Journal of General Virology (2007), **88**, 1967–1976 DOI 10.1099/vir.0.82714-0

Genetic divergence of Chikungunya viruses in India (1963–2006) with special reference to the 2005–2006 explosive epidemic

Vidya A. Arankalle, Shubham Shrivastava, Sarah Cherian, Rashmi S. Gunjikar, Atul M. Walimbe, Santosh M. Jadhav, A. B. Sudeep and Akhilesh C. Mishra

National Institute of Virology, 130/1 Sus Road, Pashan, Pune 411021, India

Correspondence
Vidya A. Arankalle
v_arankalle@yahoo.co.in

Received 14 November 2006
Accepted 19 March 2007

Re-emergence of Chikungunya (CHIK), caused by CHIK virus, was recorded in India during 2005–2006 after a gap of 32 years, causing 1.3 million cases in 13 states. Several islands of the Indian Ocean reported similar outbreaks in the same period. These outbreaks were attributed to the African genotype of CHIK virus. To examine relatedness of the Indian isolates (IND-06) with Reunion Island isolates (RU), full-genome sequences of five CHIK virus isolates representative of different Indian states were determined. In addition, an isolate obtained from mosquitoes in the year 2000 (Yawat-2000), identified as being of the African genotype, and two older strains isolated in 1963 and 1973 (of the Asian genotype), were sequenced. The IND-06 isolates shared 99.9 % nucleotide identity with RU isolates, confirming involvement of the same strain in these outbreaks. The IND-06 isolates shared 98.2 % identity with the Yawat-2000 isolate. Of two crucial substitutions reported for RU isolates in the E1 region, M269V was noted in the Yawat-2000 and IND-06 isolates, whereas D284E was seen only in the IND-06 isolates. The A226V shift observed with the progression of the epidemic in Reunion Island, probably associated with adaptation to the mosquito vector, was absent in all of the Indian isolates. Three unique substitutions were noted in the IND-06 isolates: two (T128K and T376M) in the Nsp1 region and one (P23S) in the capsid protein. The two Asian strains showed 99.4 % nucleotide identity to each other, indicating relative stability of the virus. No evidence of recombination of the Asian and African genotypes, or of positive selection was observed. The results may help in understanding the association, if any, of the unique mutations with the explosive nature of the CHIK outbreak.

INTRODUCTION

Chikungunya (CHIK) virus has recently re-emerged as an important pathogen causing epidemics of the disease in several countries. Epidemic resurgence of CHIK was recorded in 2000 in the Democratic Republic of Congo (DRC) (Pastorino *et al.*, 2004), in Indonesia during 2001–2003 (Laras *et al.*, 2005) and in India during 2005–2006 (Yergolkar *et al.*, 2006), after gaps of 39, 20 and 32 years, respectively. CHIK is emerging as an important infection in South-East Asia and the Pacific region (Thaikruea *et al.*, 1997; Mackenzie *et al.*, 2001; Kit, 2002). Recently, massive outbreaks of CHIK have been reported from many islands of the Indian Ocean (WHO, 2006).

CHIK virus infection is characterized by fever, headache, rash, nausea, vomiting, myalgia and arthralgia. The virus was first isolated during an outbreak in Tanzania in 1952 (Ross, 1956). CHIK outbreaks in east, south, west and central Africa have been documented. The virus appears to have spread from Africa to other parts of the world and caused epidemics in the Asian tropics (Powers *et al.*, 2000). In Africa, the virus is maintained in a sylvatic cycle involving wild primates and many species of *Aedes* mosquito. *Aedes albopictus* is considered to be the vector in Reunion and other islands of the Indian Ocean. Although both *Aedes aegypti* and *A. albopictus* mosquitoes are prevalent in India, the former is the main vector (Yergolkar *et al.*, 2006).

CHIK virus belongs to the genus *Alphavirus* of the family *Togaviridae*. Phylogenetic analyses based on E1 gene sequences grouped CHIK viruses isolated worldwide into three genotypes: Asian, east/central/south African (ECSA) and west African (Powers *et al.*, 2000; Schuffenecker *et al.*, 2006). The complete nucleotide sequence for the African prototype strain, S27, was determined and the presence of an internal polyadenylation [I-poly(A)] site and repeated sequence elements within the 3′ non-translated region

The GenBank/EMBL/DDBJ accession numbers for the full genome sequences of eight CHIK virus isolates determined in this study are EF027134–EF027141.

Supplementary tables are available with the online version of this paper.

(3′-NTR) was observed (Khan *et al.*, 2002). Full-genome sequences for the Ross and Senegal strains have been determined. However, similar data for Asian strains are not available. Recently, several isolates from Reunion and other islands have been sequenced (Schuffenecker *et al.*, 2006).

In India, the first CHIK outbreak was recorded in 1963 in Kolkata (Calcutta) (Shah *et al.*, 1964), followed by epidemics in eastern coastal areas, namely Chennai (Madras), Pondicherry and Vellore in 1964, Visakhapatnam, Rajmundry and Kakinada in 1965 (Rao, 1966), Nagpur in 1965 (Rodrigues *et al.*, 1972) and Barsi in 1973 (Padbidri & Gnaneswar, 1979). In Chennai alone, nearly 400 000 cases were recorded in 1964. In Nagpur, the incidence in certain wards was as high as 40–70 %. In view of the long absence of CHIK epidemics, it was postulated that CHIK virus had disappeared from India and South-East Asia (Burke *et al.*, 1985; Pavri, 1986). Serological surveys supported this view (Neogi *et al.*, 1995).

Several states in India experienced massive outbreaks of CHIK during 2005–2006. Initially, two southern states [Andhra Pradesh (AP) and Karnataka] and one western state (Maharashtra) were affected (Yergolkar *et al.*, 2006). The outbreak continued with reports of a large number of cases from several other states (Rajasthan, Gujarat, Tamilnadu, Orissa and Madhya Pradesh).

We determined the full-genome sequences of eight CHIK virus isolates: one from each of five states affected during the 2005–2006 episode, one isolate from mosquitoes in 2000 (Yawat, Maharashtra state) and two strains isolated during the epidemics in 1963 and 1973. An attempt has been made to determine the association, if any, of mutations in the genome with the increased transmissibility of the virus, leading to a large epidemic affecting 13 states of the country within a year.

