty of Science at the Universidad Nacional de Colombia.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Region	Countries (n seqs)
Africa	Uganda (10), Cape verde (3), Cetral African Republic (3), Guinea (1), Nigeria (1), Senegal (6)
Asia_Cont	China (20), Japan (4), South Korea (1), Taiwan (2), India (1)
Asia_Southeast	Cambodia (2), Indonesia (1), Malaysia (3), Philippines (1), Singapore (57), Thailand (14)
Bl_Brazil	Brazil (54), Argentina (1), Ecuador (3)
Bl_Caribbean	Cuba (2), Dominican Republic (12), USA (34), French Guiana (2), Canada (2), Haiti (11), Guadeloupe (7), Puerto Rico (16), Suriname (3), Martinique(1)
Bl_Colombia	Colombia (40), Panama (10), Peru (2)
Bl_Mexico	Mexico (35), Honduras (14), Nicaragua (16), Guatemala (1)

Oceania	Australia (1), French Polynesia (13)

Table A1. Origin countries of the sequences included in their respective geographical subregions found in the study.

References

1. Wang, L.; Valderramos, S.; Wu, A.; Ouyang, S.; Li, C.; Brasil, P.; Bonaldo, M.; Coates, T.; Nielsen-Saines, K.; Jiang, T.; Ali-yari, R.; Cheng, G. From Mosquitos to Humans: Genetic Evolution of Zika Virus. Cell Host & Microbe 2016, 19, 561-565. doi:10.1016/j.chom.2016.04.006
2. Oehler, E.; Watrin, L.; Larre, P.; Leparc-Goffart, I.; Lastère, S.; Valour, F.; Baudouin, L.; Mallet, H.; Musso, D.; Ghawche, F. Zika virus infection complicated by Guillain-Barré syndrome – case report, French Polynesia, December 2013. Eu-ro-surveillance 2014, 19. doi:10.2807/1560-7917. ES2014.19.9.20720
3. Fauci, A.; Morens, D. Zika Virus in the Americas — Yet Another Arbovirus Threat. New England Journal of Medicine 2016, 374, 601-604. DOI: 10.1056/NEJMp1600297
4. Ventura, C.; Maia, M.; Bravo-Filho, V.; Góis, A.; Belfort, R. Zika virus in Brazil and macular atrophy in a child with mi-crocephaly. The Lancet 2016, 387, 228.doi: 10.1016/S0140-6736(16)00006-4.
5. Metsky, H.; Matranga, C.; Wohl, S.; Schaffner, S.; Freije, C.; Winnicki, S.; West, K.; Qu, J.; Baniecki, M.; Gladden-Young, A.; Lin, A.; Tomkins-Tinch, C.; Ye, S.; Park, D.; Luo, C.; Barnes, K.; Shah, R.; Chak, B.; Barbosa-Lima, G.; Delatorre, E.; Vieira, Y.; Paul, L.; Tan, A.; Barcellona, C.; Porcelli, M.; Vasquez, C.; Cannons, A.; Cone, M.; Hogan, K.; Kopp, E.; Anzinger, J.; Garcia, K.; Parham, L.; Ramírez, R.; Montoya, M.; Rojas, D.; Brown, C.; Hennigan, S.; Sabina, B.; Scotland, S.; Gangavarapu, K.; Grubaugh, N.; Oliveira, G.; Robles-Sikisaka, R.; Rambaut, A.; Gehrke, L.; Smole, S.; Halloran, M.; Villar, L.; Mattar, S.; Lorenzana, I.; Cerbino-Neto, J.; Valim, C.; Degrave, W.; Bozza, P.; Gnirke, A.; Andersen, K.; Isern, S.; Michael, S.; Bozza, F.; Souza, T.; Bosch, I.; Yozwiak, N.; MacInnis, B.; Sabeti, P. Zika virus evolution and spread in the Americas. Nature 2017, 546, 411-415. doi: 10.1038/nature22402
6. Weaver, S.; Costa, F.; Garcia-Blanco, M.; Ko, A.; Ribeiro, G.; Saade, G.; Shi, P.; Vasilakis, N. Zika virus: History, emergence, biology, and prospects for control. Antiviral Research 2016, 130, 69-80. DOI:10.1016/j.antiviral.2016.03.010.
