

# Phylogenetic and evolutionary analyses of St. Louis encephalitis virus genomes

Gregory J. Baillie<sup>a,\*</sup>, Sergios-Orestis Kolokotronis<sup>a</sup>, Eric Waltari<sup>a</sup>, Joseph G. Maffei<sup>b</sup>,  
Laura D. Kramer<sup>b</sup>, Susan L. Perkins<sup>a</sup>

<sup>a</sup> *Division of Invertebrate Zoology, and Sackler Institute for Comparative Genomics, American Museum of Natural History, Central Park West at West 79th Street, New York, NY 10024, USA*

<sup>b</sup> *The Arbovirus Laboratories, Wadsworth Center, New York State Department of Health, 5668 State Farm Road, Slingerlands, NY 12159, USA*

Received 15 August 2007; revised 24 January 2008; accepted 14 February 2008  
Available online 29 February 2008

## Abstract

St. Louis encephalitis virus belongs to the Japanese encephalitis virus serocomplex of the genus *Flavivirus*, family *Flaviviridae*. Since the first known epidemic in 1933, the virus has been isolated from a variety of geographical, temporal, and host origins. We have sequenced 10,236 nucleotides of the open reading frame (93.6% of the full-length genome) of 23 of these strains, and have used the sequences to conduct phylogenetic analyses, in order to investigate the forces shaping the evolution of St. Louis encephalitis virus. Contrary to previous reports, we found little evidence for recombination in these isolates. Most of the amino acid sites in the SLEV polyprotein appeared to be under negative selection, with some sites evolving neutrally, and a small number under positive selection. The strongest signal for positive selection was evident in the *N*-linked glycosylation site of the envelope protein. Intra-strain sequence variability within strains was observed at this site, and analyses suggested that it is under selection *in vitro*. Furthermore, using heterochronous sequence data, we estimated the most recent expansion of St. Louis encephalitis virus in North America to have happened towards the end of the 19th century.

© 2008 Elsevier Inc. All rights reserved.

**Keywords:** St. Louis encephalitis virus; *Flavivirus*; Phylogeny; Positive selection; Recombination; *N*-linked glycosylation; Divergence times

## 1. Introduction

*St. Louis encephalitis virus* (SLEV) belongs to the Japanese encephalitis virus serocomplex within the genus *Flavivirus* (Family: *Flaviviridae*) (Lindenbach and Rice, 2001). Its known distribution includes parts of North America, Central America, South America, and islands of the Caribbean, and its primary invertebrate and vertebrate hosts are mosquitoes of the genus *Culex* and various species of wild birds, respectively (Reisen, 2003). Mam-

mals, including humans, are also occasionally infected, but are considered dead-end hosts (Reisen, 2003). The virus is named after an epidemic that occurred in St. Louis, Missouri in 1933, and there have been at least 41 outbreaks in North America since that time (Day, 2001; Reisen, 2003).

For the last 74 years, samples of SLEV from a variety of geographical, temporal, and host origins have been collected and stored. These strains show wide diversity in the degree of viremia, neurovirulence, and severity of symptoms induced in both avian and mammalian hosts, with some correlation between geographic origin and these phenotypes (Bowen et al., 1980; Monath et al., 1980), although isolates made after 1977 were not included in these analyses. Oligonucleotide fingerprinting (Trent et al., 1980) and phylogenetic analysis of the envelope gene (Kramer

\* Corresponding author. Fax: +44 1223 764667.  
E-mail address: [gb354@cam.ac.uk](mailto:gb354@cam.ac.uk) (G.J. Baillie).

<sup>1</sup> Present address: Cambridge Infectious Diseases Consortium, Department of Veterinary Medicine, University of Cambridge, Madingley Road, Cambridge CB3 0ES, UK.

and Chandler, 2001) show that, in general, geographically proximal strains are also genetically similar.

The genome of SLEV is, like that of all members of the genus *Flavivirus*, a non-segmented, single-stranded, positive-sense RNA, approximately 11 kb in length. It encodes a single 3429 amino acid polyprotein, which is co- and post-translationally cleaved by viral and cellular proteases into three structural (C, PrM/M, E) and seven non-structural (NS1, NS2A, NS2B, NS3 [protease/helicase], NS4A, NS4B, NS5 [polymerase]) proteins (Lindenbach and Rice, 2001) (Fig. 1).

The E protein is the primary determinant of cell receptor binding and immune recognition (Lindenbach and Rice, 2001). In some strains of SLEV, the E protein is glycosylated at an asparagine at position 154 of the protein, when it is part of a Asn-X-Ser/Thr (NXS/T) tripeptide, whereas in other strains this site is lacking or is variably glycosylated (Vorndam et al., 1993). Glycosylation does not appear to affect formation or release of viral particles, although virions lacking glycosylated E infect SW-13 (human adeno-

carcinoma) and CRE (hamster) cells significantly less efficiently than those with glycosylated E (Vorndam et al., 1993). No correlation is observed between glycosylation of E and virulence (Vorndam et al., 1993). In other flaviviruses, the glycosylation state of the envelope protein is also heterogeneous (Shirato et al., 2004), can be altered by passage of viruses in cell culture (Chambers et al., 1998; Halevy et al., 1994; Shirato et al., 2004), affects virus replication *in vitro* (Li et al., 2006), and can influence viremia, neuroinvasiveness, and neurovirulence *in vivo* (Beasley et al., 2005; Chambers et al., 1998; Halevy et al., 1994; Hanna et al., 2005; Li et al., 2006; Scherret et al., 2001; Shirato et al., 2004).

Recent studies have highlighted the potential roles of recombination and selection in the ongoing evolution of flaviviruses. Twiddy and Holmes (2003) report evidence of recombination in dengue virus, Japanese encephalitis virus, and SLEV. Their results for SLEV, based on previously published (Kramer and Chandler, 2001) envelope gene sequences from 61 isolates, suggest that the recombi-

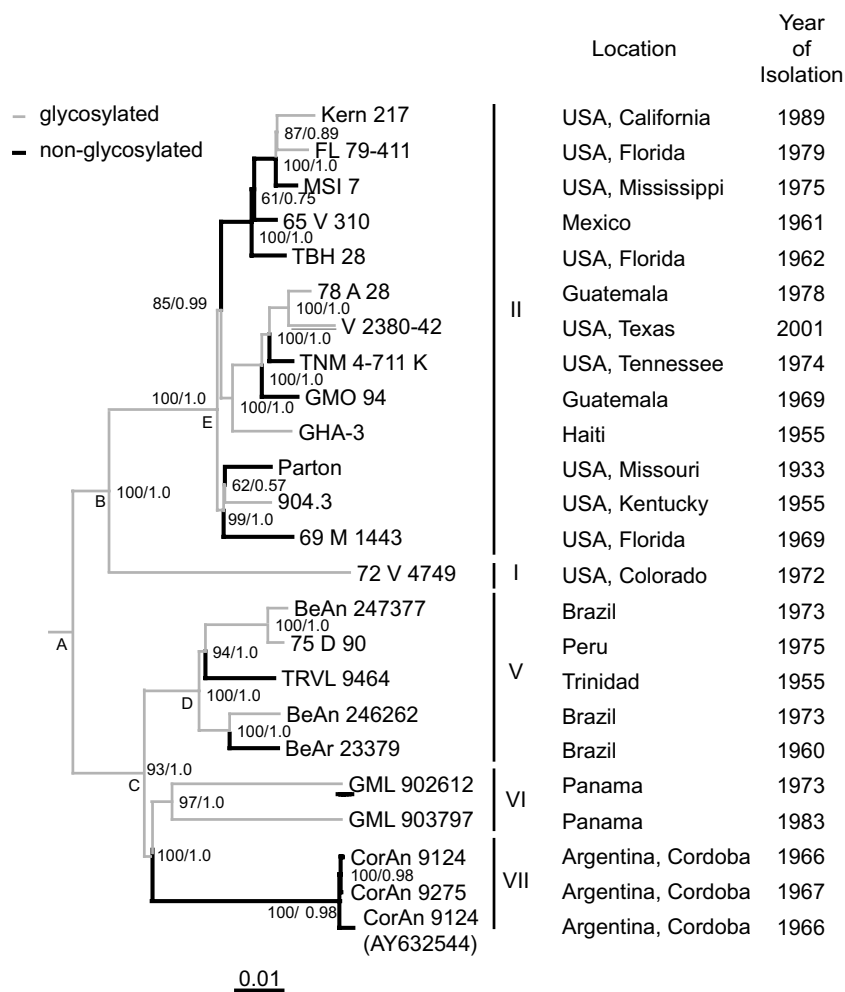


Fig. 1. Maximum likelihood tree based on open reading frame sequences. The tree is mid-point rooted for purposes of illustration. Glycosylation state at position 154–156 of E protein, as estimated by ML analysis, is indicated. Numbers at nodes indicate bootstrap support in percentage of 1000 pseudoreplicates for the ML analysis and posterior probabilities for the Bayesian inference of phylogeny. The scale bar indicates the number of substitutions per site.

