Journal of General Virology (2004), **85**, 1971–1980 DOI 10.1099/vir.0.79850-0

Sequences of flavivirus-related RNA viruses persist in DNA form integrated in the genome of *Aedes* spp. mosquitoes

Sandrine Crochu,[1] Shelley Cook,[2] Houssam Attoui,[1] Remi N. Charrel,[1] Reine De Chesse,[1] Mourad Belhouchet,[1] Jean-Jacques Lemasson,[1] Philippe de Micco[1] and Xavier de Lamballerie[1]

[1]Unité des Virus Emergents, Faculté de Médecine de Marseille, IFR48-IRD UR034, 27 boulevard Jean Moulin, 13005 Marseille, France

[2]Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK

Correspondence
Xavier de Lamballerie
Xavier.de-Lamballerie@medecine.univ-mrs.fr

Flavivirus-related sequences have been discovered in the dsDNA genome of *Aedes albopictus* and *Aedes aegypti* mosquitoes, demonstrating for the first time an integration into a eukaryotic genome of a multigenic sequence from an RNA virus that replicates without a recognized DNA intermediate. In the *Aedes albopictus* C6/36 cell line, an open reading frame (ORF) of 1557 aa with protease/helicase and polyprotein processing domains characteristic of flaviviruses was identified. It is closely related to NS1–NS4A genes of the *Cell Fusing Agent* and *Kamiti River virus* and the corresponding mRNAs were detected. Integrated sequences homologous to the envelope, NS4B and polymerase genes of flaviviruses were identified. Overall, approximately two-thirds of a flavivirus-like genome were characterized. In the *Aedes aegypti* A20 cell line, a 492 aa ORF related to the polymerase of the *Cell Fusing Agent* and *Kamiti River virus* was identified. These flavivirus-related integrated DNA sequences were detected in laboratory-bred and wild *Aedes albopictus* and *Aedes aegypti* mosquitoes, demonstrating that their discovery is not an artefact resulting from the manipulation of mosquito cell lines, since they exist under natural conditions. This finding has major implications regarding evolution, as it represents an entirely different mechanism by which genetic diversity may be generated in eukaryotic cells distinct from accepted processes.

Received 25 November 2003
Accepted 23 February 2004

INTRODUCTION

The capture of genomic information from non-retroviral RNA viruses by eukaryotic cells is a potential evolutionary mechanism that has not been comprehensively addressed to date. As originally suggested (Zhdanov, 1975), one theoretical process could allow such capture based on an intracellular reverse transcriptase activity capable of copying viral RNA into a DNA form prior to integration into the host genome. Zhdanov and collaborators reported the presence, in infected cells, of DNA forms of numerous RNA viruses such as Tick-borne encephalitis virus (family *Flaviviridae*; Drynov *et al.*, 1981), Measles virus (family *Paramyxoviridae*; Zhdanov & Parfanovich, 1974), Sindbis virus (family *Togaviridae*; Zhdanov & Azadova, 1976) and Lymphocytic Choriomeningitis virus (LCMV, family *Arenaviridae*; Gaidamovich *et al.*, 1978). These findings could not be confirmed by other groups except in one case: Klenerman *et al.* (1997) described the presence of LCMV sequences present as DNA in mice over 200 days after infection. *In vitro* studies demonstrated that DNA synthesis from viral RNA was mediated by endogenous reverse transcriptase activity, but the possibility of integration of the viral sequence into the cellular genome was not investigated. In addition to this particular case, limited observations in the scientific literature support the hypothesis that horizontal transfer of genetic information from RNA viruses to eukaryotes is not impossible; for example, Morvan *et al.* (1999) detected DNA sequences of the glycoprotein gene of Ebola virus (family *Filoviridae*, RNA genome) in the spleen of an infected rodent by using PCR amplification. The issue of whether this DNA fragment was integrated or extrachromosomal could not be resolved. Also, Malik *et al.* (2000) identified regions of high similarity to the G2 envelope protein of phleboviruses (i.e. RNA viruses belonging to the family *Bunyaviridae*) in the genome of *Caenorhabditis elegans*. These envelope-like regions were inserted into the sequences of two BEL-like LTR-retrotransposons of the Pao/Ninja lineage.

The sequences reported in this article have been deposited in the DDBJ/EMBL/GenBank databases under accession numbers AF411835, AY223844, AY223845, AY223846, AY223847 and AY223848.

We report here a novel dataset that definitively establishes the existence of genetic transfer from RNA viruses to eukaryotic cells. Our observations involve flavivirus-related RNA viruses. Flaviviruses are enveloped viruses, the genome of which is a ssRNA molecule of positive polarity encoding a polyprotein secondarily cleaved to form structural (capsid- and envelope-associated) and non-structural (NS1–NS5) proteins. The evolutionary lineage includes more than 30 arthropod-borne viruses pathogenic for humans (e.g. Yellow Fever, Dengue, West Nile, Tick-borne encephalitis virus) but also non-vectored viruses and viruses only isolated in mosquitoes, such as Cell Fusing Agent virus (CFAV) (Cammisa-Parks *et al.*, 1992) and Kamiti River virus (KRV) (Sang *et al.*, 2003; Crabtree *et al.*, 2003). The presence of DNA sequences related to CFAV and KRV in the genome of *Aedes* mosquitoes is reported and evolutionary implications are discussed.

METHODS

Cell culture. C6/36 *Aedes albopictus* cells, A20 *Aedes aegypti* cells and *Aedes w-albus* cells were grown at 28 °C in Leibowitz L15 medium supplemented with 5 % fetal bovine serum and 10 % tryptose phosphate broth, in the presence of penicillin (100 U ml^{-1}) and streptomycin (100 μg ml^{-1}) (all reagents from Invitrogen).

Preparation of clonal C6/36 cell lines. Serial dilutions in culture medium of a suspension of C6/36 cells were prepared. One hundred microlitre aliquots of the dilution containing an average of 1 cell per ml were distributed in 96-well microplates and incubated at 28 °C for 7 days. Wells containing a single colony were selected, cells were transferred to 11 cm^2 wells and grown until cellular confluence.

Extraction of nucleic acids. Cultured cells were centrifuged at 2500 ***g*** and the resultant pellet was washed twice with Hanks' balanced salts solution. Individual mosquitoes were crushed using a mixer mill MM300 (Qiagen) in 200 μl Hanks' solution. DNA and RNA were extracted using the proteinase K/phenol/chloroform method (Sambrook *et al.*, 1989) and the RNA Now kit (Biogentex), respectively.

PCR protocols

Flavivirus-like NS3 sequence in the genome of C6/36 cells. PCR was performed using DNA template that had been extracted as described above, under standard conditions and without a reverse transcription step using the X1 (5′-YIRTIGGIYTITAYGGIWWYGG-3′)/X2 (5′-RTTIGCICCCATYTCISHDATRTCIG-3′) primer set and a recombinant *Taq* polymerase (Invitrogen). The X1/X2 primer set was designed to hybridize with NS3 nucleotide patterns conserved in most of the flavivirus sequences available to date.

