Infection, Genetics and Evolution 10 (2010) 129–136

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

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

Evolution of the sequence composition of *Flaviviruses*

Alyxandria M. Schubert [a], Catherine Putonti [a,b,c,*]

[a] Department of Bioinformatics, Loyola University Chicago, Chicago, IL, USA
[b] Department of Biology, Loyola University Chicago, Chicago, IL, USA
[c] Department of Computer Science, Loyola University Chicago, Chicago, IL, USA

ARTICLE INFO

Article history:
Received 13 August 2009
Received in revised form 26 October 2009
Accepted 3 November 2009
Available online 13 November 2009

Keywords:
RNA viruses
Flaviviruses
Sequence composition
Codon usage

ABSTRACT

The adaption of pathogens to their host(s) is a major factor in the emergence of infectious disease and the persistent survival of many of the infectious diseases within the population. Since many of the smaller viral pathogens are entirely dependent upon host machinery, it has been postulated that they are under selection for a composition similar to that of their host. Analyses of sequence composition have been conducted for numerous small viral species including the *Flavivirus* genus. Examination of the species within this particular genus that infect vertebrate hosts revealed that sequence composition proclivities do not correspond with vector transmission as the evolutionary history of this species suggests. Recent sequencing efforts have generated complete genomes for many viral species including members of the *Flavivirus* genus. A thorough comparison of the sequence composition was conducted for all of the available *Flaviviruses* for which the complete genome is publicly available. This effort expands the work of previous studies to include new vector-borne species as well as members of the insect-specific group which previously have not been explored. Metrics, including mono-, di-, and tri-nucleotide abundances as well as N_C values and codon usage preferences, were explored both for the entire polyprotein sequence as well as for each individual coding region. Preferences for compositions correspond to host-range rather than evolutionary history; species which infect vertebrate hosts exhibited particular preferences similar to each other as well as in correspondence with their host's preferences. *Flaviviruses* which do not infect vertebrate hosts, however, did not show these proclivities, with the exception of the Kamiti River virus suggesting its recent (either past or present) infectivity of an unknown vertebrate host.

1. Introduction

Given the compact nature of the genomes of many viral pathogens, acquisition of host-specific coding regions cannot be supported. Since many of these viral pathogens are entirely dependent upon host machinery, it has been postulated that they are under selection for a composition similar to that of their host. Host species exhibit compositional preferences as a consequence of many factors including: nucleotide abundances, DNA stacking energies, methylation, modification, replication, and repair mechanisms (Karlin, 1998; Xia and Yuen, 2005). Examination of mono-, di-, tri- and tetra-nucleotide usage within small viral genomes has revealed correspondence between the pathogen and host species. For example, it has been observed that CpG dinucleotides are underrepresented for the majority of these species, regardless of the nature of their genome (Karlin et al., 1994; Rima and McFerran, 1997; Auewarakul, 2004; Shackelton et al., 2006; Sewatanon et al., 2007; Greenbaum et al., 2008; Tao et al., 2009). Moreover, the correspondence between the codon usage preferences (Sharp and Li, 1986) of eukaryotic viruses and their host species has been the focus of many studies including a wide variety of DNA-based viruses, RNA-based viruses and retro-transcribing viruses (e.g. Karlin et al., 1990; Levin and Whittome, 2000; Jenkins and Holmes, 2003; Zhao et al., 2003; Adams and Antoniw, 2004; Gu et al., 2004; Zhou et al., 2005; Shackelton et al., 2006; Tsai et al., 2007; van Hemert et al., 2007; Jiang et al., 2008; Tao et al., 2009). Studies have supported both translational and mutational selection as the primary force shaping codon bias in these viral species (Jenkins and Holmes, 2003). The correspondence of longer subsequences, four (Pride et al., 2006) or 17–26 nucleotide long sequences (Kerr and Boschetti, 2006), within pathogen and host genomes has also been considered, but with mixed conclusions.

The *Flavivirus* genus, one of the three genera within the *Flaviviridae* family, consists of over 70 different known species. These single-stranded positive sense RNA viruses (~11,000 bases) encoding for three structural – capsid (C), membrane (prM/M), envelope (E) – and seven non-structural proteins – NS1, NS2A, NS2B, NS3, NS4A, NS4B and NS5 – are spread world-wide, some regions having several different species coexisting within a single

* Corresponding author at: Department of Computer Science, Loyola University Chicago, 1032 W. Sheridan Rd., Chicago, IL 60660, USA. Fax: +1 773 508 3646.
E-mail address: cputonti@luc.edu (C. Putonti).