nant strain from Guatemala (GMO 94) arose by recombination between an isolate from Argentina (CorAn9124/Arg66) and another closely related to that from Memphis, Tennessee (TNM4-711K) (Twiddy and Holmes, 2003). Flaviviruses, particularly vector-borne flaviviruses, appear to be predominantly constrained by purifying selection (Holmes, 2003; Woelk and Holmes, 2002), but evidence is beginning to emerge that some flavivirus proteins are subject to positive selection. Twiddy et al. (2002) used a maximum likelihood (ML) approach to detect adaptive evolution in the four serotypes of dengue virus. Positive selection appeared to be acting primarily on the envelope gene of dengue virus, on amino acids involved in immune recognition, cell receptor binding, and membrane fusion (Twiddy et al., 2002). Recently, Brault et al. (2007) identified a single amino acid substitution in the NS3 protein which confers high virulence on otherwise low virulence strains of *West Nile virus* (WNV).

To date, most studies on the evolution of flaviviruses have focused on one or a few genes, most often the envelope gene. Here, we employ a whole genome approach to the investigation of the roles of selection and recombination in the evolution of SLEV. We have sequenced almost the entire open reading frame (ORF) from 23 strains of SLEV, and find no evidence for recombination in these strains, but some evidence for positive selection, particularly on the envelope gene.

## 2. Materials and methods

### 2.1. Virus strains and growth

We sequenced SLEV strains from throughout the previously published phylogenetic tree based on E gene sequences (Kramer and Chandler, 2001) (Table 1). Strains were obtained from World Reference Center for Emerging Viruses and Arboviruses at the University of Texas Medical Branch (UTMB), Galveston, Texas, United States, and Center for Disease Control (CDC). They were passed once on African green monkey kidney (Vero) cells, one strain at a time, and RNA was isolated from cells using Qiagen RNeasy spin columns (Qiagen Inc., Valencia, CA) and eluted in sterile distilled water at The Arbovirus Laboratories, Wadsworth Center, New York State Department of Health.

### 2.2. RT-PCR and sequence analysis

RNA was reverse-transcribed using BD Sprint PowerScript PrePrimed (Clontech, Mountain View, CA), and PCR-amplified in overlapping fragments using Platinum Taq High Fidelity polymerase (Invitrogen) and primers (Table 2) designed from the two previously known near-full-length SLEV sequences (Billoir et al., 2000; Kuno and Chang, 2005). To enable amplification in the presence of primer-template mismatches, a touchdown PCR protocol (Don et al., 1991) was employed: initial denaturation/

Taq activation of 94 °C/2 min; 35 cycles of 94 °C for 15 s, 52–0.2 °C per cycle for 15 s, 68 °C for 2 min; final extension of 68 °C for 5 min. PCR products were purified using AMPure reagent (Agencourt, Beverly, MA), and were sequenced directly using combinations of terminal and internal primers (Table 2), BigDye Terminator v3.1 cycle sequencing chemistry (Applied Biosystems, Foster City, CA), and an Applied Biosystems 3730xl capillary sequencer.

Sequences were assembled using Sequencher 4.6 (Gene Codes Corporation, Ann Arbor, MI), and are deposited in GenBank under Accession Numbers EF158048–EF158070. The sequences of the open reading frames (ORFs) were manually aligned. The alignment is available from the authors upon request.

### 2.3. Recombination detection

Recombination analysis was performed using two software packages. We employed the RDP, GENECONV, Bootscan, MaxChi, and Chimaera methods within the RDP2 v Beta 0.8 program (Martin et al., 2005a). For each method, a range of window and step sizes was tested. We also used the GARD recombination-detection method (Kosakovsky-Pond et al., 2006) within the HyPhy package (Kosakovsky-Pond et al., 2005) or on the Datamonkey server (Kosakovsky-Pond and Frost, 2005a). Comparison of phylogenies derived from either side of putative breakpoints was performed using the Kishino–Hasegawa test (Kishino and Hasegawa, 1989), as implemented in HyPhy.

### 2.4. Phylogenetic analysis

Phylogenetic relationships were evaluated using maximum likelihood (ML) and Bayesian methods. ML trees were generated with PHYML v2.4.4 (Guindon and Gascuel, 2003), following testing for appropriate models using Modeltest (Posada and Crandall, 1998) and PAUP\* v4b10 (Swofford, 2003), using the GTR substitution model with I (proportion of invariable sites) and  $\Gamma_4$  (gamma-distributed rates in four categories) (Lanave et al., 1984; Rodriguez et al., 1990; Yang, 1994). Bayesian inference of phylogeny was performed in MrBayes v3.1.2 (Altekar et al., 2004; Ronquist and Huelsenbeck, 2003). The analysis was run twice with four chains for 10 million MCMC steps discarding 10% as burn-in using the GTR +  $\Gamma_4$  substitution model. We used relative Bayes factors (RBF) to select the best fit partitioning scenario given the data by dividing  $2\ln\text{BF}$  by the difference in free parameters between two competing scenarios (Newton and Raftery, 1994). We compared the following partitioning strategies: (1) no partitioning, (2) by 1st, 2nd, and 3rd codon position separately, (3) by 1st and 2nd codon position together and 3rd position separately, and (4) by gene; in all cases substitution models were allowed to evolve independently (“unlink” command in MrBayes). The harmonic mean of the sampled likelihoods was calculated in MrBayes and

Table 1  
St. Louis encephalitis strains used in this study

Strain	Country	State/ Province	City/Community	Host (common)	Host (Latin)	Year	Mouse virulence <sup>a</sup>	Avian adult virulence <sup>b</sup>	Avian nestling virulence <sup>b</sup>	Passage history <sup>c</sup>
COR AN 9124	Argentina	Cordoba	Unknown	Rodent	<i>Calomys musculus</i>	1966	Low	n.d.	Low	SM7, Vero
COR AN 9275	Argentina	Cordoba	Unknown	Rodent	<i>Mus musculus</i>	1967	Low	n.d.	n.d.	SM7, Vero
BE AN 246262	Brazil	Para	Itaituba	Opossum	<i>Didelphis marsupialis</i>	1973	High	High	n.d.	?+DE1
BE AN 247377	Brazil	Para	Itaituba	Bird	<i>Hylophilax poecilonota</i>	1973	n.d.	n.d.	n.d.	
BE AR 23379	Brazil	Para	Road Belem-Brasilia, Km 94	Mosquito	<i>Sabethes belisarioi</i>	1960	Low	Intermediate	Low	?+SM2
78 A 28	Guatemala	Unknown	La Avellana	Mosquito	Unknown	1978	n.d.	n.d.	n.d.	
GMO 94	Guatemala	Unknown	Unknown	Mosquito	<i>Culex nigripalpus</i>	1969	High	High	High	SM2
GHA-3	Haiti	Unknown	Unknown	Bird (green heron)	<i>Butorides virescens</i>	1955	High	n.d.	n.d.	SM3, Vero
65 V 310	Mexico	Unknown		Bird (green heron)	Unknown	1961	Intermediate	n.d.	Intermediate	SM5, Vero
GML 902612	Panama	Panama	Bayano	Mosquito (pool)	<i>Haemagogus equinus</i>	1973	n.d.	n.d.	n.d.	SM1, Vero
GML 903797	Panama	Panama	Bayano	Sentinel chicken	Unknown	1983	n.d.	n.d.	n.d.	Vero 4
75 D 90	Peru	Unknown	Muniches Airport Road, Yuri?	Mosquito	Unknown	1975	High	n.d.	Intermediate	SM3, Vero
TRVL 9464	Trinidad	Unknown	Melajo Forest	Mosquito	<i>Psorophora ferox</i>	1955	Intermediate	Low	Intermediate	SM10
Kern 217	USA	California	Kern Co.	Mosquito	<i>Culex tarsalis</i>	1989	n.d.	n.d.	n.d.	Vero 2
72 V 4749	USA	Colorado	Washington Co.	Mosquito	<i>Culex tarsalis</i>	1972	High	Intermediate	n.d.	SM1
69 M 1143	USA	Florida		Raccoon	<i>Procyon lotor</i>	1969	Low	n.d.	Low	SM4, Vero
FL 79-411	USA	Florida	Lee County	Mosquito	<i>Culex nigripalpus</i>	1979	n.d.	n.d.	n.d.	C6/36 1
TBH 28	USA	Florida	Tampa Bay	Human	<i>Homo sapiens</i>	1962	n.d.	n.d.	n.d.	Unknown
904.3	USA	Kentucky	Calvert City	Bird (Flicker)	<i>Colaptes auratus</i>	1955	High	High	n.d.	SM3, Vero
MSI 7	USA	Mississippi	Unknown	Bird (sparrow)	<i>Passer domesticus</i>	1975	High	Intermediate	High	SM2
Parton	USA	Missouri	St. Louis City and County	Human (brain)	<i>Homo sapiens</i>	1933	n.d.	n.d.	n.d.	Unknown
TNM 4-711 K	USA	Tennessee	Memphis	Mosquito	<i>Culex pipiens</i>	1974	High	n.d.	High	SM1, Vero
V 2380-42	USA	Texas	Houston	Mosquito (pool)	<i>Culex quinquefasciatus</i>	2001	n.d.	n.d.	n.d.	

