JOURNAL OF VIROLOGY, Sept. 1995, p. 5773–5780
0022-538X/95/$04.00+0

Vol. 69, No. 9

Nucleotide Sequence Variation of the Envelope Protein Gene Identifies Two Distinct Genotypes of Yellow Fever Virus

GWONG-JEN J. CHANG,[1]* BRUCE C. CROPP,[1] RICHARD M. KINNEY,[1]
DENNIS W. TRENT,[2] AND DUANE J. GUBLER[1]

Division of Vector-Borne Infectious Diseases, National Center for Infectious Diseases, Centers for Disease Control and Prevention, Public Health Service, U.S. Department of Health and Human Services, Fort Collins, Colorado 80522,[1] *and Division of Viral Products, Center for Biologics Evaluation and Research, Food and Drug Administration, Bethesda, Maryland 20892*[2]

Received 3 April 1995/Accepted 19 June 1995

The evolution of yellow fever virus over 67 years was investigated by comparing the nucleotide sequences of the envelope (E) protein genes of 20 viruses isolated in Africa, the Caribbean, and South America. Uniformly weighted parsimony algorithm analysis defined two major evolutionary yellow fever virus lineages designated E genotypes I and II. E genotype I contained viruses isolated from East and Central Africa. E genotype II viruses were divided into two sublineages: IIA viruses from West Africa and IIB viruses from America, except for a 1979 virus isolated from Trinidad (TRINID79A). Unique signature patterns were identified at 111 nucleotide and 12 amino acid positions within the yellow fever virus E gene by signature pattern analysis. Yellow fever viruses from East and Central Africa contained unique signatures at 60 nucleotide and five amino acid positions, those from West Africa contained unique signatures at 25 nucleotide and two amino acid positions, and viruses from America contained such signatures at 30 nucleotide and five amino acid positions in the E gene. The dissemination of yellow fever viruses from Africa to the Americas is supported by the close genetic relatedness of genotype IIA and IIB viruses and genetic evidence of a possible second introduction of yellow fever virus from West Africa, as illustrated by the TRINID79A virus isolate. The E protein genes of American IIB yellow fever viruses had higher frequencies of amino acid substitutions than did genes of yellow fever viruses of genotypes I and IIA on the basis of comparisons with a consensus amino acid sequence for the yellow fever E gene. The great variation in the E proteins of American yellow fever virus probably results from positive selection imposed by virus interaction with different species of mosquitoes or nonhuman primates in the Americas.

Yellow fever (YF) virus, an arthropod-borne virus in the *Flavivirus* genus of the family *Flaviviridae*, is the etiologic agent of YF, a viral hemorrhagic fever that occurs in six tropical South American countries and much of sub-Saharan Africa (29). The primary transmission cycle involves nonhuman primates and tree-hole-breeding mosquitoes. Mosquitoes of the genera *Haemagogus* in tropical America and *Aedes* in equatorial Africa are the principal vectors responsible for year-round virus transmission (28). Humans, when they intrude into this cycle, are exposed to the infective mosquitoes, and transmission may occur by sylvatic vectors. The urban mosquito, *Aedes aegypti*, is a very efficient vector and may transmit the virus to humans, resulting in explosive outbreaks of urban YF (28).

A safe and effective YF vaccine, 17D, has been available since 1937. However, it is not used effectively, and the disease continues to occur in Africa and South America, where 21 and 3 outbreaks, respectively, were recorded between 1940 and 1987 (29). A major public health concern in Africa has resurfaced as a result of a devastating YF outbreak in Nigeria and surrounding countries between 1986 and 1993 (8, 36) and the first recorded YF outbreak in the East African country of Kenya in 1992 and 1993 (26).

YF virus has a single-stranded, positive-polarity RNA genome of 10,862 nucleotides (nt) which encodes three structural proteins (C, PrM, and E) and seven nonstructural proteins (NS1, NS2A, NS2B, NS3, NS4A, NS4B, and NS5) in a single open reading frame (33). Antigenic differences between South American and West African strains of YF virus have been shown by antibody absorption techniques (6). Strains can also be differentiated on the basis of mouse virulence (2, 16). Four genetic topotypes of YF virus have been distinguished by RNA oligonucleotide fingerprinting (9), and three have been identified by analysis of limited envelope (E) protein gene sequences (23).

The E protein is important in receptor binding, hemagglutination of erythrocytes at acid pH, induction of the protective immune response, and involvement in an intraendosomal, acid-catalyzed fusion step necessary for infection (18). Because of its biologic importance, we have sequenced the entire YF E protein gene of 20 epidemiologically important viruses. Phylogenetic analysis of the 1,479-nt E gene sequences identified two E genotypes of YF viruses. The data genetically link YF viruses isolated during outbreaks that occurred from 1986 to 1993 in Nigeria and from 1992 to 1993 in Kenya to previously isolated YF viruses. Amino acid sequence comparisons of the E proteins of antigenically distinct YF viruses have provided an explanation for the observed antigenic variation among these viruses.

MATERIALS AND METHODS

Virus strains and reverse transcription (RT)-PCR amplification of viral RNA. YF viruses sequenced in this study were obtained from the Division of Vector-Borne Infectious Diseases, National Center for Infectious Diseases, Centers for

* Corresponding author. Mailing address: Division of Vector-Borne Infectious Diseases, National Center for Infectious Diseases, Centers for Disease Control and Prevention, P.O. Box 2087, Fort Collins, CO 80522. Phone: (970) 221-6400. Fax: (970) 221-6476.

