Evidence for homologous recombination in Chikungunya Virus

Pablo E. Casal a, Diego Chouhy a,b, Elisa M. Bolatti b, Germán R. Perez a, Emma J. Stella b, Adriana A. Giri a,b,*

aVirology Area, School of Biochemistry and Pharmaceutical Sciences, National University of Rosario, Suipacha 531, 2000 Rosario, Argentina
bHuman Virology Group, IBR-CONICET/National University of Rosario, Suipacha 570, 2000 Rosario, Argentina

A R T I C L E   I N F O

Article history:
Received 4 November 2014
Revised 27 January 2015
Accepted 28 January 2015
Available online 18 February 2015

Keywords:
Chikungunya Virus
Bayesian phylogeny
Complete ORFs/E1 gene
Nucleotide identity classification
Recombination
East/Central/South African clade

A B S T R A C T

Chikungunya Virus (CHIKV), a mosquito-transmitted alphavirus, causes acute fever and joint pain in humans. Recently, endemic CHIKV infection outbreaks have jeopardized public health in wider geographical regions. Here, we analyze the phylogenetic associations of CHIKV and explore the potential recombination events on 152 genomic isolates deposited in GenBank database. The CHIKV genotypes [West African, Asian, East/Central/South African (ECSA)], and a clear division of ECSA clade into three sub-groups (I–II–III), were defined by Bayesian analysis; similar results were obtained using E1 gene sequences. A nucleotide identity-based approach is provided to facilitate CHIKV classification within ECSA clade. Using seven methods to detect recombination, we found a statistically significant event (p-values range: 1.14 × 10⁻⁷–4.45 × 10⁻²⁴) located within the nsP3 coding region. This finding was further confirmed by phylogenetic networks (PHI Test, p = 0.004) and phylogenetic tree incongruence analysis. The recombinant strain, KJ679578/India/2011 (ECSA III), derives from viruses of ECSA III and ECSA I. Our study demonstrates that recombination is an additional mechanism of genetic diversity in CHIKV that might assist in the cross-species transmission process.

© 2015 Elsevier Inc. All rights reserved.

1. Introduction

Chikungunya Virus (CHIKV; Togaviridae: Alphavirus) is an arthropod-borne virus transmitted by Aedes (Ae.) mosquitoes which was first isolated from a Tanzanian outbreak in 1952 (Ross, 1956). The virus transmission cycle requires infection of female mosquitoes via a viraemic blood meal taken from a susceptible vertebrate host and, following a suitable extrinsic incubation period, transmission to another vertebrate host during subsequent feeding (Schuffenecker et al., 2009). After an incubation period, most patients suffer from polyarthralgia and myalgia, with a significant impact on their quality of life. In some patients, minor hemorrhagic signs such as epistaxis or gingivorrhagia have also been reported (Schuffenecker et al., 2006).

In Africa, the virus is maintained in a mainly sylvatic cycle, being spread among wild primates by forest dwelling mosquitoes (Powers and Logue, 2007). In contrast, in Asia the virus spread between humans and the primary vector Ae. aegypti (Chevillon et al., 2008). CHIKV was first registered in Thailand in 1958 (Ng and Hapuarachchi, 2010). Large epidemics were recorded throughout Asia in countries including Cambodia, Vietnam, Burma, Sri Lanka, India, Indonesia and the Philippines, before the virus virtually disappeared following the 1973 outbreak in India (Chevillon et al., 2008). A large outbreak in Kenya in 2004 initiated a resurgence of the virus leading to widespread infection in the Indian Ocean islands of the Comoros, Seychelles, Mauritius and the French islands of Mayotte and La Reunion. The epidemiology of the virus changed, with the major vector on La Reunion identified as Ae. albopictus (Ng and Hapuarachchi, 2010).

CHIKV is an enveloped virus with an 11.8 kb genome consisting of a 5’ capped positive sense single-strand RNA that harbors a poly (A) tail in its 3’ end. The genome is composed of two open reading frames (ORFs) embedded between 5’ and 3’ untranslated regions (UTR). Enclosed between the ORFs there is another UTR sequence, the Junction region (J). The ORF located at the 5’ end of the genome encodes a polyprotein precursor of nonstructural proteins (nsP1, nsP2, nsP3, nsP4) with replicative and proteolytic activities. The second ORF encodes the polyprotein precursor of the structural proteins (C, E1, E2) (Chevillon et al., 2008).

