Arch Virol (2007) 152: 687–696
DOI 10.1007/s00705-006-0903-z
Printed in The Netherlands

Archives of Virology

Full-length sequencing and genomic characterization of Bagaza, Kedougou, and Zika viruses

G. Kuno and **G.-J. J. Chang**

Arbovirus Diseases Branch, Division of Vector-Borne Infectious Diseases, National Center for Zoonotic, Vector-Borne, and Enteric Diseases, Centers for Disease Control and Prevention, Fort Collins, Colorado, U.S.A.

Received August 8, 2006; accepted November 22, 2006; published online January 3, 2007

Summary

Many members of the genus *Flavivirus* are the agents of important diseases of humans, livestock, and wildlife. Currently, no complete genome sequence is available for the three African viruses, Bagaza, Zika, and Kedougou viruses, each representing a distinct virus subgroup according to the latest virus classification. In this study, we obtained a complete genome sequence of each of those three viruses and characterized the open reading frames (ORFs) with respect to gene sizes, cleavage sites, potential glycosylation sites, distribution of cysteine residues, and unique motifs. The sequences of the three viruses were then scanned across the entire length of the ORF against available sequences of other African flaviviruses and selected reference viruses for genetic relatedness. The data collectively indicated that Kedougou virus was close to dengue viruses but nonetheless distinct, while Bagaza virus shared genetic relatedness with West Nile virus in several genomic regions. In the non-coding regions, it was found that a particular organizational pattern of conserved sequences in the 3′ terminal region generally correlated with the current virus grouping.

Author's address: Dr. G. Kuno, Arbovirus Diseases Branch, Division of Vector-Borne Infectious Diseases, National Center for Zoonotic, Vector-Borne, and Enteric Diseases, Centers for Disease Control and Prevention, P.O. Box 2087, Fort Collins, CO 80522-2087, U.S.A. e-mails: gok1@cdc.gov; G.-J. J. Chang: gxc7@cdc.gov

Introduction

In the past two decades, we witnessed an accelerated global dispersal of vector-borne flaviviruses beyond the traditional range of geographic distribution. This situation is illustrated in the increased cases of tick-borne encephalitis in Scandinavian countries, outbreaks of dengue in Nepal, Argentina, Hawaii, and the Easter Islands in the Pacific, in the introduction of West Nile virus into the New World and of Usutu virus into Europe, and in the emergence of a genotype (Alkhurma virus) of tick-borne Kyasanur Forest disease virus in Saudi Arabia. For monitoring the introduction of unusual flaviviruses in any location, prior in-depth characterization of non-indigenous viruses is essential not only to better understand the disease-causing and transmission potentials in new environments but also to develop diagnostic reagents for improved surveillance.

According to the latest virus classification [11], three mosquito-borne viruses of Africa, Bagaza virus (BAGV), Kedougou virus (KEDV), and Zika virus

(ZIKV), represent distinct subgroups. For the Ntaya and Spondweni subgroups, which include BAGV and ZIKV, respectively, there is no full genome sequence available for any of its members. On the other hand, for KEDV, which was recently classified for the first time as a member of the dengue virus (DENV) subgroup [11], its genome characterization is essential for providing information on how closely it is related to the four dengue serotypes. Genome characterization of BAGV is also critically important with respect to epizootic concerns, because in our previous study [14] we identified this virus as synonymous with Israel turkey meningoencephalitis virus (ITV), a serious avian pathogen of economic importance in the Middle East and southern Africa [2].

In this study, we obtained the full-length genome sequences and examined the genomic traits of BAGV, KEDV, and ZIKV to provide the basic information necessary for improved disease surveillance and other virologic investigations. The basic data thus obtained are useful for the discussion of flaviviral classification and phylogenetics as well.

Materials and methods

Viruses

The three viruses (in suckling mouse brain passage levels ranging from 1 to 3) obtained from the WHO Collaborating Center in the Division of Vector-Borne Infectious Diseases of CDC (Fort Collins, Colorado, USA) are BAGV (strain DakAr B209), KEDV (strain DakAar D1470), and ZIKV (strain MR-766).

Table 1. Primers for initial RT-PCR amplification and sequencing

Genomic location	Amino acid motif	Direction	Base sequence (5′→3′)
5′NCR*		F**	ATG(A/T) CTAA(A/G) AAACCAGGA
		F	TCAATATGCT(A/G) AAACGCGG
Envelope	DRGWGNGC	F	GAYMGWGGVTGGGGHAAYGGVTG
	GLFGKGS	F	GGMYTKTTYGGDAARGGRAGC
	GHLKCRV	F	GGMCAYSTBWMYTGTMGVSTG
	PFGDSYIV	F	CCSTTYGGWGAYTCNTACATHGT
		R**	TCCYCTKCCHACYACDATGTA
	DTAWDFGS	F	GAYACDGCBTGGGAYTTTGGMTC
NS1	GCWYGMEI	F	GGHTGTTGGTAYGGMATGGA
	YGMEIRP	F	TAYGGMATGGARAYHMGRCC
NS3	GTSGSPI	F	GGMACDTCNGGHTCNCCHAT
	GLYGNG	F	GGNCTNTAYGGNAAYGG
	LAPTRVV	F	YTRGCDCCNACNMGRGTNGT
	DVMCHATF	F	GAYGTSATGTGYCAYGCHAC
	MDEAHF	F	ATGGAYGARGCHCAYTT
	SIAARGY	F	AGYATMGCNGCHMGAGG
	MTATPPG	F	ATGACNGCVACNCCNCCNGG
	ISEMGAN	F	ATHTCNGARATGGGDGCVAA
	SAAQRRGR	F	WSYGCWGCBCARAGRMGRGGVMG
NS5	DLGCGRG	F	GAYCTNGGNTGYGGNMGNGG
	SRNSTHEMY	F	TCAAGGAACTCCACACATGAGATGTACT
	NMMGKREKK	F	TACAACATGATGGGAAAGAGAGAGAA
	ADDTAGWDT	F	GCTGATGACACCGCCGGCTGGGACAC
		R	GTGTCCCAGCCGGCGGTGTCATCAGC
	WMTTEDML	R	AGCATGTCTTCCGTGGTCATCCA
3NCR			
CS2***		R	GGGTCTCCTCTAACCTCTAG