1567-1348/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.meegid.2009.11.004

area. *Flaviviruses* are most often transmitted by arthropods (primarily mosquitoes and ticks) to a variety of vertebrate hosts. Moreover, the *flavivirus* genus also includes species with no known vector (NKV) as well as species which do not infect vertebrates, henceforth referred to as the insect-specific group. Phylogenies derived from sequence alignments of individual coding regions and the whole genomic sequences of available *flavivirus* species suggest that species within one vector group are typically more closely related than species of the other vector groups (Billoir et al., 2000; Gaunt et al., 2001; Cook and Holmes, 2006). Analysis of the base composition and codon usage of the *flavivirus* genus has been previously conducted based upon partial NS5 gene sequences and the complete polyprotein sequence of 13 species (Jenkins et al., 2001). This study found relationships between base compositions and vector specificity which the authors attribute to differences in mutational biases among *flaviviruses* (Jenkins et al., 2001). Furthermore, this study revealed that *flaviviruses* exhibit dinucleotide and codon biases, but these biases do not covary with arthropod association (Jenkins et al., 2001).

Since the study conducted by Jenkins et al. (2001), additional isolates have been sequenced as well as new species totaling well over 2000 complete *flavivirus* sequences. Moreover, new species within the insect-specific group have been discovered and their complete genomes have been sequenced and annotated, including Cell Fusing Agent virus (Cammisa-Parks et al., 1992), Kamiti River virus (Crabtree et al., 2003), Culex flavivirus (Hoshino et al., 2007) and Quang Binh virus (Crabtree et al., 2009). This prompted us to revisit the study of Jenkins et al. (2001) and expand analysis to all of the publicly available annotated *flavivirus* species. Furthermore, because the individual protein products produced from the transcribed polyprotein vary in the number required to form the mature virion as well as in their interaction with the host, it is likely that the pressures of selection for a composition similar to that of their host varies from gene to gene. Thus, in addition to examining the compositional properties of the NS5 gene, the compositional properties of each of the 10 *flavivirus* genes were examined.

2. Materials and methods

2.1. Flavivirus genome sequences

A total of 37 *flavivirus* genomes were used in this study (Table 1). These sequences were obtained from NCBI's RefSeq

Table 1
RefSeq viral genomes examined.