<sup>a</sup> Mouse virulence data is from Monath et al. (1980).

<sup>b</sup> Avian virulence data is from Bowen et al. (1980).

<sup>c</sup> Passage history is indicated by cell-line or host followed by number of passages. SM, suckling mouse; Vero, Vero (African green monkey kidney) cell culture; DE, duck embryo cell culture; C6/36, C6/36 (mosquito) cell culture.

Table 2  
Primers used in this study

Primer name	Primer sequence (5' → 3')	Use	Primer strand	Primer position (5' → 3')	Primer gene location
SLEV_16F	GTGAGCGGAGAGGAAACAGA	PCR, sequencing	Sense	16–35	5' UTR
SLEV_357F	TGACCTCGGATCCATGCTAGA	Sequencing	Sense	357–377	Capsid
SLEV_451F	CTGGCGAGTTCATTGCAGTT	Sequencing	Sense	451–470	Signal peptide-Pr
SLEV_626R	ACTGGGCACAGGTATGTGATGT	Sequencing	Antisense	626–605	Pr
SLEV_986F	ACAGGGACTTTGTGCGAGGGA	Sequencing	Sense	986–1005	E
SLEV_1061F	TGATGGCACCAGAGAAACCAA	Sequencing	Sense	1061–1081	E
SLEV_1317R	CTTAGCGCATGTGTCAATGCT	Sequencing	Antisense	1317–1297	E
SLEV_1500F	CATGGGCGAGTATGGAACAGT	PCR, sequencing	Sense	1500–1520	E
SLEV_1505F	GCGAGTATGGAACAGTTACCA	PCR, sequencing	Sense	1505–1525	E
SLEV_1653R	TGTGGCAGGGCTCGTCCA	PCR, sequencing	Antisense	1653–1636	E
SLEV_1876R	ACATGCCATATGTCTGTTCCCTT	PCR, sequencing	Antisense	1876–1855	E
SLEV_2003F	TGGACTTGACACCGTTGGAA	Sequencing	Sense	2003–2023	E
SLEV_2010F	GACACCGGTTGGAAGATTGGT	Sequencing	Sense	2010–2030	E
SLEV_2251R	ATCCAAAGTCCCACGCTGTGT	Sequencing	Antisense	2251–2231	E
SLEV_2511F	TGGAGGAGGCATCTTCGTGTA	PCR, sequencing	Sense	2511–2531	NS1
SLEV_2684R	CTCTGTATGTTCTCCACATCAA	PCR, sequencing	Antisense	2684–2662	NS1
SLEV_2745R	ATCTTCTTGCACTACAACACTGACAA	Sequencing	Antisense	2745–2722	NS1
SLEV_3033F	AGGAGACAGAGCTGTTTACA	PCR, sequencing	Sense	3033–3052	NS1
SLEV_3314R	GTTCTGGACAGTAGTCAAAGT	PCR, sequencing	Antisense	3314–3293	NS1
SLEV_3498F	GCTTGTGAAATCTCGAGTGACA	PCR, sequencing	Sense	3498–3519	NS1
SLEV_3711R	AGTTCCAACAAGGACCAGGTA	PCR, sequencing	Antisense	3711–3691	NS2A
SLEV_3988F	TGTGCTTTGACTCCAGGAATGA	Sequencing	Sense	3988–4009	NS2A
SLEV_4295R	ACCACCATTGAGGTTTCTCTCAA	Sequencing	Antisense	4295–4274	NS2B
SLEV_4473F	CCCTGGAGCTCCAGTGCA	PCR, sequencing	Sense	4473–4490	NS2B
SLEV_4516F	CTGGGCTTGAGTGCACGAT	PCR, sequencing	Sense	4516–4534	NS2B
SLEV_4755R	AGTTGCTGCCAATTGTGT	PCR, sequencing	Antisense	4755–4736	NS3
SLEV_4923F	CCAGACAACACCAGGAGTGTT	PCR, sequencing	Sense	4923–4943	NS3
SLEV_5155R	TCTTCCGGAGCATTTCTCAT	Sequencing	Antisense	5155–5135	NS3
SLEV_5403R	TTTCTGTGTGAGCGTGGCAT	PCR, sequencing	Antisense	5403–5384	NS3
SLEV_5631F	GAGCACTGGGTATGAGTGGAT	PCR, sequencing	Sense	5631–5651	NS3
SLEV_6007R	CATCTCCAATCTCGAAGGAT	PCR, sequencing	Antisense	6007–5987	NS3
SLEV_6048R	ATCATGGTCTGCTCATTTGGT	PCR, sequencing	Antisense	6048–6028	NS3
SLEV_6192F	CTTTGTTGAGCTGATGCGCAA	PCR, sequencing	Sense	6192–6212	NS3
SLEV_6209F	GCAACGGTGACCTCCCAGT	Sequencing	Sense	6209–6227	NS3
SLEV_6490R	TTGGCATTCTACCATGACCT	PCR, sequencing	Antisense	6490–6470	NS4A
SLEV_6539R	AGGAGATACAGTGTGTCTGCA	Sequencing	Antisense	6539–6519	NS4A
SLEV_6660F	TCTCATGAACAGGAAAGGAT	PCR, sequencing	Sense	6660–6681	NS4A
SLEV_7027R	CGGTGGCTGGTTTGATGCA	PCR, sequencing	Antisense	7027–7008	NS4B
SLEV_7058R	ACCATTGTTGCTGCCGCATA	PCR, sequencing	Antisense	7058–7039	NS4B
SLEV_7285F	CCAGGCTGGCAGGCTGAA	PCR, sequencing	Sense	7285–7302	NS4B
SLEV_7530R	CGTCACAAGGGCTGCTGAT	Sequencing	Antisense	7530–7512	NS4B
SLEV_7530R	CGTCACAAGGGCTGCTGA	Sequencing	Antisense	7530–7513	NS4B
SLEV_7710F	GTC AAGACTGAATCAGCTGACA	PCR, sequencing	Sense	7710–7731	NS5
SLEV_8035R	TCCATCGATAGCTTTGCATGA	PCR, sequencing	Antisense	8035–8015	NS5
SLEV_8311F	AACTCGACACATGAGATGTACT	PCR, sequencing	Sense	8311–8332	NS5
SLEV_8534R	TCCCTCAATCTCCGTATCCTT	PCR, sequencing	Antisense	8534–8514	NS5
SLEV_8572R	GGTTGTTGTTGTCATACGTCCA	PCR, sequencing	Antisense	8572–8551	NS5
SLEV_8768F	CCGAACCACCACTAGGAGT	PCR, sequencing	Sense	8768–8786	NS5
SLEV_9079R	CCTTCCCAAATTCTCCAGTCT	PCR, sequencing	Antisense	9079–9059	NS5
SLEV_9261F	AAAGATGTACGCAGATGACACA	PCR, sequencing	Sense	9261–9282	NS5
SLEV_9527R	GCCAGGTTGGTGAAGGTGTT	PCR, sequencing	Antisense	9527–9508	NS5
SLEV_9853F	ACCATCGTGGTGCCATGCA	Sequencing	Sense	9853–9871	NS5
SLEV_10392R	CTACAGCACTCCTCTACCA	PCR, sequencing	Antisense	10392–10373	NS5

