Infection, Genetics and Evolution 12 (2012) 309–314

Contents lists available at SciVerse ScienceDirect

Infection, Genetics and Evolution

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

Comparative evolutionary epidemiology of dengue virus serotypes

Raquel L. Costa [a,1], Carolina M. Voloch [b,1], Carlos G. Schrago [b,*]

[a] Laboratório Nacional de Computação Científica, Petrópolis, Brazil
[b] Departamento de Genética, Universidade Federal do Rio de Janeiro, Brazil

ARTICLE INFO

Article history:
Received 18 August 2011
Received in revised form 29 November 2011
Accepted 12 December 2011
Available online 27 December 2011

Keywords:
Dengue
Bayesian skyline plot
Timescale
Positive selection

ABSTRACT

Evolutionary studies on dengue virus have frequently focused on intra-serotype diversity or on specific epidemics. In this study, we compiled a comprehensive data set of the envelope gene of dengue virus serotypes and conducted an extensive comparative study of evolutionary molecular epidemiology. We found that substitution rates are homogeneous among dengue serotypes, although their population dynamics have differed over the past few years as inferred by Bayesian coalescent methods. On a global scale, DENV-2 is the serotype with the highest effective population size. The genealogies also showed geographical structure within the serotypes. Finally, we also explored the causes of dengue virus serotype diversification by investigating the plausibility that it was driven by adaptive changes. Our results suggest that the envelope gene is under significant purifying selection and the hypothesis that dengue virus serotype diversification was the result of stochastic events cannot be ruled out.

1. Introduction

Dengue is an arthropod-borne disease that affects 50–100 million people per year in tropical and subtropical regions, resulting in death rates between 0.03% and 1.4% (WHO, 2005). Dengue virus (DENV) is a positive-sense single-stranded RNA (ssRNA+) *Flavivirus* transmitted by mosquitos of the genus *Aedes*. Dengue virus exists as four antigenically different serotypes (DENV-1–4), with uncertain evolutionary origins (Holmes and Burch, 2000). Within the serotypes, several authors have also characterized genotypes, based on genetic diversity and geographical distribution (Araujo et al., 2009; Klungthong et al., 2008; Kukreti et al., 2009; Rico-Hesse, 1990; Vasilakis et al., 2007).

Evolutionary studies of DENV are abundant. However, works have focused on the inference of substitution rates and the intra-serotype phylogeny (Goncalvez et al., 2002; Lanciotti et al., 1997, 1994; Twiddy et al., 2002a). It has been shown that evolutionary lineages within serotypes generally present characteristic geographical structure that reflects the spatial dynamics of epidemics (Holmes, 2008; Zhang et al., 2005). Dengue virus substitution rates, mainly measured from the envelope protein gene (E), have been estimated using several methodological approaches and are comparable to the evolutionary rates of other RNA viruses (Twiddy et al., 2002b; Wang et al., 2000).

Epidemiological surveys of dengue were conducted throughout the last century and, until recently, these were the only data available for obtaining historical information about the spread of the dengue virus (Gubler, 1998). Such surveys are essential for designing effective public health policies. However, the genetic history of the virus is as important as traditional epidemiological surveillance for understanding the full picture of epidemics (Bennett et al., 2010). This is because genetic material harbors information that can be used to reconstruct the history of epidemics in time and space. Moreover, recent coalescence techniques have also permitted the inference of population parameters such as growth rate and the effective number of infections (Drummond et al., 2003).

Several works have applied these methods to tackle DENV evolution (Bennett et al., 2010; Mondini et al., 2009; Villabona-Arenas and Zanotto, 2011). However, these studies focused on specific serotypes and an extensive comparative analysis is still lacking. A comparative analysis of dengue serotypes is important because there is increasing evidence that their epidemic potentials are different and that episodes of secondary infection with different serotypes frequently result in higher morbidity rates caused by the severe forms of the infection (Guzman et al., 2007).

Moreover, the causes of dengue serotype diversification are still unclear (Twiddy et al., 2002a). It has been suggested that the four serotypes originated from independent zoonotic passage to humans (Vasilakis et al., 2011). If this was so, we would expect that the DENV lineages have adapted to the new host and, consequently, there would be signatures of positive selection along the DENV genome. However, studies involving the estimation of d_N/d_S ratios on the envelope gene revealed extensive purifying selection (Holmes, 2003; Klungthong et al., 2004; Zhang et al., 2006).

* Corresponding author. Address: Departamento de Genética, A2-092, Inst. Biologia, CCS, Universidade Federal do Rio de Janeiro, Ilha do Fundão, Rio de Janeiro, RJ 21941-617, Brazil. Tel.: +55 21 4063 8278; fax: +55 21 2562 6397.
E-mail address: guerra@biologia.ufrj.br (C.G. Schrago).
[1] These authors contributed equally to this work.

1567-1348/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.meegid.2011.12.011

The present study aimed to perform a comprehensive comparative analysis of the history of dengue virus serotypes as inferred from a large number of publicly available sequences from the last few decades. By using coalescent-based approaches in a Bayesian framework, we assessed the demographic dynamics of epidemics throughout the last few centuries and estimated evolutionary parameters relevant for understanding the evolutionary aspects of DENV serotype emergence. We have also investigated the occurrence of adaptive molecular evolution on the emergence of DENV serotypes as well as the heterogeneity of selective forces along the envelope gene.