METHODS

Viruses. The details of the eight CHIK viruses sequenced during the present study are presented in Table 1. These include two strains of the Asian genotype isolated in 1963 from Kolkata, eastern India (IND-63-WB1), and in 1973 from Barsi, western India (IND-73-MH5), along with one 2005–2006 epidemic isolate from each of five states in India, i.e. three from the states affected earlier, IND-06-KA15, IND-06-AP3 and IND-06-MH2, and two from the states affected 3 months later, IND-06-TN1 and IND-06-RJ1. One isolate recovered from a mosquito in 2000 from western India (ECSA genotype, IND-00-MH4) was also sequenced. The older isolates were obtained from the virus repository of the National Institute of Virology, Pashan, India, reconstituted and used. CHIK viruses were isolated according to protocols described previously (Yergolkar *et al.*, 2006). Supplementary Table S1, available in JGV Online, gives details of the additional CHIK virus strains included in the study.

RNA extraction, RT-PCR and sequencing. Methods for extraction, amplification and sequencing of viral RNA have been described previously (Yergolkar *et al.*, 2006). Briefly, viral RNA was isolated by using a QIAamp Viral RNA Mini kit (Qiagen) according to the manufacturer's instructions, followed by the SuperScript II protocol (Invitrogen). Amplified fragments were visualized by ethidium bromide agarose gel staining, extracted from the gels and both strands were sequenced by using a BigDye Terminator Cycle Sequencing kit (Applied Biosystems).

The nucleotide sequence of the S27 strain (GenBank accession no. AF369024) (Khan *et al.*, 2002) was used for primer designing. Supplementary Table S2, available in JGV Online, provides a list of primers used for PCR/sequencing.

Sequence and phylogenetic analyses. Phylogenetic analysis based on the available full-genome and E1 gene (1044 nt) sequences of CHIK viruses was performed by using MEGA version 3.1 (Kumar *et al.*, 2004). CLUSTAL_X version 1.83 (Thompson *et al.*, 1997) was used to perform multiple nucleotide and amino acid sequence alignments. For the construction of phylogenetic trees, the neighbour-joining algorithm and the Kimura two-parameter distance model were utilized. The reliability of the analysis was evaluated by a bootstrap test with 1000 replications.

Recombination analysis. To search for recombination events between strains, within or between genotypes, SimPlot (Salminen *et al.*, 1995) was used. The structural and non-structural genes of the CHIK virus strains for which full-genome sequences were available, and the E1 gene sequences of several other strains, were used as datasets. For graphical detection of conflicting phylogenetic signals, SimPlot was used, wherein isolates were examined by using sliding-window diversity and bootscan plots. The pairwise percentage difference between the query sequences and other sequences in the

Table 1. Viruses sequenced during the present study

Strain*	Year	Place of origin	Passage history	Length of 5′- (3′-) NTR (nt)	Total sequence length (nt)	GenBank accession no.
IND-06-KA15	2006	Karnataka	Two in C6/36	45 (447)	11 729	EF027135
IND-06-AP3	2006	Andhra Pradesh	One in C6/36	57 (485)	11 779	EF027134
IND-06-MH2	2006	Maharashtra	One in C6/36	62 (501)	11 800	EF027136
IND-06-TN1	2006	Tamilnadu	One in mosquitoes, two in mice	63 (450)	11 750	EF027138
IND-06-RJ1	2006	Rajasthan	Two in C6/36	62 (468)	11 767	EF027137
IND-00-MH4 (Yawat-2000)	2000	Yawat, Maharashtra	Three in mosquitoes, three in mice	77 (500)	11 814	EF027139
IND-63-WB1	1963	Kolkata, West Bengal	Six in mice, one in C6/36	66 (481)	11 784	EF027140
IND-73-MH5	1973	Barsi, Maharashtra	One in mice, one in C6/36	62 (506)	11 805	EF027141

*All strains were obtained from humans except for IND-00-MH4 (Yawat-2000), which was obtained from a mosquito.

alignment was determined by sliding a window of 400 and 200 bp along the alignment in 3 and 10 bp increments. The diversity profiles were used to determine which of the other sequences were related most closely to the putative recombinant and could therefore be used as parental sequences. Putative recombination break points detected were assessed for significance by reconstructing a separate maximum-likelihood (ML) tree for each region.

Molecular sequence evolution and diversifying selection analyses. The modified method of Nei and Gojobori (Nei & Gojobori, 1989; Suzuki & Gojobori, 1999) as implemented in the MEGA package was used to calculate the synonymous (dS) and non-synonymous (dN) substitution rates and the dN/dS ratio (ω) across all amino acid sites in pairwise comparisons between nucleotide sequences. This ratio, if >1, is used as evidence for positive, diversifying selection or adaptive evolution. On the other hand, if $\omega<1$, it is inferred as negative, deleterious or purifying selection, whilst $\omega=1$ in the case of neutral substitutions. To test whether the sequences are under selection pressure and also whether there are specific amino acids affected by diversifying selection, ML models of codon substitution that allow for heterogeneous selection pressures among sites were implemented (Yang *et al.*, 2000), using the CODEML program in the PAML package (Yang, 1997). Among the various codon-substitution models, only M1, M2, M7 and M8 were employed in the present study, as comparisons of M1 (neutral) with M2 (selection) and M7 (β) with M8 (β and ω) are specific tests for positive selection. The likelihood-ratio test (LRT) was applied to compare the null models (M1 and M7) with alternative ones (M2 and M8, respectively) that account for sites under positive selection. M1 divides codons into two categories, representing the proportion (p_0) of conserved sites with $\omega=0$ and the proportion of neutral sites (p_1) with $\omega=1$. M2 accounts for positive selection by including a third category of codons (p_2) with $\omega 2$ that can take any value, including >1, as estimated from the data. M7 and M8 are more complex models; M7 uses a discrete β distribution ($0<\omega<1$), whilst M8 also uses a β distribution, where $\omega>1$ is incorporated.

The F3×4 model, which computes equilibrium codon frequencies from the nucleotide frequencies at the three codon positions, was used to account for codon-usage bias. Branch lengths of the phylogeny (measured as the expected number of nucleotide substitutions per codon along a branch) and the transition-to-transversion ratio (κ) were estimated by using ML. Further, the Bayes theorem (Yang *et al.*, 2005) as implemented in CODEML was used to calculate the posterior probability that a particular amino acid site belongs to a given selection class (neutral, deleterious or advantageous). Sites with a high posterior probability of being from the class with $\omega>1$ were deemed more likely to be under diversifying selection.