* *NCR* Noncoding region.
** *F* Forward; *R* reverse.
*** VD8 of Pierre et al [19].

RT-PCR and sequencing

Viral RNA was extracted directly from infected mouse brain suspension with the QIAmpViral RNA Mini Kit (Qiagen, Valencia, CA). cDNA was prepared by first incubating 14 µl viral RNA and 1 µl reverse primer and then rapidly cooling on ice. For sequencing the genomic region between the 5′- end of the genome and the conserved sequence (CS2) in the 3′-noncoding region (3′-NCR), primer (VD8) [19] (Table 1) was used for cDNA synthesis. After reverse transcription, polymerase chain reaction (PCR) was performed using an Expand Long Template PCR System kit (Roche Applied Science, Indianapolis, IN). The thermocycling program set up in a Gene Amp PCR System 9600 thermocycler (Perkin-Elmer, Norwalk, CT) was 1 cycle of 94 °C for 1 min/50 °C for 1 min/68 °C for 5 min; 3 cycles of 94 °C for 20 sec/50 °C for 1 min/68 °C for 4 min; 10 cycles of 94 °C for 20 sec/50 °C for 30 sec/68 °C for 4 min with an increment of 20 sec per cycle; and 1 cycle of extension at 68 °C for 7 min. Most of the primers used (Table 1) were designed primarily based on the conserved amino acid motifs among mosquito-borne flaviviruses [5].

5′- and 3′-ends of the genomes were amplified using 5′-RACE and 3′-RACE kits and the necessary enzymes and other reagents (Invitrogen, Carlsbad, CA – formerly Life Technologies, Rockville, MD), respectively. For 5′-RACE, 13 µl viral RNA was reverse transcribed in a reaction mixture (total volume 25 µl) according to the kit's instructions and using the kit's reagents and a reverse primer. The reverse primer used was selected approximately within 250 bases from the 5′ terminus. To 16.5 µl of the cDNA thus prepared were added 5 µl tailing buffer, 2.5 µl 100 mM dCTP, and 1 µl terminal deoxynucleotide transferase, and the mixture was incubated at 37 °C for C-tailing of the 5′ terminus. The kit's Abridged Anchor poly-G primer was used for PCR amplification of this terminal segment, following the PCR protocol in the kit. For 3′-RACE, 36 µl viral RNA was mixed in a reaction mixture containing poly-A buffer, 1 µl 12.5 mM ATP, and 2 µl poly-A polymerase in a total volume of 50 µl, and the mixture was incubated at 37 °C for 11 min for poly-A tailing the 3′ terminus. After column purification, poly-A-tailed RNA was amplified by RT-PCR using a forward primer selected approximately within 100 bases upstream of CS2 and as a reverse primer either the kit's Adaptor Primer (5′-GGCCACGCGTCGACTAC[T_{17}]-3′) supplied in the kit or one of the following reverse primers: 5′-GCATG CGGCCGC[T_{18}]AGT-3′; 5′-GCATGCGGCCGC[T_{18}]AGA-3′; 5′-GCATGCGGCCGC[T_{18}]AGC-3′; 5′-GCATGCGGCC GC[T_{18}]AG-3′.

Complete genomes were sequenced in both directions by primer walking. For a speedy sequencing of several viruses simultaneously, we adopted the following strategy. In the first phase, a few to several non-overlapping genomic regions (preferably less than 3.5 kb in length) were amplified using the selected, paired primers shown in Table 1. The same amplification primers were also used at first for sequencing the amplicons they generated. In the second phase, from the sequences obtained in the first phase, internal primers were selected for primer walking in both directions and for bridging the gaps between amplified regions. Amplicons were purified using a Centricep column (Princeton Separations, Adelphia, NJ), and aliquots of approximately 60–160 ng of the purified DNA templates were used for direct cycle sequencing using a PRISM DNA sequencing kit (Big Dye) for dye terminator cycle sequencing with Ampli-Taq FS enzyme (ABI, Foster City, CA), as described previously [14], and CEQ 8000 Genetic Analysis System (Beckman Coulter, Inc., Fullerton, CA).

Sequence alignment

The full-length genome sequences of the three African viruses were deposited in GenBank as revised versions (R1) of the ORF sequences deposited previously (BAGV: AY632545; KEDV: AY632540; and ZIKV: AY632535). The open reading frames (ORFs) of those viruses were aligned first by Clustal X [23], followed by manual adjustment with BioEdit (version 5.0.0) [10]. We then applied the "ReGap DNA project" function in the GeneDoc program [17] to generate a properly aligned nucleotide sequence file.