CHIKV was classified into three lineages according to phylogenetic analyses based on partial E1 gene (Powers et al., 2000) and full-genome sequences of the virus (Volk et al., 2010; Cui et al., 2011): West African (WA), Asian and East/Central/South African (ECSA) (Powers et al., 2000; Schuffenecker et al., 2006). Nevertheless, most of the reported phylogenetic analyses were based on limited numbers of CHIKV genomes or partial sequences.
Many recently emerged human diseases are caused by RNA viruses that display active recombination with a major impact on their evolution, emergence and epidemiology. Indeed, one of the best examples of recombination and viral emergence is found in the Alphavirus genus: the Western equine encephalitis virus, which was generated through recombination between a Sindbis-like virus and an Eastern equine encephalitis virus-like virus (Hahn et al., 1988). However, the attempts to find evidence of recombination in CHIKV have failed until now (Volk et al., 2010; Cui et al., 2011; Wanlapakorn et al., 2014).

Here we analyze the phylogenetic associations and explore the potential recombination events on 152 genomic isolates deposited in GenBank database.

2. Materials and methods

2.1. Sequence dataset

A total of 161 nucleotide (nt) sequences of CHIKV complete genomes were obtained from the GenBank, dated on September 2014.Cloning vector strains, vaccine sequences, as well as high-passage strains were excluded from the analysis (GenBank: AF490259; GenBank: EF452494; GenBank: HM045795; GenBank: HM045798; GenBank: HM045797; GenBank: KC149887; GenBank: KC149888; GenBank: KC149889; GenBank: L37661). This led to a dataset of 152 CHIKV sequences. Supplementary Table summarizes the complete information regarding strain names, isolation source, collection dates and isolation countries. Strains in the text and figures are labeled as follows: GenBank Accession Number/Isolation country/Collection Date (year).

2.2. Phylogenetic analysis

Due to the ambiguous alignments in the 3′ and 5′ UTR, only the concatenated ORFs were used in further analysis. Multiple-sequence alignments were created using CLUSTAL W (Larkin et al., 2007) according to amino acid sequence alignments (Gonnet protein weight matrix) to preserve codon homology, and manually edited in MEGA v6 (Tamura et al., 2013). Phylogenetic trees were inferred by the Bayesian method using the BEAST package v1.7.5 (Drummond and Rambaut, 2007). To do this, Markov Chain Monte Carlo (MCMC) simulations were performed during 10^7 generations, sampling one state every 1000 generations, with a burnin of 10%. The JModelTest v2 (Darriba et al., 2012) determined that GTR +1 +I was the best fit evolutionary substitution model for each run. Statistical convergence of MCC was assessed by calculating the effective sample size using TRACER v1.4 (available from http://beast.bio.ed.ac.uk/Tracer). The maximum clade credibility tree across all of the plausible trees generated by BEAST was then computed using the TreeAnnotator program available in the BEAST package.

2.3. Recombination analysis

Determination of potential recombinant events, parental sequences and localization of recombinant breakpoints were performed triplet-by-triplet using the Recombination Detection Program (RDP): 4.1 (Martin et al., 2010; Martin and Rybicki, 2000), and the algorithms embedded in it: GENECOV (Padidan et al., 1999), BootScan (Martin et al., 2005), MaxChi (Smith, 1992), CHIMAERA (Posada and Crandall, 2001), SILScan (Gibbs et al., 2000), and 3Seq (Boni et al., 2007). The sequences were set to linear. p-values ≤ 0.05 were regarded statistically significant. All other parameters were set as default RDP settings.

Phylogenetic networks and phylogenetic tree incongruence analyses were performed in order to confirm the potential recombinant events obtained with RDP program. Unrooted phylogenetic networks and Pairwise Homoplasy Index (PHI) Test (Bruen et al., 2006) were obtained using the SplitsTree software version 4.13.1 (Huson and Bryant, 2006). For tree incongruence analyses, Bayesian method was performed as described above.

2.4. Computational resources

This research was performed in part through the computational resources provided by the High Performance Computing (HPC) Cluster available at the “Centro Científico Tecnológico CONICET-Rosario” (http://cluster.rosario-conicet.gov.ar), Rosario City, Santa Fe province, Argentina.

