Infection, Genetics and Evolution 49 (2017) 164–173

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Research paper

New bunya-like viruses: Highlighting their relations

Alexandro Guterres [a,b,*], Renata Carvalho de Oliveira [a], Jorlan Fernandes [a], Elba Regina Sampaio de Lemos [a], Carlos Guerra Schrago [b]

[a] Laboratório de Hantaviroses e Rickettsioses, Instituto Oswaldo Cruz, Fundação Oswaldo Cruz, Rio de Janeiro, RJ, Brazil
[b] Departamento de Genética, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil

ARTICLE INFO

Article history:
Received 3 November 2016
Received in revised form 16 January 2017
Accepted 18 January 2017
Available online 19 January 2017

Keywords:
Bunyavirus supergroup
Arenaviruses
Bunyaviruses
Phleboviruses
Orthobunyaviruses
Tospoviruses
Evolutionary history

ABSTRACT

The standard virus classification scheme for arenaviruses and bunyaviruses shifted dramatically when several groups reported the detection and isolation of divergent groups of viruses in a variety of insect collections. Although these viral families can differ in terms of morphology, structure and genetics, recent findings indicate these viruses may have a shared evolutionary origin. To determine the phylogenetic relations among these families, we inferred phylogenetic trees using three methods. The Maximum Likelihood and Bayesian trees were rooted as suggested by the (molecular clock-rooted) BEAST tree. Our results highlight a noteworthy relation among these viral supergroups of different genome organizations. Our study suggests that the best scenario is the existence of at least three monophyletic supergroups, all of them well supported. The recent data indicate that these viruses are evolutionarily and genetically interconnected. While these supergroups appear to be closely related in our phylogenetic analysis, other viruses should be investigated in future research. In sum, our results also provide insights into the classification scheme, thereby providing a new perspective about the fundamental questions of family origins, diversity and genome evolution.

1. Introduction

Most human pathogenic bunyaviruses and arenaviruses cause severe hemorrhagic fevers with a high rate of fatalities. Increasing number of outbreaks and the possibility of cases spreading over international borders has led to increased interest in these viruses and their relations. The ongoing threat of emerging hemorrhagic diseases has made the search for reservoir species with a history of coevolution, for example, with the mammarenaviruses, and hantaviruses, a priority (Charrel and de Lamballerie, 2010; de Oliveira et al., 2014; Zapata and Salvato, 2013). The origin, phylogenetic relationships and evolutionary history of viral genomes is a classic problem that has inspired a long series of questions and hypotheses in evolutionary biology. Recently, sequence analyses of emerging viruses have shown that genes from arenaviruses are potentially homologous to other negative-strand viruses such as bunyaviruses and filoviruses (Carter et al., 2012; Gallaher et al., 2001).

Viruses in the family *Bunyaviridae* (the bunyaviruses) share several molecular characteristics. Based on their differences, the International Committee on Taxonomy of Viruses (ICTV) has classified them into five genera: *Orthobunyavirus*, *Hantavirus*, *Nairovirus*, *Phlebovirus* and *Tospovirus* (Elliott, 2009). In 1975, this family was proposed to encompass viruses with morphological and structural similarities, but with diverse life cycles. The *Bunyaviridae* genome is comprised of three negative-sense RNA segments (large, medium and small) that employ a variety of coding strategies to generate a limited set of structural and non-structural proteins (Schmaljohn and Nichol, 2007). The large (L) RNA of these viruses codes for the transcriptase and replicase proteins and large RNA-dependent RNA polymerase (RdRp or L protein). Glycoproteins are coded by the medium (M) RNA, which generates a polyprotein that is proteolytically processed. In viruses of some genera, a non-structural protein (NSm) of unknown function is also included. The small (S) RNA codes for the nucleocapsid protein and also for a non-structural protein in viruses of several genera (Elliot and Schmaljohn, 2013; Elliott, 2014). For instance, the L and S segments of the tick-borne Crimean Congo hemorrhagic fever virus (family *Bunyaviridae*, genus *Nairovirus*) encode a polymerase and a nucleocapsid with strong similarity to the Lassa virus (family *Arenaviridae*, genus *Mammarenavirus*) (Carter et al., 2012).