we slightly modified the criterion of [Kass and Raftery \(1995\)](#) for scenario selection ( $2 \ln BF > 30$  suggests strong support). Convergence was assessed by examining the stationarity of the  $\ln$ -likelihood and the effective sample size ( $ESS > 200$ ) of each estimated parameter in Tracer v1.4 ([Rambaut and Drummond, 2005](#)), along with the inspection of the standard deviation of split frequencies ( $< 0.003$ ) and successful state swap frequencies (in the range  $[0.25, 0.85]$ ).

The heating parameter was fine-tuned so as to lead to better mixing in the range  $[0.005, 0.05]$ .

### 2.5. Selection detection

Global and branch-specific non-synonymous to synonymous rate ratios were estimated in HyPhy, using the ML and Bayesian phylogeny, and the MG94xREV codon

model. Tests for positive selection were performed using four methods—single-likelihood ancestor counting (SLAC), fixed effects likelihood (FEL), internal FEL (IFEL), and random effects likelihood (REL) (Kosakovsky-Pond and Frost, 2005b)—as implemented in HyPhy/Datamonkey. All of these methods estimate selection in a phylogenetic context. The SLAC method infers the ancestral states throughout a phylogeny, calculates the expected numbers of synonymous and non-synonymous substitutions at each site, and compares these with the observed numbers of synonymous and non-synonymous changes at those sites. The FEL method also estimates selection on a site-by-site basis, by fitting synonymous and non-synonymous substitution rates at each site under two models, one where the synonymous and non-synonymous rates are independent, and the other where the synonymous and non-synonymous rates are constrained to be equal. A LRT is then used to determine whether the independent model is significantly better than the constrained model. Finally, the REL method employs a codon substitution model to determine independent general discrete distributions (GDDs) for both the global synonymous and non-synonymous substitution rates, and then uses Empirical Bayes analysis to determine whether each site is under selection using the GDDs as priors.

To test the hypothesis that selection was acting on terminal branches of the phylogeny, we used (1) the IFEL method (which is similar to the FEL method but only measures substitutions on internal branches) on the Datamonkey server, and (2) the TestBranchDNDS method in HyPhy, with the MG94xREV model of codon substitution, general discrete distributions of synonymous and non-synonymous substitutions with 3 categories each, and the polarity/charge/hydrophobicity (PCH) amino acid class model.

### 2.6. Divergence times, substitution rates, and population size changes

We made use of the heterochronous nature of our sequences—specifically, the year of isolation for each of the 23 analyzed strains—to estimate substitution rates and divergence times in SLEV. We investigated the scenario that showed the lowest diminishing returns following the RBF analysis: partitions with 1st and 2nd codon positions together and the 3rd position by itself with unlinked substitution models. The GTR +  $\Gamma_4$  substitution model was applied along with a strict clock. We estimated the time to the most recent common ancestor (TMRCA) for all phylogenetic clades, as well as substitution rates per partition and overall. We applied different demographic scenarios (constant population size, exponential population growth, and Bayesian skyline plot) and compared their relative fit to the data using Bayes factors. The harmonic mean of the sampled likelihoods from the posterior, an estimator of the marginal likelihood, was calculated following Suchard et al. (2001). Changes in effective population

size over time were examined using the Bayesian skyline plot (Drummond et al., 2005). We used default priors in these analyses and they were carried out in BEAST v1.4.6 (Drummond and Rambaut, 2006) for 30 million generations sampling every 1000 steps and discarding the first 10% as burn-in (50% for demographic model comparison), while convergence was assessed in Tracer.

## 3. Results

As part of an effort to understand the roles that recombination and selection play in flavivirus evolution, we sequenced nucleotides 99–10334 (corresponding to codons 1–3412) of 23 strains of SLEV from wide geographic and temporal origin.

### 3.1. Limited evidence for recombination in SLEV strains

Numerous computer programs are available for the detection of recombination in nucleic acid sequences. Evaluation of these programs on simulated and empirical data sets showed that all methods are sensitive to factors such as the degree of sequence divergence, the amount of recombination, and the nucleotide substitution rate distribution, that some methods perform better than others under different conditions, and that multiple methods should be used to interrogate sequences for potential recombination (Posada, 2002; Posada and Crandall, 2001; Posada et al., 2002). Using four of the methods in RDP2—RDP (Martin and Rybicki, 2000), GENECONV (Padidam et al., 1999), Bootscan (Martin et al., 2005b), and MaxChi (Smith, 1992)—with a range of window and step sizes, analysis of the sequences for our 23 strains showed no evidence for recombination. A fifth method, Chimaera (Posada and Crandall, 2001), suggested two potential recombination events (data not shown). However, the multiple-comparison-corrected probabilities of these events were borderline significant ( $p = 0.038$  and  $p = 0.028$ ), and were not supported by the other methods, thus we consider these putative recombination events to be questionable. GARD detected a putative recombination breakpoint at or near nucleotide 6291, with a change in Akaike Information Criterion ( $\Delta c$ -AIC) of 11.85, which suggests poor support for the non-recombination model compared to the recombination model. The single breakpoint (SBP) algorithm in HyPhy identified a potential recombination breakpoint at nucleotide 6176, with a  $\Delta c$ -AIC of 35.0194, although the Bayesian Information Criterion supported a non-recombination model. Bayesian phylogenies constructed using the sequences either side of these supposed breakpoints had only minor differences in topology. Although the tree derived from the sequences 5' to the SBP fitted the 5' sequences significantly better ( $p = 0.0006$ ) than that derived from the tree made with the sequences 3' to the breakpoint, as assessed by a Kishino–Hasegawa test in HyPhy, neither tree was a significantly better fit than the other for the 3' sequences ( $p = 0.065$ ). In addition, there

were few differences in the number and location of sites identified as being under positive or negative selection when the full sequences were used, compared to when the sequences on either side of the putative breakpoints were analyzed separately (see below).

Twiddy and Holmes (2003) reported a likely recombinant strain of SLEV—GMO 94—based on envelope sequences published by Kramer and Chandler (2001). Our analyses revealed no evidence of recombination in strain GMO 94. The likely reason for this discrepancy is differences between our sequence for strain GMO 94 and that reported by Kramer and Chandler (2001): there are 32 nucleotide differences between the sequences (Table 3), all but one of which are 5' of the recombination breakpoint reported by Twiddy and Holmes (2003). This difference is also evident in ML and Bayesian trees constructed using sequences either side of the putative breakpoint identified by Twiddy and Holmes (2003) (data not shown). Our sequence for strain GMO 94 falls into a group with strains V 2380-42 (Texas, USA, 2001), 78 A 28 (Guatemala, 1978), and TNM 4-711 K (Tennessee, USA, 1974), in the trees based on sequences both 5' and 3' of the putative breakpoint. The Kramer and Chandler (2001) sequence from 3' of the breakpoint is also located in this group, but falls into the group with CorAn9124 (Cordoba, Argentina, 1966) and CorAn9275 (Cordoba, Argentina, 1967) when the sequence 5' of the breakpoint is used, as reported by Twiddy and Holmes (2003) (data not shown).

