Taylor & Francis
Taylor & Francis Group

Phylogenesys and homology modeling in Zika virus epidemic: food for thought

Silvia Angeletti, Alessandra Lo Presti, Marta Giovanetti, Alba Grifoni, Massimo Amicosante, Marco Ciotti, Luiz-Carlos J. Alcantara, Eleonora Cella & Massimo Ciccozzi

View supplementary material

Accepted author version posted online: 10 Sep 2016.
Published online: 27 Sep 2016.

Submit your article to this journal

Article views: 66

View related articles

View Crossmark data

Phylogenesys and homology modeling in Zika virus epidemic: food for thought

Silvia Angeletti[1], Alessandra Lo Presti[2], Marta Giovanetti[3,4], Alba Grifoni[5], Massimo Amicosante[5,6], Marco Ciotti[7], Luiz-Carlos J. Alcantara[3], Eleonora Cella[2,8], Massimo Ciccozzi[1,2]

[1]Unit of Clinical Pathology and Microbiology, University Campus Bio-Medico, Rome, Italy, [2]Department of Infectious, Parasitic, and Immune-Mediated Diseases, Epidemiology Unit, Reference Centre on Phylogeny, Molecular Epidemiology, and Microbial Evolution (FEMEM), National Institute of Health, Rome, Italy, [3]Oswaldo Cruz Foundation (FIOCRUZ), Salvador, Brazil, [4]Department of Biology, University of Rome 'Tor Vergata', Rome, Italy, [5]ProxAgen Ltd, Sofia, Bulgaria, [6]Department of Biomedicine and Prevention, University of Rome "Tor Vergata", Rome, Italy, [7]Laboratory of Molecular Virology, Polyclinic Tor Vergata Foundation, Rome, Italy, [8]Department of Public Health and Infectious Diseases, Sapienza University of Rome, Rome, Italy

Zika virus (ZIKV) is an emerging Flavivirus that have recently caused an outbreak in Brazil and rapid spread in several countries. In this study, the consequences of ZIKV evolution on protein recognition by the host immune system have been analyzed. Evolutionary analysis was combined with homology modeling and T-B cells epitope predictions. Two separate clades, the African one with the Uganda sequence, as the most probable ancestor, and the second one containing all the most recent sequences from the equatorial belt were identified. Brazilian strains clustered all together and closely related to the French Polynesia isolates. A strong presence of a negatively selected site in the envelope gene (*Env*) protein was evidenced, suggesting a probable purging of deleterious polymorphisms in functionally important genes. Our results show relative conservancy of ZIKV sequences when envelope and other non-structural proteins (NS3 and NS5) are analyzed by homology modeling. However, some regions within the consensus sequence of NS5 protein and to a lesser extent in the envelope protein, show localized high mutation frequency corresponding to a considerable alteration in protein stability. In terms of viral immune escape, envelope protein is under a higher selective pressure than NS5 and NS3 proteins for HLA class I and II molecules. Moreover, envelope mutations that are not strictly related to T-cell immune responses are mostly located on the surface of the protein in putative B-cell epitopes, suggesting an important contribution of B cells in the immune response as well.

Keywords: Zika, Evolutionary analysis, Homology modeling, T-B cells epitope predictions

Introduction

Zika virus (ZIKV) is an emerging threat in South and Central America and the Caribbean, rapidly spreading around novel regions. First discovered in a Ugandan forest in 1947, ZIKV is the newest discovery among the different Flaviviridae viruses.[1–3] Being part of Arboviruses, ZIKV is transmitted by its main vector *Aedes aegypti* mosquito even though *Aedes albopictus* mosquitoes might be capable to transmit the virus as well. ZIKV epidemic in Brazil has currently estimated 440,000–1,300,000 cases and it spread further to more than 13 additional countries.[4,5] ZIKV infected people are generally asymptomatic. When symptomatic disease occurs symptoms last between a few days and one week and they are characterized by fever, maculopapular rash, arthralgia, or non-purulent conjunctivitis. In addition to these symptoms, in the current Brazil outbreak, ZIKV has shown potential correlation with infants' microcephaly and fetal losses in infected pregnant women as well as Guillain-Barre syndrome.[2,6–9]