Flavivirus-like NS5 sequence in the genome of *Aedes aegypti*. PCR was performed using DNA template that had been extracted as described above, under standard conditions and without a reverse transcription step using the PF1S (5′-TGYRTBTAYAACATGATGG-G-3′)/PF2R (5′-GTGTCCCADCCDGCDGTRTC-3′) primer set and a recombinant *Taq* polymerase. The PF1S/PF2R primer set was designed to hybridize with NS5 nucleotide patterns conserved in most of the flavivirus sequences available to date.

Specific detection of cell silent agent (CSA) inserts in the genome of individual *Aedes albopictus* mosquitoes and in C6/36 subclones. Conducted under similar conditions with primer sets CSA_NS3_S1/R1 (NS3 region: CSA_NS3_S1, 5′-GATCATCGTGCG-CAGCTTTATGG-3′; CSA_NS3_R1, 5′-CCTTGGTTTCAGAAACAA-TGACC-3′) and CSA_seq#2_S1/R1 (NS5 C-terminal region: CSA_seq#2_S1, 5′-AATTAGCAAGGAAGACTTGC-3′; CSA_seq#2_R1, 5′-GTGAGGTTCTTTCCTCAAGA-3′).

Detection of CSA2 inserts in the genome of individual *Aedes aegypti* mosquitoes. Conducted under similar conditions using the primer set PF1S/PF2R.

PCR of the cytochrome oxidase 1 (CO1) gene of *Aedes* mosquitoes. DNA was extracted from cell lines or individual mosquitoes as described above and amplified under standard conditions using primers COI-1S (5′-TACACAAGAAAGWGGAAAAAAGGAA-3′) and COI-1R (5′-GTAATTCTGAATAASTATGTTCTGC-3′).

Digestion by nucleases. DNA extracted from C6/36 cells was digested at 37 °C for 2 h in the appropriate buffer using either 50 U bovine pancreas DNase I, 10–40 U of restriction enzymes (*Pvu*II or *Alu*I) or 10 μg bovine pancreas RNase A ml^{-1} (all enzymes from Roche Molecular Biochemicals).

Genome walking. This was performed using the Universal GenomeWalker kit and the Advantage Genomic Polymerase mix (Clontech Laboratories).

Sequencing. Sequencing of both DNA strands of amplicons was performed using PCR primers (and additional primers deduced from the sequence for long fragments), dRhodamine DNA sequencing kit and an ABI Prism 377 sequence analyser (Perkin Elmer).

RT-PCR identification of mRNAs. RNA was extracted from C6/36 cells using the RNA Now kit (Biogentex) and incubated at 37 °C for 2 h with 50 U bovine pancreas DNase I. mRNAs were captured, washed and eluted using the PolyAtract system 1000 kit (poly T magnetic beads; Promega). mRNAs were reverse transcribed at 42 °C for 90 min in the presence of 200 U MuMLV Superscript II RNase H^- reverse transcriptase (Invitrogen) and random hexaprimers (Roche Molecular Biochemicals). PCR was performed using overlapping sets of primers spanning the ORF of CSA (primers available upon request to the corresponding author). Direct PCR using the same primers was used as a control to exclude the presence of contaminating DNA.

Southern blot. DNA was extracted from C6/36 and *Aedes w-albus* cells as described above, digested by either *Xba*I, *Not*I or *Sma*I restriction enzymes and run on a 0·8 % agarose gel (overnight, 40 V in TAE buffer). The gel was subsequently treated with 0·24 M HCl (10 min) and a denaturation buffer (0·5 M NaOH, 1·5 M NaCl, 2 × 15 min). Transfer on to a nylon membrane (Hybond-N; Amersham Biosciences) was performed overnight using the capillary transfer method and the denaturation buffer. After neutralization (0·5 M Tris/HCl, pH 7·2, 1·5 M NaCl, 5 min), DNA was fixed (80 °C for 1 h, followed by UV cross-linking). A ^{32}P-labelled probe was prepared by the nick-translation method using a 599 nt PCR product located at the NS1–NS2 junction of CSA and hybridized to the immobilized DNA [overnight, at 67 °C in CHURCH buffer (7 % SDS, 0·5 M $NaHPO_4$, pH 7·2)]. Washings and exposure to X-ray film were performed as described elsewhere (Sambrook *et al.*, 1989).

Sequence analysis. The CFAV genomic sequence was obtained from GenBank (accession no. NC_001564) and the KRV sequence was provided by M. B. Crabtree (Centers for Disease Control and Prevention, Fort Collins, CO, USA). This sequence is now available in GenBank under AY149904. Other sequences of flaviviruses used for phylogenetic reconstruction are the same as reported previously (de Lamballerie *et al.*, 2002).

Sequences reported here were deposited in GenBank under accession numbers, AF411835 (sequence #1), AY223844 (#2), AY223845 (#3), AY223846 (#4), AY223847 (#5) and AY223848 (#6).

Alignments and phylogenetic reconstruction. Alignments of nucleotide or amino acid sequences were generated with the help of the CLUSTAL W (v1.74) program (Thompson *et al.*, 1994) and pairwise genetic distances were estimated with the program MEGA v2.0 (Kumar *et al.*, 2001).

A phylogenetic tree was reconstructed using the alignment of the CSA ORF amino acid sequence with the corresponding sequences of representative member viruses of the flavivirus lineage, with the pairwise distance algorithm and the neighbour-joining method implemented in MEGA. The robustness of the resulting branching patterns was tested by bootstrap analysis with 1000 replications.

BLAST. The relatedness of newly characterized sequences to sequences deposited in databases was assessed by the Basic Local Alignment Search Tool (Altschul *et al.*, 1990) implemented via the National Center for Biotechnology Information website (www.ncbi.nlm.nih.gov/blast/) against the complete GenBank database. The BLASTN (nucleotide query–nucleotide database comparison), BLASTP (protein query–protein database comparison) and BLASTX (nucleotide query–protein database comparison) algorithms were used. Relatedness to KRV sequence was tested using a local database of flavivirus polyproteins and the BLAST program implemented in the DNATools (v5.2.014) software program.

Sequence repeats. These were sought with DNATools (v5.2.014).

Hydropathy plots. Profiles of the protein encoded by CSA ORF and the homologous region of KRV polyprotein were produced in Microsoft Excel using the amino acid hydropathy values determined by Kyte & Doolittle (1982), and a sliding window of 11 aa. Aligned sequences were exported with all alignment-generated gaps maintained (with a hydropathy value of zero). This permitted the comparison of hydropathy plots of sequences of unequal lengths.

Evaluation of divergence dates for mosquitoes and viruses. The putative age of origin of the Diptera [$\sim$225 to 280 millions of years ago (mya)] has been previously evaluated from fossils, biogeographical history or molecular clock work (Simmons & Weller, 2001; Gaunt & Miles, 2002). This corresponds to a $\sim$60 % nucleotide divergence for the CO1 gene sequences of the most divergent diptera (see Fig. 4). If we put forward the hypothesis that evolution of the CO1 gene within this order conforms approximately to a molecular clock, the observed $\sim$9 % divergence between the CO1 sequences of *Aedes albopictus* and *Aedes w-albus* (this study, accession no. AY223849) indicates a date of divergence $\sim$34–42 mya.