TABLE 1. YF viruses used to analyze the E protein gene sequence

Country	Yr	Designation	Strain	Source of isolation	Genotype	Accession no. (reference)
Sudan	1940	SUDAN40	M-112	Human	I	U23577
Ethiopia	1961	ETHIOP61	Kouma	Human	I	U23576
Uganda	1964	UGANDA64	SE7445	Human	I	U23578
Central African Republic	1977	CAFR77	ArB8883	*Aedes africanus*	I	U23571
Central African Republic	1985	CAFR85	DaHB1504	Human	I	U23573
Kenya	1993	KENYA93H	7914	Human	I	U23569
Kenya	1993	KENYA93M	KE93-477	*Aedes africanus*	I	U23575
Ghana	1927	GHANA27	Asibi	Human	IIA	(17)
Senegal	1927	SENEGAL27	FVV	Human	IIA	L02865 (21)
Senegal	1965	SENEGAL65	HD1279	Human	IIA	U23574
Nigeria	1986	NIGERIA86	BA-55	Human	IIA	U23572
Nigeria	1991	NIGERIA91	56205	Human	IIA	U23567
Trinidad	1979	TRINID79A	788379	*Haemagogus* sp.	IIA	U23568
Trinidad	1979	TRINID79B	T790882	*Haemagogus* sp.	IIB	U23579
Trinidad	1979	TRINID79C	T797984	*Haemagogus* sp.	IIB	NA[a]
Colombia	1979	COLOMB79	V-528A	*Haemagogus* sp.	IIB	U23580
Brazil	1978	BRAZIL78	AR350397	*Haemagogus* sp.	IIB	U23570
Peru	1977	PERU77	1362	Human	IIB	U23565
Peru	1981	PERU81	1899/81	Human	IIB	D14458 (1)
Ecuador	1981	ECUADO81	1914/81	Human	IIB	U23566

[a] NA, not available.

Disease Control and Prevention (Fort Collins, Colo.). Of the 20 YF virus sequences analyzed, 3 were obtained from previous publications (1, 17, 21). Designation, country of origin, year of isolation, E genotype, and GenBank accession number for the E gene sequence of each virus are summarized in Table 1.

Viral RNA was extracted directly from virus stock, as previously described (24). The YF virus-specific oligonucleotides used in the RT-PCR amplification of viral RNA and the sequencing of the E gene are shown in Table 2. The RT-PCR to convert single-stranded YF RNA to cDNA and to amplify double-stranded cDNA was performed in a reaction volume of 100 μl containing 10 mM Tris-HCl (pH 8.5), 1.5 mM $MgCl_2$, 50 mM KCl, 0.01% gelatin, 1 μM (each) deoxynucleotide triphosphate, 5 mM dithiothreitol, 100 pmol (each) of two amplimers, 0.15 U of RAV-2 reverse transcriptase (Amersham, Arlington Heights, Ill.), 40 U of RNasin (Promega, Madison, Wis.), 2.5 U of Amplitaq polymerase (Perkin-Elmer Corp., Norwalk, Conn.), and 5 to 15 μl of viral RNA. This single-tube RT-PCR was incubated at 50°C for 1 h, denaturated at 94°C for 4 min, and then subjected to 30 cycles of denaturation at 94°C for 30 s, primer annealing at 50°C for 30 s, and primer extension at 72°C for 5 min and terminated with a final extension step at 72°C for 20 min before being held at 4°C in a thermocycler GeneAmp PCR System 9600 (Perkin-Elmer).

Nucleotide sequencing. Double-stranded PCR-amplified DNAs from multiple amplification reactions were purified with a Geneclean II kit (Bio 101, Inc., La Jolla, Calif.). Sequencing reactions were performed with Taq DyeDeoxy Terminator cycle sequence kits (Applied Biosystems, Foster City, Calif.). Sequences were resolved with an ABI 374 sequencer (Applied Biosystems).

Nucleic acid and protein sequence analysis. The YF E gene sequences and deduced amino acid sequences were analyzed with a computer program, HIBIO DNASIS/PROSIS (Hitachi Software Engineering, Brisbane, Calif.). A strict consensus sequence of the YF E gene, YFCONSEQ, was obtained with the computer program CONP (22). This program calculated the frequency of occurrence of each nucleotide at each E gene position and applied the majority rule method to assign a specific nucleotide to this position in the consensus sequence.

The nucleotide sequences of the YF E genes were aligned with YFCONSEQ or the E gene sequences of dengue 1, dengue 2, dengue 3, dengue 4, Japanese encephalitis, St. Louis encephalitis, Kunjin, tick-borne encephalitis, and cell-fusing-agent viruses by the Clustal V alignment program (3, 7, 11, 19, 27, 31, 32, 34, 37, 38). Phylogenetic trees were constructed with the PHYLIP version 3.5c program package (14, 15). The most parsimonious tree was generated with the uniformly weighted parsimony program of DNAPARS. The significance of the phylogenetic trees was subjected to statistical bootstrap analysis with the programs SEQBOOTS (to generate 100 reiterated data sets), DNAPARS (randomize input order option "on" and 10 replicas), and CONSENSE or with the programs SEQBOOTS, DNADIST, FITCH, and CONSENSE. The resulting phylogram was drawn with the program DRAWGRAM.

Nucleotide and amino acid signature pattern analysis. Nucleotide and amino acid signature pattern analysis based on phylogenetically informative characters was applied to examine the genetic similarity of YF virus strains in each genotype. Unique signature sites within the E gene of each genotype were identified

TABLE 2. Oligonucleotide primers used to amplify and sequence YF virus E protein genes

Primer[a]	Sequence	Melting temp (°C)	Utility[b]
YF791	5′ CGGCAAGAAAAATGGATGACTGGAAGAATGGG	66.4	Amp and seq
YF1196	5′ CCCAGCACTGGAGAGGCCC	61.3	Seq
cYF1196	5′ GGGCCTCTCCAGTGCTGGG	61.3	Seq
YF1268	5′ AGAGGCTGGGGCAATGGCTGTGG	68.7	Seq
YF1402	5′ GCATGTAGGGGCCAAGCAGGAAA	64.8	Seq
YF1459	5′ CAGTTACATCGCTGAGATGGAAACAGAGAGCTGGA	71.6	Amp and seq
cYF1459	5′ TCCAGCTCTCTGTTTCCATCTCAGCGATGTAACTG	71.6	Seq
cYF1623	5′ CCCCGCCACTTCCACTCTGCC	66.5	Amp and seq
YF1823	5′ AGAGTGAAACTGTCAGCTTTAACACTCAAGGG	66.8	Seq
cYF1823	5′ CCCTTGAGTGTTAAAGCTCACAGTTTCACTCT	64.6	Seq
YF1921	5′ CACTGTTGTGATGCAGGTGAAAGTGT	61.0	Seq
cYF1921	5′ ACACTTTCACCTGCATCACAACAGTG	61.0	Seq
cYF2470	5′ CAACTTTGGCAAGAGAGAGCTCAAGTGCGGAG	71.7	Amp and seq
cYF2649	5′ CTCATCTGCCCTGCTTCTCCACATCT	64.5	Amp and seq

[a] The number indicates the genomic position of the oligonucleotide primer. A "c" indicates an antigenomic-sense oligonucleotide primer.
[b] Amp, amplification; seq, sequencing.