Cleavage site determination

Most of the cleavage sites were identified by following the proteolytic processing cascade scheme for the flavivirus ORFs previously revealed by Rice and Strauss [22]. Junctions of intracellular capsid and premembrane (Ci/prM), prM and envelope (prM/E), and E and nonstructural protein 1 (E/NS1) processed by the host cellular signalase were determined on the basis of the highest cleavage potential score using the computer program SignalP-NN (http://www.cbs.dtu.dk/services/) [6].

Secondary structure in 3′-noncoding region (NCR) and genome cyclization

The secondary structure in the 3′-NCR and cyclization between 3′- and 5′-terminal regions were investigated by using the mfold program [27], based on the first 200 nucleotides in the 5′-terminal sequence of the genome and the entire length of 3′-terminal sequence after the stop codon of the NS5 gene, similar to the study by Khromykh et al. [13].

Bootscanning of the open reading frame (ORF)

The Bootscan program in the SimPlot package was used to examine sequence relatedness of the African flaviviruses across their open reading frames (ORFs) [7, 21]. ORFs of selected viruses were scanned against a query virus with a window size of 600 nt to obtain 100 phylogenetic tree replicas using SEQBOOT and DNAPARS of the PHYLIP program [7]. This procedure was repeated by sliding 10 nt per

step for the entire ORF. The bootstrap supports, indicating the degree of phylogenetic relatedness among reference viruses and the query virus, were tabulated and plotted for each step. The degree of bootstrap support is expressed on the Y-axis as % permuted trees across the entire length of the ORF plotted on the X-axis. Besides the three African virus sequences obtained in this study, we also used ORF sequences obtained from GenBank, including other African mosquito-borne viruses (DENV-4 [M14931], West Nile virus [AF196835], yellow fever virus [X03700]), a tick-borne virus (Powassan virus [L06436]), three vertebrate viruses without a vector relationship (Apoi virus [AF16093], Entebbe bat virus [NC012380], Tamana bat virus [AF285080]), an insect virus (Kamiti River virus [NC005064]), and Sepik virus (AY632543) from Papua New Guinea.

Results

Full-length genome

The lengths of the genomes of BAGV, KEDV, and ZIKV are 10,941, 10,723, and 10,794 nucleotides, respectively, and the ORFs encode 3426, 3408, and 3419 amino acids (aa), respectively (Table 2). It should be noted that, although KEDV is currently classified as a member of the dengue virus (DENV) subgroup, the ORF length (3408 aa) of KEDV is much longer than those (3387–3392, except for one strain of DENV-1 with 3396 aa) of the 4 DENV serotypes, which, as a group, have the shortest ORF length among the mosquito-borne viruses fully sequenced thus far. The genome length difference among the three viruses was further examined by breaking down the genome into individual genes and NCRs. As shown in Table 2, the gene lengths

Table 2. Comparison of the genes or genomic regions of three African viruses

Gene or genomic region	BAGV	KEDV (% Amino acid identity with DENV-4)	ZIKV
5′-NCR	94 nt	106 nt	106 nt
Capsid	122 aa	114 aa (37.9)	122 aa
PrM	177 aa	178 aa (48.3)	178 aa
Envelope	501 aa	501 aa (50.2)	500 aa
NS1	342 aa	342 aa (47.5)	342 aa
NS2A	226 aa	224 aa (20.8)	226 aa
NS2B	132 aa	130 aa (31.5)	130 aa
NS3	619 aa	616 aa (63.8)	617 aa
NS4A	126 aa	127 aa (37.8)	127 aa
2K	23 aa	23 aa (34.8)	23 aa
NS4B	253 aa	252 aa (49.5)	252 aa
NS5	905 aa	901 aa (63.3)	902 aa
3′-NCR	566 nt	390 nt	428 nt
Total length of genome	10941 nt	10723 nt	10794 nt

nt Nucleotide; *aa* amino acid.

Table 3. Proposed polyprotein cleavage sites of the three African flaviviruses

Virus	Cv/Ci	Ci/prM	pr/M
BAGV	RGKKKR/GGTTVG	LGVAQA/IKIGSL	ARRSRR/SITVHH
KEDV	INKRKR/SPVNWF	FGVLTA/VKIGDY	PRRSRR/SVSLPP
ZIKV	KERKRR/GADTSI	LTTAMA/AEITRR	ARRSRR/AVTLPS
	M/E	E/NS1	NS1/NS2A
BAGV	IAPAYS/FNCLGM	ATNVHA/DTGCAV	KSRVTA/YDGAGM
KEDV	VAPAYS/IRCIGV	ATSVNG/DQGCAM	KARVSA/GSGSGV
ZIKV	IAPAYS/IRCIGV	STAVSA/DVGCSV	RSMVTA/GSTDHM
	NS2A/NS2B	NS2B/NS3	NS3/NS4A
BAGV	KPSNRR/GWPVSE	THSPKR/SGAIWD	FACGKR/SAIGVS
KEDV	TGEKRR/SWPPTE	DKKAKR/SGALWD	FAAGRR/GATAGV
ZIKV	TRSGKR/SWPPSE	VKTGKR/SGALWD	FAAGKR/GAALGV
	NS4A/2K	2K/NS4B	NS4B/NS5
BAGV	EPERQR/SQTDSH	VGTVAS/NEMGWL	NGSMRR/GGGKGR
KEDV	EPERQR/SVQDNY	LGLVAA/NEAGLL	VSSTKR/SNRGLG
ZIKV	EPEKQR/SPQDNQ	LGLITA/NELGWL	GLVKRR/GGGTGE

