

Widespread interspecific phylogenetic tree incongruence between mosquito-borne and insect-specific flaviviruses at hotspots originally identified in Zika virus

Michael W. Gaunt,^{1,*†} John H-O. Pettersson,^{2,3,†} Goro Kuno,⁴ Bill Gaunt,⁵ Xavier de Lamballerie,^{6,7,§} and Ernest A. Gould⁶

¹Department of Infection Biology London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK, ²Department of Medical Biochemistry and Microbiology, Zoonosis Science Center, Uppsala University, 751 23 Uppsala, Sweden, ³Sydney Institute for Infectious Diseases, School of Life and Environmental Sciences and School of Medical Sciences, The University of Sydney, Sydney, NSW 2006, Australia, ⁴1648 Collindale Dr., Fort Collins, CO 80525, USA Formerly, Centers for Disease Control, Fort Collins, CO 80521, USA, ⁵Aeon-Sys, MBCS 29 Kensington Road, Barnsley S75 2TU, UK, ⁶UMR "Unité des Virus Emergents", Aix-Marseille Université-IRD 190-Inserm 1207-IHU Méditerranée Infection, Marseille, France and ⁷APHM Public Hospitals of Marseille, Institut Hospitalo-Universitaire Méditerranée Infection, 13005 Marseille, France

[†]<https://orcid.org/0000-0002-0080-6210>

[†]<https://orcid.org/0000-0001-6665-7492>

[§]<https://orcid.org/0000-0001-7895-2720>

*Corresponding author: E-mail: m@genome.guru

Abstract

Intraspecies (homologous) phylogenetic incongruence, or 'tree conflict' between different loci within the same genome of mosquito-borne flaviviruses (MBFV), was first identified in dengue virus (DENV) and subsequently in Japanese encephalitis virus (JEV), St Louis encephalitis virus, and Zika virus (ZIKV). Recently, the first evidence of phylogenetic incongruence between interspecific members of the MBFV was reported in ZIKV and its close relative, Spondweni virus. Uniquely, these hybrid proteomes were derived from four incongruent trees involving an *Aedes*-associated DENV node (1 tree) and three different *Culex*-associated flavivirus nodes (3 trees). This analysis has now been extended across a wider spectrum of viruses within the MBFV lineage targeting the breakpoints between phylogenetic incongruent loci originally identified in ZIKV. Interspecies phylogenetic incongruence at these breakpoints was identified in 10 of 50 viruses within the MBFV lineage, representing emergent *Aedes* and *Culex*-associated viruses including JEV, West Nile virus, yellow fever virus, and insect-specific viruses. Thus, interspecies phylogenetic incongruence is widespread amongst the flaviviruses and is robustly associated with the specific breakpoints that coincide with the interspecific phylogenetic incongruence previously identified, inferring they are 'hotspots'. The incongruence amongst the emergent MBFV group was restricted to viruses within their respective associated epidemiological boundaries. This MBFV group was RY-coded at the third codon position ('wobble codon') to remove transition saturation. The resulting 'wobble codon' trees presented a single topology for the entire genome that lacked any robust evidence of phylogenetic incongruence between loci. Phylogenetic interspecific incongruence was therefore observed for exactly the same loci between amino acid and the RY-coded 'wobble codon' alignments and this incongruence represented either a major part, or the entire genomes. Maximum likelihood codon analysis revealed positive selection for the incongruent lineages. Positive selection could result in the same locus producing two opposing trees. These analyses for the clinically important MBFV suggest that robust interspecific phylogenetic incongruence resulted from amino acid selection. Convergent or parallel evolutions are evolutionary processes that would explain the observation, whilst interspecific recombination is unlikely.

Key words: Mosquito-borne viruses; flavivirus; phylogenetic incongruence; selection; recombination; RY-coding.

1. Introduction

The genus *Flavivirus* (family *Flaviviridae*) contains more than 50 RNA positive-stranded mosquito-borne flaviviruses (MBFV) that infect well in excess of 400 million humans annually, throughout the tropics, sub-tropics, and warmer temperate regions. They include four dengue virus serotypes, yellow fever virus (YFV), Japanese encephalitis virus (JEV), St Louis encephalitis virus (SLEV), West Nile virus (WNV), and Zika virus (ZIKV). The MBFV lineage also includes some insect-specific flaviviruses (ISFV) (Huhtamo et al. 2009, 2014; Grard et al. 2010) and Ecuador Paraiso Escondido virus (EPEV) an isolate from sandflies (Alkan et al.

2015) which does not replicate in mammalian cells even though it is placed phylogenetically in the MBFV lineage with viruses that are pathogens for mammals, including YFV and DENV. Bat-associated viruses which have no recognized arthropod vector, are also placed phylogenetically in the MBFV lineage. The genus *Flavivirus* also contains basal members of the ISFV which exist outside the MBFV lineage and are closely related to cell fusing agent virus (CFAV), possibly, a non-vectorized precursor of the genus (Cook et al. 2012).

Phylogenetic analyses robustly divided the MBFV lineage into two phylo-epidemiological groups, viz. *Aedes*-associated and

Culex-associated MBFV, using either single protein gene trees (Gaunt et al. 2001; Gould et al. 2001) or whole-genome data (Huhtamo et al. 2009, 2014; Cook et al. 2012; Kolodziejek et al. 2013; Moureau et al. 2015). The *Aedes*-associated groups included YFV and DENV represented as paraphyletic clades (Huhtamo et al. 2009; Kuno et al. 2009; Kolodziejek et al. 2013). However, the *Culex*-associated flaviviruses formed a monophyletic group that included JEV, WNV, SLEV, and Rocio virus (ROCV) (Gaunt et al. 2001; Gould et al. 2001; Grard et al. 2010; Moureau et al. 2015). Intraspecific homologous phylogenetic incongruence in the MBFV has been identified in each of the four DENV serotypes, JEV, SLEV, and African ZIKV with most studies notably focusing on the E-protein (Holmes et al. 1999; Twiddy and Holmes 2003; Aaskov et al. 2007; Weaver and Vasilakis 2009; Carney et al. 2012; Faye et al. 2014; Worobey and Holmes 1999; Worobey et al. 1999). Surprisingly, in the above studies, intraspecific tree incongruence was not detected in YFV. This might be explained by the fact that the analyses were restricted because they were based on the flavivirus envelope protein (Twiddy and Holmes 2003), limiting their analytical power. Preliminary evidence for potential recombination was also reported in the non-MBFV lineage flavivirus CFAV (Cook et al. 2012), and CFAV recombination has also been observed between structural proteins and non-structural (NS) proteins within the CFAV-associated ISFV group (Baidaliuk et al. 2020).

Phylogenetic incongruence can result from different underlying evolutionary processes, including recombination (Simon-Loriere and Holmes 2011; Thézé et al. 2015), selection, and associated parallel or convergent evolution (Sackman et al. 2017; Stern et al. 2017), as well as systematic error involving the tree-building methods, for example, where the substitution rate variation leads to 'long branch attraction' (Felsenstein 1978; Huelsenbeck 1997; Olsen 2013) or compositional bias (Phillips et al. 2004; Ishikawa et al. 2012; Breinholt and Kawahara 2013; Song et al. 2016).

Recombination is of interest because in flaviviruses it makes biological sense within the theory of protein structure-stability associated with compensatory mutations. State-of-the-art analysis assumes two residues are physically associated in protein space despite not being contiguous within the coding sequence (Jones and Kandathil 2018). For example, the flavivirus E-protein monomer comprises parallel beta-pleated sheets, and a mutation in one beta-pleated sheet could destabilize the structure unless it is accompanied by a compensatory mutation in the parallel sheet. However, recombination ensures that both alleles are transferred to the progeny. In support of this argument, compensatory mutations that were reported in hepatitis C virus recombinants were identified in separate studies (Kalinina et al. 2002; Cusick et al. 2011) and interspecific recombination was detected in the NS5B protein within the genus *Hepacivirus* of the *Flaviviridae* family (Thézé et al. 2015).