RESULTS

Full-genome sequences of all of the CHIK isolates were amplified in eight overlapping fragments. Table 1 provides the lengths of genomic RNA and 3′/5′-NTR fragments of all eight Indian CHIK virus isolates with reference to the S27 strain (genomic RNA, 11 805 nt; 5′-NTR, 76 nt; 3′-NTR, 526 nt). The sequences from the earlier cases of the outbreak were from the states of Karnataka (11 729 nt), AP (11 779 nt) and Maharashtra (11 800 nt), whereas sequences corresponding to the cases occurring 3 months later were from Tamilnadu (11 750 nt) and Rajasthan (11 767 nt). In addition, the earliest isolated strain from the 1963 Kolkata epidemic (11 784 nt), the last strain isolated during the 1973 Barsi epidemic (11 805 nt) and a strain incidentally isolated from mosquitoes in 2000 (Yawat, 11 814 nt) were also sequenced.

Phylogenetic analyses

Fig. 1(a) depicts the phylogenetic tree based on full-genome analysis. All Indian isolates from the 2005–2006 (IND-06) resurgence representing the five affected states, the isolate from Yawat in 2000 (Yawat-2000), all Reunion isolates of 2005–2006 (RU) and the S27 and Ross isolates (1952) clustered together into the ECSA genotype. The earlier Indian isolates (1963 and 1973) belonged to the Asian genotype, whereas the Senegal strain formed a distinct branch (west African genotype). Similar results were obtained when the structural and non-structural regions were analysed separately (data not shown).

As E1 gene sequences were available for several additional isolates, a separate phylogenetic tree was constructed for these (Fig. 1b). This analysis revealed that all of the RU and IND-06 isolates grouped together, whereas the Yawat-2000 and Uganda-1982 isolates, sharing 99 % nucleotide identity, constituted a separate branch. Similarly, the isolates from the DRC (2000) formed a separate cluster within the same genotype, whereas the 1996 isolate from the Central African Republic remained as a separate branch. Isolates from Thailand recovered in 1975 and 1996 and the 1965 Indian and 1985 Indonesian isolates clustered in the Asian genotype. Both Senegal isolates (1966, 1983) and a Nigerian isolate remained in the west African genotype.

Sequence comparisons

Irrespective of the place of isolation (Reunion Island or India), the 2005–2006 isolates were related very closely (99.9 % identity). These isolates differed from the S27 and Yawat-2000 isolates by 2.7 and 1.7–1.8 %, respectively. The Asian genotype differed from the ECSA and west African genotypes by 4.4–5.3 and 15.4–15.5 %, respectively. The African genotypes (ECSA versus west African) were 14.5–14.8 % divergent. Amino acid identities across the three genotypes varied from 95.2 to 99.8 % (see Supplementary Table S3, available in JGV Online).

Sequence analysis of the ECSA genotype

Non-structural region. In this region, the recent Indian and RU strains exhibited 99.85±0.06 % identitiy at the amino acid level. Compared with the S27 prototype, nine identical substitutions were present in both groups of isolates: Q488R, S589N, A1328V, Y1550H, T1670I, L1794P, P1804S, T1938A and T2117A. There were seven substitutions, i.e. L507R, H909Y, V1508I, V1664A, I1709T, S1795N and Q2363L, that were shared between the Yawat-2000, IND-06 and RU isolates. The Yawat-2000 strain also showed six unique substitutions: V326M, Q1661P, S1691P, C1768R, V1771A and M1782T. The S1691P substitution in Yawat-2000 was also observed in one of the RU isolates, RU05-209 (Table 2).