TABLE 3. Ranges of nucleotide and amino acid sequence similarities for the E protein gene of YF viruses from different geographic locations

Virus by source[a]	Similarity (%)					
	West Africa		America		East and Central Africa	
	Nucleotide	Amino acid	Nucleotide	Amino acid	Nucleotide	Amino acid
YFCONSEQ	89.8–91.6	97.0–99.4	89.7–90.3	96.6–97.4	85.9–88.4	97.4–99.0
West Africa	89.7–99.7	97.0–99.8	83.8–85.5	93.9–96.8	79.2–82.0	94.3–98.4
America			89.3–99.6	95.7–99.8	79.0–81.3	93.9–96.3
East and Central Africa					90.7–99.9	97.4–99.8

[a] YFCONSEQ is the consensus sequence of the YF virus E gene.

by comparing the alignments of genotypes I, IIA, and IIB virus E gene sequences with the nucleotide and amino acid sequences of YFCONSEQ.

RESULTS

Nucleotide and amino acid sequence analysis. A comparison of the 20 YF virus E gene sequences revealed nucleotide substitutions scattered throughout the entire gene. There were no base insertions or deletions. Viruses isolated in West Africa and East and Central Africa had nucleotide sequence similarities of 89.7 to 99.7 and 90.7 to 99.9%, respectively (Table 3). American viruses had a similarity of 89.3 to 99.6%, except for the TRINID79A isolate, which had a higher similarity to West African viruses (89.7 to 98.2%) than to American viruses (84.2 to 84.8%) (data not shown). Intergeographic similarities were lower than intrageographic similarities; however, the intergeographic similarities range was narrower than the intrageographic similarities range. A YF virus consensus sequence, YFCONSEQ, which represents the dominant genetic code in the collection of YF E gene sequences, was constructed from a total of 20 E gene sequences. This YFCONSEQ had nucleotide similarities to the West African, American, and East and Central African virus sequences in the ranges of 89.8 to 91.6, 89.7 to 90.3, and 85.9 to 88.4%, respectively (Table 3).

The ratio of silent base substitutions to substitutions resulting in an amino acid change in the encoded E protein was 10 to 1 (data not shown), which was reflected in the greater conservation among amino acid sequences than among the nucleotide sequences (Table 3). The E proteins of all YF virus strains had potential N-linked glycosylation sites at amino acid positions 309 and 470. An additional glycosylation site was present at position 269 in five of the American viruses (PERU77, PERU81, COLOMB79, EQUADO81, and TRINID79B) but was absent from the isolates TRINID79A and BRAZIL78. The Lys-to-Asn change at E-93 in the ETHIOP61 isolate created an additional potential glycosylation site at this position (Fig. 1).

The TRINID79A virus had 11 unique amino acids relative to those of the E protein of other wild-type YF viruses. Three of these unique amino acids (E-54, -153, and -249) were identical to those of the FNV vaccine virus, a derivative of the SENEGAL27-FVV isolate (Fig. 1) (21). TRINID79A also had Arg at E-331, as did SENEGAL27, FNV, and all other South American viruses. The aligned E gene nucleotide sequences of 20 YF viruses are available upon request.

Phylogenetic analysis of YF virus E protein gene sequences. The pairwise similarity measurements presented in Table 3 do not provide information about the geographic relationships among the viruses. To examine the range of relationships among viruses, we subjected the E gene nucleotide sequence data sets to phylogenetic tree analyses. The data sets were aligned with those of the E genes of other flaviviruses (dengue 1, dengue 2, dengue 3, dengue 4, Japanese encephalitis, St. Louis encephalitis, tick-borne encephalitis, and cell-fusing-agent viruses) or YFCONSEQ. Most of the parsimony trees constructed with these data sets were identical, except when data were aligned with the E gene sequence of the cell-fusing-agent virus. Use of various outgroup virus sequences, except for that of the cell-fusing-agent virus, had little effect on the outcome of phylo- genetic tree construction (data not shown). We selected YFCONSEQ to serve as the outgroup and reference sequence in the phylogenetic analysis of the YF virus E gene data.

The most parsimonious tree required 919 steps (data not shown). Two trees of equal lengths with no differences in the monophyletic grouping were generated when DNAPARS was used with the randomized input sequence order in effect. Bootstrap resampling that used either a parsimony algorithm or the distance method of Fitch produced identical phylograms. A strict consensus tree, compiled from the bootstrap analysis with the parsimony algorithm (14, 15), is presented in Fig. 2. The numbers at the forks indicate the number of times the monophyletic group consisting of the viruses to the right of that fork occurred among the 100 trees. Higher numbers indicate higher confidence levels that the monophyletic nest consists of viruses to the right of the fork. By this procedure, two major genetic lineages or genotypes of YF viruses were identified. The monophyletic grouping of East and Central African viruses, as well as that of viruses from West Africa and America, was observed in all of the 100 reiterated data sets. E genotype I included all viruses from East and Central Africa. Genotype II consisted of two monophyletic nests, one of IIA viruses from West Africa and TRINID79A virus from America and the other of IIB viruses from America. Because of the eccentric genetic relationship of TRINID79A virus to IIA viruses, the E gene of two more 1979 Trinidad virus isolates, TRINID79B and TRINID79C, was sequenced. These virus isolates were identical, but they differed from the TRINID79A isolate. Only the sequence of the TRINID79B virus was used in our analysis.