Table 3  
Sequence differences between strains sequenced in this study and in that of Kramer and Chandler (2001)

Strain	Number of nucleotide differences	Number of amino acid differences	Amino acid changes <sup>b</sup>
BeAr 23379	10	4	29, 54, <b>154</b> , <b>156</b>
904.3	0	0	
MSI 7 <sup>a</sup>	13	4	58, <b>156</b> , 316, 508
GMO 94	32	3	29, 66, <b>154</b>
BeAn 246262	0	0	
75 D 90	0	0	
TBH 28	2	2	58, <b>156</b>
TRVL 9464	2	2	94, 114
Kern 217	4	2	132, 133
65 V 310	3	3	58, <b>154</b> , <b>156</b>
GML 903797	0	0	
CorAn 9124	1	1	206
69 M 1143	0	0	
FL 79-411	0	0	
GML 902612	1	1	451
TNM 4-711 K	2	1	<b>156</b>
GHA-3	2	2	<b>156</b> , 389
CorAn 9275	0	0	
72 V 4749	0	0	
Parton	37	7	24, 70, 130, 179, 276, 400, 439

<sup>a</sup> Strain MSI 7 sequence from study by Trent et al. (1987).

<sup>b</sup> Numbering is relative to the first codon of the *env* gene. Changes in N-linked glycosylation site are in bold.

### 3.2. Phylogeny of SLEV strains using ORF sequences

Previous phylogenetic studies of SLEV were based on sequences of the envelope genes and parts of the flanking M and NS1 genes (Kramer and Chandler, 2001; Twiddy and Holmes, 2003). We constructed ML and Bayesian trees using our extended strain sequences (Fig. 1). Bayesian phylogenies were constructed using unpartitioned sequences, as well as sequences partitioned by gene or by codon position. The topologies of our ML and all Bayesian trees were identical (Fig. 1), and were very similar to those reported by Kramer and Chandler (2001) and Twiddy and Holmes (2003). As in those studies, the SLEV strains in our tree are predominantly separated along geographic origins. Basal South American and Central American groups give rise to a separate, mostly North American, group, which contains some strains from Mexico, Guatemala, and Haiti.

Despite the overall similarity of our trees and those based on envelope gene sequences, the phylogenetic positions of some strains were altered. Strain GMO 94, which was basal to the group II strains in the phylogeny of Kramer and Chandler (2001), groups with strains TNM 4-711 K, V 2380-42, and 78 A 28 in our tree. This change is almost certainly due to the sequence differences described above. The 1973 Panama strain Gml902612, which groups with the 1983 Panama strain Gml903797 in our tree and in that of Twiddy and Holmes (2003), was placed in a separate group (group IV) in the study of Kramer and Chandler (2001). Finally, strains 904.3 (Kentucky, USA) and 69 M 1443 (Florida, USA) are found in a well supported group with Parton (Missouri, USA) in our phylogeny, whereas they are in a group with a Brazilian strain (SpAn9398) and several Texan strains, none of which were included in our study, in both the Kramer and Chandler (2001) and Twiddy and Holmes (2003) phylogenies (data not shown). Several other strains—BeAr23379, MSI 7, and Parton—possessed many differences from previously published sequences (Table 3), but were only slightly changed or were unchanged in phylogenetic position in our study.

### 3.3. Predominantly purifying selection in SLEV strains

The global  $d_N/d_S$  ratio over the entire SLEV open reading frame was estimated to be 0.040, which suggests predominantly purifying selection. Estimates for separate genes ranged from 0 for the 2K protein to 0.077 for Pr (Table 4).

We used the SLAC, FEL, IFEL, and REL methods in HyPhy and on the Datamonkey web server to analyze patterns of positive and negative selection in each gene of SLEV using our strain sequences.

The majority of codons in the SLEV genome appear to be under neither strong negative selection nor strong positive selection. Of the 3412 codons we analyzed, the maximum number of codons for which there was strong evidence for negative selection was 738, using the FEL method (Table 4), and the maximum number of positively

Table 4  
Evidence for negative and positive selection using SLAC, FEL, REL, and IFEL methods

Protein	Amino acid residues	Length	Overall $d_N/d_S$	SLAC <sup>a</sup>		FEL <sup>a</sup>		REL <sup>b</sup>		SLAC + FEL + REL		IFEL <sup>b</sup>	
				–ve	+ve	–ve	+ve	–ve	+ve	–ve	+ve	–ve	+ve
Capsid	1–98	98	0.057875	3	0	10	0	2	0	1	0	1	0
gap	99–121	23	0.0683512	2	0	6	0	1	0	0	0	1	0
Pr	122–213	92	0.0767125	9	0	19	0	4	1	4	0	4	0
M	214–288	75	0.0217723	6	0	15	0	2	0	1	0	2	0
E	289–789	501	0.0308009	64	1	135	1	41	3	39	1	47	0
NS1	790–1141	352	0.0573088	24	0	53	0	15	2	15	0	12	0
NS2A	1142–1368	227	0.0351804	24	0	54	0	17	0	15	0	17	0
NS2B	1369–1499	131	0.0392701	10	0	28	0	4	0	4	0	8	0
NS3	1500–2117	618	0.0347248	56	0	149	0	25	3	24	0	32	1
NS4A	2118–2243	126	0.0225293	5	0	25	0	4	0	2	0	5	0
2K	2244–2266	23	0	1	0	3	0	1	0	1	0	1	0
NS4B	2267–2524	258	0.0622813	20	0	56	0	13	0	10	0	16	0
NS5	2525–3412	888	0.0366513	70	0	175	1	43	3	38	0	45	1
Totals			0.0399773	294	1	728	2	172	12	154	1	191	2

<sup>a</sup> Number of sites where  $p < 0.05$ .

<sup>b</sup> Number of sites where Bayes factor  $> 50$ .

selected codons was 16, using the REL method (Table 4). For the SLAC and FEL methods, the number and locations of negatively and positively selected sites were almost identical regardless of whether the analyses were performed on full-length sequences, or whether sequences from either side of the putative recombination breakpoint at nt 6176 were analyzed separately. For the REL method, 105 additional sites were identified as being under negative selection using the partitioned sequences as opposed to the full-length sequences, presumably because better fits of the discrete gamma distributions of synonymous and non-synonymous substitution rates are achieved with shorter sequences (data not shown). In summary, 155 sites were identified as being under negative selection using all three of the SLAC, FEL, and REL methods, and both the full-length and partitioned data sets.

In contrast, only one codon—site 444 of the polyprotein (site 156 of the E protein)—was identified as being under strong positive selection using all three of the SLAC ( $p = 0.040$ ), FEL ( $p = 0.0024$ ), and REL (posterior probability = 0.936284, logBF = 3.58) methods (Table 4).

We also subjected the larger data set of envelope gene sequences of (Kramer and Chandler, 2001) to SLAC, FEL, and REL analysis. The FEL ( $p = 0.022$ ) and REL (posterior probability = 0.999981, logBF = 16.3269) methods suggested strong positive selection acting on codon 156, whereas the support for selection was lower using the more conservative SLAC method ( $p = 0.11$ ).

### 3.4. Selection operating at site of N-linked glycosylation in envelope protein

Codons 154–156 of the envelope gene are a site of potential N-linked glycosylation in St. Louis encephalitis virus and other flaviviruses.

Variation in the glycosylation state of E proteins of different SLEV strains was previously observed by Vorndam

et al. (1993). In that study, some strains expressed both glycosylated and non-glycosylated variants of the E protein in culture. To determine whether the strains we sequenced might be expressing both glycosylated and non-glycosylated forms of E, we examined the electropherograms of these sequences for the presence of secondary peaks that might confer an altered glycosylation state. For ten of the strains, secondary peaks were present in the N-glycosylation site that would change glycosylation. For strain GMO, a secondary peak was located at the third position of the first codon, which would change the glycosylation-refractory lysine (K) at this site to a glycosylation-favorable asparagine (N). For seven strains, the secondary peak was located at the second position of the third codon, causing this codon to vary between TTT (encoding glycosylation-favorable serine, S) and TCT (glycosylation-refractory phenylalanine, F). For two strains (65 V 310, TNM 4-711 K), there were variants at both codons (data not shown).

Of the 14 strains that contained non-synonymous substitutions between this study and that of Kramer and Chandler (2001), seven differed at the glycosylation site 154–156 (Table 3).