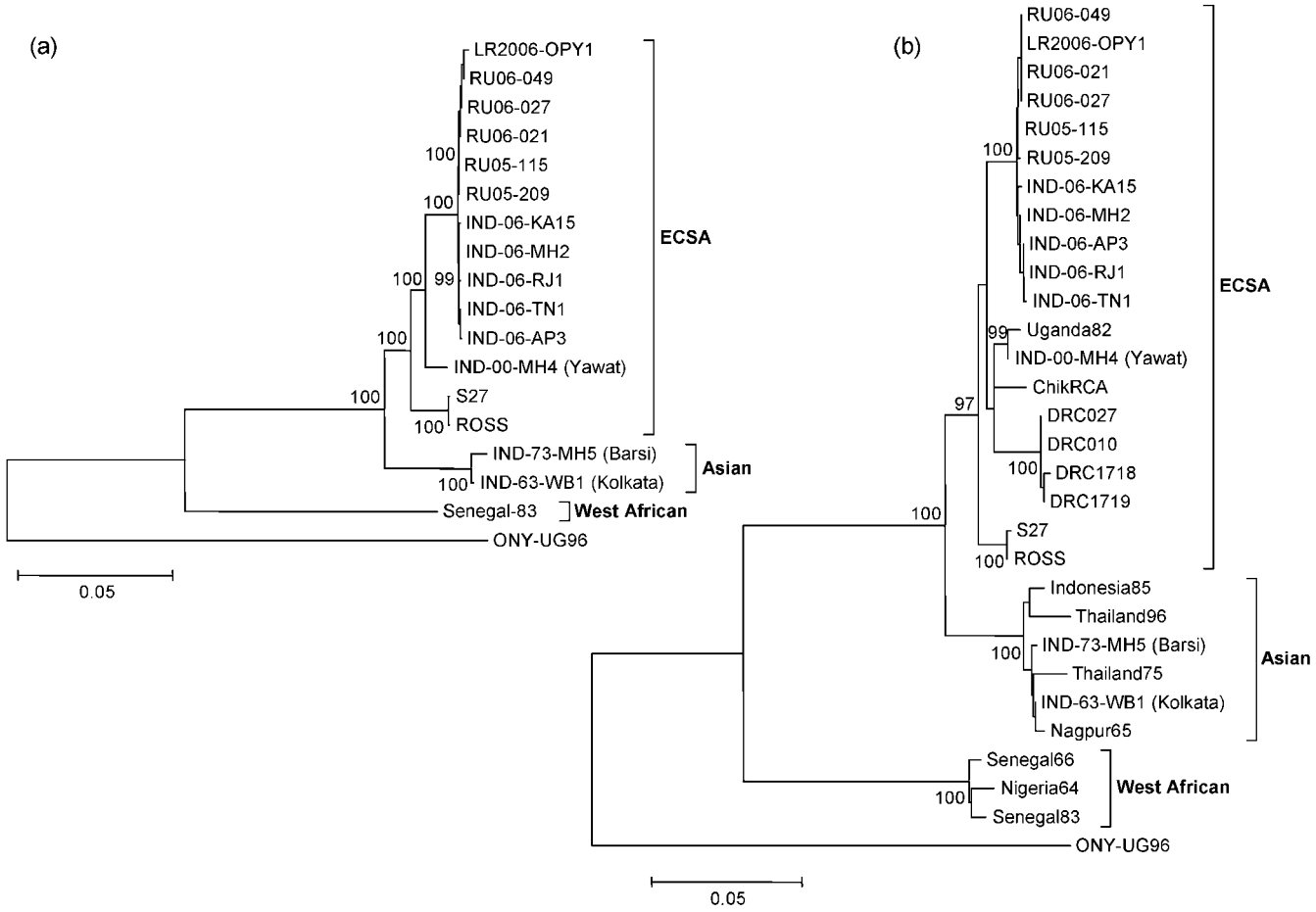

Fig. 1. Phylogenetic trees depicting the genotypic status of Indian CHIK virus isolates, based on (a) full-length genome sequences of 17 isolates and (b) E1 gene sequences (1044 nt) of 29 isolates including 17 full-genome sequences. See Table 1 for details of the isolates used for full-genome analysis. Numbers at nodes indicate bootstrap support (%). The additional isolates examined for E1 gene-based analysis included Uganda82 (GenBank accession no. AF192907), ChikRCA (AY549583), DRC027 (AY549577), DRC010 (AY549576), DRC1718 (AY549578), DRC1719 (AY549579), Indonesia85 (AF192894), Thailand96 (AF192900), Thailand75 (AF192898), Nagpur65 (AY424803), Senegal66 (AF192891) and Nigeria64 (AF192893).

Two unique substitutions were noted in all of the IND-06 isolates, both in the Nsp1 region (T128K and T376M). Except for these substitutions, the isolates from the first three states reporting CHIK cases were identical to the RU isolates. Of the two states reporting CHIK cases later, Tamilnadu exhibited one substitution (T1674M in Nsp3), whilst isolates from the state of Rajasthan exhibited two mutations (A101V in Nsp1 and T1210M in Nsp2) (Table 2).

Structural region. The IND-06 and RU isolates shared six substitutions in the structural region: E2-I536T, T637M, S700T and V711A, 6K-V756I and E1-D1093E. Further, the two groups had four substitutions (K63R, I284T, A489T and M1078V) identical to Yawat-2000. The unique mutations in the Yawat-2000 strain were A487V, V643M, I702V (in the E2 region), V828I and A1186T (in the E1 region). The mutation I702V (E2), although different from all of the IND-06 and RU isolates, was similar to the IND-63-WB1 and IND-73-MH5 strains, belonging to the Asian genotype (Table 3).

Two unique mutations in the capsid region, P23S and V27I, were noted among five and four IND-06 isolates, respectively. The V27I substitution was not recorded in the AP isolate. Further, Karnataka displayed one (K1020N, E1) and Tamilnadu two (N80D, capsid, and V1022I, E1) mutation(s) (Table 2).

Sequence analysis of the Asian genotype

CHIK viruses of the Asian genotype isolated in 1963 and 1973 showed 99.4 (99.72) and 99.39 (99.44) % nucleotide (amino acid) identities in the non-structural and structural regions, respectively. These strains exhibited several amino acid substitutions compared with the S27 strain.

Table 2. Amino acid substitutions in isolates of the ECSA genotype with respect to the S27 strain

A dot indicates a match with the amino acid of the S27 strain.

Region	Polypeptide position	Protein position	S27	RU05-209	LR2006-OPY1	IND-06-AP3	IND-06-KA15	IND-06-MH2	IND-06-RJ1	IND-06-TN1	Yawat-2000
Nsp1	101	101	A	.	.	.	.	.	V	.	.
	128	128	T	.	.	K	K	K	K	K	.
	326	326	V	.	.	.	.	.	.	.	M
	376	376	T	.	.	M	M	M	M	M	.
	488	488	Q	R	R	R	R	R	R	R	.
	507	507	L	R	R	R	R	R	R	R	R
Nsp2	589	54	S	N	N	N	N	N	N	N	.
	909	374	H	Y	Y	Y	Y	Y	Y	Y	Y
	1210	675	T	.	.	.	.	.	M	.	.
	1328	793	A	V	V	V	V	V	V	V	.
Nsp3	1508	175	V	I	I	I	I	I	I	I	I
	1550	217	Y	H	H	H	H	H	H	H	.
	1661	328	Q	.	.	.	.	.	.	.	P
	1664	331	V	A	A	A	A	A	A	A	A
	1670	337	T	I	I	I	I	I	I	I	.
	1674	341	T	.	.	.	.	.	.	M	.
	1691	358	S	P	.	.	.	.	.	.	P
	1709	376	I	T	T	T	T	T	T	T	T
	1768	435	C	.	.	.	.	.	.	.	R
	1771	438	V	.	.	.	.	.	.	.	A
	1782	449	M	.	.	.	.	.	.	.	T
	1794	461	L	P	P	P	P	P	P	P	.
	1795	462	S	N	N	N	N	N	N	N	N
	1804	471	P	S	S	S	S	S	S	S	.
Nsp4	1938	75	T	A	A	A	A	A	A	A	.
	2117	254	T	A	A	A	A	A	A	A	.
	2363	500	Q	L	L	L	L	L	L	L	L
Capsid	23	23	P	.	.	S	S	S	S	S	.
	27	27	V	.	.	.	I	I	I	I	.
	63	63	K	R	R	R	R	R	R	R	R
	80	80	N	.	.	.	.	.	.	D	.
E3	284	23	I	T	T	T	T	T	T	T	T
E2	487	162	A	.	.	.	.	.	.	.	V
	489	164	A	T	T	T	T	T	T	T	T
	536	211	I	T	T	T	T	T	T	T	.
	637	312	T	M	M	M	M	M	M	M	.
	643	318	V	.	.	.	.	.	.	.	M
	700	375	S	T	T	T	T	T	T	T	.
	702	377	I	.	.	.	.	.	.	.	V
	711	386	V	A	A	A	A	A	A	A	.
6K	756	8	V	I	I	I	I	I	I	I	.
E1	828	19	V	.	.	.	.	.	.	.	I
	1020	211	K	.	.	.	N	.	.	.	.
	1022	213	V	.	.	.	.	.	.	I	.
	1035	226	A	.	V	.	.	.	.	.	.
	1078	269	M	V	V	V	V	V	V	V	V
	1093	284	D	E	E	E	E	E	E	E	.
	1186	377	A	.	.	.	.	.	.	.	T

Table 3. Amino acid substitutions in isolates of the Asian genotype with respect to the S27 strain

A dot indicates a match with the amino acid of the S27 strain.