Amongst the MBFV lineage viruses, phylogenetic evidence of interspecific phylogenetic incongruence was reported between the most recent common ancestor of ZIKV and SPOV i.e. an *Aedes*-associated virus and more than one virus representing the *Culex*-associated virus clades (Gaunt et al. 2020). This discovery provided a novel and rational explanation for the long-standing observation of phylogenetic incongruence observed in several separate MBFV evolutionary studies. For years, it was unclear whether ZIKV and closely related SPOV belonged, topologically, in the DENV-*Aedes*-associated clade or the *Culex*-associated clade. The study provided robust phylogenetic evidence that the ancestral *Aedes*-associated lineage of ZIKV and SPOV had undergone three distinct interspecific recombination events, each of which involved a genetic

crossover with one or more *Culex*-associated viruses, resulting in hybrid proteomes.

With the evidence of robust interspecific phylogenetic incongruence established for ZIKV and SPOV, the recombined loci were observed to associate with distinct immunological properties; these genomes encompassed domains corresponding to the DENV pre-membrane (prM), envelope (E), and non-structural protein (NS1), all of which primarily induce human B-cell immune responses, plus the NS4B protein. In addition, the ZIKV and SPOV proteins that generate dominant CD8⁺ and/or CD4⁺ T-cell responses, formed three independent incongruent monophyletic groups within two *Culex*-associated clades. One of these clades comprised Australasian/Melanesian flavivirus strains with dominant HLA-associated CD8⁺ epitopes. The other clade involved the JEV-related *Culex*-associated viruses. The three incongruent *Culex* clade protein loci responsible for T-cell responses, comprised the capsid (C), NS2A to NS3 DEAD domain, the C-terminal NS3 HELICc domain to NS4A protein, inclusively, and the NS5 protein (Gaunt et al. 2020). Importantly, it has recently been demonstrated that the T-cell response following immunization with a live-attenuated JEV vaccine, produced a cross-protection against ZIKV infection in humans (Wang et al. 2020). This was phylogenetically predicted in our previous publication which identified interspecific recombination between DENV and JEV-related *Culex*-associated flaviviruses (Gaunt et al. 2020).

The discovery of interspecific phylogenetic incongruence between flaviviruses occupying different but potentially sympatric ecological niches raised three important questions, 'are the three independent occurrences of ancestral hybrids of ZIKV and SPOV isolated events within the MBFV lineage?'. Alternatively, 'do the interspecific phylogenetic incongruence breakpoints identified in the ZIKV and SPOV ancestor represent "hotspots" for other MBFV lineage viruses?' More fundamentally, given that the ZIKV and SPOV/DENV (*Aedes/Culex*) hybrid proteomes represent phylogenetic incongruence over a large genetic distance, 'could any of the other MBFV lineage viruses, including the ISFV, demonstrate phylogenetic incongruence at these breakpoints?' In this respect, the previously reported tree incongruence between CFAV-associated ISFV, which are ancestral to the MBFV lineage, is important because it could help towards understanding the evolutionary processes in viruses nearer to or at the root of the tree (Cook et al. 2012; Baidaliuk et al. 2020). Here, following an extensive programme of phylogenetic analyses we report the identification of widespread interspecific phylogenetic incongruence between MBFV lineage viruses notably for amino acid or first and second codon position alignments, whilst third position codon alignments demonstrated within loci phylogenetic incongruence against the other two data types.

2. Methods

2.1 Bayesian amino acid and nucleotide phylogenies analysed using Beast

The uncorrelated relaxed clock was run using all available models, viz. the lognormal, gamma (γ), and exponential distributions. (Suchard et al. 2018) The strict and all relaxed clock models were run under two coalescent priors constant size and Skyline plot (Drummond et al. 2005). All Bayesian MCMC calculations were performed in duplicate using two different random number seeds that were identical between data sets. Each Bayesian MCMC comprised 5×10^6 replications and a burn-in of 0.5×10^6 replications. If convergence was not observed the calculation was repeated for 1×10^7 , 2×10^7 , or 5×10^7 replications until convergence was

observed. Bayesian MCMC convergence was assessed firstly using the tools within the Beast v1.10.4 package via the effective sample size (ESS) ('loganalyser') where the burn-in was applied, the two MCMC run tree and log files combined ('logcombiner') and the following parameters assessed for an ESS > 200: marginal, likelihood, prior, population size, tree model root height, tree length, uclld/ucgd mean, coefficient of variation, covariance, tree likelihood, branch rates, and coalescent. All statistical outputs of the ESS are available on request. The duplicate MCMC runs were assessed individually for convergence against each other using the diversity of graphing tools in Tracer 1.7 (Rambaut et al. 2018).

The relaxed clock model, running under an exponential distribution, did not achieve convergence within 2×10^7 replications for all large data sets and was therefore discontinued. Hereafter, we refer to the remaining Bayesian phylogenetic calculations as the '6 models', viz. one strict and two relaxed clocks (lognormal and gamma distributions) each for two population size assumptions (constant and 'skyline').

2.1.1 *Beast and maximum likelihood amino acid phylogenies analysed using the LG matrix*

The phylogenies using all members of the MBFV were generated using amino acid data rather than nucleotides to reduce problems arising from substitutional saturation in the third codon position. Moreover, amino acid data carry more information because they report on transversion substitutions within the third codon position for a large proportion of amino acid mutations. Recent applications of the LG matrix (Le and Gascuel 2008) support this assumption (Wolf et al. 2018; Gaunt et al. 2020).

2.2 Tree visualization

Dendroscope 3 was used as the primary phylogenetic tool for visualizing different tree structures between loci under the network clustering algorithm (Huson and Scornavacca 2012). In Dendroscope 3, differences between trees are represented as reticulation, approximating to a 'tri-furcating' tree.

2.3 Phylogenetic robustness analysed non-parametric, parametric bootstrapping, and posterior probabilities

Subsequent analyses were based on the observation that there was a maximum of one phylogenetic incongruence for a given flavivirus, i.e. a robustly incompatible tree structure from an amino acid alignment identified using maximum likelihood non-parametric bootstrapping and then assessed by Bayesian MCMC posterior probabilities for all six models. These two alternative structures were examined using parametric bootstrapping and again verified for supporting evidence using posterior probabilities for all six models (below). The rationale for the two approaches is that non-parametric bootstrap analysis provides a means of data mining whereas parametric bootstrapping (SOWHAT test) enables two alternative scenarios to be precisely examined (below). Posterior probabilities provide support for both methods because they have been observed to provide higher support than non-parametric bootstrapping (Douady et al. 2003; Alfaro et al. 2003).

Phylogenetic incongruence was considered robust by non-parametric bootstrapping when nodes were supported by values greater than or equal to 80 per cent and for Bayesian analysis a posterior probability greater than 0.95 is considered robust.

2.4 Alignments

The genome-wide alignments in this study comprised seven loci representing robust interspecific phylogenetic breakpoints identified in the ZIKV-SPOV lineage are presented in Fig. 1 (Gaunt et al. 2020) together with the immunological associations for the breakpoints. It is important to note that the two combinatorial genomic loci identified by breakpoints in Fig. 1 (defined below) were combined without reference to B- or T-cell immunity, instead the aim was to obtain the most robust phylogenies with which to identify the interspecific phylogenetic incongruence that resulted in ZIKV and SPOV.