### 3.5. Selection may be an artifact of in vitro selection

We mapped the putative glycosylation states of the 154–156 sites of our sequences (as determined by the deduced amino acid sequence) onto the tree in Fig. 1. Many of the substitutions at the site of potential glycosylation appear to occur in the terminal branches of the tree. To estimate the significance of this possible bias, we employed two methods implemented in HyPhy. First, we used the IFEL method to estimate positive and negative selection acting only on internal branches of the SLEV phylogeny. Codon 156 of the E protein, which was indicated as being under positive selection by the SLAC, FEL, and REL

methods, was not identified as being under positive selection by the IFEL method (Table 4). Second, we used the TestBranchDNDS method in HyPhy to test whether overall positive and negative selection were significantly different on the terminal branches than they were on the rest of the tree. Selection acting on the terminal branches was significantly different ( $p < 0.001$ ) than that acting on the rest of the tree.

### 3.6. Substitution rates, divergence times, and population size changes

Estimates of divergence times of SLEV lineages suggested that the deepest divergence of known SLEV strains was at 283 ybp. The clade containing North American strains diverged about 116 ybp (clade II, mean = 116.22 ybp, 95% HPD: 102.57–130.65 ybp), indicating a colonization of N. America in the late 19th century (Table 5). The strains from Central and South America showed a TMRCA of 212 ybp suggesting a recent northwards colonization. Overall substitution rate was  $2.17 \times 10^{-4}$  substitutions/site/year (95% HPD:  $1.77\text{--}2.57 \times 10^{-4}$ ) with 1st + 2nd codon positions to evolve at  $5.44 \times 10^{-5}$  substitutions/site/year (95% HPD:  $4.05\text{--}7.02 \times 10^{-5}$ ) and 3rd codon positions at  $5.41 \times 10^{-4}$  substitutions/site/year (95% HPD:  $4.34\text{--}6.54 \times 10^{-4}$ ). Using a conservative burn-in of 50% for the Bayes factors calculation, we found that the BSP fit the data better (2lnBF was 103.052 and 105.99 when the BSP was compared to the constant size model and the exponential growth model, respectively). When the constant size and exponential growth models were contrasted, none was a better fit for the data (2lnBF = 2.94). The BSP showed a fourfold increase in the median effective number of infections (median at present  $N_e = 759.72$ ; 95% HPD: 252.25–4673.75) in the past 280 years, but given the overlap of 95% HPDs of present and historical  $N_e$ , it would be cautious not to dismiss an almost constant population size scenario for SLEV (Fig. 2). We posit that the better fit of the BSP over the other two demographic models is owed to its ability to capture a complex yet mild dynamic that the other two models are too “strict” to capture. This complexity is most likely owed to an increase in the effective number of infections that occurred a bit over 100 years ago, corresponding to the SLEV colonization of North America.

Table 5  
Divergence times of major SLEV lineages

Node <sup>a</sup>	Group <sup>a</sup>	Divergence date years BP (95% HPD)
A	All SLEV	283.64 (238.36–333.23)
B	I + II	245.89 (206.59–289.12)
C	V + VI + VII	212.27 (179.28–249.13)
D	V	126.49 (108.55–145.08)
E	II	116.22 (102.57–130.65)

<sup>a</sup> As indicated in Fig. 1.

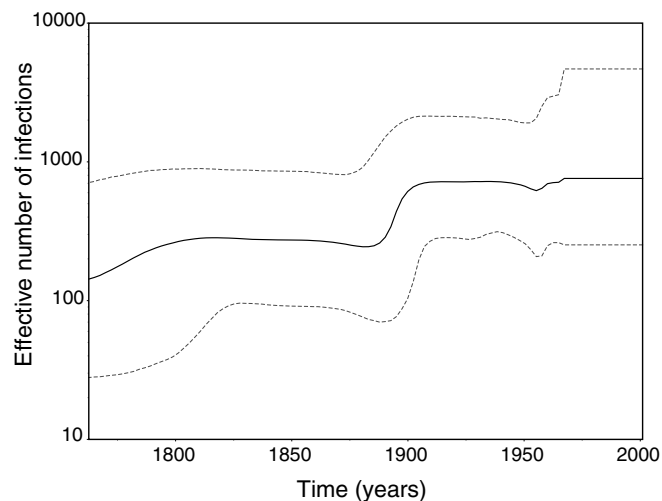


Fig. 2. Bayesian skyline plot of St. Louis encephalitis strains in the American continent sampled between 1933 and 2001. The bold line represents the median estimate of the effective number of infections through time. The dashed lines indicate the upper and lower bounds of the 95% highest posterior density.

## 4. Discussion

To improve our ability to assess the roles of recombination and selection in the evolution of St. Louis encephalitis virus and other flaviviruses, we have sequenced nearly the entire open reading frame of 23 historic strains of SLEV. Overall, the phylogeny we derived using the entire open reading frame was very similar to those derived using only the envelope gene (Kramer and Chandler, 2001; Twiddy and Holmes, 2003), with some differences.

In a study of the role of recombination in the genus *Flavivirus*, Twiddy and Holmes (2003) reported that a strain of SLEV from Guatemala, GMO 94, was likely a recombinant between an Argentine strain, CorAn9124, and a strain closely related to that from Tennessee, USA, TNM 4-711K. Our analyses, using sequences from all three of these strains, failed to support this result. It should be noted that we also detected recombination in Kramer and Chandler (2001) sequences using our methods, and the main reason for the discrepancy appears to be differences between our sequence for strain GMO 94 and the sequence for the same strain published by Kramer and Chandler (2001), upon which the Twiddy and Holmes (2003) study was based. The differences in sequences are too numerous to be explained by intra-strain variability or PCR errors. The 427 nucleotides in the 5' portion of the Kramer and Chandler (2001) sequence of strain GMO 94 are identical to those of strain CorAn9124 and the putative recombination breakpoint detected by Twiddy and Holmes (2003) is close to one of their PCR and sequencing primers; thus it appears that a laboratory or sequence assembly error may have occurred in that study. Our sequence for strain GMO 94 consistently fell within a clade including V 2380-42 (Texas, USA, 2001), 78 A 28 (Guatemala, 1978),

and TNM 4-711 K (Tennessee, USA, 1974) in all of the trees generated.

The numbers of nucleotide differences per strain between our sequences and those generated by Trent et al. (1987) and Kramer and Chandler (2001) varied greatly, and the reasons for these differences is difficult to explain. Some differences are to be expected due to variability arising during growth of the RNA viruses in culture, and introduction of errors in the reverse transcription and PCR steps. Some differences may also be due to the differences in sequencing strategies—Trent et al. (1987) used manual sequencing, whereas we used modern automated sequencing protocols. The two strains with the highest numbers of sequence differences were GMO 94 (32 differences, described above) and Parton (37 differences). The Parton strain was isolated during the first known human outbreak in St. Louis in 1933, and its passage history is largely known. Similarly, the BeAr 23379 strain (Para, Brazil, 1960; 10 differences) has a partly unknown passage history. It may be that variability has arisen within these strains over multiple passages, and that different variants dominate cultures on different occasions, or the consensus sequencing approaches used in all of the studies have failed to detect this variability. Cloning of RT-PCR products and sequencing multiple clones may answer this question.

The global  $d_N/d_S$  of 0.04 agreed with previous estimates in SLEV and other flaviviruses (Woelk and Holmes, 2002). However, analysis of negative and positive selection using ML methods suggested that the majority of codons in the SLEV ORF are neither under strongly negative nor under strongly positive selection. Of those codons that are under selection, the vast majority (154/155, 99.4% using the consensus of the SLAC, FEL, and REL methods) are under negative selection. The only site to be identified as being under positive selection, using the three methods we employed, was codon 444 of the SLEV ORF, or codon 156 of the E gene. This site corresponds to the third position of the potential *N*-linked glycosylation sequence, which is conserved in many flaviviruses (Chambers et al., 1990).

*N*-linked glycosylation occurs co-translationally, and can influence protein folding, secretion, stability, and solubility, and modulate antigenicity of proteins (Jones et al., 2005). In enveloped viruses, proteins in the membrane are often glycosylated, a modification which can modulate detection by the host's immune system, influence stability, and alter interactions with receptors on the surfaces of host cells (Vigerust and Shepherd, 2007). The glycosylation state of the E protein varies between strains of SLEV (Vorndam et al., 1993) and is variable within other *Flavivirus* species (Beasley et al., 2005; Halevy et al., 1994; Hanna et al., 2005). It can have marked effects on virus replication *in vitro*, as well as viremia, neuroinvasiveness, and neurovirulence *in vivo* (Beasley et al., 2005; Chambers et al., 1998; Halevy et al., 1994; Hanna et al., 2005; Li et al., 2006; Shirato et al., 2004).