The ZIKV genome consists of a single-stranded positive sense RNA molecule of 10794 kb in length encoding a polyprotein cleaved into capsid (C), precursor membrane (prM), envelope (E), and non-structural proteins (NS1-NS2A-NS2B-NS3-NS4A-NS4B-NS5).[10] ZIKV is closely related to Dengue virus (DENV) and serological tests may be not specific because of possible cross reactions. Moreover, molecular tests such as polymerase chain reaction can distinguish ZIKV from other arboviruses but Zika specific tests are not available yet. Although ZIKV presents several similarities to DENV infection, so far no studies have addressed the immunological aspects. A

better understanding of the viral protein structure and its modifications represents a key agenda in understanding the difference and similarities with DENV, allowing the design of ZIKV-specific immunological tests, particularly useful for large population screening, as well as for vaccination and drug design approaches.

In this report, the consequences of ZIKV evolution on protein recognition by the host immune system have been analyzed in the context of selective pressure.[11,12] To this end, phylogenetic and evolutionary analysis were combined with homology modeling analysis and T-B cells epitope predictions.

Material and methods
Phylogenetic and evolutionary analysis
Phylogenetic and evolutionary analysis were performed on Env gene as the most variable ones. ML phylogenetic tree was generated with the HKY + I + G model of nucleotide substitution, by using Phyml v 3.0. The evolutionary model was chosen as the best-fitting nucleotide substitution model in accordance with the results of the hierarchical likelihood ratio test implemented in Modeltest software version 3.7.[17] The statistical robustness and reliability of the branching order within the phylogenetic trees were confirmed by the bootstrap analysis and considering as significant statistical support a bootstrap value >70%.

Selective pressure analysis
The dN/dS rate (ω) was estimated by the ML approach implemented in the program HyPhy.[18] In particular, the global (assuming a single selective pressure for all branches) and the local (allowing the selective pressure to change along every branch) models were compared by likelihood ratio test. Two different algorithms: the fixed effects likelihood, which fits an ω rate to every site and uses the likelihood ratio to test if dN = dS; and random effect likelihood, a variant of the Nielsen–Yang approach,[19] which assumes that a discrete distribution of rates exists across sites and allows both dS and dN to vary independently site by site, as described in more detail elsewhere.[17,20] In order to select sites under selective pressure and keep our test conservative, a p value of ≤0.1 or a posterior probability of ≥0.9 as relaxed critical values was assumed.[20]

Homology modeling
To study the amino acid mutation frequency, three different Zika proteins have been analyzed: the Envelope protein and the non-structural proteins NS3 and NS5.

Three different data-sets were built for Envelope (59 sequences), NS3 (17 sequences) and NS5 (108 sequences) proteins, respectively. All the sequences were obtained from the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/). Sequences below 500 bp in length have been excluded from the study. Partial cds and complete genome sequences have been considered with an overall number of 140 (Table S1, supplemental material).

The sampling dates for the sequences in the data-set ranged from 1997 to 2014. All the sequences were aligned using ClustalX software and edited with non-conservative editing by Bio-Edit software v. 7.0, as already described.[21] Alignment has been performed using a complete nucleotide sequence of ZIKV MR766-NIID strain (Accession Number: LC002520.1) as reference sequence evaluating all the amino acid mutation frequencies and positions. Envelope, NS3, and NS5 consensus sequences generated after the alignment, have been modeled using Phyre2 server intensive mode.[22] All the resulting models obtained have been 'repaired' to obtain best protein quality and each mutation has been considered independently to evaluate protein change stability using FoldX tools implemented in YASARA.[23] B-cell epitope predictions have been performed on the modeled Envelope protein using NeutraCorp tool.[24]

T and B cells epitope prediction
T-cell epitope prediction for both HLA class I and HLA class II have been performed using IEDB MHC Binding Prediction tools and the allele reference set on Env, NS3, and NS5 consensus sequences obtained after alignment.[25–30] Each peptide is considered binding peptide with an IEDB RANK ≤ 0.5 for HLA class I and an IEDB RANK ≤ 1.0 for HLA class II. For each putative T-cell epitope, the number of mutations within the epitope and with a frequency higher than 0.1 is calculated for Envelope, NS3, and NS5 proteins. All the data are expressed as mean ± standard deviation of the mean (SD), and frequency as appropriate. Comparisons between means are made by paired test using GraphPad Prism 5.0 (San Diego, CA).