The methods and resulting alignments used to confirm interspecific phylogenetic incongruence in ZIKV-SPOV are described in Supplementary Information 1 (Forslund et al. 2004; Lemey et al. 2009; Gaunt et al. 2020). In summary, the MBFV lineage alignments of Fig. 1 and Supplementary Information 1 delineate: the proteins prM and E; a contiguous amino acid locus from NS2A to the NS3 DEAD domain (hereafter termed NS2A-NS3 DEAD); a contiguous locus from N-terminal NS3 HELIC domain and NS4A protein; NS4B protein; combined loci of NS1 and partial NS3 HELICc domains (hereafter termed NS1-NS3 HELICc); and a combined locus of C-protein and NS5 protein (hereafter termed C-NS5).

2.4.1 Amino acid alignments for maximum likelihood bootstrapping and Beast

Two variations in alignments were used between the two core approaches in this study described in Fig. 2: 'data mining' for non-parametric bootstrap analysis (70 taxa alignment) and the Bayesian analysis and 'SOWHAT' test used in the analysis (50 taxa alignment). Non-parametric bootstrapping was conducted for 1000 replications using 70 flavivirus genome sequences, comprising 50 different members of the MBFV lineage and two tick-borne virus genomes to root the tree (Supplementary Table S1). A reduced taxa alignment was required for amino acid parametric bootstrapping because overall this calculation is more intensive than non-parametric bootstrapping (SOWHAT test described below). The Bayesian and parametric bootstrap analysis used the same alignments as the non-parametric bootstrap analysis but with 50 flavivirus genomes, comprising 48 different members of the MBFV lineage and two tick-borne flaviviruses (Supplementary Table S1). Duck egg-drop syndrome virus (DEDSV) and Israel turkey meningoencephalitis virus (ITV) were discarded because genetically they were almost identical at all loci to Baiyangdian virus (BYDV) and Bagaza virus (BAGV), respectively.

2.4.2 Nucleotide alignments, RY-coding, and associated phylogenies

Nucleotide phylogenies were produced for the same loci but utilizing restricted taxa to minimize nucleotide saturation. This analysis comprised two alignments each for first and second codon positions and a separate alignment for third codon positions, hereafter called 'wobble codons': (1) *Culex* I group using the KOKV group as the outgroup; (2) YFV group using the SOKV group as the outgroup. Each MBFV group listed here is described in Fig. 2. To overcome third codon position saturation all wobble codons were purine (R) and pyrimidine (Y), RY, transformed (Phillips et al. 2004). Nucleotide analysis used both maximum likelihood (RAxML) non-parametric bootstrapping for 1000 replications (Stamatakis 2014) and Bayesian methods (Beast v1.10.4; Suchard et al. 2018) both under a general time reversible (GTR) model with 4-category γ -distribution. All further methods involved in nucleotide tree

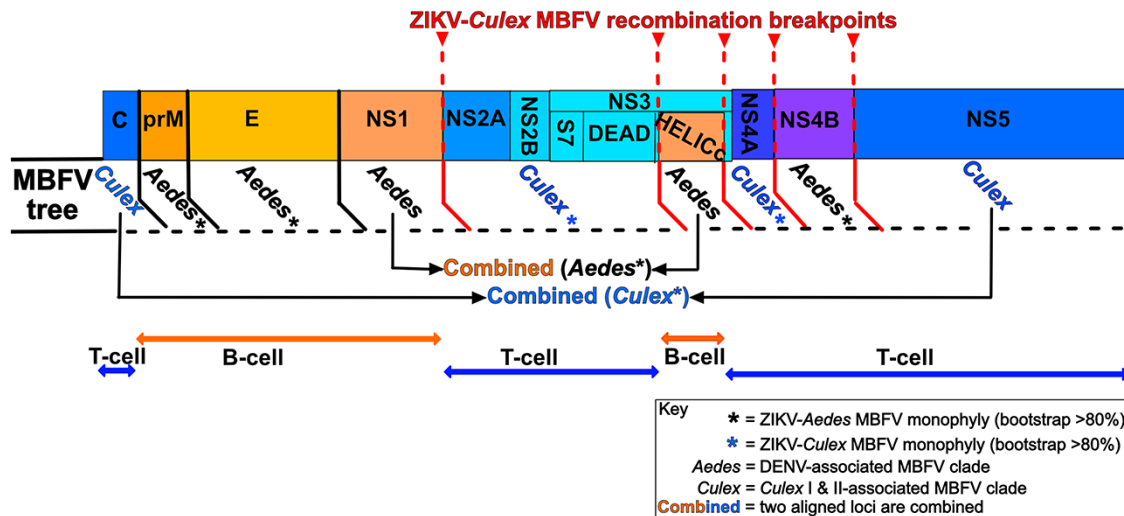


Figure 1. Schematic representation of the recombination breakpoints and resulting seven protein loci across the MBFV amino acid genome, viz. prM, E, NS1–NS3 HELICc, NS2A–NS3 DEAD, C-terminal NS3–NS4A, NS4B and C–NS5, identified in the interspecific recombination events between ZIKV–SPOV (DENV *Aedes*-associated clade) with the *Culex* I and II-associated MBFV resulting in a chimeric genome. (Supplementary Information 1). Combined locus pairs (NS1–NS3 HELICc and C–NS5) were identical in topology using the SOWHAT test prior amalgamation. The predominant immunological associations are indicated in the diagram below after the recombination breakpoints were robustly established for both contiguous and combined loci. Note, C-protein and NS5-protein are juxtaposed during replication due to genome circularisation.

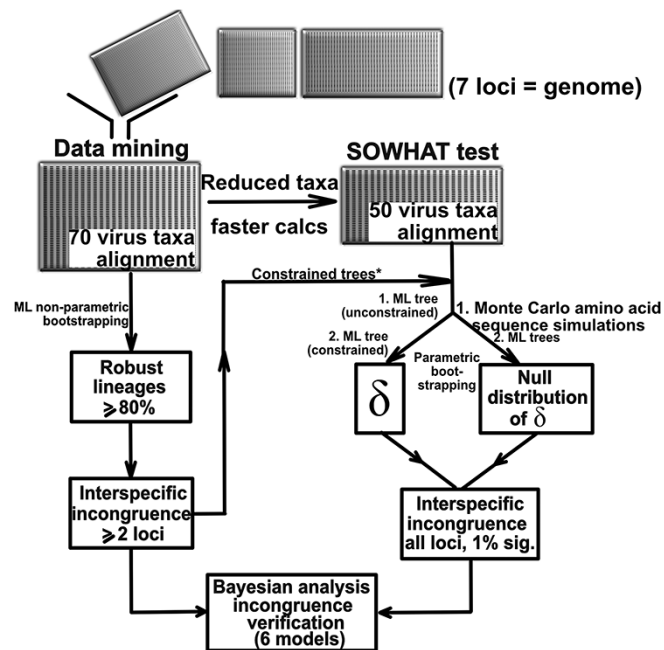


Figure 2. The analytical approach to genome-wide interspecific recombination using ‘data mining’ via non-parametric bootstrap analysis and then the SOWHAT test followed by Bayesian analysis between the MBFV described for the seven loci described in Fig. 1. The ‘data mining’ method and SOWHAT amino acid alignment datasets, differ in the number of taxa in the analysis, but not in interspecific genetic diversity. Note, the trees used for SOWHAT analysis are zero constrained.

construction were identical to the amino acid methods, with the exception that a single relaxed clock (lognormal) was run under both a constant and ‘Skyline’ (3–5 groups) population size assumption. The Bayesian MCMC could not achieve convergence in any RY-coded data set therefore phylogenetic analysis was discontinued for these data sets.