3. Results

3.1. Phylogenetic analysis of 152 CHIKV complete ORFs

Like previous findings using partial CHIKV E1 sequences (Powers et al., 2000) or CHIKV genomes (Volk et al., 2010; Cui et al., 2011), our phylogenetic tree with 152 complete ORFs included three distinct CHIKV clades (Fig. 1): ECSA, WA, and Asian. In addition, the ECSA clade showed a clear division into three subgroups: ECSA I, ECSA II and ECSA III. The recent Indian Ocean basin outbreak (ECSA III) formed a monophyletic clade descendant from ECSA II (Volk et al., 2010), and therefore, both sub-groups shared the most recent common ancestor. In turn, the ECSA II/III subgroups shared a common ancestor with ECSA I. The same tree topology was found when applying the Bayesian method to the E1 gene sequences from the same strains (Supplementary Fig. 2).

Based on phylogenetic relationships (Fig. 1) we proposed a nucleotide identity-based approach to facilitate the classification of CHIKV strains belonging to the ECSA clade. As shown in Fig. 2, sequences showing nucleotide identities ≥98.5% may belong to the same ECSA subgroup. On the other hand, sequences showing nucleotide identities ≤98.5% may belong to different ECSA subgroups. The same cut-off value can be applied for CHIKV classification using complete ORFs or E1 gene sequences, obtaining similar results.

3.2. Recombination analysis

Interestingly, the strain KJ679578/India/2011 (GenBank: KJ679578) showed a divergent pattern with respect to the other strains grouped within the ECSA III subgroup (Fig. 1). To further investigate this finding, we used RDP to determine possible events of recombination. This analysis provided statistically significant evidence for recombination (Fig. 3, top), with all methods included in RDP (p-values ranging from 1.14 × 10^{-5} to 4.45 × 10^{-24}). The CHIKV isolate KJ679578/India/2011 (GenBank: KJ679578) was recognized as the recombinant genome, being the major and minor viral parents the previously described isolates EF012359/Mauritiu/2006 (GenBank: EF012359) (ECSA III) (Edwards et al., 2007) and HM045821/Senegal/1963 (GenBank: HM045821) (ECSA I) (Volk et al., 2010), respectively. Nucleotide sequence similarities between the recombinant strain and its parents were of 97.4% (minor parent) and of 99.3% (major parent). No major amino acid differences among the recombination site were observed. The recombination breakpoints extended from nt 4640 to nt 5275 in the recombinant sequence, representing a 635 nt fragment within the nonstructural protein 3 (nsP3) coding region. The recombinant fragment is located 220 nt downstream of the nsP3 macro domain (Malet et al., 2009), in a region whose function remains unknown.
The analysis of the major parent sequence EF012359/Mauritius/2006 (GenBank: EF012359) showed the characteristic E1 envelope protein mutation A226V (E1-A226V), a first-step adaptation that has been shown to enhance replication in Ae. albopictus (Edwards et al., 2007; Tsetsarkin et al., 2007). In contrast, neither the minor parent sequence nor the recombinant strain had the E1-A226V
mutation. The second-step Ae. albopictus-adaptive substitutions (E2-K252Q and E2-L210Q) (Tsetsarkin et al., 2014) were absent in all strains.

Another useful strategy to analyze recombination events is to generate phylogenetic trees from different parts of the genome in order to visualize potential phylogenetic incongruences. Based on this, the 152 CHIKV complete ORF alignment was divided into three regions in order to obtain three independent trees: REGION 1, before the recombination event; REGION 2, between the recombination breakpoints; and REGION 3, after the recombination event. As depicted in Fig. 3 (bottom), the recombinant strain and the major parent clustered together in the ECSA III sub-group, both in REGION 1 and REGION 3. However, in REGION 2 the recombinant virus shifted from ECSA III to ECSA I sub-group, clustering with the minor parent, and confirming the close association between the recombinant virus and the minor viral parent only in this region.

Finally, evidence for recombination was provided by SplitsTree analyses using nsP3 sequence alignments. As shown in Fig. 4A, a conflicting phylogenetic signal was found between isolates included in ECSA I (HM045821/Senegal/1963 (GenBank: HM045821)) and ECSA III (EF012359/Mauritius/2006 (GenBank: EF012359)), giving rise to the recombinant isolate KJ679578/India/2011 (GenBank: KJ679578). The reticulate topology observed disappeared when the phylogenetic network was performed removing the ECSA I subgroup from the alignment (Fig. 4B). The same was true when, retaining the recombinant strain, the complete ECSA III sub-group was removed (data not shown). The PHI Test for Recombination ($p = 0.004$) confirmed that a recombinant signal was present within the sequences, and validated the visual evidence of recombination obtained from the phylogenetic network approach. Taken together, these results provide statistically significant data supporting the existence of homologous recombination in CHIKV.