There are no published data regarding the evolutionary rate of viruses related to CFAV. The most comprehensive study of the rate of evolution in dengue virus (an *Aedes*-borne flavivirus) published to date (Twiddy *et al.*, 2003) indicates that the substitution rate in the envelope gene is 2·5 to $11{\cdot}8 \times 10^{-4}$ substitutions site^{-1} year^{-1}. For other ssRNA viruses, the most commonly reported substitution rates are 10^{-3}–10^{-4} substitutions site^{-1} year^{-1} (Meyerhans & Vartanian, 1999) with extreme values at 10^{-6} substitutions site^{-1} year^{-1}. Thus, if we consider the $\sim$35 % nucleotide divergence between the CFAV, KRV and CSA sequences, the date of divergence between these viruses should be $\sim$3500 years ago (ya) (substitution rate $\sim 10^{-4}$ substitutions site^{-1} year^{-1}), with possible extremes from 350 ($\sim 10^{-4}$ substitutions site^{-1} year^{-1}) to 350 000 ya ($\sim 10^{-6}$ substitutions site^{-1} year^{-1}).

RESULTS

Detection of DNA sequences related to the NS3 of flaviviruses in C6/36 *Aedes albopictus* cells

Using consensus degenerate PCR oligonucleotides designed to amplify the NS3 region of flaviviruses and a direct PCR protocol (without a reverse transcription step), we obtained repeatedly positive amplification results from DNA extracts of uninfected cultured C6/36 *Aedes albopictus* cells. Negative results were constantly obtained in similar conditions using various mammalian cells (Vero, MRC5, BGM or Hep2 cells). Sequencing revealed that the amplified sequence was homologous to the NS3 gene of CFAV and KRV. To ensure the cellular origin of this sequence, the same cell line was obtained from other research units and from the American Tissue Culture Collection. All samples, tested independently, contained the DNA sequence. When cellular extracts were treated with RNase, DNase or restriction enzymes with cleavage sites in the amplified sequence (*Pvu*II, *Alu*I), only treatment with RNases allowed the subsequent PCR amplification of the sequence in question. This demonstrated that the molecular template was a double-stranded DNA molecule.

Extension of sequence by genome walking and comparison with flaviviral sequences

The sequence initially characterized was extended using genome walking. Results are presented in Fig. 1. A 10 623 nt sequence was obtained (*Aedes albopictus* sequence #1) that includes an ORF of 1557 aa homologous to a region of CFAV and KRV genomes (59 and 67 % identity, respectively), starting in the N-terminal part of the NS1 gene and ending shortly after the NS4A–NS4B junction.

Phylogenetic reconstruction using the complete amino acid ORF sequence and homologous sequences from viruses belonging to the flavivirus lineage confirmed the evolutionary relationship of the newly described '*Aedes albopictus* cell silent agent' (CSA) with CFAV and KRV (Fig. 2). The region of the flavivirus polyprotein homologous to the CSA ORF includes several cleavage sites for the viral serine proteinase, a series of cysteine residues (in the NS1 gene) and enzymic motifs (in the NS3 gene) responsible for serine protease and helicase/NTPase activities, which are widely conserved among flaviviruses (de Lamballerie *et al.*, 2002). These sites and motifs are present and conserved in the CSA sequence, and the hydropathy profile of the protein deduced by conceptual translation of the ORF sequence is remarkably similar to that of the homologous region of KRV polyprotein (Fig. 3).

Detection of mRNAs

The presence of functional motifs in sequence #1 suggested that the transcription of an mRNA extending over the complete ORF might lead to functional viral enzymes. C6/36 total RNAs were extracted and treated with DNase. mRNAs

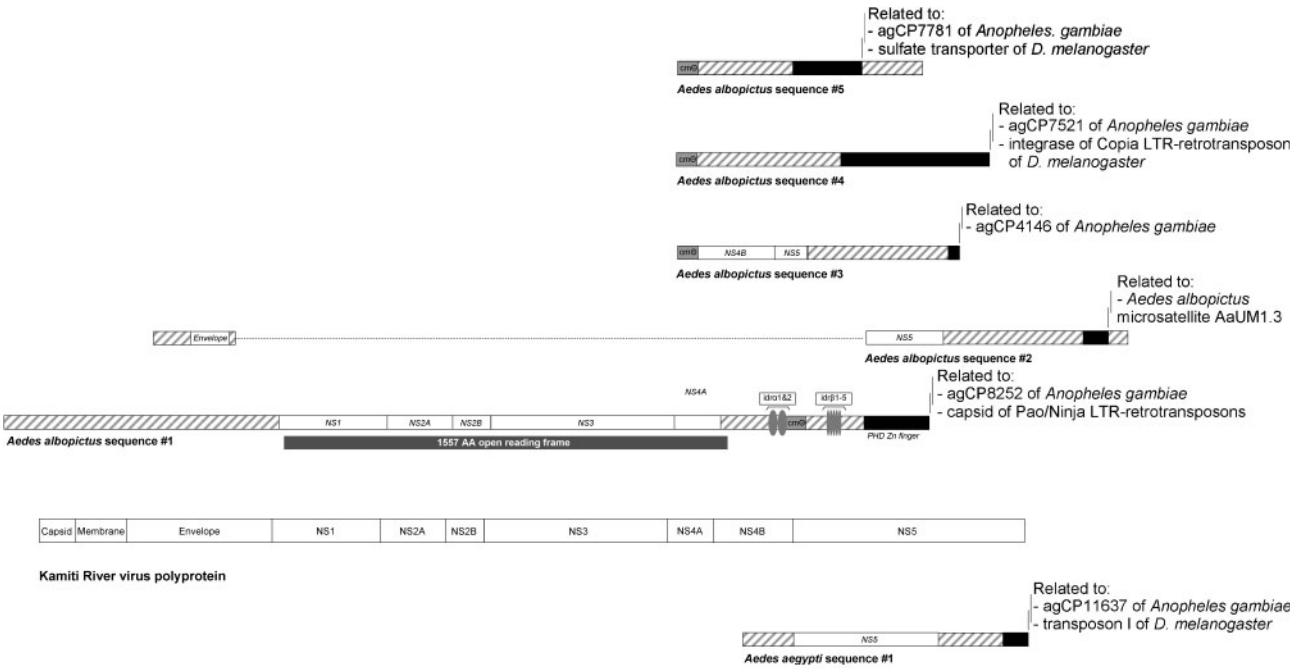


Fig. 1. Sequences determined from *Aedes* cultured cells via genome walking. Comparison of five sequences from *Aedes albopictus* (upper portion of figure) and one from *Aedes aegypti* (lower portion of figure) with the complete coding region of the Kamiti River virus. Open bars have significant similarity with the KRV polyprotein. Filled bars have significant similarity with insect genes as detailed in the figure. Hatched areas correspond to regions with no significant similarity with any known sequence. *Aedes albopictus* sequence #1 includes a series of imperfect direct repeats (idrα1 and 2; idrβ1 to 5) and the conserved motif Θ (cmΘ) shared with sequences #3, #4 and #5.