2. Materials and methods

2.1. Data set

We have downloaded all envelope gene sequences, which codes for the envelope protein, available in GenBank that contained information on the collection date, geographical location and that comprised the whole gene sequence. Then, a data set was composed of 596 sequences sampled to obtain the maximum geographical and chronological diversity (Table 1). This was achieved by excluding sequences collected in the same year from the same geographical location. Recombining sequences were excluded from the data set after inspection in the RDP3 program (Martin et al., 2010), using the RDP method (Martin and Rybicki, 2000). This sampling was performed to permit the implementation of sophisticated evolutionary analyses, which are unfeasible to be run on large amount of sequences. The data set covers a temporal range of 64 years (1944–2008) from 71 countries and consists of the largest DENV data set compiled to date. Sequences were aligned in ClustalW and inspected in MEGA 4.1. The final alignment included 1485 nucleotides. A list of access numbers is provided as supplementary information. The alignment used is available from the authors upon request. Analyses were conducted on each serotype independently and on the full data set.

2.2. Evolutionary rates and genealogical analysis

The comparative analysis of the DENV serotypes consisted of inferring the genealogy of the serotypes and the historical dynamics of the epidemics. We also investigated the action of differential selective forces along the envelope gene. Nucleotide substitution model choice was performed in HyPhy (Pond et al., 2005) and the GTR + G was selected since it significantly increased the likelihood of the data. All coalescent-based analyses were conducted in BEAST 1.6.1 (Drummond and Rambaut, 2007), which implements a Bayesian inference of evolutionary parameters via the Markov chain Monte Carlo (MCMC) method. The chronology of the dengue virus diversification was estimated using the relaxed molecular clock, in which the priors for the evolution of rates among branches were assumed to follow an uncorrelated lognormal distribution. The Bayesian skyline was the tree topology coalescent prior applied, since there is no indication whether the effective population size of DENV serotypes follow simple demographic models (i.e., constant, exponential or logistic growth). During MCMC, the parameters were visited for 50,000,000 generations and sampled every thousand cycles, resulting in 50,000 topologies and associated parameters. Of these, 10% were discarded as burn-in. Convergence of the MCMC run was checked in TRACER by inspecting the trace plots and calculating the effective sample sizes. To reconstruct the dengue virus demographic history, we employed the Bayesian skyline plot (BSP) model (Drummond et al., 2003, 2005), which generates piecewise constant population size trajectories. In order to test the robustness of the estimates of effective population sizes of serotypes to the number of sequences analyzed, we have composed an additional data set in which DENVs-1, 2 and 3 samples were reduced to 90 randomly chosen sequences. Note that this is also the number of sequences present in the smallest data set (DENV-4).

Table 1
Characteristics of DENV data set used in this study.

Serotype	Number of sequences	Number of countries	Time intervals
DENV-1	139	39	1944–2008
DENV-2	206	47	1944–2008
DENV-3	161	37	1956–2008
DENV-4	90	28	1956–2008

2.3. Differential selection along codon sites

Analysis of differential selective pressures along codon sites of the envelope gene was conducted in the CODEML program of the PAML 4.4 package (Yang, 2007) by calculating codon-specific d_N/d_S values (the ω parameter). Alignments of each DENV serotype were tested for positive selection independently, and jointly, using the M1a–M2a, the M7–M8 and the M8–M8a models (Wong et al., 2004). A qualitative measure of the distribution of ω values along codon sites in DENV serotypes was obtained by plotting the beta distribution with the parameters inferred by the M7 model (Yang et al., 2000). A comparison of the strength of selection along the envelope gene among the serotypes was conducted using the approach of Choisy et al. (2004), who used a paired Wilcoxon rank sum test to verify whether or not the weighted ω for each codon, estimated under the discrete M3 model (Yang et al., 2000), significantly differed between serotypes. Finally, the SLAC and the FEL algorithms (Pond and Frost, 2005) of the HyPhy package (Pond et al., 2005) were also used to access codons under positive selection in DENV serotypes independently and on the joint data set.

2.4. Differential selection on internal branches of DENV phylogeny

In order to investigate the action of adaptive molecular evolution on the diversification of dengue serotypes, we allowed internal branches of the phylogeny containing all DENV serotypes (Fig. 1) to evolve under different d_N/d_S ratios using the approach described in Yang (1998), implemented in the PAML 4.4 package. We have also run the IFEL test of HyPhy (Pond et al., 2006) and the newly proposed branch-site test of episodic diversifying selection (Kosakovsky Pond et al., 2011), available at www.datamonkey.org, which permits the inference of lineage-specific events of positive selection. These analyses were conducted to verify whether the hypothesis of diversification lead by genetic drift (founder effects) could be rejected.

2.5. Analysis of geographic correlation and ancestral states

In order to understand the spatial dynamics of DENV evolution comparatively, we have verified the association between phylogeny and geography for each serotype independently. This test was conducted in Befi-BaTS, v0.1.1 (Parker et al., 2008), which implements several tests of phylogeny–trait correlation considering phylogenetic error. We have inferred the parsimony score (PS), the association index (AI) and the monophyletic clade size statistics to test phylogeny–geography correlation. As geography traits, we have considered the continent where the samples were obtained. The reconstruction of the ancestral geographical states was implemented in Mesquite 2.7 (Maddison and Maddison, 2009) using the parsimony algorithm.