2.5 Parametric bootstrapping

Likelihood-based parametric tests of tree topologies have a long history in phylogenetics and the power of a parametric bootstrap over non-parametric bootstrapping resampling is well established (Goldman 1993; Hillis et al. 1996; Huelsenbeck and Bull 1996), in particular one method designated the Swofford–Olsen–Waddell–Hillis (SOWH) test (Hillis et al. 1996) being known for its rigorous verification (Goldman et al. 2000). A standard SOWH test compares the log-likelihood of the maximum likelihood (unconstrained tree) to the log-likelihood of an a priori tree hypothesis and the difference in log-likelihood scores between the unconstrained and constrained tree is the test statistic and denoted δ , i.e. $\delta \equiv L_{t_{ML}} - L_{t_1}$ or t_2 . The null distribution is calculated by parametric bootstrapping using the tree parameters and the maximum likelihood tree of the constrained tree (t_1) to generate a Monte Carlo simulation of amino acid sequence data.

In this study 100 parametric bootstrap replicates were calculated for each test using amino acid data and 500 replicates for nucleotide data both using Seq-Gen v1.3.4 (Rambaut and Grassly 1997). Seq-Gen permitted simulation under the same models used for non-parametric bootstrapping (LG matrix for amino acid data and GTR for nucleotide data both with a 4-category γ distribution). The test statistic was compared against a null distribution at 1 per cent threshold for amino acid data and 5 per cent for nucleotide data due to the increased number of replicates; thus for amino acid data if more than one value in the null distribution was greater than δ the constrained tree could not be significantly different from the unconstrained maximum likelihood tree. Furthermore if more than 10 per cent of the null distribution was greater than δ then the constrained tree was assumed to be the unconstrained maximum likelihood tree.

SOWHAT is the most recent development of SOWH (Church et al. 2015) and was used in this study. The default option was set to ensure that SOWHAT propagates the same number of insertions/deletions (‘indels’) as there are in the original

comparative sequence alignment. The SOWHAT pipeline was modified to ensure that the 'LG' variable could be inserted into both the RAxML and Seq-Gen (regex modification at line 510). SOWHAT is a Perl5 data pipeline using RAxML v8.2.12 and Seq-Gen v1.3.4, which in addition to automating SOWH incorporates more recent developments in parametric bootstrapping (Susko 2014). SOWHAT was leveraged within an optimized Docker container for nucleotide data (shchurch/sowhat). We followed these recommendations using constrained trees that are minimally resolved and hereafter we use the formal designation describing such trees as a 'zero-constrained' tree.

2.6 Phylogenetic incongruence

In this analysis, phylogenetic incongruence is a conditional probability where,

$$P(H_1 | H_0) \text{ or } P(H_0 | H_1) \leq 0.01$$

Here H_0 is tree 1 (t_1) and H_1 is tree 2 (t_2). 'True' means is not significantly different from the null distribution. In other words, one tree must be accepted where $P > 0.1$ and the alternative tree rejected if incongruence is to be identified. If $P(H_0 \text{ and } H_1) \leq 0.01$ then the tree is unresolved and if $P(H_0 \text{ and } H_1) > 0.05$ the result is not significant, most likely resulting from a lack of phylogenetic resolution. For simplicity in Results we define for example $P(H_1 | H_0) < 0.01$ as ' H_0 accepted, H_1 rejected'. The calculation is very similar to the likelihood ratio test (LRT) using parametric bootstrapping developed and used to identify genetic loci which were phylogenetically incongruent (Huelsenbeck and Bull 1996).

2.7 Positive selection analysis ω

We investigated whether selection was acting on individual amino acid residues using a maximum likelihood method to assess the nonsynonymous (amino acid-altering) to synonymous (silent) substitution rate ratio ($\omega = d_N/d_S$) specifically at branches representing phylogenetic incongruence under the branch-lineage model through 'CodeML' within PAML (Yang and Nielsen 2002). This will result in two values for ω , one concerning the phylogenetic incongruence branch under investigation and the background ω for all other branches. The incongruent branch was designated as either: (1) the branch associated with the ancestral node if the incongruence resulted in a monophyletic group or (2) the branch connecting two paraphyletic groups if the phylogenetic incongruence resulted in a paraphyly. Multiple incongruence was present in the bifurcating trees under ω analysis, therefore the incongruent lineages not under immediate analysis were presented as polytomies, in addition to the root lineages and for SLEV group NTA group viruses. A PAML (Yang 2007) pipeline was leveraged through the ETE3 suite (Huerta-Cepas et al. 2016).

All materials described in Methods, including all alignments, are available on request.

3. Results

3.1 Phylogenetic overview

A cluster network maximum likelihood consensus amino acid phylogeny was generated for 50 members of the MBFV lineage representing 70 taxa (Fig. 3). The consensus phylogeny superimposes the E- and C-NS5 protein trees onto a single tree, where differences in tree branching order (incongruence) between the E-protein and C-NS5 locus are represented in curved blue lines that

support 'floating branches'. The phylogeny summarises the common genetic pattern of interspecific tree incongruence observed across the MBFV lineage, as identified by incongruence between the E- and C-NS5 proteins, with the exception of YFV which does not show incongruence between these two loci in this specific analysis (tree 1), and the ISFV remain unresolved. The horizontal red branches, labelled trees 1–8, each highlight a putative phylogenetic incongruence, viz. YFV (Tree 1); Uganda S virus (UGSV), Banzi virus (BANV) and Bouboui virus (BOUV) (Tree 2); JEV and Usutu virus (USUV) (Tree 3); WNV and Yaounde virus (YAOV) (Tree 4); Rocio virus (ROCV) and Ilheus virus (ILHV) (Tree 5); Ecuador Paraiso Escondido virus (EPEV)(Tree 6); the 'bat MBFV lineage viruses' comprising the three members of the Sokuluk virus (SOKV) group (Tree 7); and two members of the ISFV, BJV, and NOUV, from the six members of this group (Tree 8). In summary, Trees 1–2 represent the YFV group, Trees 3–5 represent the *Culex*-associated clades, and Trees 6–8 inclusively represent the MBFV lineage viruses not associated with a mosquito-vertebrate transmission cycle. The rationale of combining the C-NS5 proteins into a single locus for comparison with the other combinatorial locus in this analysis, is summarized in Methods and was described in detail previously (Gaunt et al. 2020).

Given the precedent set by the discovery of phylogenetic incongruence involving ZIKV and SPOV (Gaunt et al. 2020), it seemed justified to ask the question, 'do the remaining genomic loci in the analysis look like the E tree (alternate hypothesis, H_1) or the C-NS5 tree (null hypothesis, H_0)'? By definitively identifying which of the other loci in the genome associate with either the 'E-protein' tree topology or a 'C-NS5 protein' tree topology, the hybrid structure of tree incongruence in the given MBFV genome could be determined initially following the procedures employed for ZIKV and SPOV. The E- versus C-NS5 trees provide a general pattern (not a rule) to conceptualize the analysis for MBFV lineage virus hybrid genomes.

3.2 Is phylogenetic incongruence widespread amongst the MBFV?

3.2.1 Data mining

In attempting to assess the extent of the putative interspecific phylogenetic incongruence amongst the MBFV lineages, the first approach was data mining as described in Fig. 2. This examined the maximum likelihood non-parametric bootstrapping for the 70 flavivirus alignments and was confirmed using Bayesian posterior probabilities as indicated in 'Phylogenetic Overview' and in Fig. 3, for each locus originally defined by the breakpoints identified for the interspecific phylogenetic incongruence of ZIKV and SPOV (Fig. 1) (Gaunt et al. 2020). Interspecific tree incongruence was a possibility if the two mirror phylogenetic trees produced different robust branching topologies for two different MBFV lineage viruses (Methods).

Robust phylogenetic incongruence identified by the non-parametric bootstrapping for Trees 1–8 is described in Fig. 4 and summarised in Table 1 using the loci described in Fig. 1. Each mirror tree in Trees 1–8 revealed two examples of interspecific incongruence between two or more loci and generated a null hypothesis (H_0) and an alternative hypothesis (H_1) which are summarised in Table 1. In all cases, C-NS5 loci were considered to be the 'null hypothesis' and E 'the alternative hypothesis', where incongruent lineages for C-NS5 or E are robust. In cases, where incongruence results in a monophyletic group against a paraphyletic group the monophyly is considered to be 'null'.