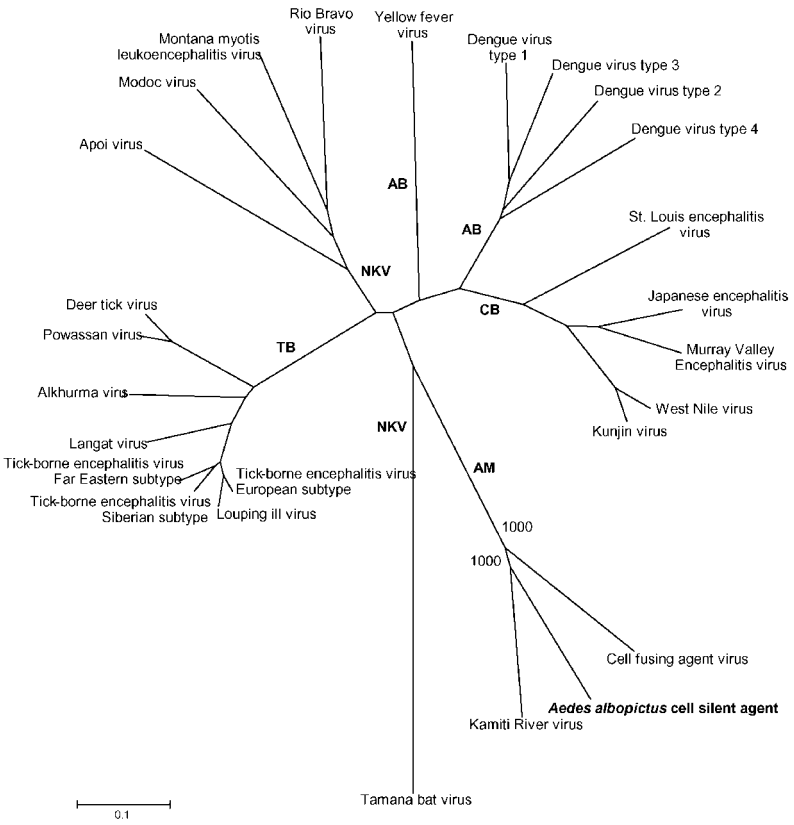
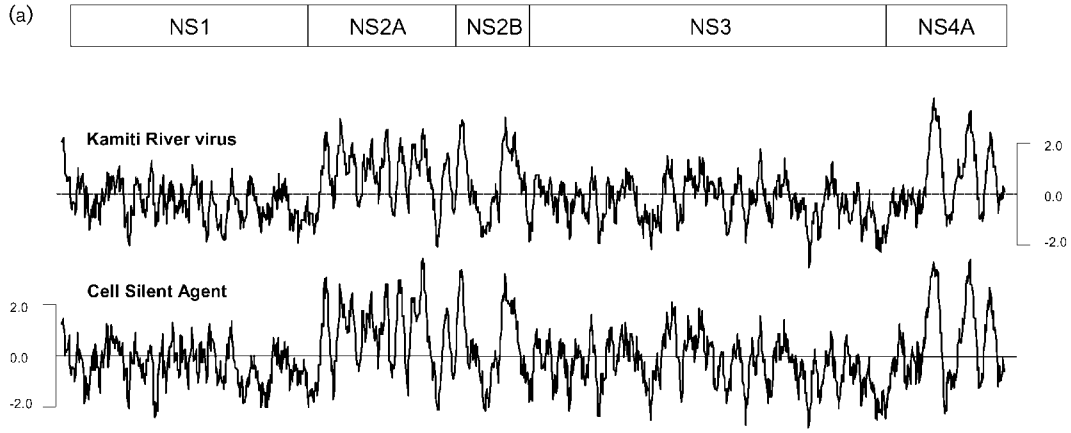


Fig. 2. Phylogenetic relationship between the *Aedes albopictus* Cell Silent Agent and representative members of the flavivirus lineage. The CSA ORF aa sequence was aligned with the corresponding sequences of viruses belonging to the flavivirus lineage, including tick- (TB), *Aedes*- (AB) and *Culex*-borne (CB) viruses, viruses with no known vector (NKV) and viruses isolated only from *Aedes* mosquitoes (AM). The pairwise distance algorithm and the neighbour-joining method were used for phylogenetic reconstruction. Bootstrap values are indicated within the AM group (for 1000 pseudo-replications).

(a)



(b)

Cleavage site	CFAV	KRV	CSA
E/NS1	FYYVRA/-DLGC	VRSVSA/-DVGC	YIRRV**A**/G**D**V**GC**
NS1/NS2A	YGKANA/QSDFR	YGKAHA/CSDFR	**YGKA**D**A**/R**SDF**H
NS2A/NS2B	WAAEKA/HQPTV	WAAERA/QQPTI	**WAAE**R**A**/EH**P**S**V**
NS2B/NS3	LTASNR/SDDLL	LSEQNR/SDDLL	**L**N**E**H**C**R/**SDDLL**
NS3/NS4A	EWETRK/VSIDF	EWDTRK/LSIEF	**EW**D**Q**R**K**/L**SI**E**M**
NS4B/NS5	FNQFRA/LEKST	FNQFRA/LEKST	**FSQFRA/LEKST**

(c)

Serine protease

	BOX 1		BOX 2		BOX 3		BOX 4
					↓		↓ ↓↓ ↓
Flaviviruses	g..HT.W**H**.T.G		**D**...Y.G.W		D...G.**S**GSP		G...GLYGNG.......y.S
TABV	G..TT.Y**H**.T.G		**D**...Y.G.W		K...G.**S**GTP		Q...GTLKPV.......V.G
CFAV	GVFHTLM**H**VTRG	//	**D**VVSYGGPW	//	DFGKGS**S**GSP	//	GEPVGFYGFGFYVNGI-YRS
KRV	GVFHTLM**H**VTRG		**D**VVSYGGPW		DFGKGS**S**GSP		GDPVGFYGFGFYVDGI-YRS
CSA	GVFHTLM**H**VTRG		**D**VVSYGGPW		DFGKGS**S**GSP		GEVVGFYGYGFYIDGI-YRS

Helicase / NTPase

	MOTIF I		MOTIF IA		MOTIF II		MOTIF III
Flaviviruses	d.HPG.**GKT**		rT..**LAPTRVV**..Em..A		N.....MDEAH..DP.S.**A**ARG		MTATPPG
TABV	V.KCG.**GKT**		LV..**LVPTRVV**..EA..V		N.....VDESH..NP.TLALHN		MYL**T**ATG
CFAV	TWHPGKG**KT**	//	RTVI**LTPTRVV**MAEVVEA	//	RVAMIIMDECHFMDPMSI**A**ARG	//	LSATPPG
KRV	TWHPG**KGKT**		RTII**LTPTRVV**MAEVIEA		RVALIIMDECHFMDPMSI**A**ARG		LSATPPG
CSA	TWHPG**KGKT**		RTII**LTPTRVV**MAEVIDA		GVTLIIMDECHFMDPMSI**A**ARG		LSA**T**PPG