Several observations suggest that the selection acting on the glycosylation site of SLEV is an artifact of *in vitro* growth in cell culture. When we compared our sequences with those of Kramer and Chandler (2001), it was apparent that many strains that were included in both studies differed in nucleotide sequence and amino acid sequence at the site of potential glycosylation. Similarly, several of the strains (MSI 7, 65 V310, 69 M 1443) with glycosylated E proteins in the study of Vorndam et al. (1993) did not encode appropriate glycosylation triplets according to our sequences. In fact, many of our sequences contained variants at the site of glycosylation, suggesting that they are in transition between glycosylation states, similar to a process that has been observed for Kunjin virus (Adams et al., 1995). When we mapped the sequences at the potential glycosylation site onto the phylogenetic tree, several of the non-synonymous mutations that altered potential glycosylation state occurred in the terminal branches. When the terminal branches were omitted from analysis, the selection operating at codon 156 of the E gene was no longer significant. Finally, numerous reports have detected *in vitro* selection acting on E glycosylation in other flaviviruses (Adams et al., 1995; Chambers et al., 1998; Lee et al., 1997; Scherret et al., 2001).

The substitutions in the terminal branches of the SLEV phylogeny suggested a transition from glycosylated to non-glycosylated E protein was favored. This observation would contradict previous results showing that a SLEV strain encoding glycosylated E grown in the presence of tunicamycin (an inhibitor of glycosylation), and a SLEV strain encoding non-glycosylated E, both showed lower infectivity on SW-13 (human adenocarcinoma) and CRE (hamster) cells than strains with glycosylated E (Vorndam et al., 1993). Similarly, *Kunjin virus* (KUN) encoding glycosylated E grew more successfully in Vero cells than non-glycosylated equivalents, and passage of several non-glycosylated KUN strains acquired glycosylation competency after passaging through Vero cells (Scherret et al., 2001). In *tick-borne encephalitis virus* (TBEV), inhibition of glycosylation or mutation of the glycosylation site to a glycosylation-refractory sequence resulted in reduced secretion of recombinant subviral particles (RSPs) (Lorenz et al., 2003). Finally, loss of E glycosylation in WNV was associated with reduced infectious viral titers in Vero cells (Beasley et al., 2005). However, it is possible that the substitution pattern indicated by the phylogeny is misleading, and that the non-glycosylated isolates are simply those that have not yet undergone transition to the glycosylated form. Indeed, 4 of the 10 strains that showed intra-strain variability at the site of *N*-linked glycosylation are shown as non-glycosylated in Fig. 1. Clarification of the effect of cell culture on glycosylation of SLEV E protein could be achieved by identifying clones encoding glycosylation-competent and glycosylation-refractory sequences at codon 156 of the E gene, and growing these clones in different cell-types. The glycosylation- status of the viruses at various time-points could then be analyzed by a combination of meth-

ods: real-time RT-PCR using glycosylation-specific primers, deep-amplicon sequencing to determine the proportions of glycosylation-encoding variants, and protein electrophoresis of purified viruses.

From a molecular demographic perspective, our estimates of SLEV nucleotide substitution rates ( $2.17 \times 10^{-4}$  substitutions/site/year overall,  $5.44 \times 10^{-5}$  substitutions/site/year at 1st + 2nd codon positions,  $5.41 \times 10^{-4}$  substitutions/site/year at 3rd codon positions) are slightly lower than earlier estimates for the SLEV *env* gene ( $3.5 \times 10^{-4}$  substitutions/site/year at all sites,  $3.7 \times 10^{-4}$  at synonymous sites) (Jenkins et al., 2002), and other flaviviruses (Bryant et al., 2007; Hanada et al., 2004; Jenkins et al., 2002). SLEV appears to have been evolving under an almost-constant population size with an increase in the effective number of infections placed in the late 19th century; this increase appears to be coinciding with the colonization of North America by SLEV. A basal diversification of South American SLEV was estimated at 177–247 ybp (Table 5), thus predating the colonization of North America and allowing us to propose a northwards route of colonization for SLEV. Additional SLEV sequences from South America, the Caribbean, and Canada, as well as from SLEV avian hosts can shed light on demographic and selection processes.

## Acknowledgments

This material is based upon work supported by, or in part by, the U.S. Army Research Laboratory and the U.S. Army Research Office under Grant Number W911NF-05-1-0271. We thank the Wadsworth Center Media and Tissue Culture facility for providing cells and media for this work, and the World Reference Center for Emerging Viruses and Arboviruses through the University of Texas Medical Branch, Galveston, supported by NIAID, for the virus strains. Part of this work was carried out by using the resources of the Computational Biology Service Unit from Cornell University, which is partially funded by Microsoft Corporation, and the CamGrid computational cluster at the University of Cambridge.

## References

- Adams, S.C., Broom, A.K., Sammels, L.M., Hartnett, A.C., Howard, M.J., Coelen, R.J., Mackenzie, J.S., Hall, R.A., 1995. Glycosylation and antigenic variation among Kunjin virus isolates. *Virology* 206, 49–56.
- Altekar, G., Dwarkadas, S., Huelsenbeck, J., Ronquist, F., 2004. Parallel metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics* 20, 407–415.
- Beasley, D., Whiteman, M., Zhang, S., Huang, C., Schneider, B., Smith, D., Gromowski, G., Higgs, S., Kinney, R., Barrett, A., 2005. Envelope protein glycosylation status influences mouse neuroinvasion phenotype of genetic lineage I West Nile virus strains. *J. Virol.* 79, 8339–8347.
- Billoir, F., de Chesse, R., Tolou, H., de Micco, P., Gould, E., de Lamballerie, X., 2000. Phylogeny of the genus flavivirus using complete coding sequences of arthropod-borne viruses and viruses with no known vector. *J. Gen. Virol.* 81, 781–790.
- Bowen, G., Monath, T., Kemp, G., Kerschner, J., Kirk, L., 1980. Geographic variation among St. Louis encephalitis virus strains in the viremic responses of avian hosts. *Am. J. Trop. Med. Hyg.* 29, 1411–1419.
- Brault, A.C., Huang, C.Y., Langevin, S.A., Kinney, R.M., Bowen, R.A., Ramey, W.N., Panella, N.A., Holmes, E.C., Powers, A.M., Miller, B.R., 2007. A single positively selected West Nile viral mutation confers increased virogenesis in American crows. *Nat. genet.* 39, 1162–1166.
- Bryant, J.E., Holmes, E.C., Barrett, A.D., 2007. Out of Africa: a molecular perspective on the introduction of yellow fever virus into the Americas. *PLoS Pathog.* 3, e75.
- Chambers, T., Hahn, C., Galler, R., Rice, C., 1990. Flavivirus genome organization, expression, and replication. *Annu. Rev. Microbiol.* 44, 649–688.
- Chambers, T., Halevy, M., Nestorowicz, A., Rice, C., Lustig, S., 1998. West Nile virus envelope proteins: nucleotide sequence analysis of strains differing in mouse neuroinvasiveness. *J. Gen. Virol.* 79, 2375–2380.
- Day, J., 2001. Predicting St. Louis encephalitis virus epidemics: lessons from recent, and not so recent, outbreaks. *Annu. Rev. Entomol.* 46, 111–138.
- Don, R.H., Cox, P.T., Wainwright, B.J., Baker, K., Mattick, J.S., 1991. ‘Touchdown’ PCR to circumvent spurious priming during gene amplification. *Nucleic Acids Res.* 19, 4008.
- Drummond, A., Rambaut, A., 2006. BEAST. University of Oxford, Oxford, UK. Available from: <<http://beast.bio.ed.ac.uk/>>.
- Drummond, A.J., Rambaut, A., Shapiro, B., Pybus, O.G., 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.* 22, 1185–1192.
- Guindon, S., Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* 52, 696–704.
- Halevy, M., Akov, Y., Ben-Nathan, D., Kobiler, D., Lachmi, B., Lustig, S., 1994. Loss of active neuroinvasiveness in attenuated strains of West Nile virus: pathogenicity in immunocompetent and SCID mice. *Arch. Virol.* 137, 355–370.
- Hanada, K., Suzuki, Y., Gojobori, T., 2004. A large variation in the rates of synonymous substitution for RNA viruses and its relationship to a diversity of viral infection and transmission modes. *Mol. Biol. Evol.* 21, 1074–1080.
- Hanna, S., Pierson, T., Sanchez, M., Ahmed, A., Murtadha, M., Doms, R., 2005. N-linked glycosylation of west nile virus envelope proteins influences particle assembly and infectivity. *J. Virol.* 79, 13262–13274.
- Holmes, E., 2003. Patterns of intra- and interhost nonsynonymous variation reveal strong purifying selection in dengue virus. *J. Virol.* 77, 11296–11298.
- Jenkins, G.M., Rambaut, A., Pybus, O.G., Holmes, E.C., 2002. Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. *J. Mol. Evol.* 54, 156–165.
- Jones, J., Krag, S., Betenbaugh, M., 2005. Controlling N-linked glycan site occupancy. *Biochim. Biophys. Acta* 1726, 121–137.
- Kass, R.E., Raftery, A.E., 1995. Bayes factors. *J. Am. Stat. Assoc.* 90, 773–795.
- Kishino, H., Hasegawa, M., 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *J. Mol. Evol.* 29, 170–179.
- Kosakovsky-Pond, S.L., Frost, S.D., 2005a. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics* 21, 2531–2533.
- Kosakovsky-Pond, S.L., Frost, S.D., 2005b. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol. Biol. Evol.* 22, 1208–1222.
- Kosakovsky-Pond, S.L., Frost, S.D., Muse, S.V., 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 21, 676–679.
- Kosakovsky-Pond, S.L., Posada, D., Gravenor, M.B., Woelk, C.H., Frost, S.D., 2006. GARD: a genetic algorithm for recombination detection. *Bioinformatics* 22, 3096–3098.