7. Musso, D.; Gubler, D. Zika Virus. Clinical Microbiology Reviews 2016, 29, 487-524. DOI: 10.1128/CMR.00072-15
8. Petersen, L.; Jamieson, D.; Powers, A.; Honein, M. Zika Virus. New England Journal of Medicine 2016, 374, 1552-1563. 10.1056/NEJMra1602113
9. Kuno, G.; Chang, G. Full-length sequencing and genomic characterization of Bagaza, Kedougou, and Zika viruses. Ar-chives of Virology 2007, 152, 687-696. https://doi.org/10.1007/s00705-006-0903-z
10. Rodenhuis-Zybert, I.; Wilschut, J.; Smit, J. Dengue virus life cycle: viral and host factors modulating infectivity. Cellular and Molecular Life Sciences 2010, 67, 2773-2786. doi: 10.1007/s00018-010-0357-z.
11. Romero-López, C.; Berzal-Herranz, A. Unmasking the information encoded as structural motifs of viral RNA genomes: a potential antiviral target. Reviews in Medical Virology 2013, 23, 340-354. doi: 10.1002/rmv.1756
12. Fernández-Sanlés, A.; Ríos-Marco, P.; Romero-López, C.; Berzal-Herranz, A. Functional Information Stored in the Con-served Structural RNA Domains of Flavivirus Genomes. Frontiers in Microbiology 2017, 08. Doi:10.3389/fmicb.2017.00546
13. Manokaran, G.; Finol, E.; Wang, C.; Gunaratne, J.; Bahl, J.; Ong, E.; Tan, H.; Sessions, O.; Ward, A.; Gubler, D.; Harris, E.; Garcia-Blanco, M.; Ooi, E. Dengue subgenomic RNA binds TRIM25 to inhibit interferon expression for epidemiological fitness. Science 2015, 350, 217-221. doi: 10.1126/science.aab3369
14. Akiyama, B.; Laurence, H.; Massey, A.; Costantino, D.; Xie, X.; Yang, Y.; Shi, P.; Nix, J.; Beckham, J.; Kieft, J. Zika virus produces noncoding RNAs using a multi-pseudoknot structure that confounds a cellular exonuclease. Science 2016, 354, 1148-1152. doi: 10.1126/science.aah3963
15. Coutard, B.; Barral, K.; Lichière, J.; Selisko, B.; Martin, B.; Aouadi, W.; Lombardia, M.; Debart, F.; Vasseur, J.; Guillemot, J.; Canard, B.; Decroly, E. Zika Virus Methyltransferase: Structure and Functions for Drug Design Perspectives. Journal of Virology 2016, 91. http://dx.doi.org/10.1128/JVI.02202-16.
16. Villordo, S.; Gamarnik, A. Genome cyclization as strategy for flavivirus RNA replication. Virus Research 2009, 139, 230-239. doi: 10.1016/j.virusres.2008.07.016.
17. Sadri Nahand, J.; Bokharaei-Salim, F.; Karimzadeh, M.; Moghoofei, M.; Karampoor, S.; Mirzaei, H.; Tabibzadeh, A.; Jafari, A.; Ghaderi, A.; Asemi, Z.; Mirzaei, H.; Hamblin, M. MicroRNAs and exosomes: key players in HIV pathogenesis. HIV Medicine 2020, 21, 246-278. https://doi.org/10.1111/hiv.12822
18. Bruscella, P.; Bottini, S.; Baudesson, C.; Pawlotsky, J.; Feray, C.; Trabucchi, M. Viruses and miRNAs: More Friends than Foes. Frontiers in Microbiology 2017, 8. doi:10.3389/fmicb.2017.00824
19. Sadri Nahand, J.; Bokharaei-Salim, F.; Karimzadeh, M.; Moghoofei, M.; Karampoor, S.; Mirzaei, H.; Tabibzadeh, A.; Jafari, A.; Ghaderi, A.; Asemi, Z.; Mirzaei, H.; Hamblin, M. MicroRNAs and exosomes: key players in HIV pathogenesis. HIV Medicine 2020, 21, 246-278. https://doi.org/10.1111/hiv.12822

20. Bernier, A.; Sagan, S. The Diverse Roles of microRNAs at the Host–Virus Interface. Viruses 2018, 10, 440. https://dx.doi.org/10.3390/v10080440
21. Scheel, T.; Luna, J.; Liniger, M.; Nishiuchi, E.; Rozen-Gagnon, K.; Shlomai, A.; Auray, G.; Gerber, M.; Fak, J.; Keller, I.; Bruggmann, R.; Darnell, R.; Ruggli, N.; Rice, C. A Broad RNA Virus Survey Reveals Both miRNA Dependence and Functional Sequestration. Cell Host & Microbe 2016, 19, 409-423. Doi:https://dx.doi.org/10.1016/j.chom.2016.02.007
22. Mishra, R.; Kumar, A.; Ingle, H.; Kumar, H. The Interplay Between Viral-Derived miRNAs and Host Immunity During Infection. Frontiers in Immunology 2020, 10. doi:10.3389/fimmu.2019.03079.