Cv Virion capsid; *Ci* C-terminal hydrophobic domain of capsid; *prM* premembrane; *NS* nonstructural protein.

are similar among the three viruses, and the region contributing most to the difference in genome length is the 3′-NCR. To further examine the relatedness between KEDV and DENVs, we compared the amino acid identities in the coding regions of KEDV and DENV-4 (representing the four DENV serotypes). As shown in Table 2, the identity in the envelope gene was only 50%. The highest identities found in the NS3 and NS5 genes were both about 63%, which were still below 69%, the minimal criterion used for classifying KEDV as a member of the dengue cluster according to our quantitative molecular classification scheme [14].

Cleavage sites

The predicted cleavage sites of the three African viruses are shown in Table 3. The N-termini of all sites expected to be cleaved by the viral serine protease (virion capsid and Ci [Cv/Ci], NS2A/NS2B, NS2B/NS3, NS3/NS4A, NS4A/2K, and NS4B/NS5) and furin (or furin-like protease) (pr/M) follow two C-terminal basic amino acids (most typically KR, RR, or QR), as were found in most other mosquito-borne flaviviruses before. In the sites cleaved by host signalase, the C-terminal and N-terminal amino acids immediately flanking M/E, E/NS1, and 2K/NS4B are generally similar. The NS1/NS2A site, which is believed to be cleaved by an unknown cellular signalase, follows the sequence V-X-A (in which X is variable), as defined by Rice and Strauss [22]. Regarding the C-terminal quadrapeptides preceding the M/E cleavage site, they are PAYS in all mosquito-borne viruses thus far examined (including KEDV), with the only exceptions being the four dengue serotypes, in which they are PS(Y or M) (G, T, or A). This is of taxonomic interest because KEDV is currently classified as a member of the DENV subgroup.

Glycosylation sites and cysteine residues

The numbers of N-linked, potential glycosylation sites in the pre-membrane (prM), envelope (E), and nonstructural protein 1 (NS1) genes in BAGV, KEDV, and ZIKV were 1-1-3, 1-0-3, and 1-1-3, in that order, respectively. No glycosylation site was found in hydrophobic domains in any virus studied. The pattern of 12 cysteine residues found in all mosquito-borne flaviviruses in the E and NS1 genes is also conserved in the three African viruses studied. The 6 cysteine residues in prM are clustered in the pr domain, as in most other flaviviruses.

Motifs

The well conserved amino acid motifs in the NS3 and NS5 genes (Table 1) as well as many others not shown in the table are also conserved in the three African viruses. The tripeptide in the domain III region of the envelope protein, corresponding to the RGD motif in Murray Valley encephalitis virus and Japanese encephalitis virus that has been speculated

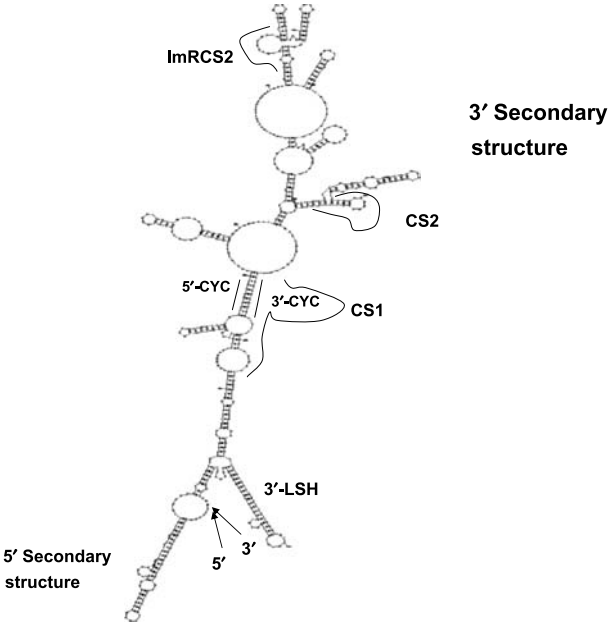

Fig. 1. A predicted secondary structure of Kedougou virus. The folding pattern was obtained using the mfold program of Zuker [27], using the first 200 nucleotides in the 5′ noncoding region (NCR) and the entire 3′-NCR. Two arrows (5′ and 3′) point to where the 5′- and 3′-termini of the genome are placed next to each other (but not linked). The 5′-NCR sequence is read downstream from the arrow clockwise along the partially double-strand structure; and the 3′-NCR sequence is read upstream counterclockwise from the arrow. *CS* Conserved sequence, *ImCS2* imperfect CS2, *3′-LSH* long stable hairpin of the 3′-terminal sequence, *3′-CYC* cycling sequence within CS1 in the 3′noncoding region, *5′-CYC* cycling sequence within the capsid gene

to be involved in adsorption to host cells, is TGE in both BAGV and ZIKV but is VGD in KEDV. Numerous amino acid motifs in the NS5 gene of the three African viruses are shared by all other mosquito-borne flaviviruses including the DENV serotypes. However, with respect to the dodecapeptide (RGSGQVVTYA) in the RNA-dependent RNA polymerase domain, four DENV serotypes differ uniquely from all other mosquito-borne viruses including KEDV. The dodeca sequence of the four DENV serotypes is RGSGQVGTYG. Thus, like in the aforementioned example of the prM/E cleavage site, in this regard, KEDV has a genomic property that differs from that of DENV.