Results
Phylogenetic analysis
Maximum likelihood (ML) phylogenetic tree showed the presence of two distinct genotypes in *Env* gene: African and Asiatic genotype (Fig. 1). Two statistically supported clades appeared. Clade I typically represented by African sequences and Clade II represented by sequences of Asiatic and Brazilian origin (Fig. 1). The Brazilian sequences were closely related to a sequence from French Polynesia. All of the clusters were statistically supported by the bootstrap values >70% (Fig. 1).

Evolutionary analysis
Selection pressure analyses performed in the Envelope gene uncovered about 60% of sites under strong negative selection (by HYPHY) indicated by $\omega < 0$ suggesting high degree conservation in the genomic region analyzed (Table S1, supplemental material). Estimates of the transition/transversion rate ratio (ts/tv) are quite homogeneous among models and thus are not shown in Table S2. The average ω ratio ranged from 0.0555 to 0.0674 among all models, suggesting that a non-synonymous mutation has only 5.55–6.74% as much chance as a synonymous mutation of being fixed in the population.
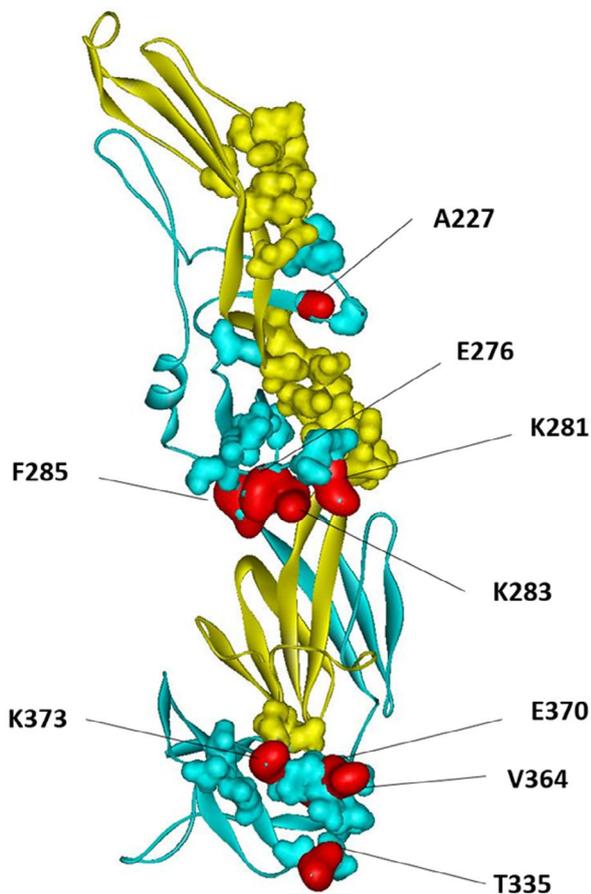
Figure 1 Mutation analysis for Envelope B-cell predicted epitopes.

Notes: Envelope protein has been modelled entirely and after B-cell prediction, the terminal portion after amino acid position 381 has been removed. In yellow are shown amino acid residues belonging to amino acid position 1–162 that have not been considered in sequence analysis. Surface VdW analysis is shown for B-cell predicted epitopes falling entirely or partially in Env amino acid positions 163–381 (for full list see Table S3 supplemental material). In red are shown mutated amino acid positions belonging to predicted B-cell epitope.

Homology modeling

In order to compare differences in Zika proteins in respect to DENV protein, the main target of DENV immune response NS5, NS3 for T-cell immune response and Envelope for B-cell immune response have been selected for further studies.[13,14] Considering the mutation frequency of the three different genes examined, NS5 protein presented the highest number of mutated amino acid positions (47 out of 236, 19.91%, Fig. 2*a*), of which 28 with capability to stabilize the protein structure and only six with destabilizing effect (Table S3 supplemental material). The Envelope protein presented a similar frequency of mutation (41 out of 219, 18.72%, Fig. 2*b*), of which 25 with stabilizing effect and only 3 with destabilizing effect on protein structure (Table S3, supplemental material). Finally, NS3 presented a lower frequency of mutation compared with the other analyzed ZIKV proteins (10 out of 256, 3.9%, Fig. 2*c*), with three presenting a stabilizing effect and

seven with a destabilizing effect on protein structure (Table S3, supplemental material).