A human isolate, KENYA93H, and a mosquito isolate, KENYA93M, both from a recent YF outbreak in Kenya, had a nucleotide sequence similarity of 99.9% and belonged to genotype I. The NIGERIA86 and NIGERIA91 viruses, isolated from same general area in Nigeria during a prolonged epidemic from 1986 to 1993, belonged to genotype IIA and had a nucleotide sequence similarity of 99.7%. SUDAN40 and ETHIOP61, isolated during epidemics in 1940 and 1961 in Sudan and Ethiopia, respectively, had a similarity of 98.2% and shared 99.1 and 98.5% nucleotide sequence similarities with UGANDA64, respectively. The remarkable similarity among these three viruses implied that the 1940 Sudan and 1961 Ethiopia epidemics were caused by a similar, UGANDA64-like virus.

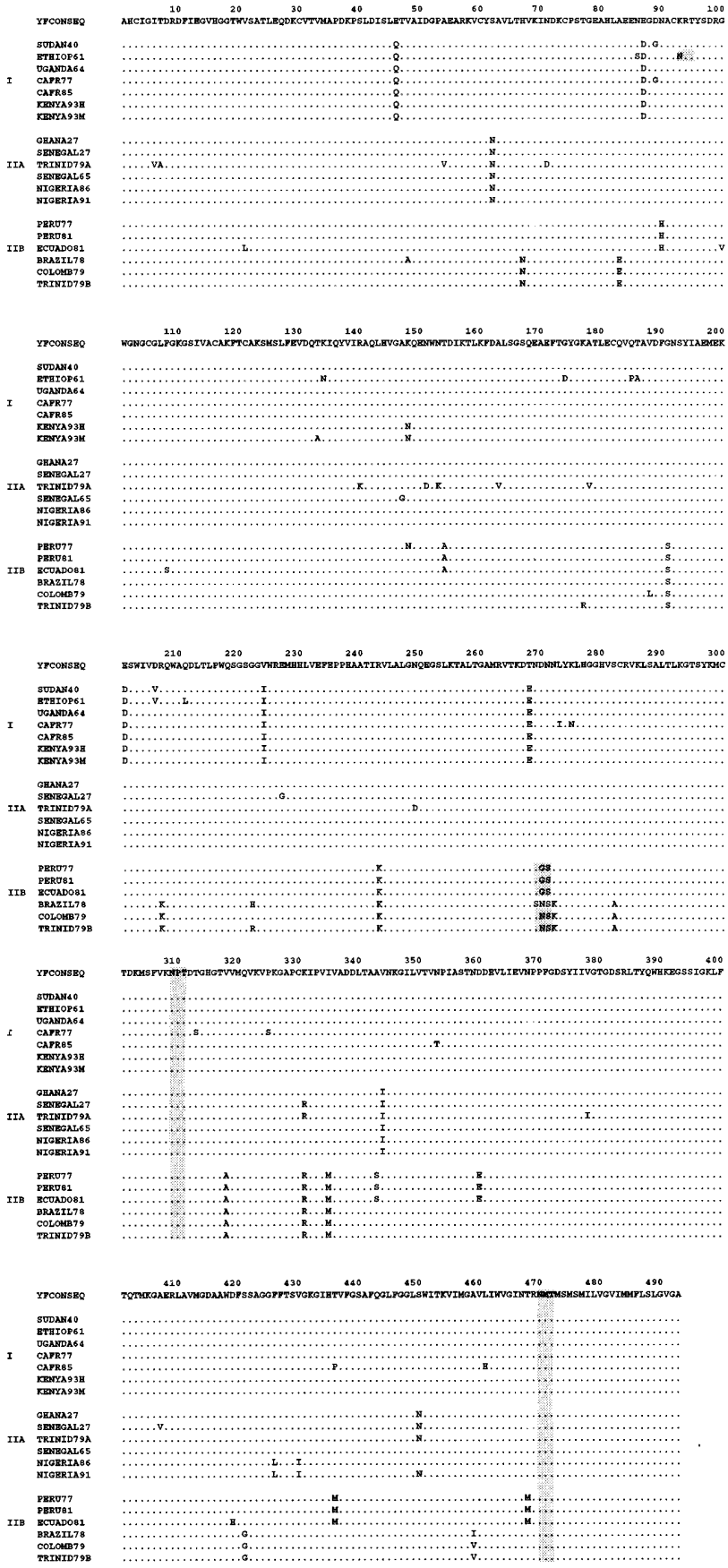

FIG. 1. Aligned deduced amino acid sequences of the E protein of YF viruses from different geographic locations and years of isolation. See Table 1 for virus isolation and designation information. The shaded areas indicate the potential N-linked glycosylation sites.

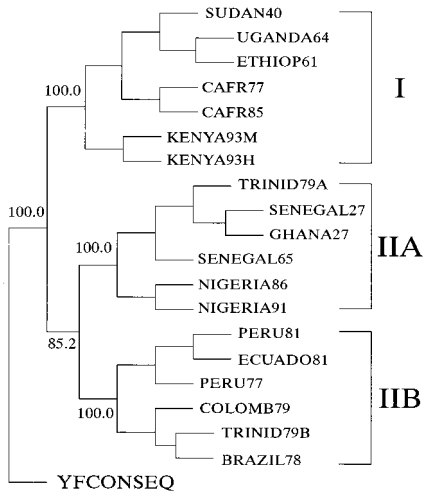

FIG. 2. Phylogram derived from the nucleotide sequences of the E gene of YF viruses, illustrating the evolutionary relationships of genotypes I and II (IIA and IIB) of YF viruses. The procedures used for tree construction are described in Materials and Methods. The numbers at the forks indicate the number of times the monophyletic group consisting of the viruses to the right of that fork occurred among the 100 trees.