5′- and 3′-NCRs

The genomic organization of the 3′-NCR and secondary structures that form as a result of the predicted cyclization between the 5′- and 3′-NCRs were studied using the mfold program of Zuker [27]. Forty-two, 37, and 27 predicted folding patterns with slightly different folding energy levels were generated for BAGV, KEDV, and ZIKV. One of the most prevalent patterns for KEDV has been selected to summarize the essential findings for all three viruses (Fig. 1). As shown in Fig. 1, the long stable hairpin structure near the 3′-terminal sequence (3′-LSH) contains the conserved pentanucleotide (CACAG) in the loop. The octanucleotide sequence (CATATTGA) in the 5′ terminal segment within CS1 and which is involved in cyclization (hereafter called 3′-CYC) forms a double-strand structure with the complementary sequence in the 5′-terminal capsid gene (hereafter called 5′-CYC) (Fig. 1). Regarding the upstream AUG region (UAR) in the 5′-NCR reported to be involved in cyclization of DENV-2 [1], currently, the functionally-homologous sequences for the other 3 dengue serotypes and KEDV remain unknown in the absence of experimental proof. A comparison of the 5′-terminal sequences between KEDV and the four dengue serotypes upstream of AUG in the 5′-NCR [1, 8, 18, 26] revealed that although the 5′-UAR of DENV-2 is identical to the comparable sequence of DENV-1, slightly different from that of DENV-3, and quite different from DENV-4, none of the dengue serotype sequence was shared by KEDV.

Regarding the organization of the 3′-NCR, the names of the structures and/or domains used in this study follow the nomenclatures established earlier [4, 9, 20]. The consensus sequences of CS1, CS2, and CS3 among YFV, DENV, and JEV complex viruses are shown in Table 4. ImCS was designated

Table 4. Conserved sequences (CSs) in the 3′-noncoding regions of Bagaza, Kedougou and Zika viruses

CS1	
Consensus sequence*	ASCATATTGACRCCWGGGAWAAGAC
Bagaza virus	AGCATATTGACACCTGGGA-GAGAC
Kedougou virus	AGCATATTGACACCTGGGA-AAGAC
Zika virus ImCS1**	AGCATATTGACG--TGGGA-AAGAC
CS2	
Consensus sequence*	GGWCTAGA-GG--TTAGW-G-GAGACC-C
Bagaza virus	GGACTAGA-GG--TTAGA-G-GAGACC-C
Kedougou virus ImRCS2**	GGACTTGAAGGACTT-GACGTCAGGCCAC
Zika virus	GGACTAGT-GG--TTAGA-G-GAGACC-C
CS3	
Consensus sequence*	YCCCAGGWGG-ACTGGGTDAMCAAAKSBR
Bagaza virus	CCCCAGGTGG-ACTGGGTAAACAAAGCCG
Zika virus ImCS3**	CCCCAGGAGAAGCTGGGAAACCAA-GCTC

* Consensus sequences of CS1, CS2, and CS3 are based on the sequences of YFV, DENV complex viruses, and JEV complex viruses. Solid line indicates a gap created artificially for alignment purposes. The following single letter codes are used; B = C, G, or T; D = A, G, or T; K = G or T; M = A or C; R = A or G; S = C or G; W = A or T; Y = C or T.