Considering only the mutations in the protein sequences with a frequency higher than 0.1, of the 16/47 mutated positions in NS5 (counting for the total of 34.0426% of the observed frequency mutations) nine determined an increased protein stability, i.e. a general protein fitness, while for the other seven positions the selective pressure might be due to different reasons (Table S3, supplemental material). Similarly in the Envelope protein, of the 10/41 mutations with frequency above 0.1 (counting for the total of 24.39% of the observed frequency mutations) seven determined an increase in the protein stability (Fig. 1) (Table S3, supplemental material). Finally, the NS3 protein is presenting 4/10 positions with frequency of mutation above 0.1 (counting for the total of 40% of the observed frequency mutations). Of these, only one is presenting a stabilizing effect on the protein structure while three positions have a destabilizing effect and the mutation is maintained for reasons other than selective.

T and B cells epitope prediction

T-cell epitope prediction analysis has been performed on the three proteins considering for each epitope, the number of mutations observed to determine a potential selective pressure exerted by the immune response. Beside the presence of a higher number of mutations in NS3 and NS5 proteins, the Envelope protein shows the highest number of mutations within the putative epitopes. This difference is significantly higher in respect to NS5 and NS3 proteins both for epitopes restricted to HLA class I and HLA class II molecules (Fig. 3).

Finally, B-cell epitope prediction has been performed in the Envelope protein to evaluate the effect of the observed mutations on B-cell epitopes. In the B-cell epitope prediction, Env protein model has been considered in its overall structure in order to avoid the prediction of an epitope that is masked by the rest of the sequence. Once the B-cell epitopes have been predicted, only epitopes that fully or partially contain amino acid positions 163–381 have been considered, those which were having an overall good quality according to the homology server used for the analysis and were in agreement with the analyzed sequences (Fig. 4 and Table S4, supplemental material). Four out of the six protein surface patches predicted to have characteristics of B-cell epitopes were included in the region analyzed for mutations. Nine out of 54 (17%) amino acid residues of the B-cell epitopes presented mutations. In particular, epitope #1 presented 1/15 mutated amino acid positions, epitope #4 4/14, epitope #5 3/13 and epitope #6 2/12 (Table S4, supplemental material).