Region	Polypeptide position	Protein position	S27	IND-63-WB1	IND-73-MH5
Nsp1	3	3	P	S	S
	34	34	P	S	S
	253	253	K	T	T
	451	451	V	M	.
	454	454	S	G	G
	473	473	S	R	R
	486	486	D	N	N
	491	491	R	Q	Q
	507	507	L	H	H
Nsp2	551	16	P	L	L
	714	179	I	T	T
	753	218	T	S	S
	765	230	S	.	P
	873	338	K	M	M
	1001	466	M	V	V
	1028	493	K	.	T
	1139	604	A	V	V
Nsp3	1402	69	V	A	.
	1509	176	V	I	I
	1557	224	T	I	I
	1571	238	S	.	N
	1616	283	S	N	N
	1667	334	A	V	V
	1670	337	T	A	A
	1682	349	V	A	A
	1686	353	I	T	T
	1700	367	L	P	P
	1709	376	I	V	V
	1714	381	S	T	T
	1716	383	T	I	I
	1770	437	V	A	A
	1782	449	M	I	I
	1791	458	A	T	T
	1792	459	T	M	M
	1816	483	N	D	D
	1850	517	S	P	P
Nsp4	1906	43	A	L	L
	1948	85	R	K	K
	1953	90	S	A	A
	1964	101	T	I	I
	2098	235	Q	R	R
	2143	280	E	D	D
	2229	366	T	A	A
	2427	564	D	.	E
	2445	582	V	A	A
Capsid	32	32	P	.	L
	37	37	Q	K	K
	78	78	Q	R	R
	81	81	T	M	M
	89	89	K	T	.
E3	284	23	I	A	A
	294	33	E	K	K
	305	44	R	S	S
	321	60	H	R	R
E2	327	2	T	I	I
	409	84	F	L	L
	443	118	S	G	G
	474	149	K	R	R
	482	157	V	A	A
	530	205	G	D	D
	571	246	A	V	V
	609	284	I	.	T
	643	318	V	R	R
	702	377	I	V	V
	709	384	M	V	V
6K	795	47	A	T	I
E1	864	55	I	.	T
	881	72	N	S	S
	907	98	A	T	T
	951	142	I	V	V
	954	145	T	S	S
	1020	211	K	E	E
	1034	225	A	S	S
	1124	315	V	A	A

Non-structural region. There were eight, six, 17 and eight unique substitutions in the Nsp1, Nsp2, Nsp3 and Nsp4 regions, respectively, in the two strains of the Asian genotype (Table 3) compared with the S27 strain. In addition, unique amino acid substitutions were recorded for the IND-63-WB1 strain (Nsp1, V451M, and Nsp3, V1402A). The IND-73-MH5 strain displayed several unique substitutions, such as S765P, K1028T (Nsp2), S1571N (Nsp3) and D2427E (Nsp4).

Thirteen mutations were shared between the IND-06, RU, Yawat-2000 and Asian isolates. These included L172V, E234K, M383L, I384L (only in IND-73-MH5), T481I, C1177Y, S1178N, P1659S, K1685E, A1715T, I2377T, V2418I and V2467I.

Structural region. As shown in Table 3, the capsid, E3, E2 and E1 regions exhibited three, four, ten and seven amino acid replacements, respectively, with reference to the S27 strain. In addition, the IND-63-WB1 strain exhibited a single substitution in the capsid (K89T) and 6K (A795T) regions, whereas the IND-73-MH5 strain showed four substitutions: capsid-P32L, E2-I609T, 6K-A795I and E1-I864T. Eleven mutations were shared between the IND-06, RU, Yawat-2000 and Asian isolates: G382K, I399M, G404E, N485T, L506M, S519G, M592R, S624N, A669T, I802V and V1131A.

5′- and 3′-NTRs. The 5′-NTR was highly conserved, whereas the 3′-NTR showed maximum divergence (10.1–17.4 % between different genotypes). Within the 3′-NTR, the Asian genotype was characterized by an insertion of

10 nt between positions 11377 and 11378, another of 11 nt between positions 11514 and 11515 and one insertion at position 11425 with respect to S27. Similarly, several deletions compared with S27 were reported in the Asian genotype. These included two deletions at positions 11465–11466, one deletion at position 11595 and another deletion at position 11629. The IND-73-MH5 strain exhibited unique deletions at positions 11436, 11743 and 11744 compared with the S27 strain. A stretch of 19 'A' nucleotides, a possible I-poly(A) site in S27 (Khan *et al.*, 2002), showed six substitutions in the Asian genotype (Fig. 2). Deletion of a stretch of 14 of the 19 'A' nucleotides reported for the RU isolates (Schuffenecker *et al.*, 2006) was maintained in all Indian isolates belonging to the ECSA genotype. In addition, the AP strain showed one insertion between positions 11579 and 11580 and two deletions at positions 11629 and 11800.

Analysis of genetic recombination

Considering that the present strains belong to the African genotype, whilst strains prevalent during earlier outbreaks during 1963–1973 are of the Asian genotype, we looked specifically for evidence of a recombination event(s) between these two genotypes and between strains within genotypes. SimPlot diversity plots showed no clear evidence for relative shifts in pairwise diversity with other strains, nor could evidence for recombination be observed from the ML break-point analysis (data not shown).

Molecular evolution

All IND-06 isolates exhibited three unique substitutions (T128K and T376M in Nsp1 and P23S in capsid). Another substitution, V27I (capsid), was displayed by all IND-06 isolates except the isolate from the AP state, which reported cases earlier. Additional substitutions were seen in the Karnataka (E1, K211N) and Tamilnadu (capsid, N80D) isolates. The isolate from the state of Rajasthan, reporting cases later, exhibited two additional substitutions (Nsp1, A101V; Nsp2, T1210M).

None of the ω values, calculated by the modified method of Nei & Gojobori (1989) for all of the non-structural as well as the structural genes, exceeded 1 (data not shown). The results of applying the tests for positive selection on all of the genes are presented in Table 4.