** *Im* Imperfect (see the definition in the text).

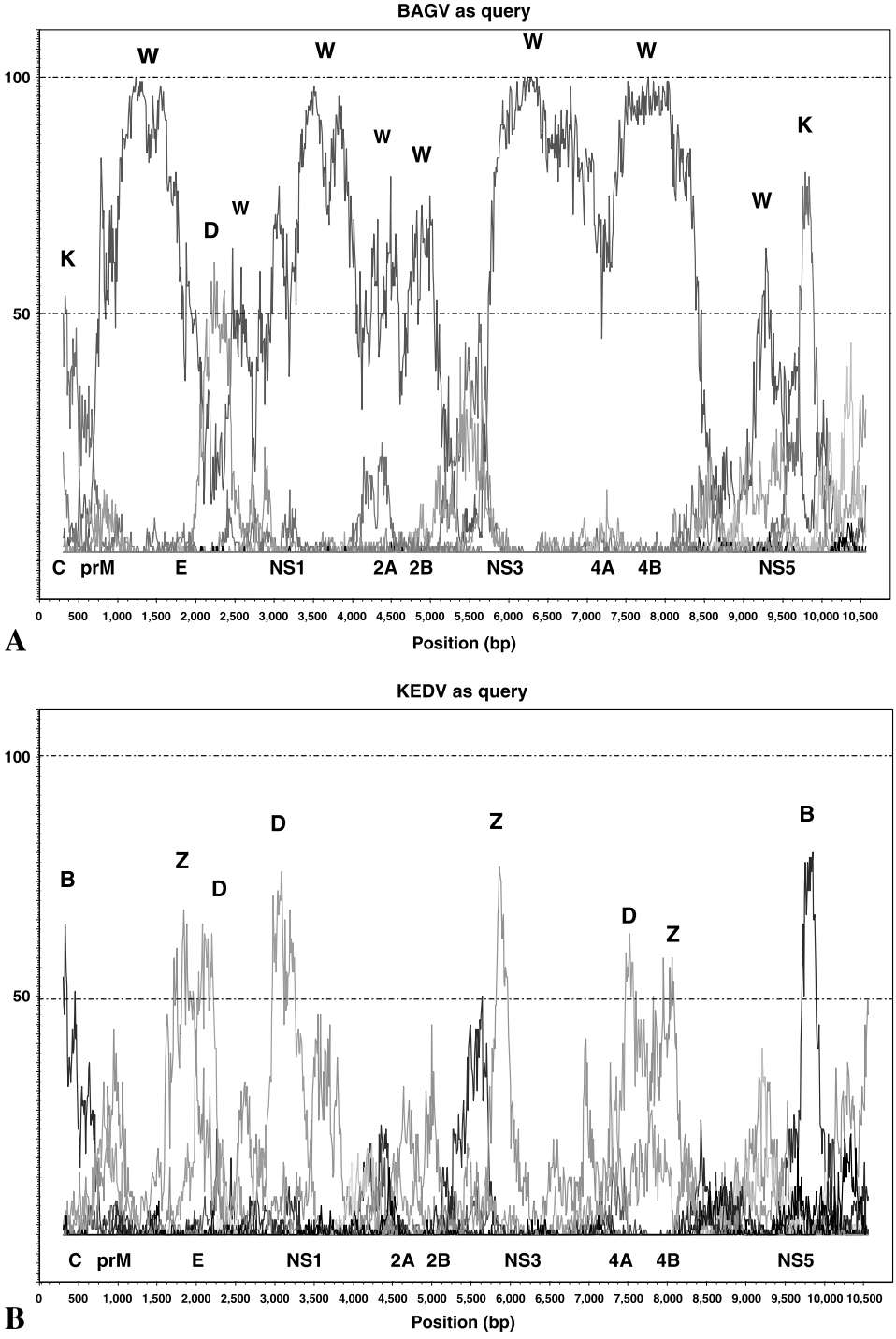

Fig. 2. Bootscanning of the complete open reading frames (ORFs) of flaviviruses. The entire ORF is scanned across the X-axis. The percentage of permutated trees (or phylogenetic relatedness) is shown on the Y-axis. The query sequence of a selected virus is an unmarked horizontal line at 100% permutatation across the ORF. Peaks above 50% permutation, labeled B, D, K, W, and Z, indicate Bagaza virus, DENV-4, Kedougou virus, West Nile virus, and Zika virus, respectively. Virus sequences used: Apoi virus, Bagaza virus, DENV-4, Entebbe bat virus, Kedougou virus, Kamiti River virus, Powassan virus, Sepik virus, Tamana bat virus, West Nile virus, yellow fever virus, Zika virus. **A.** Bagaza virus as query, **B.** Kedougou virus as query

to denote imperfectness if the conserved sequences (CS1, CS2, and CS3) of the three African viruses differed from the corresponding consensus sequences in 3 or more bases, as manifested by substitution, insertion, and/or deletion for a proper alignment. The three conserved sequences of BAGV are most identical to the corresponding consensus CSs, while KEDV and ZIKV differ more significantly in one and two CSs, respectively (Table 4). Thus, including the repeated CSs (RCSs), the CS organization (in 5′->3′ direction) in the 3′-NCR is RCS3-CS3-CS2-CS1 for BAGV, ImRCS2-CS2-CS1 for KEDV, and ImCS3-CS2-ImCS1 for ZIKV. Although the CS organizations of KEDV and DENV are identical, the CS2 sequence difference between the two viruses is considerable (Table 4). The results showed that CS2 and CS3 are also involved in secondary structure (Fig. 1). Although minor variations are observed (such as the number of bulges or of extra small loops), cyclization between 3′- and 5′-CYCs apparently constrains the number of variations in secondary structure predicted between them (the structure below the 3′-CYC and 5′-CYC in Fig. 1), as the patterns of the predicted secondary structures in the 5′-terminal and 3′-terminal sequences are more or less the same in each virus. On the other hand, the secondary structures from CS2 to CS3 (the structure above the 3′-CYC and 5′-CYC joint in Fig. 1) are highly variable among the possible folding patterns generated at different energy levels.

Bootscanning of ORFs

By arbitrarily selecting a 50% permutation on the Y-axis as the threshold of significant phylogenetic relatedness, for BAGV (query), a strong sign of relatedness (>80% permutation) with WNV is observed in the prM-E, NS1, and NS3-NS4B genes. Moderate but still significant relatedness of BAGV is observed with WNV in E-NS1, NS2A-2B, and the C-terminal region of the methyltransferase domain of the NS5 gene; with DENV-4 in the C-terminal part of E; and with KEDV in N-terminal domain of capsid and the RNA-dependent RNA polymerase (RdRp) domain of the NS5 gene (Fig. 2A). When KEDV was used as query, moderate relatedness was observed with DENV-4 in the middle portions of E, NS1, and the N-terminal part of NS4A; with ZIKV in the middle segment of E, middle NS3, and NS4B; and with BAGV in capsid and a segment of the RdRp domain of the NS5 gene (Fig. 2B). None of those viruses had a significant relatedness exceeding 80% permuted trees. When ZIKV was used as query, likewise, the profile of phylogenetic relatedness across ORF was shown with a mosaic of only moderate peaks represented by different viruses (DENV-4, KEDV, BAGV, and WNV) (data not shown).