The first discovery of a "virus of experimental lymphocytic choriomeningitis", today well known as the lymphocytic choriomeningitis virus (LCMV), occurred in 1933 (Armstrong and Lillie, 1934). However, only in 1976 was the family *Arenaviridae* established to include the genus *Arenavirus* with LCMV and Tacaribe complexes recognized. Members of the monogeneric family

* Corresponding author at: Laboratório de Hantaviroses e Rickettsioses, Instituto Oswaldo Cruz – FIOCRUZ, Pavilhão Helio e Peggy Pereira – 1° pav. Sala B115, Av. Brasil 4365 – Manguinhos, Rio de Janeiro, RJ 21045-900, Brazil.
E-mail address: guterres@ioc.fiocruz.br (A. Guterres).

http://dx.doi.org/10.1016/j.meegid.2017.01.019
1567-1348/

Arenaviridae had been reported to infect only rodents, except in one case of bats and ticks (Radoshitzky et al., 2015; Sayler et al., 2014). This scenario shifted dramatically when several groups published the detection and isolation of divergent arenaviruses in captive snakes (Bodewes et al., 2013; Stenglein et al., 2012). The new virus is characterized by ambisense coding and its L and NP genes are homologous to those found in arenaviruses (Zapata and Salvato, 2013). However, the GP sequences are homologous to filovirus envelope glycoproteins (family *Filoviridae*), and the Z gene sequences are homologous to host ubiquitin ligase (Bodewes et al., 2013; Hetzel et al., 2013; Stenglein et al., 2012).

Therefore, arenaviruses are currently classified in two different genera: *Mammarenavirus* and *Reptarenavirus*. Arenaviruses possess single-stranded bi-segmented RNA genomes. Each of the two RNA segments codes for two non-overlapping reading frames of opposite polarities: the viral RNA-dependent RNA polymerase (L protein) and a zinc-binding matrix protein (Z protein) for the large (L) genomic segment. The nucleocapsid protein (NP) and the glycoprotein precursor (GPC) are secondarily cleaved into the envelope proteins G1 and G2 for the small (S) genomic segment (Pinschewer et al., 2003; Qi et al., 2010; Shtanko et al., 2010).

The paradigmatic phylogenetic relation scheme for arenaviruses and bunyaviruses shifted dramatically when several groups independently reported the detection and in some cases the isolation of a divergent group of viruses in a diverse collection of insects, spiders and other arthropods (Frey et al., 2016; Li et al., 2015; Marklewitz et al., 2015). Notably, these new negative-sense RNA viruses were found to be spread across the major lineages of the family *Bunyaviridae*. Although these viral families differ in structure or genetics, these recent findings indicate that these viruses may have a shared evolutionary origin. In this study, we explore key aspects of the evolution of these viruses, particularly their phylogenetic relations, highlighting a perspective in the viral classification scheme, diversity and genome evolution.

2. Materials and methods

2.1. Compiled sequence data

The genomic sequences used in the study were all retrieved from the GenBank® database of NCBI (http://www.ncbi.nlm.nih.gov/nuccore/), including the protein sequences of the full genome of the *Arenaviridae* and *Bunyaviridae* families. Within the *Arenaviridae* and *Bunyaviridae* families, we retrieved representative sequences of each genus due to the very high number of species. We added novel lineages of bunyaviruses that have been discovered in insects, spiders, centipedes and other arthropods that might lead to the establishment of at least eight new genera. We retrieved two genera, also related to bunyaviruses (*Emaravirus* and *Tenuivirus*), that are recognized by the ICTV but have not yet been assigned to a family.