Population mixing among different genotypes was not observed. This was reflected by the narrower range of variation observed in the pairwise similarity comparisons of viruses of different genotypes (Table 3) and by the clustering of related viruses within genotypes established by 100% bootstrap confidence values (Fig. 2). A near linear phylogenetic relationship among genotype I and IIA viruses was observed. These viruses had average genetic drifts of 1.6×10^{-3} (from 1927 to 1991) and 1.7×10^{-3} (from 1940 to 1993) substitutions per nt per year, respectively. Genetic drift between two viruses isolated from the Central African Republic was higher, running at a rate of 3.6×10^{-3} substitutions per nt per year. The NIGERIA86 and NIGERIA91 viruses, isolated during the same epidemic from 1986 to 1993, had a drift rate of 6.8×10^{-4} substitutions per nt per year.

Nucleotide and amino acid signature patterns of YF virus E genotypes. Nucleotide and amino acid signature pattern analysis based on phylogenetically informative characters was applied to examine the genetic similarities of the genotypes. A unique signature pattern composed of 111 nt and 12 amino acids (aa) was identified within the E gene by comparing an alignment of genotype I, IIA, and IIB virus sequences with the nucleotide and amino acid sequences of YFCONSEQ (Table 4). Six signature amino acid sites were encoded by codons that contained two signature nucleotides at codon positions 136 to 138, 184 to 186, 571 to 573, 802 to 804, 952 to 954, and 1030 to 1032. Viruses from East and Central Africa had a signature pattern consisting of residues at 60 nt and 5 aa positions, and these residues differed from the residues found in those positions in West African and American viruses. West African viruses had 25 unique nt and 2 unique aa signature sites, and American viruses had 30 such nt and 5 such aa sites.

The nucleotide signature sites were scattered throughout the entire E gene. Although G-to-A transitions were the most common change observed (20.87%), transversion and transition frequencies were about the same, 50.45 and 49.56%, respectively (data not shown). The substitution matrix was highly asymmetric. T-to-A (12.17%) and T-to-G (5.22%) nucleotide transversions were more common than T-to-C (3.48%) nucleotide transitions. Nucleotide transversions occurred at 60.3, 29.6, and 44.8% of the signature sites of genotype I, IIA, and IIB viruses, respectively (data not shown).

The predicted structural characteristics of amino acid changes observed in the signature sites of the E protein sequences of viruses from different genotypes are summarized in Table 5. On the basis of the tick-borne encephalitis virus E protein model of Mandl et al. (25), 8 aa changes were in the antigenic domain A of the E protein and seven of these changes were located in the hydrophilic region. This domain is sensitive to low pH, sodium dodecyl sulfate, and the reduction of disulfide bridges and contains flavivirus cross-reactive and subtype-specific epitopes. The E protein changes at aa residues 46, 268, and 344 resulted in a predicted secondary structural alteration of the protein.

DISCUSSION

Two major evolutionary lineages or genotypes of YF viruses were identified by analysis of the E protein gene sequences of 20 virus strains isolated over a period of 67 years. The uniform weighted parsimony method was used to analyze sequence data, because we observed an equal probability of transition and transversion in the E protein genes of YF viruses. Hillis et al. (20) assessed the performance of various phylogenetic analysis methods by numerical simulation and by experimental evolution of organisms under controlled laboratory situations. The parsimony method consistently performs better than the unweighted pair-group method of averages with Kimura distances (UPGMA), the neighbor joining method with Kimura distances, or the evolutionary parsimony of Lake's invariants method. Division of the YF viruses into two major E genotypes was further supported by bootstrap analysis with the parsimony algorithm (15) (Fig. 2) and nucleotide signature analysis (Table 4).

Selection of the outgroup sequence can influence the outcome of sequence alignments and significance of phylogenetic relationships. A deviate phylogram was observed when the data set was aligned with the sequence of the cell-fusing-agent virus (data not shown). The YF virus strict consensus sequence, YFCONSEQ, generated a significant, accurate phylogram and represents a collective and dominant sequence of 20 YF viruses isolated over the past 67 years. It also has the advantage of eliminating both any sequence abnormality caused by cloning or sequencing artifacts and the inherent geographical bias of the resulting root in choosing a single viral sequence.

Recently, Nichol et al. (30) described the evolution of the vesicular stomatitis virus (VSV) as being highlighted by a stepwise evolutionary pattern outward from the tree's ancestral root to the terminal branch tips, with grossly unequal rates of change within the species. These authors attributed this observation to positive Darwinian evolution due to selection pressure by a great diversity of potential insect hosts in the different ecological zones of VSV disease activity (30). We did not observe the VSV-like stepwise evolutionary pattern in our analysis of the YF virus. The primary transmission cycle of the YF virus involves wild nonhuman primates and two genera of mosquitoes, *Haemagogus* in tropical America and *Aedes* in equatorial Africa, with humans as incidental hosts (28, 29). It is possible that the less complex and restricted virus-host relationship of YF virus is the reason for the observed differences in VSV and YF virus evolution and why only two distinct genotypes of YF virus were observed.

YF virus has evolved independently in different geographic locations, as evidenced by the monophyletic grouping of viruses from the same region (Fig. 2). The genetic variation of

TABLE 4. Nucleotide and amino acid signatures in the E protein gene of YF viruses[a]