	MOTIF IV		MOTIF V		MOTIF VI
Flaviviruses	T.W**F**v**P**S.........L		D....TDI**S**EMGAN		SAAQRRGR.GR
TABV	I.Y**F**VAS.........L		G....TNI**S**EMGAN		SKIQRRGR.GR
CFAV	TIL**F**V**P**SHNQANTLASAI	//	ALVISTDI**S**EMGAN	//	SMIQRRGRTGR
KRV	TIM**F**V**P**SHNQANALAAAI		TLVVSTDI**S**EMGAN		SMVQRRGRTGR
CSA	AIM**F**V**P**SHTQANSLAGSI		TLVVSTDI**S**EMGAN		SMIQRRGRTGR

Fig. 3. Comparison of CSA, KRV and CFAV polyproteins. (a) Hydropathy profiles of the protein encoded by the main ORF of CSA and the corresponding region of KRV polyprotein. Amino acid hydropathy values determined by Kyte & Doolittle (1982), and a sliding window of 11 aa were used. (b) Proposed cleavage sites in the CSA polyprotein. Possible cleavage sites were deduced via alignment of CSA with CFAV and KRV polyproteins and comparison with cleavage sites identified previously by Cammisa-Parks *et al.* (1992) and Crabtree *et al.* (2003). Residues in bold in the CSA column are conserved in all three polyproteins. Underlined residues are common to one other polyprotein. (c) Conserved enzymic motifs in the proteins encoded by the NS3 gene. Sequence alignments are provided that include flaviviruses, the flavivirus-related Tamana bat virus (TABV), CFAV, KRV and the newly characterized CSA. Dashes represent alignment gaps; residues in bold are conserved; arrows indicate putative substrate binding sites; shaded letters correspond to catalytic sites. First line: amino acid completely (capitals) or nearly completely (lower-case letters) conserved among flaviviruses belonging to the tick-borne group, to the mosquito-borne group and to the group with no known vector (dots stand for non-conserved amino acid positions). Second line: TABV sequence (only positions corresponding to conserved residues in flaviviruses are indicated). Third, fourth and fifth lines: amino acid sequences of CFAV, KRV and CSA.

were captured and purified by poly T magnetic beads and tested by direct PCR or RT-PCR with overlapping sets of primers in the main ORF of CSA. Results were constantly negative using direct PCR and positive using RT-PCR. This indicated the presence of RNA transcripts, which were sequenced and found to be 100 % identical to the CSA ORF sequence.

Identification of other sequences related to flaviviruses

Genome walking using primers located in motif Θ (a 117 nt sequence located in the 3′-region of sequence #1) led to the characterization of a new sequence of interest (sequence #3), which is not contiguous to sequence #1. Sequence #3 includes a 1037 nt sequence related to KRV NS4B/NS5 (70 % aa identity, BLASTP E value = $5e^{-153}$). In order to identify additional flavivirus-related sequences, primers were designed from the NS5 gene of CFAV/KRV and tested on C6/36 DNA extracts. This led to the identification of an 852 nt sequence (sequence #2) similar to the C-terminal region of the KRV NS5 (70 % nt identity using coding positions 1 + 2), which is not contiguous to sequence #3. Genome walking extension of sequence #2 showed that the NS5-like insert is contiguous to a 334 nt region homologous to the KRV envelope (57 % aa identity, BLASTP E value = $4e^{-31}$).

Evidence for integration into the cell genome

Overall, approximately two-thirds of a flavivirus-like genome were characterized, comprising a patchwork of sequences of various sizes. In order to test fully the hypothesis of integration into the cellular genome, we investigated the nature of the flanking sequences. We compared the latter with those short partial sequences that are available for *Aedes albopictus*, plus other insect sequences from GenBank. We found strong evidence that CSA sequences are present in the *Aedes albopictus* genome, at three different positions. In sequence #1 (first integration site), characterization of the 3′-non-viral flanking sequence led to the identification of the first 213 aa of an ORF with significant similarity (BLASTP E value = $7e^{-42}$) to the agCP8252 protein of *Anopheles gambiae*. Both deduced proteins contain a plant homeo domain (PHD) finger motif that folds into an interleaved type of zinc finger chelating two zinc ions. They are significantly related (BLASTP E value < $4e^{-06}$) to the capsid proteins of LTR-retrotransposons of the Pao/Ninja lineage. In sequence #2 (second integration site), the 3′-non-viral flanking sequence contains a 266 nt sequence that is 98 % identical to the *Aedes albopictus* microsatellite AaUM1.3 (BLASTN E value = $1e^{-137}$). Moreover, sequences #1 and #3 (third integration site) contain the conserved motif Θ. This was also identified in cellular sequences #4 and #5, which contain ORFs with cognate specific coding sequences in insect genomes. In sequence #4, a 602 aa ORF is closely related (BLASTP E value < $1e^{-154}$) to the agCP7521 protein of *Anopheles gambiae*. It includes a zinc-binding motif that is the central catalytic domain of an integrase related to the retrotransposon Copia of *Drosophila melanogaster*[9] (BLASTP E value = $1e^{-93}$). In sequence #5, a 241 aa ORF is related to the agCP7781 protein of *Anopheles gambiae* (BLASTP E value = $5e^{-76}$) and the AAD53951 sulfate transporter of *Drosophila melanogaster* (BLASTP E value = $1e^{-35}$).

In addition to investigations based on sequence determination and analysis, direct evidence for integration of the CSA sequence into the *Aedes albopictus* genome was provided by Southern blotting. A probe located at the NS1–NS2 junction of CSA specifically hybridized to 25, 19 and 13 kb restriction fragments of DNA extracted from *Aedes albopictus* and digested with *Not*I, *Xba*I and *Sma*I, respectively. Hybridization with DNA extracted from *Aedes w-albus* cells was not observed (Fig. 4).

Analysis of CSA inserts in subclones of C6/36 cells

Having demonstrated the presence of CSA in the genome of C6/36 cells, we investigated whether all individual cells within the cell line contained this integration. We produced 50 new clonal lines from C6/36 cells that were tested using specific primers designed from the NS3 and NS5 regions of CSA. All clones tested positive for both the NS3 and NS5 insertion sites.

Analysis of CSA inserts in *Aedes albopictus* mosquitoes

Wild *Aedes albopictus* mosquitoes from Thailand, Texas (USA), Cameroon and Madagascar and laboratory-bred mosquitoes from Japan, Thailand, Madagascar, Italy and France were tested for the presence of CSA sequences using primers specific for the NS3 and NS5 regions. Amplification and sequencing of the CO1 gene was used as a control for the quality of DNA and identification of the mosquito (Simmons & Weller, 2001). Among 130 mosquitoes assigned as valid samples, 97·7 % tested positive for at least one integration site, including those from all geographical origins tested, all stages of development (eggs, larvae, nymphs and adults) and both males and females.