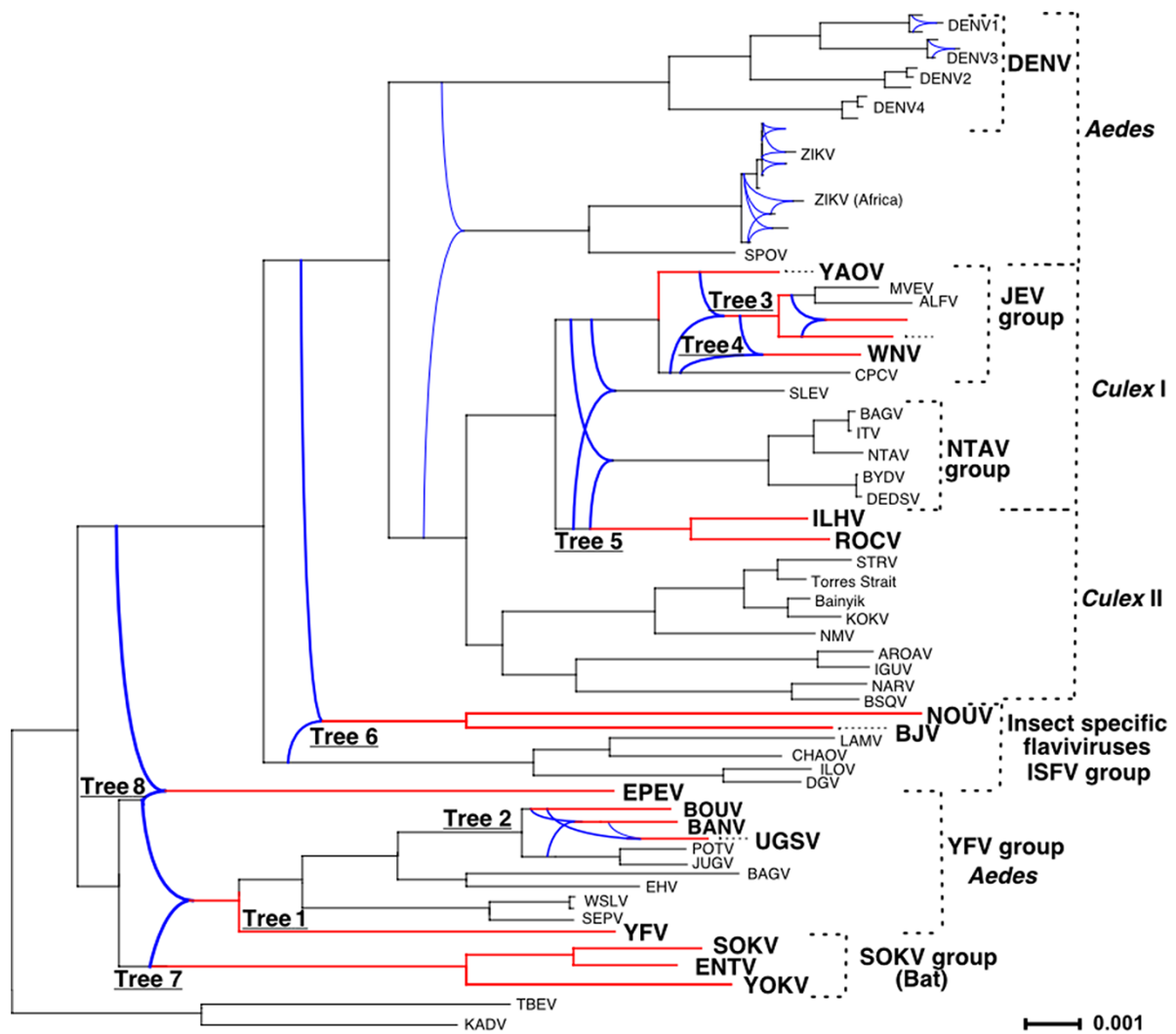


Figure 3. Amino acid maximum likelihood network consensus phylogeny (Dendroscope 3) between the E-protein and C-NS5 protein phylogenies where differences in tree branching order (incongruence) between the E-protein and C-NS5 trees are represented by curved vertical lines that support '8 floating horizontal lines' representing the MBFV members of interest to this study. 'Trees 1-8' denote the eight interspecific recombination events, better defined in Table 1.

3.2.2 Non-parametric bootstrapping and Bayesian phylogenetic analysis

All observations of possible incongruence were confirmed, using Bayesian posterior probabilities for Trees 1-5 and Tree 8 (Section 3.3.3). Trees 6 and 7 which represented the deepest nodes in the MBFV tree (Fig. 4) were not confirmed. The E-protein locus topology for Tree 6 H_1 (EPEV) was supported for a relaxed clock with a dynamic ('skyline') population size at $pp=0.96$, but $pp=0.88-0.93$ for all other models. Secondly, the NS4A topology for Tree 7 H_1 had $pp=0.951$ for a relaxed 'skyline' clock, but $pp=0.77-0.88$ for other models.

3.2.3 Non-parametric bootstrapping summary

Most loci did not show significant incongruence using non-parametric bootstrapping and are not included in Fig. 4. Non-parametric bootstrapping was insufficiently powerful specifically to identify, genome-wide interspecific phylogenetic incongruence for the viruses represented in Trees 1-8. For example, the C-NS5 protein topology Tree 1 H_1 is robustly supported by Bayesian

relaxed clock models but not by non-parametric bootstrapping (Fig. 4d).

However non-parametric bootstrap analysis was useful because H_0 and H_1 for Trees 1-8 could then be examined on a genome-wide basis for a given locus using LRT parametric bootstrapping, a more powerful phylogenetic method.

3.3 Genome recombination: LRTs via parametric bootstrapping

Parametric bootstrapping was used to resolve the null and alternative trees for all loci within the genome for the same phylogenetic test (LG amino acid substitution matrix) under the SOWHAT test (Materials and Methods). The SOWHAT test for phylogenetic incongruence specified in Materials and Methods is a conditional probability where either H_0 must be rejected and the alternative H_1 accepted or vice versa.

All results of the SOWHAT analysis are schematically represented in Fig. 5 and numerically described in Table 2 for MBFV that produce clinically apparent infections (Trees 1-5) and Table 3 for MBFV lineage viruses not associated with a mosquito-vertebrate transmission cycle (Trees 6-8).