23. Li, P.; Wei, Y.; Mei, M.; Tang, L.; Sun, L.; Huang, W.; Zhou, J.; Zou, C.; Zhang, S.; Qin, C.; Jiang, T.; Dai, J.; Tan, X.; Zhang, Q. Integrative Analysis of Zika Virus Genome RNA Structure Reveals Critical Determinants of Viral Infectivity. Cell Host & Microbe 2018, 24, 875-886.e5. https://doi.org/10.1016/j.chom.2018.10.011
24. National Center for Biotechnology Information (NCBI)[Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988]–[cited 2020 Dec 06]. Available from: https://www.ncbi.nlm.nih.gov/
25. Pickett, B.; Sadat, E.; Zhang, Y.; Noronha, J.; Squires, R.; Hunt, V.; Liu, M.; Kumar, S.; Zaremba, S.; Gu, Z.; Zhou, L.; Larson, C.; Dietrich, J.; Klem, E.; Scheuermann, R. ViPR: an open bioinformatics database and analysis resource for virology research. Nucleic Acids Research 2011, 40, D593-D598. doi: 10.1093/nar/gkr859.
26. Altschul, S.; Gish, W.; Miller, W.; Myers, E.; Lipman, D. Basic local alignment search tool. Journal of Molecular Biology 1990, 215, 403-410. doi: 10.1016/S0022-2836(05)80360-2.
27. Stajich, J. The Bioperl Toolkit: Perl Modules for the Life Sciences. Genome Research 2002, 12, 1611-1618.
28. RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/
29. Sievers, F.; Wilm, A.; Dineen, D.; Gibson, T.; Karplus, K.; Li, W.; Lopez, R.; McWilliam, H.; Remmert, M.; Söding, J.; Thompson, J.; Higgins, D. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology 2011, 7, 539. doi:10.1038/msb.2011.75
30. Okonechnikov, K.; Golosova, O.; Fursov, M. Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics 2012, 28, 1166-1167. doi:10.1093/bioinformatics/bts091
31. Charif, D; Lobry, J. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In: Bastolla U, Porto M, Roman H, Vendruscolo M (eds.), Structural approaches to se-quence evolution: Molecules, networks, populations. New York. Biological and Medical Physics, Springer Verlag; 2007. p. 207-232. ISBN : 978-3-540-35305-8. 2007
32. Schliep, K. phangorn: phylogenetic analysis in R. Bioinformatics 2010, 27, 592-593. https://doi.org/10.1093/bioinformatics/btq706
33. Paradis, E.; Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 2018, 35, 526-528. doi: 10.1093/bioinformatics/bty633
34. Guindon, S.; Dufayard, J.; Lefort, V.; Anisimova, M.; Hordijk, W.; Gascuel, O. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Systematic Biology 2010, 59, 307-321. https://doi.org/10.1093/sysbio/syq010
35. Geneious Prime.2020.available at (https://www.geneious.com)
36. Gruber, A.; Findeis, S.; Washielt, S.; Hofacker, I.; Stadler, P. RNAZ 2.0:. Biocomputing 2010 2009, 69-79. https://doi.org/10.1142/9789814295291_0009.
37. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. 2016. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org .
38. Kalvari, I.; Nawrocki, E.; Ontiveros-Palacios, N.; Argasinska, J.; Lamkiewicz, K.; Marz, M.; Griffiths-Jones, S.; Tof-fano-Nioche, C.; Gautheret, D.; Weinberg, Z.; Rivas, E.; Eddy, S.; Finn, R.; Bateman, A.; Petrov, A. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. Nucleic Acids Research 2020, 49, D192-D200. 10.1093/nar/gkaa1047
39. Grubaugh, N.; Ladner, J.; Kraemer, M.; Dudas, G.; Tan, A.; Gangavarapu, K.; Wiley, M.; White, S.; Thézé, J.; Magnani, D.; Prieto, K.; Reyes, D.; Bingham, A.; Paul, L.; Robles-Sikisaka, R.; Oliveira, G.; Pronty, D.; Barcellona, C.; Metsky, H.; Baniecki, M.; Barnes, K.; Chak, B.; Freije, C.; Gladden-Young, A.; Gnirke, A.; Luo, C.; MacInnis, B.; Matranga, C.; Park, D.; Qu, J.; Schaffner, S.; Tomkins-Tinch, C.; West, K.; Winnicki, S.; Wohl, S.; Yozwiak, N.; Quick, J.; Fauver, J.; Khan, K.; Brent, S.; Reiner, R.; Lichtenberger, P.; Ricciardi, M.; Bailey, V.; Watkins, D.; Cone, M.; Kopp, E.; Hogan, K.; Cannons, A.; Jean, R.; Monaghan, A.; Garry, R.; Loman, N.; Faria, N.; Porcelli, M.; Vasquez, C.; Nagle, E.; Cummings, D.; Stanek, D.; Rambaut, A.; Sanchez-Lockhart, M.; Sabeti, P.; Gillis, L.; Michael, S.; Bedford, T.; Pybus, O.; Isern, S.; Palacios, G.; Andersen, K. Genomic epidemiology reveals multiple introductions of Zika virus into the United States. Nature 2017, 546, 401-405. https://doi.org/10.1038/nature22400
40. Faria, N.; Quick, J.; Claro, I.; Thézé, J.; de Jesus, J.; Giovanetti, M.; Kraemer, M.; Hill, S.; Black, A.; da Costa, A.; Franco, L.; Silva, S.; Wu, C.; Raghwani, J.; Cauchemez, S.; du Plessis, L.; Verotti, M.; de Oliveira, W.; Carmo, E.; Coelho, G.; Santelli, A.; Vinhal, L.; Henriques, C.; Simpson, J.; Loose, M.; Andersen, K.; Grubaugh, N.; Somasekar, S.; Chiu, C.; Muñoz-Medina, J.; Gonzalez-Bonilla, C.; Arias, C.; Lewis-Ximenez, L.; Baylis, S.; Chieppe, A.; Aguiar, S.; Fernandes, C.; Lemos, P.; Nasci-mento, B.; Monteiro, H.; Siqueira, I.; de Queiroz, M.; de Souza, T.; Bezerra, J.; Lemos, M.; Pereira, G.; Loudal, D.; Moura, L.; Dhalia, R.; França, R.; Magalhães, T.; Marques, E.; Jaenisch, T.; Wallau, G.; de Lima, M.; Nascimento, V.; de Cerqueira, E.; de Lima, M.; Mascarenhas, D.; Neto, J.; Levin, A.; Tozetto-Mendoza, T.; Fonseca, S.; Mendes-Correa, M.; Milagres, F.;

Segu-rado, A.; Holmes, E.; Rambaut, A.; Bedford, T.; Nunes, M.; Sabino, E.; Alcantara, L.; Loman, N.; Pybus, O. Establishment and cryptic transmission of Zika virus in Brazil and the Americas. Nature 2017, 546, 406-410. doi: 10.1038/nature22401.