Among the non-structural genes, Nsp1 identified a very small class (0.5 %) of positively selected sites, with strength of selection $\omega \cong 8.3$ under both M2 and M8 models. The Bayesian method assigned two sites (384I and 507L) to the positively selected class with 81 % probability under M8. In the Nsp2 gene, M8 identified 0.2 % of the sites to be under positive selection pressure with $\omega = 10.9$. Here, site 642C was identified with a posterior probability of 82 %. It was noted that, for all of the non-structural proteins, neither of the LRTs, between M2 (selection) and M1 (neutral) or between the more complex models M8 and M7, allowed the model of neutrality to be rejected in favour of positively selected sites (Table 4).

```
S27         CTAATAATC- ---------T GTAGATCAAA GGGCTATATA ACCCCTGAAT
RU05-209    .A.......- ---------. A......... ......CGC. ..........
LR2006-OPY1 .A.......- ---------. A......... ......CGC. ..........
IND-06-AP3  .A.......- ---------. A......... ......CGC. ..........
IND-00-MH4  .........- ---------. A......... ......C.C. ..........
IND-63-WB1  ....G....A ATAGATAAG. A......... .....GA.C. ..........
IND-73-MH5  ....G.C..A ATAGATAAG. A......... .....GA.C. ..........

S27         AGTAACAAAA TACAAAA-TC ACTAAAAATT ATAAAAAAAA AAAAAAAAAA
RU05-209    .......... .......-.. .......... .......--- ----------
LR2006-OPY1 .......... .......-.. .......... .......--- ----------
IND-06-AP3  .......... .......-.. .......... .......--- ----------
IND-00-MH4  .......... .......-.. .......... .......--- ----------
IND-63-WB1  .......... ..T....A.. .A......A. CAT....T.G ....CC.G..
IND-73-MH5  .......... ..T....A.. .A......-. CAT....T.G ....CC.G..

S27         ACAGAAAAAT ATATAAATAG GTATACGTGT CCCCTAAGAG ACACATTGTA
RU05-209    -......... .C........ .......... .......... ..........
LR2006-OPY1 -......... .C........ .......... .......... ..........
IND-06-AP3  -......... .C........ .......... .......... ..........
IND-00-MH4  -......... .C........ .......... .......... ..........
IND-63-WB1  ......G--. .GG...GA.. .....T.... .......... .....CCA..
IND-73-MH5  ......G--. .GG...GA.. .....T.... .......... .....CCA..

S27         TGTAGGT--- --------GA
RU05-209    .......--- --------..
LR2006-OPY1 .......--- --------..
IND-06-AP3  .....T.--- --------..
IND-00-MH4  .......--- --------..
IND-63-WB1  .A...C.AAG AATCAATA..
IND-73-MH5  .A...C.AAG AATCAATA..
```

Fig. 2. Alignment of 3′-NTR sequences of CHIK viruses, showing the region from nucleotide position 11369 to 11516 (S27 numbering). Dots indicate a match with the reference S27 strain; dashes indicate the absence of a nucleotide at that position.

Table 4. Likelihood values, parameter estimates and sites identified as being under positive selection by four models as applied to each gene

Gene	Model	ln λ	Parameter estimate	Positively selected sites*
Nsp1	M1 (neutral)	−2782.90	p_0=0.93, p_1=0.07	
	M2 (selection)	−2782.28	p_0=0.995, p_1=0, p_2=0.005, ω_2=8.27	384I, 507L
	M7 (β, neutral)	−2782.96	p=0.053, q=0.461	
	M8 (β, selection)	−2782.29	p_0=0.995, p_1=0.005, p=8.452, q=99, ω=8.26	**384I, 507L**
Nsp2	M1 (neutral)	−3969.05	p_0=0.975, p_1=0.025	
	M2 (selection)	−3968.59	p_0=0.942, p_1=0.01, p_2=0.048, ω_2=10.92	642C
	M7 (β, neutral)	−3969.12	p=0.038, q=0.643	
	M8 (β, selection)	−3968.60	p_0=0.998, p_1=0.002, p=4.42, q=99, ω=10.90	**642C**
Nsp3	M1 (neutral)	−2928.03	p_0=0.889, p_1=0.111	
	M2 (selection)	−2928.03	p_0=0.889, p_1=0.049, p_2=0.062, ω_2=0.999	337T, 358S, 376I, 449M
	M7 (β, neutral)	−2927.89	p=0.083, q=0.523	337T, **358S**, 376I, 449M
	M8 (β, selection)	−2927.89	p_0=0.974, p_1=0.026, p=0.084, q=0.64, ω=1.0	
Nsp4	M1 (neutral)	−3207.92	p_0=0.98, p_1=0.02	
	M2 (selection)	−3207.92	p_0=0.98, p_1=0.002, p_2=0.018, ω_2=0.999	43A
	M7 (β, neutral)	−3207.89	p=0.144, q=2.049	
	M8 (β, selection)	−3207.89	p_0=1, p_1=0, p=0.144, q=2.049, ω=1.77	**43A**
Capsid	M1 (neutral)	−1323.29	p_0=0.918, p_1=0.082	
	M2 (selection)	−1322.90	p_0=0.944, p_1=0, p_2=0.056, ω_2=1.88	27V, 63K
	M7 (β, neutral)	−1323.45	p=0.005, q=0.070	
	M8 (β, selection)	−1322.90	p_0=0.944, p_1=0.056, p=0.005, q=99, ω=1.88	**27V**, 63K
E3	M1 (neutral)	−335.19	p_0=0.837, p_1=0.163	
	M2 (selection)	−335.03	p_0=0.948, p_1=0.0, p_2=0.052, ω_2=3.2	23I
	M7 (β, neutral)	−335.27	p=0.01, q=0.53	
	M8 (β, selection)	−335.03	p_0=0.948, p_1=0.052, p=5.655, q=99, ω=3.2	**23I**
E2	M1 (neutral)	−2331.79	p_0=0.816, p_1=0.184	
	M2 (selection)	−2331.75	p_0=0.885, p_1=0, p_2=0.115, ω_2=1.326	57G, 194S, 211I, 318V, 377I
	M7 (β, neutral)	−2331.80	p=0.027, q=0.106	
	M8 (β, selection)	−2331.75	p_0=0.886, p_1=0.114, p=5.938, q=98.99, ω=1.329	57G, 194S, 211I, **318V**, 377I
6K	M1 (neutral)	−301.70	p_0=0.706, p_1=0.294	
	M2 (selection)	−299.55	p_0=0.924, p_1=0, p_2=0.076, ω_2=8.815	8V, **47A**
	M7 (β, neutral)	−301.70	p=0.005, q=0.012	
	M8 (β, selection)	−299.55	p_0=0.924, p_1=0.076, p=0.005, q=1.82, ω=8.815	
E1	M1 (neutral)	−2291.45	p_0=0.978, p_1=0.022	
	M2 (selection)	−2291.45	p_0=0.978, p_1=0.01, p_2=0.016, ω_2=0.999	211K
	M7 (β, neutral)	−2291.43	p=0.247, q=2.778	
	M8 (β, selection)	−2291.43	p_0=1, p_1=0, p=0.247, q=2.778, ω=1.00	**211K**

*Sites with a posterior probability of >50 % of having ω>1. Bold type indicates a posterior probability of >75 %. Sites are labelled according to amino acid position in the individual proteins.