Discussion

For a comprehensive characterization of vector-borne flaviviruses, both biologic and genomic traits need to be studied. Unfortunately, for many viruses in the tropics, gaps still exist in our knowledge regarding the biology of vectors and vertebrates as well as the natural transmission mechanism. As for genomic characterization, full genome sequences provide useful information for an accurate viral classification, improved phylogenetics, and designing more specific and/or broadly-reactive molecular diagnostic primers and probes used in disease surveillance. Also, a dataset of full genome sequence provides unique opportunities to examine more comprehensively natural recombination between different viruses or within genotypes of a virus.

Among the three viruses sequenced, ZIKV has been recognized to be a cause of febrile illness in humans in Africa and Southeast Asia [12]. Currently, among the mosquito-borne flaviviruses that have been fully sequenced, this is the only virus whose CS organization in the 3′-NCR has the CS3-CS2-CS1 pattern. This virus represents a distinct virus group (Spondweni) for which little has been known before about the CS organization in the 3′-NCR. The BAGV is an African virus that represents another distinct group (Ntaya) from which complete genome sequence has not been available. Previously, based on partial sequence of the NS5 gene, we demonstrated that this virus was synonymous with ITV [14]. This phylogenetic finding was corroborated by a considerable two-way cross neutralization result between the two viruses [3]. Because

ITV is a serious pathogen for domesticated birds not only in the Middle East but also in southern Africa [2], the BAGV sequence serves as a useful reference for a critically important study to determine the exact relationship between the two viruses and possibly for the development of a better veterinary vaccine for avian protection. Taxonomically, BAGV is currently classified as a member of the Ntaya virus group. However, the result of this study revealed that, in terms of CS organization in the 3′-NCR, it closely resembles the JEV complex viruses, except that BAGV does not have RCS2. Furthermore, our bootscanning revealed that this virus shares high levels of phylogenetic relatedness with WNV in many regions of the ORF. The closer relationship of BAGV with the JEV complex viruses was previously demonstrated in a phylogenetic tree [15].

Not much is known about human infection by KEDV in Africa. However, its recent classification as a member of the DENV group [11] prompted us to fully characterize its genome for an in-depth genomic comparison. Previously, KEDV was found to be closer to the DENV clade based on partial NS5 gene sequencing, but this virus was not included in the clade because of the genetic distance [14]. In the complete NS5 gene tree, the significant segregation of this virus from DENVs was more evident [15]. Bootscanning of the ORF of KEDV in this study confirmed only moderate levels of genetic relatedness with multiple viruses, each in a different genomic region. Furthermore, as described earlier, at least a few amino acid motifs uniquely found only among 4 serotypes of DENV were not found in KEDV. In addition, unlike DENVs, the length of the ORF of KEDV is much longer. Although KEDV and DENV share a similar CS organization in the 3′-NCR, the degree of sequence difference of ImRCS2 of KEDV is considerable. Finally, the 16-base sequence upstream of AUG in the 5′-NCR of KEDV is quite different from the 5′-UAR of DENV-2 or the comparable sequences of all other dengue serotypes. Thus, collectively, multiple results support our previous phylogenetic classification of KEDV in a phylogenetic cluster placed closer to, but separated from, the DENV cluster [14, 15].

The functional roles of the 3′-NCR of flaviviruses have been elucidated more recently regarding genome cyclization, viral replication and translation, and defects in certain critical regions in the conserved region that have a profound impact on the survival and/or pathogenicity of the virus. Currently, the presence of CS2 or CS3 (and their repeats) is not considered absolutely essential for viral replication, since spontaneous loss of those CSs is known to have occurred during laboratory cultivation and because infectious clones with deletion of CS2 and/or CS3 are nonetheless replication-competent even if they may demonstrate reduced rates of replication. However, as far as wild-type viruses are concerned, the accumulated data suggest a potential utility of the CS pattern for sub-grouping viruses within the mosquito-borne group. For example, all members of the JEV complex thus far fully sequenced (Alfuy virus, Murray Valley encephalitis virus, JEV, SLEV, Usutu virus, and WNV) uniquely share the RCS3-CS3-CS2-CS1 pattern that is not shared by other subgroups. Similarly, the RCS2-CS2-CS1 pattern is found only among the members of the DENV complex and KEDV. Thus, further studies are warranted to determine if the CS pattern is a useful, supplementary marker for taxonomic subgrouping of viruses within the mosquito-borne group.

The predicted cleavage sites of the three African viruses studied basically follow the patterns established for other mosquito-borne viruses, and, likewise, cysteine residues are well conserved. Regarding the functional significance of potential N-linked glycosylation sites, contrasting opinions have been presented. Generally, carbohydrate does not play a major role in the antigenic properties of flaviviruses, since deglycosylated viruses maintain the same antigenicity [25], and glycosylation does not alter epitope recognition [24]. On the other hand, others have found an important role in replication and maturation [16]. Infectious clones of the three African viruses that can be generated based on the full genome sequencing reported in this study will be useful for providing answers to some research questions raised in the discussion, including the functional role of glycosylation.