Discussion

ZIKV, was accidentally discovered in Uganda in 1947, during a course of surveillance on mosquitoes and primates and until now remained an obscure virus confined to a narrow
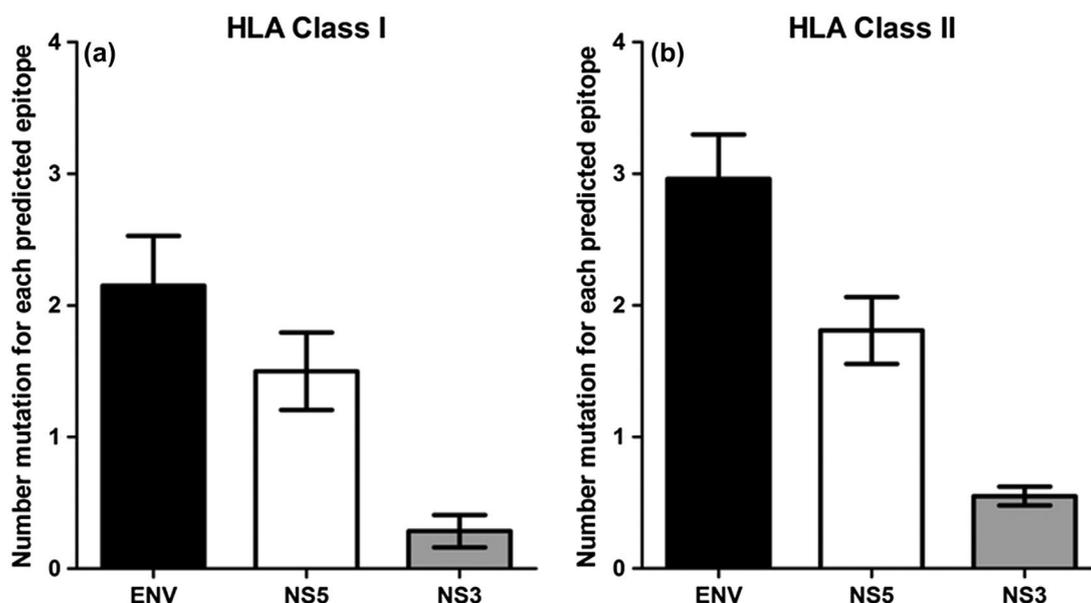
Figure 2 Mutation analysis for Envelope (ENV), NS5, and NS3 T-cell predicted epitopes in the context of HLA class I (a) and HLA class II (b) molecules.
Notes: Predictions' system has been made using consensus sequences and allele reference set provided by IEDB. For each putative epitope, the number of mutation has been considered.
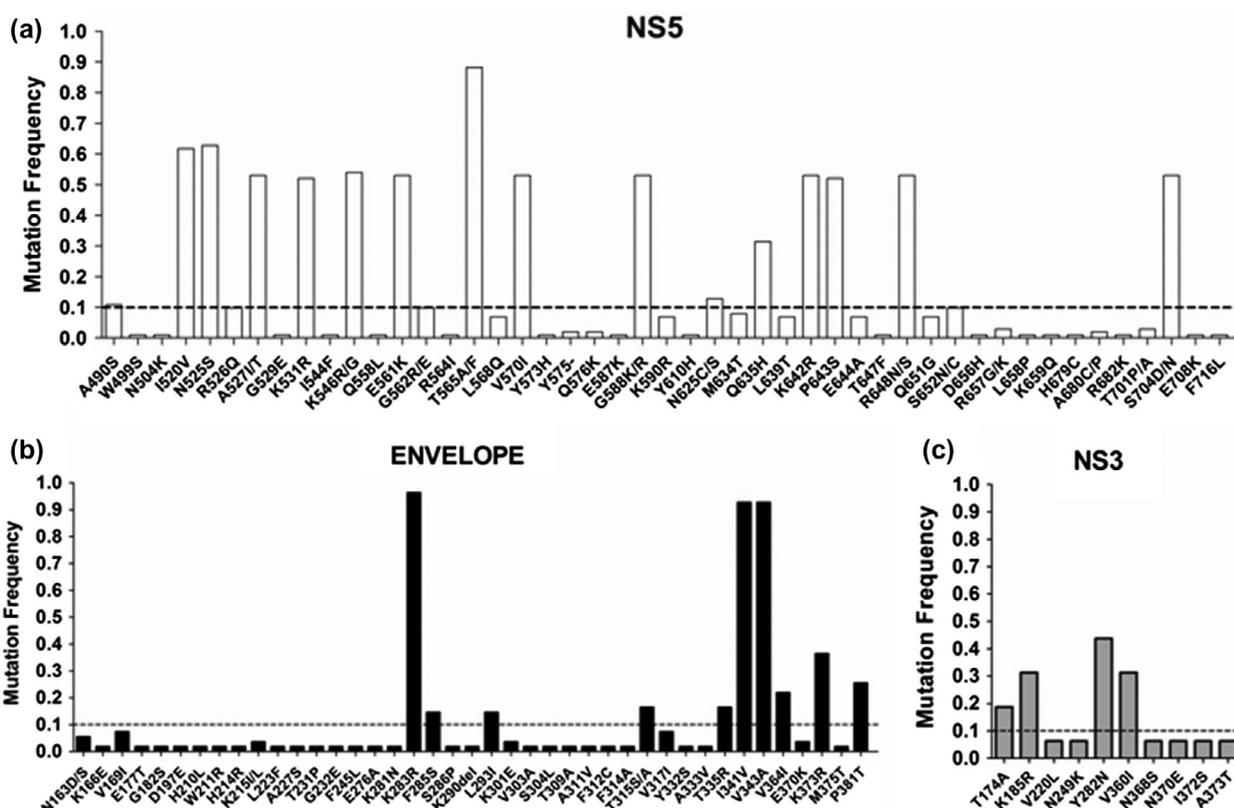

Figure 3 Mutation frequency analysis for Zika proteins NS5 (a), Envelope (b), and NS3 (c).
Notes: For each protein, mutated positions and mutation frequency as well as amino acid change in that positions are shown. All mutation frequency that are above 0.1 (see dotted line) are considered to be significant.