Among the structural proteins, for the capsid gene, M8 identified 5.6 % of sites with ω>1. A single site, 27V, was identified by Bayesian analysis to be under positive selection with a probability of 79 %. In the E3 gene, 5.2 % of sites were identified to be under positive selection with $\omega \cong 3.2$ under both the M2 and M8 models. Site 23I was detected to be in the positively selected class with a probability of 82 % under M8. The strength of positive selection in the case of the E2 gene under model M2, as well as under M8, was low, with $\omega \cong 1.3$, and the proportion of positively selected sites was estimated to be about 11.1 %. Of the five sites, i.e. 57, 194, 211, 318 and 377, identified in the E2 gene by using Bayesian analysis, one, 318V, was found with a probability of 88 % under M8. On the other hand, both the M2 and M8 models estimated a higher strength of positive selection ($\omega \cong 8.8$) in the 6K gene, with the proportion of positively selected sites estimated as being 7.6 %. The 6K region was found to have two sites, 8V and 47A, under positive selection, with a higher probability of 86 % under M8. Again, for all of the structural genes, the models for neutrality (M1 and M7) fitted the data better than those for positive selection (M2 and M8).

Among the positively selected sites with a posterior probability of >75 %, Nsp1-507L, Nsp3-358S, E2-318V, 6K-8V, E1-211K and capsid-27V represent non-conserved sites (Tables 2 and 4).

DISCUSSION

After a long absence of 32 years, CHIK has re-emerged in an explosive epidemic form in India, with estimates of over 1.3 million cases (http://www.nvbdcp.gov.in/Chikun-cases.html). As we reported previously, there was also a shift of genotype from Asian to ECSA (Yergolkar *et al.*, 2006). In Asia, although epidemic re-emergence of CHIK was noted after 20 years during 2001–2003 in Indonesia (Laras *et al.*, 2005), no sequence-based phylogenetic analysis was performed and hence evidence, if any, of a genotypic shift from Asian to African is not available. These data would be crucial to understanding the mode of genotypic shift in the continent. As the ECSA genotype was implicated in the current epidemics in the RU islands, it was envisaged that the same genotype was introduced into India from the Indian Ocean islands.

The key issues are understanding of (i) the origin and spread of the current CHIK virus strain and (ii) the association of mutations in the viral genome with increased transmissibility/virulence of the virus. As far as the first issue is concerned, sequence similarity of 99.9 % between the RU and Indian strains at the full-genome level implies circulation of the same strains in both countries. It also indicates a possibility of spread of the current strain from Indian Ocean islands to India, leading to an explosive epidemic of the ECSA genotype and not the Asian genotype that circulated earlier. However, the presence of the ECSA genotype in India in the year 2000 without any epidemic of the disease is noteworthy. The origin of this strain is not clear. As is evident from Fig. 1(b), the Yawat-2000 strain had highest nucleotide identity (99.62 %) with a strain from Uganda isolated in 1982. This strain was related more closely (98.2 %) to the IND-06 strains than to the S27 strain (97.3 %). It is pertinent to note here that, in the year 2000, the DRC experienced epidemic resurgence after 39 years. However, the IND-06 strains are related more closely to Yawat-2000 than to the DRC strains. Further studies would be required to understand the genesis of IND-06 strains with special reference to the Yawat-2000 strain. Sequence comparison (Table 2) of Yawat-2000 with the IND-06 and RU isolates showed that, with reference to the S27 strain, seven and four mutations in the non-structural and structural polyprotein, respectively, were present in all of these isolates. The Yawat-2000 strain exhibited six and five unique substitutions in the non-structural and structural regions, respectively. Association of these mutations with the increased transmissibility of the virus needs to be determined.

Taking into consideration the co-circulation of the Asian and ECSA genotypes in India, the possibility of generation of the current strain as a result of recombination events between the two genotypes was examined. However, no evidence for this was noted. Thus, the current strain is not a recombination product of the ECSA genotype with the Asian (present study) or west African (Schuffenecker *et al.*, 2006) genotype. Further, no statistically significant evidence for positive selection was obtained and all of the genes in the CHIK virus genome were under purifying selection. It was further noted that the positively selected sites with a posterior probabability of >75 % represented some of the non-conserved sites.

An interesting observation of the evolution of CHIK viruses with the progression of the outbreak in Reunion includes a shift, A226V, during the latter period (beyond September 2005). However, among all of the Indian isolates, including Yawat-2000, 226A was maintained. The relationship of this change with the increase in the rate of transmission postulated for RU isolates does not hold true in the Indian scenario. In the present study, a few additional mutations at different locations were recorded in viruses isolated later during the epidemic and no definite pattern emerged. Of the seven amino acid substitutions (A164T, T312M, S375Tand V386A in E2, V8I in 6K, M269V and D284E in E1) in the structural polyprotein recorded as being unique for the RU isolates (Schuffenecker *et al.*, 2006), all were present in the IND-06 isolates. Significantly, Yawat-2000 showed two of these seven substitutions: A164T in the E2 ectodomain and M269V in the E1 protein, suggesting importance of the other five mutations to the increased transmissibility of the virus.

Similar to the RU isolates (Schuffenecker *et al.*, 2006), we confirm the presence of an opal stop codon at position 1857 in all of the IND-06 isolates. However, an arginine residue (as in S27) was maintained in Yawat-2000 and in both isolates belonging to the Asian genotype. As seen from Table 1, none of these isolates have been passaged extensively and, hence, this substitution does not seem to be related to the number of *in vitro* passages, as postulated for the S27 strain.

Similar to the RU isolates, we confirm the deletion of 14 of the 19 'A' nucleotides in the I-poly(A) site in all of the IND-06 isolates and the Yawat-2000 isolate. All of the above isolates were sequenced at very early passage levels (Table 1). As speculated by Khan *et al.* (2002), the possibility of introduction of the I-poly(A) site during passaging (>50 in the C6/36 cell line) needs to be confirmed by repeated passaging of virus(es) lacking the same. Sequences of two strains of the Asian genotype isolated 10 years apart showed distinct insertions/deletions in the 3′-NTR (Fig. 2). The I-poly(A) stretch of 19 nt reported for the S27 isolate showed several mutations, leading to three stretches of four 'A' nucleotides interspersed with seven substitutions. Insertions of 10 nt (after position 11377 with respect to S27) and 11 nt (after 11514) are other characteristic features of this genotype. Considering 99.4 % nucleotide identity at the full-genome level over a period of 10 years between the two Asian strains sequenced, it may be concluded that the virus is relatively stable.

Understanding the association of genotypic and epidemiological changes is particularly important. The earlier outbreaks (1963–1973, Asian genotype) affected urban areas, whereas the current resurgence (African genotype) affected

mainly rural areas for almost 8–10 months and later penetrated some of the large cities. Interestingly, although the African genotype was responsible for the resurgence of CHIK in the DRC in 2000, urban areas were affected for the first time, compared with earlier observations of the disease being essentially endemic to rural areas of tropical Africa and caused by virus belonging to the same genotype.