Multiple sequence alignment (MSA) were performed using MAFFT version 7 employing the E-INS-i algorithm and TCOFFEE version 11 applying the PSI-Coffee algorithm (Katoh and Standley, 2013; Notredame et al., 2000). Additionally, we used COBALT, which is a protein multiple sequence alignment tool that finds a collection of pairwise constraints derived from a conserved domain database, the protein motif database, and sequence similarity using RPS-BLAST, BLASTP and PHI-BLAST (Papadopoulos and Agarwala, 2007). The sequence alignment was limited to conserved domains, with ambiguously aligned regions removed using TrimAl (Capella-Gutiérrez et al., 2009). We measured alignment confidence based on a Transitive Consistency Score (TCS) web server. The TCS makes it possible to estimate the local reliability of protein MSAs using the TCS index. The purpose of an alignment reliability index is to discriminate between correctly and incorrectly aligned residues. This evaluation can be used to identify the aligned positions most likely to contain structurally analogous residues, as judged from BAliBASE and PREFAB structure-based reference alignments, and is also most likely to support an accurate phylogenetic reconstruction (Chang et al., 2015; 2014).

2.2. Phylogenetic analyses

We estimated phylogenetic relations of the protein sequences from the three major open reading frames (RdRp, NP and glycoprotein) using (a) ML phylogenetic inference as implemented in PhyML 3 (Guindon and Gascuel, 2003) under the LG + G + I model of sequence evolution, and (b) a Bayesian Markov Chain Monte Carlo (MCMC) method as implemented in MrBayes v3.2.5 (Ronquist et al., 2012). The MCMC settings consisted of two simultaneous independent runs with four chains each that were run for 10 million generations and sampled every 100th generation, yielding 100,000 trees. After eliminating 10% of the samples as burn-in, a consensus tree was built. Statistical support of the clades was measured by a heuristic search with 1000 bootstrap replicates in PhyML (Anisimova and Gascuel, 2006) and the Bayesian posterior probabilities in MrBayes. For the Bayesian analyses, we used a mixed aa model of evolution with γ-shaped distribution of rates across sites. This model allows selection to be integrated across all best-fit models. The best-fit evolutionary model was determined using MEGA version 6.06, using the Bayesian Information Criterion (Tamura et al., 2013).

A rooted timetree (relaxed molecular clock) of amino acid (aa) sequences was inferred using the Bayesian MCMC method available in the BEAST v1.8.4 package. (Drummond and Rambaut, 2007) We modeled the evolutionary rate evolution along branches by an uncorrelated lognormal prior, and the topologies by the Bayesian Skyline tree prior. Two independent runs were undertaken with sampling every 1000 generations. In BEAST, the same evolutionary model was employed as described above. We used Tracer v.1.6 to check for convergence and adequate mixing (i.e., an estimated sample size >200 for all relevant parameters). The TreeAnnotator program was used to generate a Maximum Clade Credibility (MCC) tree after eliminating the first 10% of the sampled trees as burn-in. Because the MCC tree is automatically rooted in the assumption of a molecular clock, it is possible to determine which viral lineages are most likely to be the stem lineage. The stem clade estimated by the BEAST tree was then used as an outgroup to root the phylogenetic trees inferred in the ML and Bayesian phylogenetic analyses.

2.3. Criteria for demarcation of the supergroup

The most universally used sequence-based virus classification tool is phylogenetic analysis. About 70% of the families and floating genera described in the Ninth Report of the International Committee on the Taxonomy of Viruses (ICTV) are supported by phylogenetic trees (King et al., 2012). The chief characteristics of members are presented with phylogenetic analyses of selected genes to support their relations. To help define suitable phylogenetic criteria for relation demarcation, each supergroup was considered to form phylogenetic group when it was clustered into a statistically supported monophyletic clade stem (BEAST/ML/MrBayes, >0.9/>90/>0.9).

3. Results

3.1. Database and multiple sequence alignment

We retrieved a total of 397 sequences from GenBank®: 131 NPs, 119 glycoproteins and 147 L RdRp sequences (Table S1). The aa sequences were aligned using COBALT, MAFFT and TCOFFEE multiple sequence alignment (MSA) methods, and each alignment estimated the positions most likely to contain structurally analogous residues. Starting with the transitive consistency score (TCS) scheme, we selected the MSA to be performed in the MAFFT algorithm.