References

1. Alvarez DE, Lodeiro MF, Luduena SJ, Pietrasanta LI, Gamarnik AV (2005) Long-range RNA-RNA interactions circularize the dengue virus genome. J Virol 79: 6631–6643
2. Barnard BJ, Buy SB, Du Preez JH, Greyling SP, Venter HJ (1950) Turkey meningo encephalitis in South Africa. Onderstepoort J Vet Res 47: 89–94
3. Calisher CH, Karabatsos K, Dalrymple JM, Shope RE, Porterfield JS, Westaway EG, Brandt WE (1989) Antigenic relationship between flaviviruses as determined by cross-neutralization tests with polyclonal antisera. J Gen Virol 70: 37–43
4. Chambers TJ, Hahn CS, Galler R, Rice CM (1990) Flavivirus genome organization, expression, and replication. Annu Rev Microbiol 44: 649–688
5. Chang G-JJ (1997) Molecular biology of dengue viruses. In: Gubler DJ, Kuno G (eds) Dengue and dengue hemorrhagic fever. CAB International, Wallingford, UK, pp 175–198
6. Chang G-JJ, Hunt AR, Davis B (2000) A single intramuscular injection of recombinant plasmid DNA induces protective immunity and prevents Japanese encephalitis in mice. J Virol 74: 4244–4252
7. Felsenstein J (1995) PHYLIP, version 3.57c. Department of Genetics, University of Washington, Seattle, WA
8. Fu JL, Tan BH, Yap EH, Chan YC, Tan YH (1992) Full-length cDNA sequence of Dengue type 1 virus (Singapore strain S275/90). Virology 188: 953–958
9. Hahn CS, Hahn YS, Rice CM, Lee E, Dalgarno L, Strauss EG, Strauss JH (1987) Conserved elements in the 3′ untranslated region of flavivirus RNAs and potential cyclization sequences. J Mol Biol 198: 33–41
10. Hall TA (2001) BioEdit. Department of Microbiology, North Carolina State University, Raleigh, NC
11. ICTV (International Committee on the Taxonomy of Viruses) (2005) In: Fauquet CM, Mayo MA, Maniloff J, Desselberger U, Ball LA (eds) Virus taxonomy: classification and nomenclature of viruses. Elsevier, San Diego
12. Karabatsos N (1985) International catalogue of arboviruses, 3rd edn. American Society of Tropical Medicine and Hygiene, San Antonio, TX
13. Khromykh AA, Meka H, Guyatt KJ, Westaway EG (2001) Essential role of cyclization sequences in flavivirus RNA replication. J Virol 75: 6719–6728
14. Kuno G, Chang G-JJ, Tsuchiya KR, Karabatsos N, Cropp CB (1998) Phylogeny of the genus *Flavivirus.* J Virol 72: 73–83
15. Kuno G, Chang G-JJ (2005) Biological transmission of arboviruses: reexamination of and new insights into components, mechanisms, and unique traits as well as their Evolutionary trends. Clin Microbiol Rev 18: 608–637
16. Li J, Bhuvanakantham R, Howe J, Ng M-L (2006) The glycosylation site in the envelope protein of West Nile virus (Sarafend) plays an important role in replication and maturation processes. J Gen Virol 87: 613–622
17. Nicholas KB, Nicholas HB, Deerfield DW (1997) GeneDoc: analysis and visualization of genetic variation. EMBONet News 4: 14–18
18. Osatomi K, Sumiyoshi H (1990) Complete nucleotide sequence of dengue type 3 virus genome RNA. Virology 176: 643–647
19. Pierre V, Drouet M-T, Deubel V (1994) Identification of mosquito-borne flavivirus sequences using universal primers and reverse transcription/polymerase chain reaction. Res Virol 145: 179–188
20. Proutski V, Gould EA, Holmes EC (1997) Secondary structure of the 3′ untranslated region of flaviviruses: similarities and differences. Nucleic Acids Res 25: 1194–1202
21. Ray SC (1999) Simplot. School of Medicine, Johns Hopkins University, Baltimore, MD
22. Rice CM, Strauss JH (1990) Production of flavivirus polypeptides by proteolytic processing. Semin Virol 1: 357–367
23. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DJ (1997) The Clustal X window interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25: 4876–4882
24. Vorndam V, Mathews JH, Barrett ADT, Roehrig JT, Trent DW (1993) Molecular and biological characterization of a non-glycosylated isolate of St. Louis encephalitis virus. J Gen Virol 74: 2653–2660
25. Winkler G, Heinz FX, Kunz C (1987) Studies on the glycosylation of flavivirus E proteins and the role of carbohydrates in antigenic structure. Virology 159: 237–243
26. Zhao BT, Mackow E, Buckler-White A, Markoff L, Chanock RM, Lai CJ, Makino Y (1986) Cloning full-length dengue type 4 viral DNA sequences: analysis of genes coding for structural proteins. Virology 155: 77–88
27. Zuker M, Mathews DH, Turner DH (1999) Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide. In: Barciszewski J, Clark BFC (eds) RNA biochemistry and biotechnology. Kluwer Dardsecht, pp 11–43