Analysis of other *Aedes* species

The presence of CSA sequences in other *Aedes* species was examined in the closest relatives of *Aedes albopictus* (Singh & Bhat, 1971) [species *Aedes w-albus* (Shouche & Patole, 2000) and *Aedes aegypti*] with CSA-specific NS3 and NS5 primers and NS3 and NS5 degenerate primers (X1/X2 and PF1S/PF2R). In *Aedes w-albus* cell culture, all results were negative. However, when laboratory-bred *Aedes aegypti* mosquitoes were tested with flavivirus NS5 consensus primers, a ‘CSA2’ sequence distinct from CSA but related to CFAV and KRV was identified in the ten individuals tested. It was also found in all five wild specimens from Senegal that were tested and in the A20 *Aedes aegypti* cell line. The sequence was extended by genome walking and consisted of a 1479 nt long flavivirus-like region (*Aedes*

aegypti sequence #1), encoding a 492 aa ORF (61 % aa identity to KRV, BLASTP E value = 0·0) related to the NS5 of flaviviruses and flanked by non-viral sequences (Fig. 1). The 3′-flanking region is related to the agCP11637 protein of *Anopheles gambiae* (BLASTP E value = $2e^{-9}$) and transposon I of *Drosophila melanogaster* (BLASTX E value = $5e^{-5}$), indicating integration into the genome of *Aedes aegypti*.

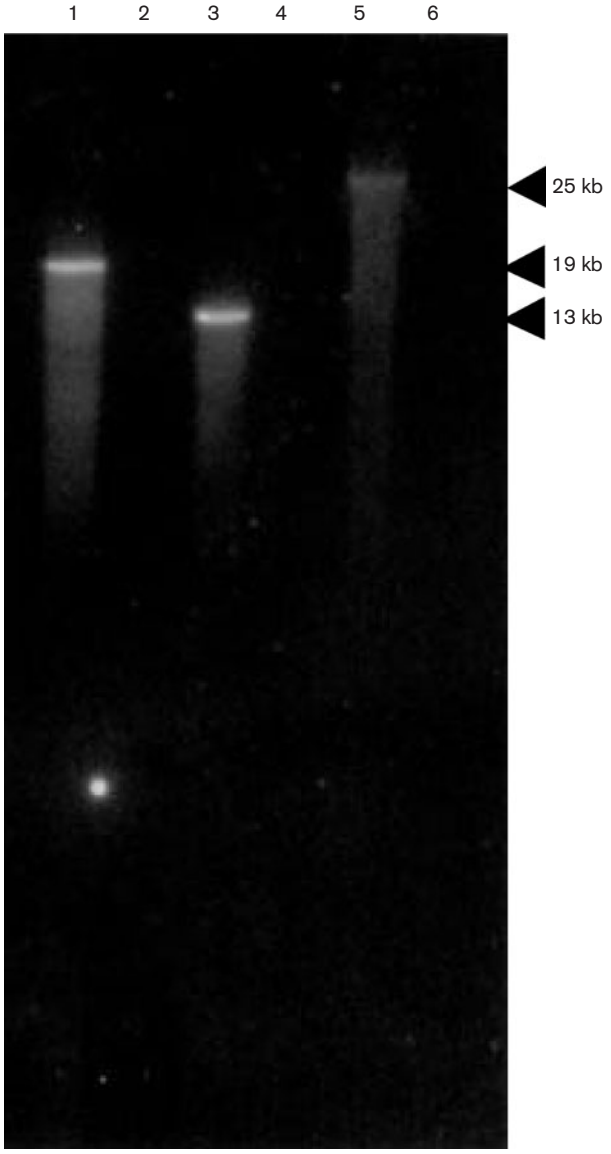


Fig. 4. Hybridization of radiolabelled probe to nucleic acids immobilized on nylon membrane. A 599 nt ^{32}P-labelled probe located at the NS1–NS2 junction was hybridized to restriction fragments of DNA extracted from *Aedes albopictus* (lanes 1, 3 and 5) and *Aedes w-albus* (lanes 2, 4 and 6) cells. Genomic DNA was digested by *Xba*I (lanes 1 and 2), *Sma*I (lanes 3 and 4) or *Not*I (lanes 5 and 6).

DISCUSSION

The PCR identification of a DNA sequence related to the NS3 of flaviviruses in uninfected C6/36 *Aedes albopictus* cells was unexpected. New cells, new primers and independent investigators working in a different location confirmed this result and ruled out the possibility of PCR contamination. The probability of such contamination was, however, very low, as the identified sequence clearly originated from a new uncharacterized virus of the CFAV group, distinct from CFAV and KRV (which had never been manipulated in our unit). The precise nature of the nucleic acid that was amplified was investigated by using nucleases specific for RNA (bovine pancreas RNase A), DNA (bovine pancreas DNase) or dsDNA (restriction enzymes). As expected, the action of RNase on C6/36 nucleic acid extracts did not prevent subsequent PCR amplification, while treatments by both DNase and restriction enzymes provided negative PCR results. This indicated that the PCR template was a dsDNA molecule. We examined whether this DNA sequence belonged to a DNA virus silently infecting C6/36 cells or was present in the cells in an episomal or integrated form. In the absence of an available genomic library for *Aedes albopictus*, this was investigated by the genome-walking method and allowed us (i) to extend the viral sequence, (ii) to discover new viral sequences, which were not contiguous to that originally identified, and (iii) to characterize non-viral flanking sequences. The latter sequences clearly belonged to the cell genome and allowed us to demonstrate the integration of the viral sequences at three different insertion points within the genome of C6/36 cells.

The main insert includes an ORF that corresponds to approximately one-half of the flavivirus polyprotein. Its integration within the cellular genome was confirmed by Southern blotting. The encoded 1557 aa polyprotein is homologous to the NS1, NS2A, NS2B, NS3 and NS4A of flaviviruses. This region includes the well-characterized enzymic domains of the viral helicase and serine protease. These motifs are conserved and the expression of the polyprotein is likely to lead to functional enzymes. This is a different situation to that observed in sequence #2, in which the sequence homologous to the flavivirus polymerase is truncated and includes numerous stop codons, formally excluding the possibility that a functional enzyme is generated. The subsistence of the long NS1–NS4B ORF along the evolution of *Aedes albopictus* and the conservation of enzymic domains suggests that the corresponding proteins are expressed. This hypothesis is reinforced by the detection of the corresponding mRNAs in C6/36 cells. Although the presence of the corresponding proteins has not yet been fully investigated and is necessary for definitive confirmation, the presence of mRNAs constitutes strong evidence that the captured genes are likely to be used by host cells.

An important point was to determine what proportion of cells contained the viral inserts. A large number of subclones of the cell line were produced and all tested positive

for the presence of inserts. Because the C6/36 cell line is clonal (Igarashi, 1978), this implied the presence of insertions in the original parent cell (and therefore in the mosquito larvae used for establishing the cell line). This led us to test laboratory-bred and wild *Aedes albopictus* mosquitoes, for which positive results were obtained with mosquitoes from all geographical origins tested, all stages of development and both males and females. Only 2·3 % tested negative for all integration sites; however, these results may represent technical artefacts because they were obtained from dried insects rather than fresh material. Nevertheless, the presence of integrated sequences in the majority of individuals is of great interest because it demonstrates that the integration of CSA sequences is not an artefact due to the manipulation of the C6/36 cell line and exists in wild mosquitoes. It is likely to be associated with vertical transmission of the sequence through generations of mosquitoes rather than current infection of *Aedes albopictus* mosquitoes by the viral form of CSA.