Position[b]		YFCONSEQ signature		Genotype signature[c]					
				I		IIA		IIB	
nt	aa	nt	aa	nt	aa	nt	aa	nt	aa
3	—	T		A		·		·	
6	—	C		T		·		·	
12	—	T		·		·		A	
27	—	G		A		·		·	
42	—	G		T		·		·	
54	—	A		C		·		·	
81	—	G		·		A		·	
90	—	T		C		·		·	
114	—	G		·		·		A	
132	—	A		G		·		·	
136	**46**	**G**	**E**	**C**	**Q**	·	*	·	*
138	**46**	**A**	**E**	**G**	**Q**	·	*	·	*
147	—	C		A		·		·	
150	—	T		A		·		·	
177	—	G		T		·		·	
185	**62**	**G**	**S**	·	*	**A**	**N**	·	*
186	**62**	**T**	**S**	**C**	*	·	**N**	·	*
189	—	A		G		·		·	
193	—	C		·		·		T	
207	—	G		A		·		·	
237	—	G		·		·		A	
252	—	A		G		·		·	
255	—	A		·		G		·	
261	**87**	**A**	**E**	**T**	**D**	**A/G**	*	**A/G**	*
264	—	T		·		G		A	
288	—	C		·		T		·	
379	—	C		·		T		·	
390	—	T		G		·		·	
420	—	G		·		A		·	
426	—	G		·		A		·	
450	—	G		·		A		·	
462	—	A		·		C		T	
474	—	T		A		·		·	
495	—	A		T		·		·	
513	—	G		A		·		·	
525	—	T		·		·		C	
528	—	G		·		A		·	
531	—	A		·		·		G	
543	—	G		·		A		·	
571	**191**	**G**	**G**	·	*	·	*	**A**	**S**
573	**191**	**C**	**G**	·	*	**C/T**	*	·	**S**
588	—	A		·		T		·	
603	**201**	**G**	**E**	**C**	**D**	·	*	·	*
615	—	G		T		·		·	
619	—	A		C		·		·	
645	—	A		·		G		·	
648	—	A		·		·		T	
666	—	C		T		·		·	
670	—	G		A		·		·	
672	**224**	**G**	**V**	**A**	**I**	·	*	·	*
678	—	A		G		·		·	
699	—	A		G		·		·	
717	—	C		·		·		T	
728	**243**	**G**	**R**	·	*	·	*	**A**	**K**
750	—	A		·		G		·	
753	—	A		G		·		·	
787	—	C		·		A		·	
801	—	C		·		·		T	
802	**268**	**A**	**T**	**G**	**E**	·	*	·	*
803	**268**	**C**	**T**	**A**	**E**	·	*	·	*
812	**271**	**A**	**D**	·	*	·	*	**G**	**S**
834	—	T		·		·		G	
840	—	T		·		·		C	
900	—	C		·		·		T	
921	—	C		G		·		·	
924	—	G		A		·		·	
927	—	C		·		·		T	
933	—	T		G		·		·	
939	—	T		A		·		·	
953	**318**	**T**	**V**	·	*	·	*	**C**	**A**
954	**318**	**T**	**V**	·	*	·	*	**A/C**	**A**
972	—	G		C		·		·	
975	—	A		C		·		·	
987	—	C		A		·		·	
1005	**335**	**A**	**I**	**A/T**	*	·	*	**G**	**M**
1020	—	A		·		T		C	
1030	**344**	**G**	**V**	·	*	**A**	**I**	·	*
1032	**344**	**C**	**V**	**G**	*	·	**I**	·	*
1035	—	C		·		T		·	
1044	—	T		A		·		·	
1053	—	A		T		·		·	
1062	—	C		·		·		A	
1065	—	A		·		C		·	
1089	—	G		C		·		·	
1090	—	C		·		·		T	
1092	—	G		·		A		·	
1098	—	G		A		·		·	
1150	—	C		A		·		·	
1179	—	A		·		·		C	
1191	—	G		·		A		·	
1194	—	G		A		·		·	
1209	—	C		A		·		·	
1218	—	C		A		·		·	
1227	—	C		T		·		·	
1230	—	G		T		·		·	
1242	—	T		·		A		G	
1257	—	C		·		T		·	
1269	—	T		A		·		·	
1293	—	G		·		·		A	
1302	—	T		·		·		A	
1320	—	T		A		·		·	
1323	—	C		T		·		·	
1377	—	T		·		G		A	
1386	—	A		C		·		·	
1405	—	A		C		·		·	
1416	—	A		C		·		·	
1426	—	A		T		·		·	
1427	—	G		C		·		·	
1434	—	C		T		·		·	
1446	—	G		·		·		C	
1479	—	G		A		·		·	
Total		111	12	60	5	25	2	30	5

[a] Boldface indicates amino acid (aa) changes at that nucleotide (nt) position.
[b] Positions are numbered from the 5′ end of the E protein gene. Amino acid sites 46, 62, 191, 268, 318, and 344 are encoded by codons containing 2 nt signature sites. —, no amino acid change.
[c] Dots and asterisks indicate nucleotides and amino acids identical to those of YFCONSEQ, respectively.

RNA viruses is the result of at least two distinct processes (for a review, see reference 12). An array of mutations generated by the error-prone nature of RNA polymerase (referred to as the mutation rate) during viral RNA replication must be checked by the competitive ability of mutant viruses to arise as a viable virus (referred to as the mutation frequency). Population equilibrium is represented by a dynamic spectrum of mutants. Wild-type viruses exist as a quasispecies composed of a self-perpetuating population of diverse, related entities which may include one or several dominant, master sequences that act as a whole (12, 13). The narrow intrageographic similarities range

TABLE 5. Characteristics of amino acid changes observed in the signature sites of YF virus E protein genes

Signature[a]				Antigenic domain[b]	Predicted characteristic	
YFCONSEQ (position)	I	IIA	IIB		Hydrophobicity[c]	Secondary structure[d]
E (46)	Q	*	*		+	H→S
S (62)	*	N	*	A	+	S
E (87)	D	*	*	A	−	T
G (191)	*	*	S	A	−	C
E (201)	D	*	*	A	−	H
V (224)	I	*	*	A	−	H
R (243)	*	*	K	A	−	S
T (268)	E	*	*	A	−	S→T
D (271)	*	*	S	A	−	C
V (318)	*	*	A	B	+	S
I (335)	*	*	M	B	+	S
V (344)	*	I	*	B	+	H→S

[a] *, same amino acid as that of YFCONSEQ.
[b] From Mandl et al. (25). The domain for position 46 is not known.
[c] +, hydrophobic region; −, hydrophilic region.
[d] H, α-helix; S, β-sheet; T, turn; C, random coil.

obtained by pairwise comparison can be viewed as a well-established equilibrium of YF virus quasispecies resulting from positive selection pressures in the ecosystem to maintain genotypic homogeneity.