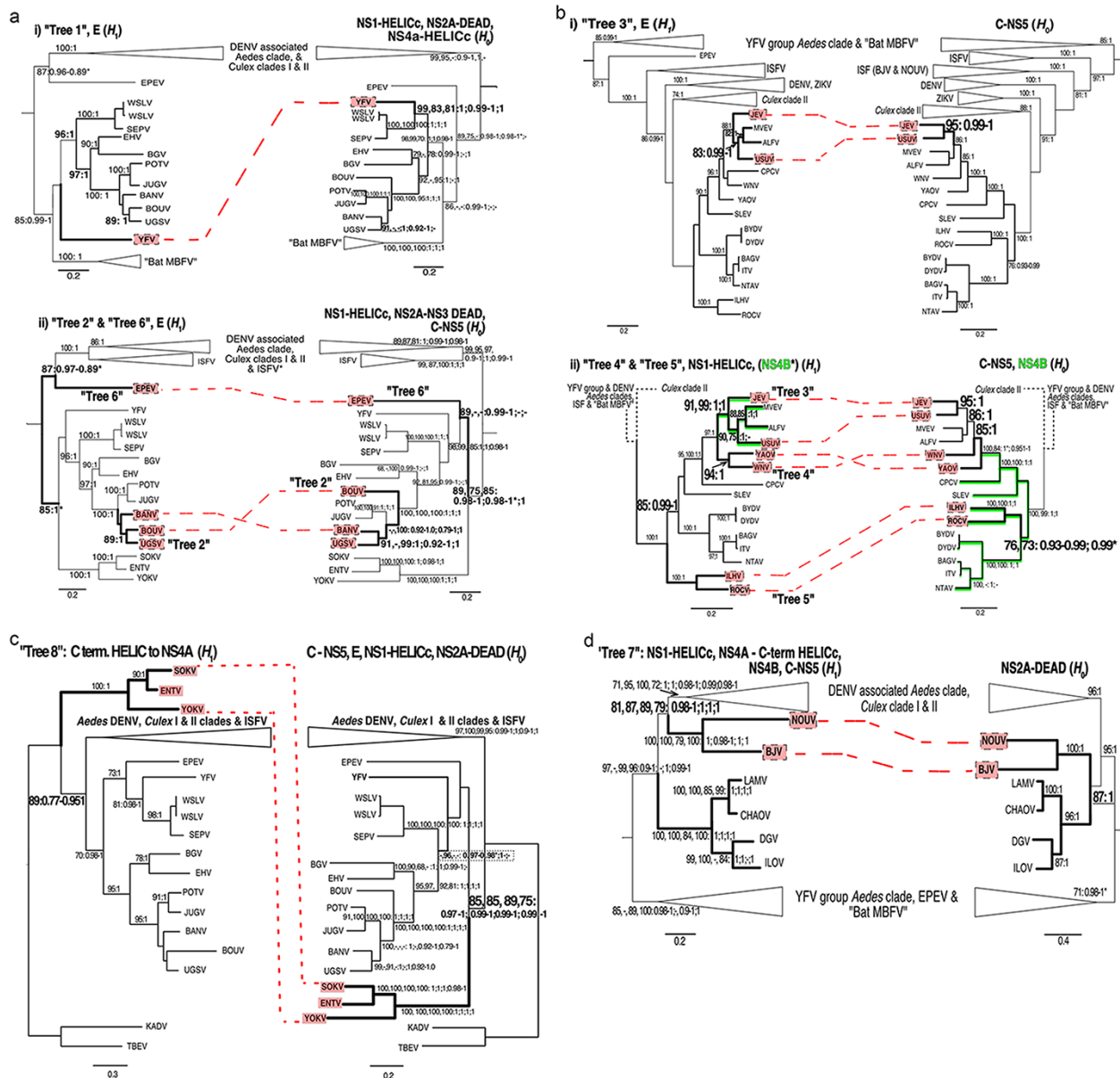


Figure 4. An extensive programme of phylogenetic analyses, involving 50 recognized MBFV lineage members was instigated to look for evidence of widespread interspecific phylogenetic incongruence amongst the MBFV. The figures present amino acid maximum likelihood analyses and posterior probabilities of phylogenetic incongruence for all non-ZIKV MBFV and ISFV using the breakpoints for ZIKV described recently (Gaunt et al. 2020). Numbers above each node are percentage bootstrap values which in most cases are above 75 per cent. The order of the bootstraps and posterior probabilities along each internal branch represents the separate protein trees that identify the protein loci defined immediately above each tree. For example, in Fig 4a i), the left-hand tree represents the E-protein, whilst the right-hand mirror tree depicts separate protein alignments for NS1–NS3 HELICc (bootstrap 'a'), NS2A–NS3 DEAD (bootstrap 'b') and NS4A–NS3 HELICc (bootstrap 'c') and the respective bootstrap scores a, b, c, are indicated above the major branches followed by the posterior probabilities a; b, c. This protocol is adopted for all trees presented in Fig 4. The 'Bat MBFV' in Fig. 4c refers to bat-borne flaviviruses within the MBFV lineage.

3.3.1 Summarized observations of parametric bootstrapping

The SOWHAT technique robustly supported the phylogenetic incongruence hypothesis defined for Trees 1–8 (Table 1). Overall, a total of 47 out of 55 loci could be defined as either tree H_0 or tree H_1 at 1 per cent significance (Fig 5; Tables 2 and 3). Therefore, all SOWHAT observations support and significantly extend the non-parametric bootstrap analysis across each flavivirus genome that revealed interspecific hybrid genomes.

3.3.2 Genome resolution

The parametric bootstrap analysis resulted in 4 of the 8 phylogenetic incongruence topologies being fully resolved as defined in Table 1 across their respective flavivirus genomes for Tree 3 (USUV–JEV), Tree 6 (EPEV), Tree 7 (SOKV group), and Tree 8 (ISFV).

3.3.3 Comparison of robustness between parametric bootstrapping and posterior probabilities

An assessment of parametric bootstrap robustness with posterior probabilities for all 6 Bayesian models deployed (Methods)

Table 1. The null and alternate hypothesis trees defined by robust non-parametric bootstrap analysis described in Fig. 4 in addition to further topologies defined by Bayesian amino acid trees.

	Phylogenetic incongruence	Null hypothesis (H_0) tree	Alternate hypothesis (H_1) tree
Tree 1	YFV ^a	Monophyly with WSLV & SEPV	Basal to the YFV group ^d
Tree 2	UGSV	Sister group with BANV	Sister group with BOUV
Tree 3	USUV	Sister group with JEV	Paraphyly with JEV
Tree 4	WNV ^{b,c}	Paraphyly with YAOV ancestral	Monophyly with YAOV
Tree 5	ILHV & ROCV	Basal to the NTAV group ^d	Basal to the Culex I group ^d
Tree 6	EPEV	Basal to the YFV group	Basal to DENV, ISFV & Culex group
Tree 7	SOKV group ^c	Basal to YFV group & EPEV	Basal to MBFV
Tree 8	BJV & NOUV	Monophyly of all ISFV ^c	Paraphyly with other ISFV & basal to DENV & Culex group

^aA third tree extended Tree 1 H_0 : the YFV, WSLV, and SEPV monophyly included both EHV and BGV.

^bA third tree extending Tree 4 H_0 and is denoted H_2 ; H_2 described WNV, CPCV and YAOV as monophyletic.

^cA further tree extending Tree 4 H_0 is denoted H_2 ; WNV and YAOV remain paraphyletic however WNV defined the immediate ancestral node rather than YAOV.

^dYFV, Culex, SOKV, DENV and ISFV groups are defined in Fig. 3.

showed a high level of congruence between the amino acid tree methods (Fig. 5). The posterior probabilities confirmed phylogenetic incongruence defined by Trees 1–5 and for one locus in Tree 8 that involved the NS2A locus (H_1 ; $pp = 0.99–1.0$). The Bayesian approach was equivocal for the topological incongruence described in Trees 6 and 7 (Section 3.2.2, viz ‘Non-parametric bootstrapping and Bayesian phylogenetic analysis’) and lacked support for the E-protein phylogenetic incongruence for Tree 8.

Overall, the 6 Bayesian models resulted in a total of 40 out of 55 loci being robustly defined as either an H_0 or H_1 tree (Fig. 5), hereafter simplified to ‘robust tree loci’: 38 of these robust tree loci were congruent amongst the 47 robust tree loci defined by parametric bootstrapping; and 2 of these robust tree loci (prM, Tree 1 and NS4A Tree 3) corresponded to loci that could not be resolved by parametric bootstrapping. The two disparities between the methods were due to a third phylogeny that was neither H_0 nor H_1 for the prM locus that involved Tree 1 (YFV; $pp = 0.97–1.0$), as well as the NS4B locus that involved Tree 4 H_0 (WNV; $pp = 1.0$) and was described in the footnotes of Table 1. Most loci lacking posterior probability support were small proteins either prM or NS4A.