4. Discussion

In less than 10 years, CHIKV has spread from the coast of Kenya throughout the Indian Ocean and has emerged in new areas, causing millions of cases of disease in over 50 countries (Ng and Hapuarachchi, 2010). Phylogenetic analysis showed that an ECSA genotype virus was responsible for the epidemics on the islands in the Indian Ocean (Schuffenecker et al., 2006). Despite an overall low mutation rate, the gradual accumulation of important genetic changes in ECSA genotype viruses has mainly caused the changed disease epidemiology (Ng and Hapuarachchi, 2010). To further complicate the scenario, we report the first evidence of homologous recombination between viruses of different subgroups within the ECSA clade as an additional mechanism of genetic diversity of CHIKV.

In agreement with Cui et al. (2011), three subgroups within the ECSA clade were found, with most of the epidemic strains being included in the ECSA III subgroup (Fig. 1). E1 gene trees obtained with the Maximum Likelihood method have previously shown different topologies with respect to those based on complete ORFs, especially for the ECSA clade (Volk et al., 2010). We obtained the same results when applying the Maximum Likelihood method on the E1 gene from 152 sequences (data not shown). However, when the E1 gene tree was generated by the Bayesian method (Supp Fig. 3), no significant differences were observed with respect to the tree obtained with the CHIKV complete ORFs. Therefore, the E1 gene is adequate to resolve the phylogenetic history of CHIKV with acceptable precision if generated by the Bayesian method. This is not a minor issue if we consider that many CHIKV strains have been characterized using the E1 region as target gene in epidemiological studies (Chusri et al., 2014; Das et al., 2010; Powers et al., 2000). In line with this, the cut-off value provided in this work by the nucleotide identity-based approach (Fig. 2) could be a useful tool to facilitate and speed up CHIKV classification, using E1 gene or complete ORF sequences within the ECSA clade.

Because recombinants are the progeny of at least two parental strains, traditional phylogenetic trees are insufficient to represent the evolutionary history of, or evolutionary relationships among, isolates including recombinants. Exploring potential events of recombination in the 152 sequence dataset with RDP (Fig. 3, top),
phylogenetic incongruence analyses (Fig. 3, bottom), and SplitsTree (Fig. 4) we found a statistically significant event located within the nsP3 coding region. The strain KJ679578/India/2011 (GenBank: KJ679578) was the recombinant virus candidate and was isolated from the serum of a patient visiting Calcutta (India) according to the GenBank datasheet. Previously, three research groups looked for homologous recombination in CHIKV without success (Volk et al., 2010; Cui et al., 2011; Wanlapakorn et al., 2014), mainly because the recombinant strain sequence here identified was released by the GenBank in July 2014.

Fig. 3. Characterization of a recombination event in CHIKV. RDP analysis of recombinant strain KJ679578/India/2011 (GenBank: KJ679578) (top). The recombination methods used, with the corresponding obtained p-values, are indicated above. The major and minor viral parents are depicted with red and green lines, respectively. The region between the recombination breakpoints is highlighted. Phylogenetic incongruence analysis (bottom). Bayesian phylogenetic trees from REGION 1 (nt positions 67–4639), REGION 2 (nt positions 4640–5275) and REGION 3 (nt positions 5276–11,303). The recombinant strain is indicated with an arrow; the major and minor parents are indicated with a square and a triangle, respectively. For clarity, WA and Asian clades were collapsed, as well as ECSA II and some parts of ECSA III sub-groups. WA: West African; ECSA: East/Central/South African. Posterior probability values are depicted beside the major clades. The complete phylogenetic trees can be found in Supplementary Fig. 3. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
Interestingly, the recombinant region in strain KJ679578/India/2011 (GenBank: KJ679578) was located within the nsP3 which is required, together with nsP1, nsP2 and nsP4, for the viral replication process. In particular, nsP3 has an RNA-binding capacity and is involved in negative strand RNA synthesis by facilitating the recruitment of RNA by other non-structural proteins (Solignat et al., 2009). Higher replication rates and increase in viral pathogenesis have been linked to the specific amino acid changes identified in nsP1, nsP2, and nsP3 in CHIKV (Schuffenecker et al., 2006) and other alphaviruses (Shin et al., 2012). If the recombinant CHIKV strain with a mosaic nsP3 has acquired an advantageous phenotype for spreading or adapting to new hosts, it certainly deserves further investigation.