It has been suggested that YF virus originated in Africa (5, 35) and was disseminated to South America during slave trading. The genetic relatedness of American YF to West African YF viruses in our study supports this notion and suggests that the dissemination originated from West Africa. Virus isolate TRINID79A is more closely related to West African than to American viruses. The presence of both West African and American YF virus strains in Trinidad confirms and extends the report of Cane and Gould (4) in which a Trinidadian strain (YF 4205) more closely resembled African strains of YF virus than South American strains, as determined by immunoblots. An identical virus strain, 788379 (TRINID79A), was grouped as the American topotype on the basis of an RNase T1 oligonucleotide fingerprint of viral RNA (Table 1) (9). Our result with TRINID79A is probably not the result of contamination during virus propagation, viral RNA extraction, or PCR amplification. All viral RNAs in this study were extracted directly from stock seed viruses. An identical result was obtained by repeating the viral RNA extraction, RT-PCR amplification, and sequencing. It had Arg at E-331, as do SENEGAL27, FNV, and all other South American viruses. Three unique amino acids at E-54, E-153, and E-249 occurred in both TRINID79A and FNV viruses (21). Most importantly, 9 aa were strictly unique to the TRINID79A virus (E-6, -7, -71, -140, -151, -163, -177, -378, and -407) (Fig. 1). Although there is no evidence for the introduction of FNV virus into nature, it is possible that TRINID79A evolved from FNV, originating in a vaccinated viremic patient, and may represent a recent introduction of West African YF virus to the Americas. The TRINID79A-like virus may have had an evolutionary disadvantage because of its close relationship to the FNV virus and may have been eliminated from natural circulation, since the two other Trinidad viruses analyzed conformed to the other American viruses in genotype IIB.

Recently, the geographic distribution and evolution of 22 YF viruses have been studied on the basis of the sequencing of limited E gene regions (23). The UPGMA method and 10% nucleotide divergence were subjectively chosen to separate YF viruses into three geographically distinct topotypes. Because of the inconsistency of the UPGMA method (20) and the inclusion of only a single American virus in this study, we reanalyzed our data and those of Lepiniec et al. (23) by limiting the analysis to the E gene region between nt 871 and nt 1218 and using bootstrap resampling with the parsimony algorithm (data not shown) (14). By phylogenetic analysis of this limited 348-bp region, West African viruses were not consistently separated into a distinctive lineage. American viruses were closely related to East and Central African viruses and were grouped together in 68.9 out of 100 observations. Use of a different analysis method (parsimony versus UPGMA) and larger numbers of American viruses (six versus one) and application of an outgroup sequence (YFCONSEQ versus none) may contribute to the differences between our reanalyzed results and those of Lepiniec et al. (23). A limited number of phylogenetically informative sites obtained by the limited sequencing may be the major contributor to this dissimilar result (Fig. 2) (23).

Changes in virus-host relationships imposed by different species of mosquitoes or nonhuman primates may have altered the quasispecies of YF virus in America. More unique amino acid changes relative to the number of nucleotide substitutions in the E protein gene region have accumulated in American viruses than in viruses of YF virus genotypes I and IIA (Table 3 and 4). This evolution has also altered mouse virulence, antigenic character, and monoclonal antibody reactivity (2, 5, 10, 16). The 5 unique aa signatures in the American viruses may be the result of positive selection pressure by virus-host interactions (Table 4).

REFERENCES

1. **Ballinger-Crabtree, M. E., and B. R. Miller.** 1990. Partial nucleotide sequence of South American yellow fever virus strain 1899/81: structural proteins and NS1. J. Gen. Virol. **71:**2115–2121.
2. **Barrett, A. D. T., and E. A. Gould.** 1986. Comparison of neurovirulence of different strains of yellow fever virus in mice. J. Gen. Virol. **67:**631–637.
3. **Cammisa-Parks, H., L. A. Cisar, A. Kane, and V. Stollar.** 1992. The complete nucleotide sequence of cell fusing agent (CFA): homology between the nonstructural proteins encoded by CFA and the nonstructural proteins encoded by arthropod-borne flaviviruses. Virology **189:**511–524.
4. **Cane, P. A., and E. A. Gould.** 1989. Immunoblotting reveals differences in the accumulation of envelope protein by wild-type and vaccine strains of yellow fever virus. J. Gen. Virol. **70:**557–564.
5. **Carter, H. R.** 1931. Yellow fever: an epidemiological and historical study of its place of origin. The Williams & Wilkins Co., Baltimore.
6. **Clarke, D. H.** 1960. Antigenic analysis of certain group B arthropod-borne viruses by antibody absorption. J. Exp. Med. **111:**21–32.
7. **Coia, G., M. D. Parker, G. Speight, M. E. Byrne, and E. G. Westaway.** 1988. Nucleotide and complete amino acid sequences of Kunjin virus: definitive gene order and characteristics of the virus-specific proteins. J. Gen. Virol. **69:**1–21.
8. **De Cock, K. M., T. P. Monath, A. Nasidi, P. M. Tukei, J. Enriquez, P. Lichfield, R. B. Craven, A. Fabiyi, B. C. Okafor, and C. Ravaonjanahary.** 1988. Epidemic yellow fever in eastern Nigeria, 1986. Lancet **i:**630–633.
9. **Deubel, V., J.-P. Digoutte, T. P. Monath, and M. Girard.** 1986. Genetic heterogenecity of yellow fever virus strains from Africa and the Americas. J. Gen. Virol. **67:**209–213.
10. **Deubel, V., R. M. Kinney, and D. W. Trent.** 1988. Nucleotide sequence and deduced amino acid sequence of the nonstructural proteins of dengue type 2 virus, Jamaica genotype: comparative analysis of the full-length genome. Virology **165:**234–244.
11. **Deubel, V., J. J. Schlesinger, J.-P. Digoutte, and M. Girard.** 1987. Comparative immunochemical and biological analysis of African and American yellow fever viruses. Arch. Virol. **94:**331–338.
12. **Domingo, E., J. Diez, M. A. Martinez, J. Hernández, A. Holguin, B. Borrego, and M. G. Mateu.** 1993. New observations on antigenic diversification of RNA viruses. Antigenic variation is not dependent on immune selection. J. Gen. Virol. **74:**2039–2045.
13. **Eigen, M.** 1993. Viral quasispecies. Sci. Am. **(7):**42–49.
14. **Felsenstein, J.** 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39:**783–791.
15. **Felsenstein, J. (Department of Genetics, University of Washington, Seattle).** 1993. PHYLIP (Phylogeny Inference Package) version 3.5c.
16. **Fitzgeorge, R., and C. J. Bradish.** 1980. The *in vivo* differentiation of strains of yellow fever virus in mice. J. Gen. Virol. **46:**1–13.