The state of Maharashtra represents a unique situation. This state reported epidemics of the disease in 1965 (Nagpur, Asian genotype), 1973 (Barsi, Asian genotype) and 2006 (almost all districts of the state, African genotype), as well as isolation of the African genotype from mosquitoes in 2000. It would be interesting to study the age-stratified prevalence of anti-CHIK virus antibodies in this state over a period of time as a measure of CHIK virus activity.

Overall, the full-genome analysis documented that an enormous number of cases in Reunion and India are caused by the same strain and are not the result of recombination of Asian and ECSA genotypes. The diversifying selection analysis suggests that the CHIK virus genome is under purifying selection, molecular divergence in the species being driven by random fixation of selectively neutral and few non-synonymous mutations. Future studies should endeavour to characterize transmissibility/virulence of various isolates of different genotypes at the molecular level.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge financial support provided by ICMR, Ministry of Health and Family Welfare, Government of India. S. S. acknowledges ICMR for providing a Junior Research Fellowship.

REFERENCES

Burke, D. S., Nisalak, A. & Nimmannitya, S. (1985). Disappearance of Chikungunya virus from Bangkok. *Trans R Soc Trop Med Hyg* **79**, 419–420.

Khan, A. H., Morita, K., del Carmen Parquet, M., Hasebe, F., Mathenge, E. G. & Igarashi, A. (2002). Complete nucleotide sequence of chikungunya virus and evidence for an internal polyadenylation site. *J Gen Virol* **83**, 3075–3084.

Kit, L. S. (2002). Emerging and re-emerging diseases in Malaysia. *Asia Pac J Public Health* **14**, 6–8.

Kumar, S., Tamura, K. & Nei, M. (2004). MEGA3: integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform* **5**, 150–163.

Laras, K., Sukri, N. C., Larasati, R. P., Bangs, M. J., Kosim, R., Djauzi, Wandra, T., Master, J., Kosasih, H. & other authors (2005). Tracking the re-emergence of epidemic chikungunya virus in Indonesia. *Trans R Soc Trop Med Hyg* **99**, 128–141.

Mackenzie, J. S., Chua, K. B., Daniels, P. W., Eaton, B. T., Field, H. E., Hall, R. A., Halpin, K., Johansen, C. A., Kirkland, P. D. & other authors (2001). Emerging viral diseases of Southeast Asia and the Western Pacific. *Emerg Infect Dis* **7** (*Suppl.*), 497–504.

Nei, M. & Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* **3**, 418–426.

Neogi, D. K., Bhattacharya, N., Mukherjee, K. K., Chakraborty, M. S., Banerjee, P., Mitra, K., Lahiri, M. & Chakravarti, S. K. (1995). Serosurvey of chikungunya antibody in Calcutta metropolis. *J Commun Dis* **27**, 19–22.

Padbidri, V. S. & Gnaneswar, T. T. (1979). Epidemiological investigations of chikungunya epidemic at Barsi, Maharashtra state, India. *J Hyg Epidemiol Microbiol Immunol* **23**, 445–451.

Pastorino, B., Muyembe-Tamfum, J. J., Bessaud, M., Tock, F., Tolou, H., Durand, J. P. & Peyrefitte, C. N. (2004). Epidemic resurgence of Chikungunya virus in Democratic Republic of the Congo: identification of a new central African strain. *J Med Virol* **74**, 277–282.

Pavri, K. M. (1986). Disappearance of Chikungunya virus from India and South East Asia. *Trans R Soc Trop Med Hyg* **80**, 491.

Powers, A. M., Brault, A. C., Tesh, R. B. & Weaver, S. C. (2000). Re-emergence of chikungunya and o'nyong-nyong viruses: evidence for distinct geographical lineages and distant evolutionary relationships. *J Gen Virol* **81**, 471–479.

Rao, T. R. (1966). Recent epidemics caused by Chikungunya virus in India, 1963–1965. *Sci Cult* **32**, 215–220.

Rodrigues, F. M., Patankar, M. R., Banerjee, K., Bhatt, P. N., Goverdhan, M. K., Pavri, K. M. & Vittal, M. (1972). Etiology of the 1965 epidemic of febrile illness in Nagpur city, Maharashtra State, India. *Bull World Health Organ* **46**, 173–179.

Ross, R. W. (1956). A laboratory technique for studying the insect transmission of animal viruses, employing a bat-wing membrane, demonstrated with two African viruses. *J Hyg (Lond)* **54**, 192–200.

Salminen, M. O., Carr, J. K., Burke, D. S. & McCutchan, F. E. (1995). Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning. *AIDS Res Hum Retroviruses* **11**, 1423–1425.

Schuffenecker, I., Iteman, I., Michault, A., Murri, S., Frangeul, L., Vaney, M. C., Lavenir, R., Pardigon, N., Reynes, J. M. & other authors (2006). Genome microevolution of chikungunya viruses causing the Indian Ocean outbreak. *PLoS Med* **3**, e263.

Shah, K. V., Gibbs, C. J., Jr & Banerjee, G. (1964). Virological investigation of the epidemic of haemorrhagic fever in Calcutta: isolation of three strains of chikungunya virus. *Indian J Med Res* **52**, 676–683.

Suzuki, Y. & Gojobori, T. (1999). A method for detecting positive selection at single amino acid sites. *Mol Biol Evol* **16**, 1315–1328.

Thaikruea, L., Charearnsook, O., Reanphumkarnkit, S., Dissomboon, P., Phonjan, R., Ratchbud, S., Kounsang, Y. & Buranapiyawong, D. (1997). Chikungunya in Thailand: a re-emerging disease? *Southeast Asian J Trop Med Public Health* **28**, 359–364.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25**, 4876–4882.

WHO (2006). Outbreak news. *Wkly Epidemiol Rec* **81**, 105–116.

Yang, Z. H. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13**, 555–556.

Yang, Z., Nielsen, R., Goldman, N. & Pedersen, A. M. K. (2000). Codon-substitution models for heterogenous selection pressure at amino acid sites. *Genetics* **155**, 431–449.

Yang, Z., Wong, W. S. & Nielsen, R. (2005). Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol Biol Evol* **22**, 1107–1118.

Yergolkar, P. N., Tandale, B. V., Arankalle, V. A., Sathe, P. S., Sudeep, A., Gandhe, S. S., Gokhle, M. D., Jacob, G. P., Hundekar, S. L. & Mishra, A. C. (2006). Chikungunya outbreaks caused by African genotype, India. *Emerg Infect Dis* **12**, 1580–1583.