- Kramer, L., Chandler, L., 2001. Phylogenetic analysis of the envelope gene of St. Louis encephalitis virus. *Arch. Virol.* 146, 2341–2355.
- Kuno, G., Chang, G., 2005. Biological transmission of arboviruses: reexamination of and new insights into components, mechanisms, and unique traits as well as their evolutionary trends. *Clin. Microbiol. Rev.* 18, 608–637.
- Lanave, C., Preparata, G., Saccone, C., Serio, G., 1984. A new method for calculating evolutionary substitution rates. *J. Mol. Evol.* 20, 86–93.
- Lee, E., Weir, R.C., Dalgarno, L., 1997. Changes in the dengue virus major envelope protein on passaging and their localization on the three-dimensional structure of the protein. *Virology* 232, 281–290.
- Li, J., Bhuvanakantham, R., Howe, J., Ng, M., 2006. The glycosylation site in the envelope protein of West Nile virus (Sarafend) plays an important role in replication and maturation processes. *J. Gen. Virol.* 87, 613–622.
- Lindenbach, B., Rice, C., 2001. Flaviviridae: the viruses and their replication. In: Knipe, D., Howley, P. (Eds.), *Fields Virology*. Lippincott-Raven Publishers, Philadelphia, pp. 991–1041.
- Lorenz, I.C., Kartenbeck, J., Mezzacasa, A., Allison, S.L., Heinz, F.X., Helenius, A., 2003. Intracellular assembly and secretion of recombinant subviral particles from tick-borne encephalitis virus. *J. Virol.* 77, 4370–4382.
- Martin, D., Rybicki, E., 2000. RDP: detection of recombination amongst aligned sequences. *Bioinformatics* 16, 562–563.
- Martin, D., Williamson, C., Posada, D., 2005a. RDP2: recombination detection and analysis from sequence alignments. *Bioinformatics* 21, 260–262.
- Martin, D.P., Posada, D., Crandall, K.A., Williamson, C., 2005b. A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. *AIDS Res. Hum. Retroviruses* 21, 98–102.
- Monath, T., Cropp, C., Bowen, G., Mitchell, C., Gardner, J., 1980. Variation in virulence for mice and rhesus monkeys among St. Louis encephalitis virus strains of different origin. *Am. J. Trop. Med. Hyg.* 29, 948–962.
- Newton, M.A., Raftery, A.E., 1994. Approximate Bayesian inference with the weighted likelihood bootstrap. *J. R. Stat. Soc. Ser. B* 56, 3–48.
- Padidam, M., Sawyer, S., Fauquet, C.M., 1999. Possible emergence of new geminiviruses by frequent recombination. *Virology* 265, 218–225.
- Posada, D., 2002. Evaluation of methods for detecting recombination from DNA sequences: empirical data. *Mol. Biol. Evol.* 19, 708–717.
- Posada, D., Crandall, K., 1998. MODELTEST: testing the model of DNA substitution. *Bioinformatics* 14, 817–818.
- Posada, D., Crandall, K.A., 2001. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proc. Natl. Acad. Sci. USA* 98, 13757–13762.
- Posada, D., Crandall, K.A., Holmes, E.C., 2002. Recombination in evolutionary genomics. *Annu. Rev. Genet.* 36, 75–97.
- Rambaut, A., Drummond, A., 2005. Tracer v1.3. University of Oxford, Oxford, UK.
- Reisen, W., 2003. Epidemiology of St. Louis encephalitis virus. *Adv. Virus Res.* 61, 139–183.
- Rodriguez, F., Oliver, J.L., Marin, A., Medina, J.R., 1990. The general stochastic model of nucleotide substitution. *J. Theor. Biol.* 142, 485–501.
- Ronquist, F., Huelsenbeck, J., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19, 1572–1574.
- Scherret, J., Mackenzie, J., Khromykh, A., Hall, R., 2001. Biological significance of glycosylation of the envelope protein of Kunjin virus. *Ann. NY Acad. Sci.* 951, 361–363.
- Shirato, K., Miyoshi, H., Goto, A., Ako, Y., Ueki, T., Kariwa, H., Takashima, I., 2004. Viral envelope protein glycosylation is a molecular determinant of the neuroinvasiveness of the New York strain of West Nile virus. *J. Gen. Virol.* 85, 3637–3645.
- Smith, J.M., 1992. Analyzing the mosaic structure of genes. *J. Mol. Evol.* 34, 126–129.
- Suchard, M.A., Weiss, R.E., Sinsheimer, J.S., 2001. Bayesian selection of continuous-time Markov chain evolutionary models. *Mol. Biol. Evol.* 18, 1001–1013.
- Swofford, D., 2003. PAUP\*. *Phylogenetic Analysis Using Parsimony (\* and Other Methods)*. Sinauer Associates, Sunderland, MA.
- Trent, D., Kinney, R., Johnson, B., Vorndam, A., Grant, J., Deubel, V., Rice, C., Hahn, C., 1987. Partial nucleotide sequence of St. Louis encephalitis virus RNA: structural proteins, NS1, ns2a, and ns2b. *Virology* 156, 293–304.
- Trent, D., Monath, T., Bowen, G., Vorndam, A., Cropp, C., Kemp, G., 1980. Variation among strains of St. Louis encephalitis virus: basis for a genetic, pathogenetic, and epidemiologic classification. *Ann. NY Acad. Sci.* 354, 219–237.
- Twiddy, S., Holmes, E., 2003. The extent of homologous recombination in members of the genus *Flavivirus*. *J. Gen. Virol.* 84, 429–440.
- Twiddy, S., Woelk, C., Holmes, E., 2002. Phylogenetic evidence for adaptive evolution of dengue viruses in nature. *J. Gen. Virol.* 83, 1679–1689.
- Vigerust, D.J., Shepherd, V.L., 2007. Virus glycosylation: role in virulence and immune interactions. *Trends Microbiol.* 15, 211–218.
- Vorndam, V., Mathews, J., Barrett, A., Roehrig, J., Trent, D., 1993. Molecular and biological characterization of a non-glycosylated isolate of St. Louis encephalitis virus. *J. Gen. Virol.* 74, 2653–2660.
- Woelk, C., Holmes, E., 2002. Reduced positive selection in vector-borne RNA viruses. *Mol. Biol. Evol.* 19, 2333–2336.
- Yang, Z., 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* 39, 306–314.