3.3.4 Comparison of phylogenetic incongruence between amino acid and nucleotide data

3.3.4.1 Nucleotide analysis on first and second codon position. To assess whether interspecific recombination was the evolutionary process of phylogenetic incongruence it is reasonable to assume the neutral phylogeny would be congruent with the amino acid tree for the same locus. Therefore, restricted taxa nucleotide alignments were produced based on the YFV group and the SLEV groups to minimize substitution saturation as described in Methods (Tree 1–5).

Phylogenetic analysis on first and second codon trees for each of the seven loci (Fig. 1) showed robust congruence with the amino acid trees, Tree 1 (YFV) and Tree 2 (YFV group) being particularly notable examples (Fig. 5). The exceptions were two loci for Tree 3, supported singly by Bayesian posterior probabilities and one locus for Tree 4 supported by non-parametric bootstraps and posterior probabilities described below (Fig. 5). All trees, non-parametric bootstraps, and posterior probabilities are available on request.

Wobble codon alignments of the SLEV and separately the YFV groups were firstly RY-coded (Phillips et al. 2004) (Methods) because both alignments were close to saturation for this codon position with an average p -distance of 0.61–0.64 for all loci (minimum of 0.49 and most maxima <0.72, with minority <0.76 all associated with NMV [Fig. 6]), with the exception of the C-NS5 locus which had an average p -distance of 0.57–0.59 (minimum 0.49 and maximum 0.65). The RY transformed wobble codon data sets were concatenated into single alignments to boost the phylogenetic signal based on whether a given locus supported H_0 or H_1 under the parametric bootstrapping analysis for amino acid data (Fig. 5) and are described in Table 4.

Parametric bootstrapping of RY-coded third codon data produced a clear and striking result for the clinically important MBFV because it did not support phylogenetic incongruence and no alignment for Trees 1–5 unequivocally supported H_1 (Table 4). Genome-wide unequivocal support was observed singly comprising H_0 for Tree 2 genomes, and H_2 for Tree 5 genomes (Table 4). The genome-wide Tree 5 H_2 topology is further supported by non-parametric bootstrap analysis (Fig. 6). Furthermore, the genomes of Tree 1 and Tree 3 in Table 1 represent a genome-wide H_0 topology because the equivocal locus in both cases was H_0 under amino acid analysis (Table 4). In addition, Tree 3 H_0 is supported by non-parametric bootstrapping (Fig. 6). The Tree 4 genomes supported H_2 , defined in Table 1, and only Tree 4 C-NS5 locus supported either H_1 or H_2 , but importantly rejected H_0 which was the amino acid tree.

Thus, the RY-coded third position codon alignments are incongruent with amino acid phylogenies concerning the same loci for either a sizeable region of a given genome for Tree 1–3 or else the entire genome for Tree 4 and 5. This strongly infers that the evolutionary process resulting in phylogenetic incongruence for amino acid data does not involve recombination (Discussion).

The effectiveness of RY-coding for recovering phylogenetic signal from wobble codon data is demonstrated for the C-NS5 locus Tree 4 because the tree is H_1 for both first and second codon position alignments using Bayesian and non-parametric bootstrapping analysis (not shown), H_1 for RY-coded third codon alignments using parametric bootstrapping (Table 4), but no phylogenetic signal was recovered for all three codon positions without RY-coding for any alignment in the analysis (not shown).

3.3.4.2 Amino acid positive selection. Amino acid selection analysis was performed as described in Methods Section 2.7 to assess lineage specific positive selection for Trees 1–5. This analysis is limited by the saturation of transition substitutions, however, this scenario for this precise positive selection analysis has been described as robust via simulation (Gharib and Robinson-Rechavi 2013).

Positive selection where $\omega \gg 1$ is evident in all lineages specifically associated with amino acid incongruence for Trees 1–5 (Table 5), whilst the minima and maximum background ω is 0.002–0.04 across all analyses. If the RY wobble codon tree (H_0 for Trees 1–3 and H_2 for Trees 4–5) was the ‘neutral tree’ and amino acid tree incongruence was driven by selection acting on individual amino

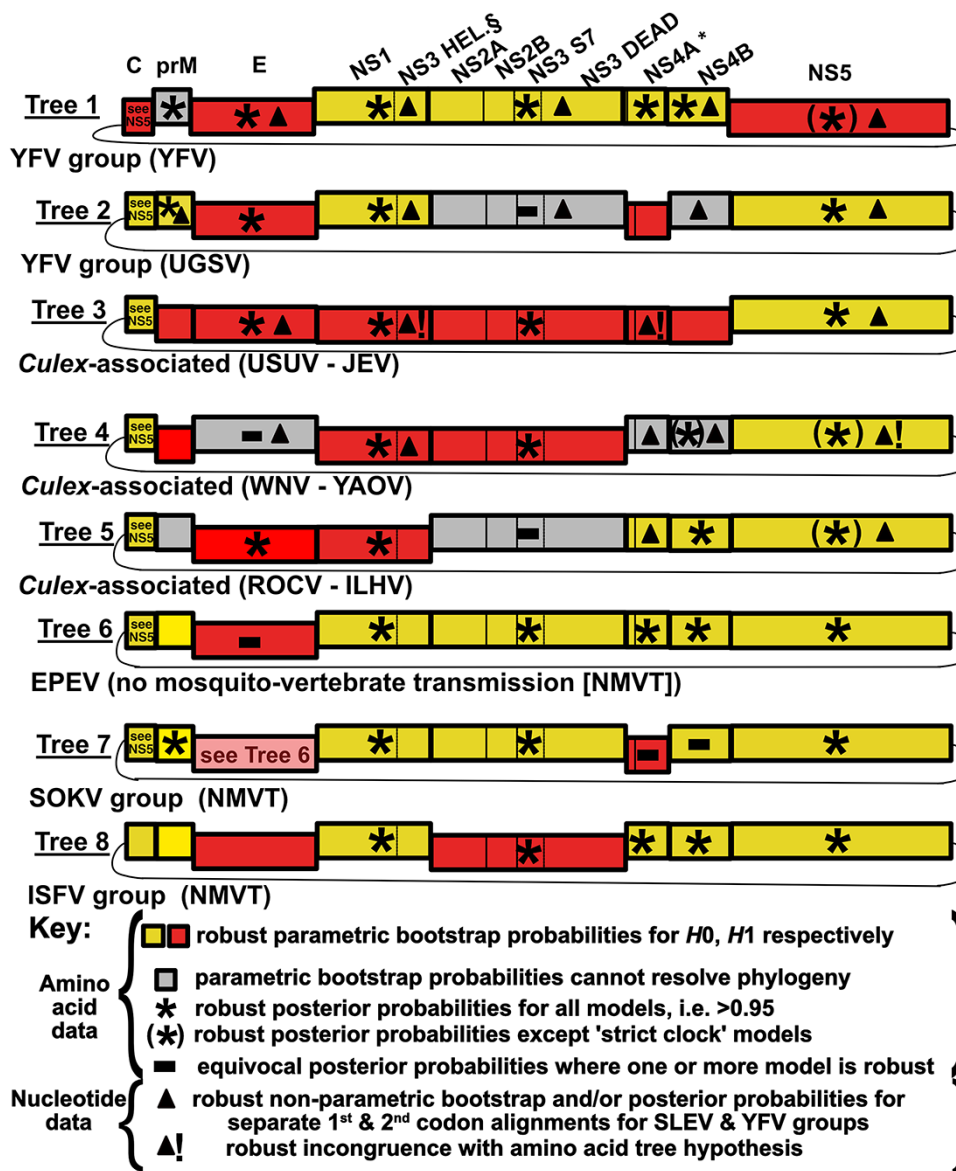


Figure 5. Schematic diagram of the MBFV lineage genome showing phylogenetic incongruence using the SOWHAT test (Tables 2 and 3), amino acid Bayesian posterior probabilities (models 1–6) and first and second codon trees for SLEV and YFV group alignments (Methods). NS3 HEL§ refers to NS3 HELICc and NS4A* refers to the C-terminal NS3 HELICc as well as NS4A. The line below each genome denotes the circularisation during replication connecting C to NS5. All non-parametric bootstraps and posterior probabilities are available on request.