41. Finol, E.; Ooi, E. Evolution of Subgenomic RNA Shapes Dengue Virus Adaptation and Epidemiological Fitness. SSRN Electronic Journal 2018. doi: 10.2139/ssrn.3295648
42. Göertz, G.; Abbo, S.; Fros, J.; Pijlman, G. Functional RNA during Zika virus infection. Virus Research 2018, 254, 41-53. doi: 10.1016/j.virusres.2017.08.015
43. Jerzak, G.; Bernard, K.; Kramer, L.; Ebel, G. Genetic variation in West Nile virus from naturally infected mosquitoes and birds suggests quasispecies structure and strong purifying selection. Journal of General Virology 2005, 86, 2175-2183. doi: 10.1099/vir.0.81015-0
44. Filomatori, C. A 5' RNA element promotes dengue virus RNA synthesis on a circular genome. Genes & Development 2006, 20, 2238-2249. doi: 10.1101/gad.1444206
45. Lodeiro, M.; Filomatori, C.; Gamarnik, A. Structural and Functional Studies of the Promoter Element for Dengue Virus RNA Replication. Journal of Virology 2008, 83, 993-1008. doi: 10.1128/JVI.01647-08
46. Zhou, Y.; Ray, D.; Zhao, Y.; Dong, H.; Ren, S.; Li, Z.; Guo, Y.; Bernard, K.; Shi, P.; Li, H. Structure and Function of Flavivirus NS5 Methyltransferase. Journal of Virology 2007, 81, 3891-3903. doi: 10.1128/JVI.02704-06
47. Zhang, B.; Dong, H.; Zhou, Y.; Shi, P. Genetic Interactions among the West Nile Virus Methyltransferase, the RNA-Dependent RNA Polymerase, and the 5′ Stem-Loop of Genomic RNA. Journal of Virology 2008, 82, 7047-7058. doi: 10.1128/JVI.00654-08
48. Bustos-Arriaga, J.; Gromowski, G.; Tsetsarkin, K.; Firestone, C.; Castro-Jiménez, T.; Pletnev, A.; Cedillo-Barrón, L.; Whitehead, S. Decreased accumulation of subgenomic RNA in human cells infected with vaccine candidate DEN4Δ30 increases viral susceptibility to type I interferon. Vaccine 2018, 36, 3460-3467.https://doi.org/10.1016/j.vaccine.2018.04.087
49. Filomatori, C.; Carballeda, J.; Villordo, S.; Aguirre, S.; Pallarés, H.; Maestre, A.; Sánchez-Vargas, I.; Blair, C.; Fabri, C.; Morales, M.; Fernandez-Sesma, A.; Gamarnik, A. Dengue virus genomic variation associated with mosquito adaptation defines the pattern of viral non-coding RNAs and fitness in human cells (accessed Feb 23, 2021). http://dx.doi.org/10.1371/journal.ppat.1006265
50. Guedes, D.; Paiva, M.; Donato, M.; Barbosa, P.; Krokovsky, L.; Rocha, S.; Saraiva, K.; Crespo, M.; Rezende, T.; Wallau, G.; Barbosa, R.; Oliveira, C.; Melo-Santos, M.; Pena, L.; Cordeiro, M.; Franca, R.; Oliveira, A.; Peixoto, C.; Leal, W.; Ayres, C. Zika virus replication in the mosquito Culex quinquefasciatus in Brazil. Emerging Microbes & Infections 2017, 6, 1-11. doi:10.1038/emi.2017.59
51. Kawai, Y.; Nakayama, E.; Takahashi, K.; Taniguchi, S.; Shibasaki, K.; Kato, F.; Maeki, T.; Suzuki, T.; Tajima, S.; Saijo, M.; Lim, C. Increased growth ability and pathogenicity of American- and Pacific-subtype Zika virus (ZIKV) strains compared with a Southeast Asian-subtype ZIKV strain. PLOS Neglected Tropical Diseases 2019, 13, e0007387. https://doi.org/10.1371/journal.pntd.0007387
52. Schnettler, E.; Sterken, M.; Leung, J.; Metz, S.; Geertsema, C.; Goldbach, R.; Vlak, J.; Kohl, A.; Khromykh, A.; Pijlman, G. Noncoding Flavivirus RNA Displays RNA Interference Suppressor Activity in Insect and Mammalian Cells. Journal of Virology 2012, 86, 13486-13500. doi: 10.1128/JVI.01104-12