The discovery of the CSA2 sequences in the genome of *Aedes aegypti* mosquitoes provides the first example of flavivirus-like sequence detected in the dsDNA of mosquitoes. The NS5-like sequence characterized is different from that of *Aedes albopictus* and does not include the functional motifs of a polymerase. As observed in the case of *Aedes albopictus*, the analysis of non-viral flanking sequences provides strong evidence for integration in the cellular genome and the flavivirus-like sequence is present in laboratory-bred and wild mosquitoes. The implications of the presence of different flavivirus-related sequences in different mosquitoes are important from an evolutionary point of view and are discussed further below.

The presence of flaviviral-like sequences in mosquito genomes could be explained by one of two hypotheses: either (i) these sequences were integrated into the genomes of *Aedes* spp. mosquitoes following infection by the corresponding RNA viruses or (ii) flaviviruses originate from the genome of *Aedes* spp. The latter hypothesis is contradicted by several observations. Firstly, the organization of these genes as a unique ORF in the genomes of flaviviruses is not observed in insect genomes, where they constitute independent inserts. In addition, with the remarkable exception of *Aedes albopictus* sequence #1, these genes are truncated (*Aedes albopictus* sequences #2 and #3 and *Aedes aegypti* sequence #1) or contain multiple stop codons (*Aedes albopictus* sequence #2). Also, there is a complete absence of overlap between sequences of the different CSA inserts identified, which strongly suggests a single original integration, followed by several reshuffling events. Secondly, our current knowledge of flaviviruses and mosquito evolution suggests that the split between *Aedes albopictus*, *Aedes aegypti* and *Aedes w-albus* (~34–42 mya; Fig. 5) is significantly more ancient than the most recent common ancestor of CSA, CFAV and KRV (probably around ~3500 ya and certainly less than 350 000 ya). Therefore, a cellular origin of flaviviruses would require the conservation of an intact ORF within *Aedes* spp. genomes throughout several million years, with a sudden disappearance from the genomes of both *Aedes albopictus* and *Aedes aegypti* mosquitoes during the recent millennia. This is a very improbable

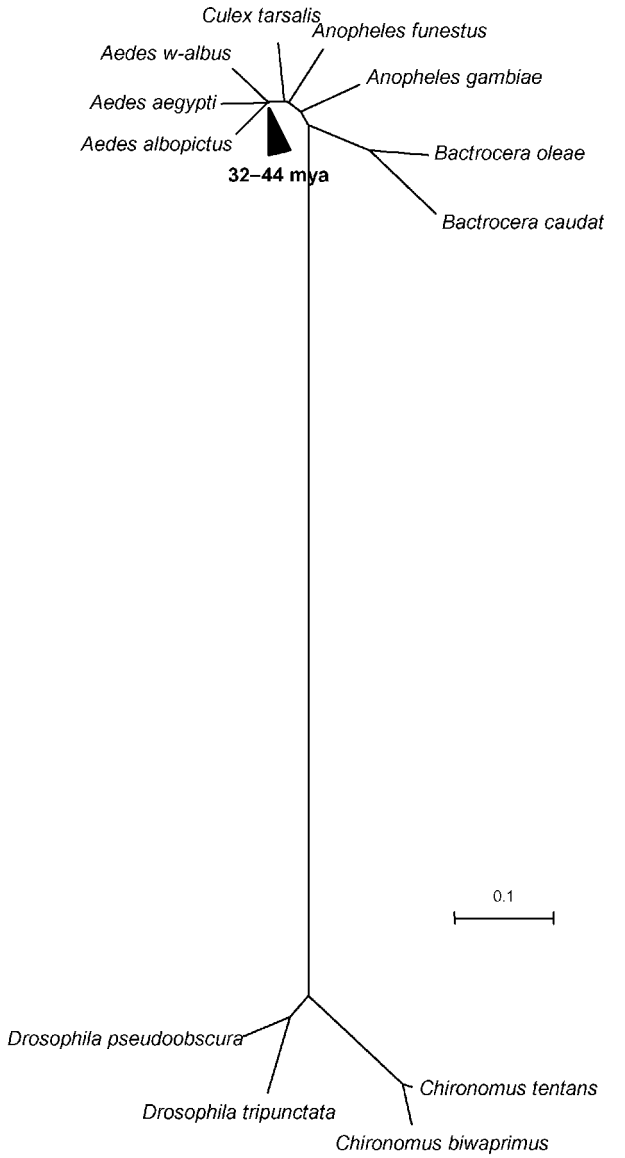


Fig. 5. Evolutionary relationship within the order Diptera based on nucleotide sequence analysis of the CO1 gene. Members of families Drosophilidae (*Drosophila* sp.), Chironomidae (*Chironomus* sp.), Culicidae (*Anopheles* sp.), Tephritidae (*Bactrocera* sp.) and of genera *Culex* (*Culex* sp.) and *Aedes* (*Aedes* sp.) were included. The genetic distances between *Aedes w-albus* and *Aedes albopictus* or *Aedes aegypti* are similar (~9 %). The corresponding polytomy observed in the tree is probably a soft multifurcation due to the low resolution of branching. Dating of the common ancestor of the three species is 32–44 mya and that of the complete order is 225–280 mya (Simmons & Weller, 2001; Meyerhans & Vartanian, 1999).

evolutionary scenario. Thirdly, the polymerases of numerous viruses that specifically infect mammalian cells (pestiviruses, hepaciviruses and Tamana bat virus) are phylogenetically related to those of flaviviruses (de Lamballerie *et al.*, 2002) and there is little chance that these viruses originated from the genome of mosquitoes in recent times; for example, it has been suggested that the common ancestor of hepaciviruses GB virus A and C existed 35 mya in primates (Charrel *et al.*, 1999).

The first hypothesis is clearly the most plausible, namely that *Aedes* spp. mosquitoes have been infected by flavivirus-like viruses prior to integration events. In the case of CSA2, this scenario is supported by the fact that *Aedes aegypti* is a mosquito of African origin, and CFAV and KRV were isolated from African *Aedes aegypti* and *Aedes macintoshi* mosquitoes, respectively. However, in the case of CSA, there is a discrepancy between the geographical origin of the mosquitoes and the virus. To date, CFAV and KRV have only been isolated from African mosquitoes, while *Aedes albopictus*, the 'Asian tiger mosquito', originates from Asia. It was propagated worldwide in the last few decades but the integration event probably occurred in Asia prior to its dispersal, since all the Asian specimens that could be tested contained CSA sequences. No virus related to CFAV and KRV has been isolated in Asia to date, but this does not rule out the possibility that such viruses exist. In addition, the times to the most recent common ancestor of the *Aedes* spp. (Fig. 4) and that of CSA, CFAV and KRV differ by orders of magnitude $\sim$35 mya and $\sim$3500 ya, respectively). This indicates that *Aedes albopictus* and *Aedes aegypti* did not inherit the CSA and CSA2 sequences from a common ancestor. Therefore, at least two independent integration events occurred. This is important to consider, as it suggests that integration events are not exceptional and might be reproduced experimentally.