equatorial belt. Zika has now circled the globe, arriving in the Americas and in September, in the country of Cape Verde in West Africa.[4] ZIKV is the last of the long list of emerging Flaviviruses that have recently received global attention due to the outbreak in Brazil and the rapid spread in several countries after the first detection in May 2015 in Brazil.[1–3]

The phylogenetic results indicate two separate clades, the African one (clade I) containing the Uganda sequence as the most probable ancestor of ZIKV, and the clade II which contains all the most recent sequences isolated from the equatorial belt. In this last clade, it is evident how in the recent past Cambodian strains diverged from the
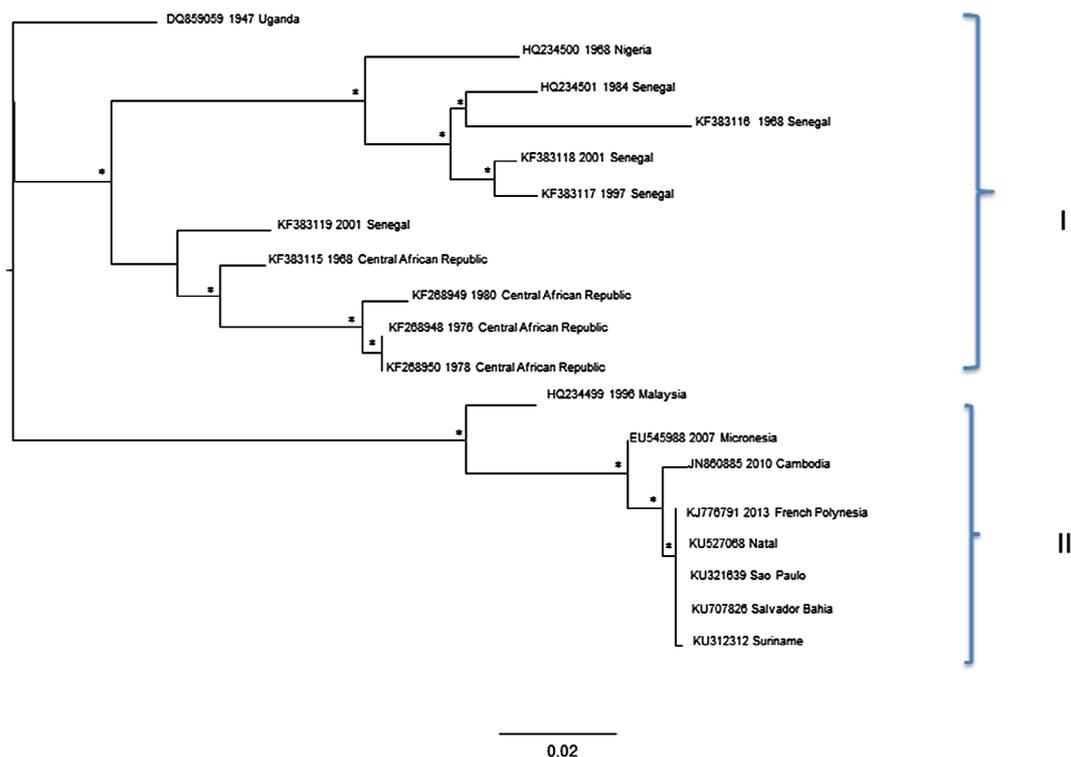
Figure 4   Phylogenetic trees of Env gene sequences of ZIKV. Notes: Phylogenetic analysis of the Env gene sequences of Zika virus. The tree was mid-point rooted. Scale bar is in units of nucleotide substitutions per site. Asterisks represent bootstrap values >90%. The GenBank accession number, year of isolation, and country of origin of each isolate are indicated on the tips of the trees for all strains except for those identified in 2015 and 2016.

Malaysian ones. It is probable that the Cambodian strain was recently introduced or circulated in the past in this region but remained undetected until 2010, the date of isolation. Interestingly, the Brazilian isolates clustered all together and appeared closely related to the French Polynesia isolates, indicating a probable relationship with strains causing outbreaks in other regions but in the same tropical belt. It has been hypothesized that the virus was introduced in Brazil during the World Cup Soccer Tournament, even though none of the teams participating in the event were from Pacific countries. An alternative hypothesis is that the virus was introduced during the Championship canoeing race in Rio de Janeiro in 2014 where the French Polynesia team participated.