Several studies have shown that the E1-A226V substitution increases the virus replication and dissemination rates in Ae. albopictus, shortening the extrinsic incubation period (Schuffenecker et al., 2006; Tsetsarkin et al., 2007). It has been recently demonstrated that the initial E1-A226V Ae. albopictus-adaptive CHIKV substitution has been followed by several second-step adaptive mutations in its envelope glycoprotein genes that further increased CHIKV fitness in this mosquito vector (Tsetsarkin et al., 2014). Our data indicate that the recombinant virus lacks the first- (E1-A226V) and second-step (E2-K252Q and E2-L210Q) Ae. albopictus-adaptive CHIKV substitutions, suggesting that the strain KJ679578/India/2011 (GenBank: KJ679578) has been transmitted mainly by Ae. aegypti mosquitoes.

Copy choice recombination occurs when two or more viruses infect the same host cell, and the RNA polymerase dissociates from the template strand and switches to a homologous template at a different genomic position during synthesis, while the nascent string remains (Lai, 1992). This process leads to the formation of chimeric molecules from parental genomes of mixed origin. In the case reported here, the temporal and geographical context in which the viral parents gave rise to the recombinant strain KJ679578/India/2011 (GenBank: KJ679578) is unknown. In fact, as shown in Supplementary Table, the major and minor parental strains were isolated at different times and from different hosts located in geographically separated regions. Nevertheless, our results indicate that the recombinant virus was originated from mixed infections produced by viruses of the ECSA I and III subgroups.
subgroups. Therefore, the strains here identified as the minor and major viral parents are not necessarily the ones that participated in the recombination process, but are closely related to them. Indeed, the same authors that submitted the recombinant sequence, have deposited a second isolated (GenBank: KJ679577) from the same place and collected around the same date, that shares 97.5% nucleotide identity with the recombinant strain. This sequence, however, had 99.9% nucleotide identity with the sequence here identify as the minor parent (HM045821/Senegal/1963), clustering together within the ECSA I subgroup. Due to the geographical area and the collection date, KJ679577 may be a more suitable candidate as the minor parent. Nevertheless, a more comprehensive study, with a larger number of isolates from the same place and similar collection dates is necessary to draw a more accurate conclusion.

Yet, the possibility that the detected recombinant could be a sequence assembly artefact cannot be excluded. Such artefacts are relatively common in sequences that are available in the Genbank, particularly in those that have been assembled either from next generation sequencing data or from directly sequenced PCR products (O’Rawe, 2015). Future efforts are needed to provide definitive evidence that the detected recombinant actually exists in the wild, by either the re-isolation of the recombinant strain from the original sample, or by the isolation of another recombinant sequence from a different host, that shares the same mosaic structure and clusters phylogenetically with the strain detected here.

Despite the fact that CHIKV has been infecting millions of people since its discovery, the number of CHIKV genome sequences isolated from natural infections of humans, vectors and non human hosts up to date are scarce, limited to less than 200 records. This situation reflects the critical lack of molecular data necessary to track the epidemiological and genetic path of the virus in both the enzootic and urban cycles.

5. Conclusions

We provide, for the first time, statistically significant data supporting the existence of homologous recombination in CHIKV. Considering that we found one recombinant virus out of a total of 152 genomes analyzed, we can speculate that the recombination process is more frequent than thought in this virus. Thus, recombination could assist CHIKV in the cross-species transmission process, allowing the virus to increase the likelihood of finding a genetic configuration that facilitates host adaptation. We also demonstrate that phylogenetic analysis of the E1 gene using the Bayesian method is a feasible tool to facilitate the CHIKV classification.
Lastly, the proposed nucleotide identity-based approach may aid in increasing the genetic information of this virus in order to elucidate the complexity of the vector-virus-environment interactions that play a key role in CHIKV evolution and therefore in the epidemiology of the disease.

Acknowledgments

We thank María Robson, Mariana de Sanctis, Geraldine Raimundo, Carola Perret and Romina Ricardo (Departamento de Inglés, Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario) for language correction and style.

Elisa M. Bolatti is supported by a PhD fellowship of CONICET, Germán R. Perez is supported by a PhD fellowship of ANPCyT (PICT 2012-0652) and Emma J. Stella is supported by a Post-Doctoral fellowship of ANPCyT (PICT 2008-0420).

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ympev.2015.01.016.

References