Virus	Isolate	Strain origin	Principal host species	Vector group	Length (bp)	Accession no.
Cell fusing agent (CFAV)*	–	*A. aegypti*[a]	*A. aegypti*	I	10,695	NC_001564
Culex *flavivirus* (CxFV)	Tokyo	*C. pipiens*[a]	*C. pipiens*	I	10,837	NC_008604
Kamiti River (KRV)*	SR-82	*A. macintoshi*[a]	*A. macintoshi*	I	11,375	NC_005064
Quang Binh (QBV)	VN180	*C. tritaeniorhyncus*[a]	*C. tritaeniorhyncus*	I	10,865	NC_012671
Bagaza (BAGV)	DakAr B209	Mosquito[b]	Unknown	M	10,941	NC_012534
Bussuquara (BSQV)	BeAn 4073	*A. belzebul*[U]	Rodents, human	M	10,290	NC_009026
Dengue virus type 1 (DENV-1)*	45AZ5	Human[c,d]	Human	M	10,735	NC_001477
Dengue virus type 2 (DENV-2)*	16681	Human[e,f,g,h,d,a]	Human	M	10,723	NC_001474
Dengue virus type 3 (DENV-3)*	D3/H/IMTSSA-SRI/2000/1266	Human[i,a]	Human	M	10,707	NC_001475
Dengue virus type 4 (DENV-4)*	rDEN4	*A. pseudoscutellaris*[f,a,j]	Human	M	10,649	NC_002640
Entebbe bat (ENTV)	UgIL-30	*T.* (*C.*) *limbata*[b]	Bat	M	10,510	NC_008718
Iguape (IGUV)	SPAn 71686	Sentinel mouse[U]	Rodents, birds	M	10,251	NC_009027
Ilheus (ILHV)*	Original	*Aedes* and *Psorophora*[U]	Bird	M	10,275	NC_009028
Japanese encephalitis (JEV)*	JaOArS982	Mosquito[a]	Bird, pig, human	M	10,976	NC_001437
Kedougou (KEDV)	DakAar D1470	Mosquito[b]	Unknown	M	10,723	NC_012533
Kokobera (KOKV)	AusMRM 32	*C. annulirostris*[U]	Kangaroos, horses	M	10,233	NC_009029
Murray Valley encephalitis (MVEV)*	Australia 1951	Human[a,j]	Bird	M	11,014	NC_000943
Sepik (SEPV)	MK7148	*M. septempunctata*[b]	Unknown	M	10,793	NC_008719
St. Louis encephalitis (SLEV)	Kern217	*C. tarsalis*[j]	Bat	M	10,940	NC_007580
Usutu (USUV)*	Vienna 2001	Blackbird[j]	Bird	M	11,066	NC_006551
West Nile virus lineage I (WNV-1)*	NY99	*B. scandiacus*[j]	Bird	M	11,029	NC_009942
West Nile virus lineage II (WNV-2)*	956	Human[b,k,j,a]	Bird	M	10,962	NC_001563
Yellow fever (YFV)*	17D vaccine strain	Human[l,m,j,k]	Monkeys	M	10,862	NC_002031
Yokose (YOKV)	Oita 36	Bat[b,j]	Bat	M	10,857	NC_005039
Zika (ZIKV)	MR 766	Sentinel monkey[b]	Monkeys	M	10,794	NC_012532
Apoi (APOIV)*	ApMAR	*Apodemus*[b,j]	Rodents	NKV	10,116	NC_003676
Modoc (MODV)*	M544	*P. maniculatus*[j]	Bat	NKV	10,600	NC_003635
Montana myotis leukoencephalitis (MMLV)*	Montana 1958	*M. lucifugus*[j]	Bat	NKV	10,690	NC_004119
Rio Bravo (RBV)*	RiMAR	Bat[b,j]	Bat	NKV	10,140	NC_003675
Tamana bat (TABV)*	Tr127154	*P. parnellii*[n,j]	Bat	NKV	10,053	NC_003996
Alkhurma (AHFV)*	1176	Human[b,j,o,n]	Camels, sheep	T	10,685	NC_004355
Karshi (KSIV)	LEIV 2247	*H. asiaticum asiaticum*[j]	Rodents	T	10,653	NC_006947
Langat (LGTV)*	TP21	*Ixodides*[j]	Rodents	T	10,943	NC_003690
Louping ill (LIV)*	369/T2	*I. ricinus*[b]	Sheep	T	10,871	NC_001809
Omsk hemorrhagic fever (OHFV)*	Bogoluvovska	*D. marginatus*[j]	Muskrats	T	10,787	NC_005062
Powassan (POWV)*	LB	Human[b,p]	Small mammals	T	10,839	NC_003687
Tick-borne encephalitis (TBEV)*	Neudoerfl	Tick[p]	Rodents	T	11,141	NC_001672

I indicates those species which do not infect a vertebrate host as the insect-specific group, M indicates those transmitted by a mosquito-vector, T indicates those transmitted by a tick vector, NKV indicates those with no known vector. Asterisk denotes genomes for which individual gene annotations were available. Passage history of isolate includes: [a]C6/36 cells, [b]suckling mouse brain, [c]rhesus lung cells (FRhL), [d]PDK cells, [e]BS-C-1 cells, [f]LLC-MK$_2$, [g]*Rhesus macaque* monkey, [h]*T. amboinensis* mosquitoes, [i]white blood cells, [j]vero cells, [k]BHK-21 cells, [l]mouse embryonic tissue, [m]monkey serum in tyrode, [n]mouse, [o]sheep, [p]chick embryo, [U]unknown. Isolate and passage information and primary host information were ascertained from genome sequence publications (Baillie et al., 2008; Bakonyi et al., 2004; Billoir et al., 2000; Cammisa-Parks et al., 1992; Campbell and Pletnev, 2000; Charlier et al., 2002; Charrel et al., 2001; Coimbra et al., 1993; Crabtree et al., 2009; de Lamballerie et al., 2002; Durbin et al., 2001; Gao et al., 1994; Gritsun et al., 1997; Hoshino et al., 2007; Hurrelbrink et al., 1999; Kinney et al., 1997; Kuno and Chang, 2006, 2007; Laemmert and Hughes, 1947; Leyssen et al., 2002; Lin et al., 2003; Mackenzie and Williams, 2009; Mandl et al., 1989, 1993; Peyrefitte et al., 2003; Puri et al., 1997; Rice et al., 1985; Sang et al., 2003; Sumiyoshi et al., 1987; Tajima et al., 2005; Theiler and Smith, 1937; Turell et al., 2008; Weissenböck et al., 2009; Yamshchikov et al., 2001) and NCBI genome files.