In summary, we have reported for the first time the presence of a multigenic sequence from a non-retroviral RNA virus in a eukaryotic cell, and have demonstrated that it takes the form of DNA integrated into the cellular genome itself. Our observations in mosquitoes shed new light upon previous studies, which have found DNA forms of RNA viruses in cells from species as wide-ranging as mammals and nematodes, suggesting that this phenomenon is in fact more common than previously thought. Previous findings described viruses with a ssRNA genome of negative polarity (*Filoviridae*) or with a segmented ssRNA genome of negative or ambisense polarity (*Arenaviridae*, *Bunyaviridae*), while flavivirus genomes are ssRNAs of positive polarity. RNA viruses are characterized by their extreme genetic diversity and their propensity to evolve quickly, and thus may represent a source of evolution for eukaryotic cells, which are genetically much more stable. In the case of CSA, the integration into the majority of mosquitoes tested, the perfect conservation of enzymic motifs and the detection of mRNAs suggests that an advantage may be conferred upon *Aedes albopictus* from the expression of the viral proteins. This finding could have major implications in terms of evolutionary theory since it represents an entirely different method apart from the accepted processes for the generation of genetic diversity in eukaryotic cells. In addition, it will be of major importance to determine whether RNA viruses infecting humans are able to transmit genes to infected somatic cells and what consequences this could have. The mechanisms of reverse transcription and integration remain to be determined precisely. They will be discussed in another article in preparation, together with the report of experimental data demonstrating that generation of cDNAs can be observed following the infection of mosquito cells by present-time flavivirus-related viruses such as CFAV.

ACKNOWLEDGEMENTS

The authors thank Dr M. B. Crabtree for providing the KRV sequence prior to publication, Drs R. Bellini, A. Failloux, J. P. Gonzalez, G. Lanzaro, V. Robert, F. Schaffner and F. Simard for providing *Aedes* mosquitoes, Dr Y. S. Shouche for providing an *Aedes w-albus* cell line and Dr M. Mitchell for his help during hybridization assays. The 'Unité des Virus Emergents' is an associated research unit of the Institut de Recherche pour le Développement (IRD). This study was supported in part by the IRD.

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol* **215**, 403–410.

Cammisa-Parks, H., Cisar, L. A., Kane, A. & Stollar, V. (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.

Charrel, R. N., De Micco, P. & de Lamballerie, X. (1999). Phylogenetic analysis of GB viruses A and C: evidence for cospeciation between virus isolates and their primate hosts. *J Gen Virol* **80**, 2329–2335.

Crabtree, M. B., Sang, R. C., Stollar, V., Dunster, L. M. & Miller, B. R. (2003). Genetic and phenotypic characterization of the newly described insect flavivirus, Kamiti River virus. *Arch Virol* **148**, 1095–1118.

de Lamballerie, X., Crochu, S., Billoir, F., Neyts, J., de Micco, P., Holmes, E. C. & Gould, E. A. (2002). Genome sequence analysis of Tamana bat virus and its relationship with the genus *Flavivirus*. *J Gen Virol* **83**, 2443–2454.

Drynov, I. D., Uryvaev, L. V., Nosikov, V. V. & Zhdanov, V. M. (1981). Integration of the genomes of the tick-borne encephalitis virus and of the cell in chronic infection due to this virus and SV40. *Dokl Akad Nauk SSSR* **258**, 1000–1002 (in Russian).

Gaidamovich, S. Y., Cherednichenko, Y. N. & Zhdanov, V. M. (1978). On the mechanism of the persistence of lymphocytic choriomeningitis virus in the continuous cell line Detroit-6. *Intervirology* **9**, 156–161.

Gaunt, M. W. & Miles, A. (2002). An insect molecular clock dates the origin of the insects and accords with palaeontological and biogeographic landmarks. *Mol Biol Evol* **19**, 748–761.

Igarashi, A. (1978). Isolation of a Singh's *Aedes albopictus* cell clone sensitive to Dengue and Chikungunya viruses. *J Gen Virol* **40**, 531–544.

Klenerman, P., Hengartner, H. & Zinkernagel, R. M. (1997). A non-retroviral RNA virus persists in DNA form. *Nature* **390**, 298–301.

Kumar, S., Tamura, K., Jakobsen, I. B. & Nei, M. (2001). MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics* **17**, 1244–1245.

Kyte, J. & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J Mol Biol* **157**, 105–132.

Malik, H. S., Henikoff, S. & Eickbush, T. H. (2000). Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. *Genome Res* **10**, 1307–1318.

Meyerhans, A. & Vartanian, J.-P. (1999). The fidelity of cellular and viral polymerases and its manipulation for hypermutagenesis. In *Origin and Evolution of Viruses*, pp. 87–114. Edited by E. Domingo, R. Webster & J. Holland. London: Academic Press.

Morvan, J. M., Deubel, V., Gounon, P. & 9 other authors (1999). Identification of Ebola virus sequences present as RNA or DNA in organs of terrestrial small mammals of the Central African Republic. *Microbes Infect* **1**, 1193–1201.

Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989). *Molecular Cloning: a Laboratory Manual*, 2nd edn. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory.

Sang, R. C., Gichogo, A., Gachoya, J., Dunster, M. D., Ofula, V., Hunt, A. R., Crabtree, M. B., Miller, B. R. & Dunster, L. M. (2003). Isolation of a new flavivirus related to cell fusing agent virus (CFAV) from field-collected flood-water *Aedes* mosquitoes sampled from a dambo in central Kenya. *Arch Virol* **148**, 1085–1093.

Shouche, Y. S. & Patole, M. S. (2000). Sequence analysis of mitochondrial 16S ribosomal RNA gene fragment from seven mosquito species. *J Biosci* **25**, 361–366.

Simmons, R. B. & Weller, S. J. (2001). Utility and evolution of cytochrome *b* in insects. *Mol Phylogenet Evol* **20**, 196–210.

Singh, K. R. & Bhat, U. K. (1971). Establishment of 2 mosquito cell lines from larval tissues of *Aedes w-albus*. *Experientia* **27**, 142–143.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673–4680.

Twiddy, S. S., Holmes, E. C. & Rambaut, A. (2003). Inferring the rate and time-scale of dengue virus evolution. *Mol Biol Evol* **20**, 122–129.

Zhdanov, V. M. (1975). Integration of viral genomes. *Nature* **256**, 471–473.

Zhdanov, V. M. & Azadova, N. B. (1976). Integration and transfection of an arbovirus by mammalian cells. *Mol Biol (Mosk)* **10**, 1296–1302 (in Russian).

Zhdanov, V. M. & Parfanovich, M. I. (1974). Integration of measles virus nucleic acid into the cell genome. *Arch Gesamte Virusforsch* **45**, 225–234.