The selective pressure analysis performed in this study, revealed a strong presence of negatively selected sites in the Env gene protein suggesting a probable purging of deleterious polymorphisms in functionally important genes. On the other hand, the contemporary lack of presence of positively selected sites also suggests and enforces the idea of highly adapted phenotypes. From a functional point of view, the alternation cycle between arthropod vector and human and no human hosts may impose a sort of barrier to non-synonymous mutations in important genes.

Among the factors that may influence ZIKV infection, pre-existing immunity induced by cross-reactive viruses and antigenic variability of the virus may have an important role. The latter one increases the virus's capability to

escape the human immune response. Amino acid sequence conservancy within ZIKV sequences can be considered in combination with potential immune escape amino acid mutations. Our results show relative conservancy of ZIKV sequences when Envelope and other non-structural proteins (NS3 and NS5) are analyzed by homology modeling. However, some regions within the consensus sequence of NS5 protein and to a lesser extent in the Envelope protein show localized high mutation frequency corresponding to a considerable alteration in protein stability. In terms of viral immune escape, the Envelope protein is under a higher selective pressure than NS5 and NS3 proteins for HLA class I and II molecules. Moreover, Envelope mutations that are not strictly related to T-cell immune response are mostly located on the surface of the protein in putative B-cell epitopes, suggesting an important contribution of B cells in the immune response as well.

ZIKV has a lot of shadow points of concern. The first one is that Zika epidemics sometimes follows chikungunya epidemics as happened in 2013 when a Zika epidemic followed the chikungunya spread from west to east. The second one is that ZIKV could change vector (*A. aegypti*) as described for chikungunya virus. Actually, in the recent past, the non-synonymous mutation A226V allowed chikungunya virus[15] to change vector from *A. aegypti* to *A. albopictus*. If a similar eventuality happened for ZIKV this could lead to the spread of the virus in regions not previously affected. Furthermore, the mild or unapparent

dengue-like disease does not help to recognize the infection and this could facilitate the spread of the virus. Indeed, serological diagnosis of ZIKV infection remains unreliable due to cross reactions with other flaviviruses. Thus, specific, validated, CE-IVD marked real-time PCRs are urgently needed.

A massive increase in the incidence of fetuses with microcephaly has been noticed in regions with a high number of ZIKV infections.[16] It is important to confirm the eventual association between ZIKV infection in pregnant women and microcephaly with appropriate epidemiological and case control studies in order to avoid unjustified alarm and psychological damage among pregnant women living in endemic regions.

Finally, vaccine development is another important issue. To the best of our knowledge, no ZIKV vaccines are now under development. Immune-informatic studies like ours highlighting the main aspects of ZIKV sequence variability might be used to select appropriate conserved regions to use in vaccine development.

Supplemental material

Supplemental data for this article can be accessed http://dx.doi.org/10.1080/20477724.2016.1235337.

Disclosure statement

The authors report no declaration of interest.

References

1 Calvet GA, Filippis AM, Mendonça MC, Sequeira PC, Siqueira AM, Veloso VG, et al. First detection of autochthonous Zika virus transmission in a HIV-infected patient in Rio de Janeiro, Brazil. J Clin Virol. 2016;74:1–3.

2 Gatherer D, Kohl A. Zika virus: a previously slow pandemic spreads rapidly through the Americas. J Gen Virol. 2016;97(2):269–73.

3 Kelser EA. Meet dengue's cousin. Zika. 2016;18(3):163–6.

4 Campos GS, Bandeira AC, Sardi SI. Zika virus outbreak, Bahia, Brazil. Emerg Infect Dis. 2015;21(10):1885–6.

5 Rodriguez-Morales AJ. Zika: the new arbovirus threat for Latin America. J Infect Dev Countries. 2015;9(6):684-5.

6 Higgs S. Zika virus: emergence and emergency. Vector Borne Zoonotic Dis. 2016;16(2):75–6.

7 Petersen EE, Staples JE, Meaney-Delman D, Fischer M, Ellington SR, Callaghan WM, et al. Interim guidelines for pregnant women during a Zika virus outbreak – United States, 2016. MMWR Morb Mortal Wkly Rep. 2016;65:30–3.