17. **Hahn, C. S., J. M. Dalrymple, J. H. Strauss, and C. M. Rice.** 1987. Comparison of the virulent Asibi strain of yellow fever virus with the 17D vaccine strain derived from it. Proc. Natl. Acad. Sci. USA **84:**2019–2023.
18. **Heinz, F. X., and J. T. Roehrig.** 1990. Immunochemistry of viruses. II. The basis for serodiagnosis and vaccines, p. 289–305. *In* M. H. V. van Regenmortel and A. R. Neurath (ed.), Flaviviruses. Elsevier Science Publishers, Oxford, United Kingdom.
19. **Higgins, D. G., and P. M. Sharp.** 1989. Fast and sensitive multiple sequence alignments on a microcomputer. Comput. Appl. Biosci. **5:**151–153.
20. **Hillis, D. M., J. P. Huelsenbeck, and C. W. Cunningham.** 1994. Application and accuracy of molecular phylogenies. Science **264:**671–677.
21. **Jenning, A. D., J. E. Whitby, P. D. Minor, and D. T. Barrett.** 1993. Comparison of the nucleotide and deduced amino acid sequences of the envelope protein genes of the wild-type French viscerotropic strain of yellow fever virus and the live vaccine strain, French neurotropic vaccine, derived from it. Virology **192:**692–695.
22. **Kinney, R. M. (Centers for Disease Control and Prevention).** 1994. Personal communication.
23. **Lepiniec, L., L. Dalgarno, V. T. Q. Huong, T. P. Monath, J.-P. Digoutte, and V. Deubel.** 1994. Geographic distribution and evolution of yellow fever viruses based on direct sequencing of genomic cDNA fragments. J. Gen. Virol. **75:**417–423.
24. **Lewis, J. G., G.-J. Chang, R. S. Lanciotti, and D. W. Trent.** 1992. Direct sequencing of large flavivirus PCR products for analysis of genome variation and molecular epidemiological investigations. J. Virol. Methods **38:**11–24.
25. **Mandl, C. W., F. Guirakhoo, H. Holzmann, F. Heinz, and C. Kinz.** 1989. Antigenic structure of the flavivirus envelope protein E at the molecular level, using tick-borne encephalitis virus as a model. Virology **63:**564–571.
26. **Marfin, A. A., P. M. Tukei, N. N. Agata, E. G. Sanders, J. W. Den Boer, I. P. Reiter, C. B. Cropp, P. S. Moore, and D. J. Gubler.** 1993. Epidemiologic aspects of a yellow fever outbreak in northwest Kenya, 1992–1993. Am. J. Trop. Med. Hyg. **49:**185.
27. **Mason, P. W., P. C. McAda, T. Mason, and M. J. Fournier.** 1987. Sequence of the dengue-1 virus genome in the region encoding the three structural proteins and the major nonstructural protein NS1. Virology **161:**262–267.
28. **Monath, T. P.** 1986. Yellow fever, p. 139–231. *In* T. P. Monath (ed.), The arboviruses: epidemiology and ecology, vol. V. CRC Press, Boca Raton, Fla.
29. **Monath, T. P.** 1990. Flaviviruses, p. 763–814. *In* B. N. Fields and D. M. Knipe (ed.), Virology, vol. 1. Raven Press, New York.
30. **Nichol, S. T., J. E. Rowe, and E. M. Fitch.** 1993. Punctuated equilibrium and positive Darwinian evolution in vesicular stomatitis virus. Proc. Natl. Acad. Sci. USA **99:**10424–10428.
31. **Osatomi, K., I. Fuke, D. Tsuru, T. Shiba, Y. Sakaki, and H. Sumiyoshi.** 1988. Nucleotide sequence of dengue type 3 virus genomic RNA encoding viral structural proteins. Virus Genes **2:**99–108.
32. **Pletnev, A. G., V. F. Yamshchikov, and V. M. Blinov.** 1990. Nucleotide sequence of the genome and complete amino acid sequence of the polyprotein of tick-borne encephalitis virus. Virology **174:**250–263.
33. **Rice, C. M., E. M. Lenches, S. R. Eddy, S. R. Shin, R. L. Sheets, and J. H. Strauss.** 1985. Nucleotide sequence of yellow fever virus: implications for flavivirus gene expression and evolution. Science **229:**726–733.
34. **Sumiyoshi, H., C. Mori, I. Fuke, K. Morita, S. Kuhara, J. Kondou, Y. Kukushi, H. Nagamatu, and A. Igarashi.** 1987. Complete nucleotide sequence of the Japanese encephalitis virus genome RNA. Virology **161:**497–510.
35. **Tayor, R. M.** 1951. Epidemiology, p. 529–533. *In* G. K. Strode (ed.), Yellow fever. McGraw-Hill, New York.
36. **Tomori, O., A. Nasidi, and R. Spiegel.** 1993. Yellow fever in Nigeria, 1986–1993: considerations on epidemic preparedness and control. Am. J. Trop. Med. Hyg. **49:**185.
37. **Trent, D. W., R. M. Kinney, B. J. B. Johnson, A. V. Vorndam, J. A. Grant, V. Deubel, C. M. Rice, and C. Hahn.** 1987. Partial nucleotide sequence of St. Louis encephalitis virus RNA: structural protein, NS1, ns2a, and ns2b. Virology **156:**293–304.
38. **Zhao, B., E. Mackow, A. Buckler-White, L. Markoff, R. M. Chanock, C.-J. Lai, and Y. Makino.** 1986. Cloning full-length dengue type 4 viral DNA sequences: analysis of genes coding for structural proteins. Virology **155:**77–88.