8 Tetro JA. Zika and microcephaly: causation, correlation, or coincidence? Microbes Infect. 2016;18(3):167–8.

9 Ventura CV, Maia M, Bravo-Filho V, Góis AL, Belfort R, Jr. Zika virus in Brazil and macular atrophy in a child with microcephaly. Lancet. 2016;387(10015):228.

10 Faye O, Freire CC, Iamarino A, Faye O, de Oliveira JV, Diallo M, et al. Molecular evolution of Zika virus during its emergence in the 20(th) century. PLoS Negl Trop Dis. 2014;8(1):e2636.

11 Ashfaq UA, Ahmed B. De novo structural modeling and conserved epitopes prediction of Zika virus envelop protein for vaccine development. Viral Immunol. 2016 Jul 20. Epub ahead of print.

12 Cox BD, Stanton RA, Schinazi RF. Predicting Zika virus structural biology: challenges and opportunities for intervention. Antivir Chem Chemother. 2016 Jun 13. pii: 2040206616653873. Epub ahead of print.

13 Weiskopf D, Angelo MA, de Azeredo EL, Sidney J, Greenbaum JA, Fernando AN, et al. Comprehensive analysis of dengue virus-specific responses supports an HLA-linked protective role for CD8+ T cells. Proc Natl Acad Sci USA. 2013 May 28;110(22):E2046–53.

14 Wahala WM, Silva AM. The human antibody response to dengue virus infection. Viruses. 2011 Dec;3(12):2374–95.

15 Lo Presti A, Ciccozzi M, Cella E, Lai A, Simonetti FR, Galli M, et al. Origin, evolution, and phylogeography of recent epidemic CHIKV strains. Infect Genet Evol. 2012;12(2):392–8.

16 Bell BP, Boyle CA, Petersen LR. Preventing Zika virus infections in pregnant women: an urgent public health priority. Am J Public Health. 2016;106(4):589–90.

17 Azarian T, Lo Presti A, Giovanetti M, Cella E, Rife B, Lai A, et al. Impact of spatial dispersion, evolution, and selection on Ebola Zaire virus epidemic waves. Sci Rep. 2015;5:10170.

18 Pond SL, Frost SD, Muse SV. HyPhy: hypothesis testing using phylogenies. Bioinformatics. 2005;21(5):676–9.

19 Nielsen R, Yang Z. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics. 1998;148(3):929–36.

20 Kosakovsky Pond SL, Frost SD. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol. 2005;22(5):1208–22.

21 Ciccozzi M, Vujošević D, Lo Presti A, Mugoša B, Vratnica Z, Lai A, et al. Genetic diversity of HIV type 1 in Montenegro. AIDS Res Hum Retroviruses. 2011;27(8):921–4.

22 Kelley LA, Sternberg MJ. Protein structure prediction on the web: a case study using the Phyre server. Nat Protoc. 2009;4(3):363–71.

23 Van Durme J, Delgado J, Stricher F, Serrano L, Schymkowitz J, Rousseau F. A graphical interface for the FoldX forcefield. Bioinformatics. 2011;27(12):1711–2.

24 Patronov A, Pirrò S, Grifoni A, Ilieva KM, Gallo C, Del Duca A, et al. 2016. NeutraCorp: a new tool for cross-model B-cell epitope prediction.

25 Bui HH, Sidney J, Peters B, Sathiamurthy M, Sinichi A, Purton KA, et al. Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications. Immunogenetics. 2005;57(5):304–14.

26 Nielsen M, Lundegaard C, Lund O. Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. BMC Bioinformatics. 2007;8:238–249.

27 Nielsen M, Lundegaard C, Worning P, Lauemøller SL, Lamberth K, Buus S, et al. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. Protein Sci. 2003;12(5):1007–17.

28 Peters B, Sette A. Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method. BMC Bioinformatics. 2005;6:132.

29 Salimi N, Fleri W, Peters B, Sette A. The immune epitope database: a historical retrospective of the first decade. Immunology. 2012;137(2):117–23.

30 Sturniolo T, Bono E, Ding J, Raddrizzani L, Tuereci O, Sahin U, et al. Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. Nat Biotechnol. 1999 Jun;17(6):555–61.