acids then the positive selection in Table 5 would provide an explanation of phylogenetic incongruence for Trees 2, 3, and 4. Under this hypothesis a lack of positive selection ($\omega \ll 1$) is acceptable where the hypothesis topology tested is congruent with the amino acid tree and would explain the pattern of lower ω . However, $\omega \ll 1$ for Tree 1 H_1 topology for NS1 to NS4B and Tree 5 H_2 topology for prM to NS2A–NS3 DEAD, even though both should show $\omega \gg 1$. This is consistent with the prediction of false negatives for ω under simulation conditions of saturation (Gharib and Robinson-Rechavi 2013) and these given branches represent the deepest nodes in this specific analysis.

4. Discussion

Intraspecific (homologous) tree incongruence within a single flavivirus species or serotype has been a focus of virological investigations for decades and was first reported in DENV. Prior to its

recognition, it was believed that genetic diversity, as seen in DENV, was clonal and arose from the accumulation of substitutions (Blok et al. 1992). Subsequently, it was generally believed that tree incongruence was due to recombination and represented an efficient mechanism for generating genetic diversity (Worobey and Holmes 1999). However, it then took another 20 years to demonstrate that interspecific tree incongruence potentially represents an evolutionary process for shaping the epidemiology of flaviviruses. This followed the discovery that the Zika virus genome is a hybrid representing phylogenetic incongruence between *Aedes*-associated MBFV and flaviviruses transmitted primarily by *Culex* species, i.e. related to Japanese encephalitis virus (Gaunt et al. 2020). This discovery explained why different authors had produced conflicting MBFV lineage phylogenies or incongruence (Kuno et al. 1998, 2009; Gaunt et al. 2001; Huhtamo et al. 2009; Grard et al. 2010; Kolodziejek et al. 2013; Moureau et al. 2015; Kuno 2020). Moreover, whilst *Aedes* and/or *Culex* species mosquitoes

Table 2. Interspecific genomic incongruence for the MBFV using the likelihood ratio test performed using parametric bootstrapping to determine between two alternative interspecific trees for the clinically important *Aedes*- and *Culex*-associated viruses, including YFV, UGSV, JEV/USUV, WNV and ROCV.

Tree no.	MBFV Tree	Locus	Likelihood ratio test (SOWHAT)						Phylo. Incon-gruent Gen.?
			Null (H ₀)			Alt. (H ₁)			
			Δ	P	δ	P	δ	P	
Aedes-associated	H ₁ : YFV-basal YFV grp H ₀ : YFV-WSLV/SEP monophyly	E	13.3	<0.01*	0	0.31	0	0.31	Yes
		C-NS5	4.9	<0.01*	6.0e ⁻⁶	0.1	6.0e ⁻⁶	0.1	Unre-solved
		prM	1.9	<0.01*	3.2	<0.01*	3.2	<0.01*	Yes
		NS1-NS3 HELICc	0	0.31	14.9	<0.01*	14.9	<0.01*	Yes
	H ₁ : UGSV-BOUV H ₀ : UGSV-BANV	NS2A-NS3 DEAD	1.0e ⁻⁶	0.99	7.9	<0.01*	7.9	<0.01*	Yes
		3'HELICc-NS4A	7.6e ⁻⁵	0.09	16.5	<0.01*	16.5	<0.01*	Unre-solved
		NS4B	2.6e ⁻³	0.21	1.9	<0.01*	1.9	<0.01*	Yes
		E	22.2	<0.01*	0	0.53	0	0.53	Yes
		3'HELICc-NS4A	1.2	0.057	2.4e ⁻⁵	0.3	2.4e ⁻⁵	0.3	Unre-solved
		NS4B	3.5	<0.01*	2.9	<0.01*	2.9	<0.01*	Yes
Culex-associated	H ₁ : USUV-ALFV/MVEV monophyly i.e. JEV-USUV paraphyly H ₀ : JEV-USUV H ₁ : WNV-YAOV monophyly	NS2A-NS3 DEAD	0.5	<0.01*	3.2	<0.01*	3.2	<0.01*	Yes
		prM	0	0.4	6.7	<0.01*	6.7	<0.01*	Yes
		NS1-NS3 HELICc	0	0.48	19.6	<0.01*	19.6	<0.01*	Yes
		C-NS5	1e ⁻⁶	0.2	47.9	<0.01*	47.9	<0.01*	Yes
	H ₀ : WNV-YAOV paraphyly H ₁ : ROCV/ILHV-basal Culex I	E	7.3	<0.01*	0	0.4	0	0.4	Yes
		3'HELICc-NS4A	2.1	<0.01*	1.6e ⁻⁵	0.2	1.6e ⁻⁵	0.2	Unre-solved
		NS1-NS3 HELICc	17.1	<0.01*	0	0.46	0	0.46	Yes
		NS2A-NS3 DEAD	5.2	<0.01*	2e ⁻⁶	0.41	2e ⁻⁶	0.41	Unre-solved
		NS4B	2.1	0.01	9e ⁻⁶	0.38	9e ⁻⁶	0.38	Yes
		prM	0.36	0.01	0.56	0.13	0.56	0.13	Yes
H ₀ : ROCV/ILHV-NTAV grp (Culex I) monophyly	C-NS5	2.9e ⁻⁵	0.13	27.3	<0.01*	27.3	<0.01*	Yes	
	prM	4.9	<0.01*	7e ⁻⁶	0.6	7e ⁻⁶	0.6	Yes	
	NS1-NS3 HELICc	16.8	<0.01*	0	0.39	0	0.39	Unre-solved	
	NS2A-NS3 DEAD	8.6	<0.01*	3e ⁻⁶	0.29	3e ⁻⁶	0.29	Yes	
	E	3.8	<0.01*	4.6	<0.01*	4.6	<0.01*	Unre-solved	
	3'HELICc-NS4A	3.1	<0.01*	3.84	<0.01*	3.84	<0.01*	Yes	
	NS4B	9.6	<0.01*	10.18	<0.01*	10.18	<0.01*	Yes	
	C-NS5	1e ⁻⁶	0.3	13.5	<0.01*	13.5	<0.01*	Yes	
	E	4.3	<0.01*	0	0.33	0	0.33	Yes	
	H ₀ : ROCV/ILHV-NTAV grp (Culex I) monophyly	NS1-NS3 HELICc	7.9	<0.01*	0	0.39	0	0.39	Not sig.
prM		4.4	<0.01*	0.09	0.04	0.09	0.04	Unre-solved	
NS2A-NS3 DEAD		3.2	<0.01*	9.6	<0.01*	9.6	<0.01*	Yes	
3'HELICc-NS4A		4e ⁻⁶	0.27	5.4	<0.01*	5.4	<0.01*	Yes	
NS4B		0.001	0.2	12.9	<0.01*	12.9	<0.01*	Yes	
C-NS5	7e ⁻⁵	<0.01*	10.2	<0.01*	10.2	<0.01*	Yes		

Note: '.' replaces '-' for brevity.
* = significant at 1 per cent.

Table 3. Interspecific genomic incongruence for the MBFV lineage as described in Table 1 to distinguish between two alternative interspecific trees for all members which are not associated with a dual-host-affiliated insect-specific flavivirus.

	Tre. no.	MBFV Tree	Locus	Likelihood ratio test (SOWH)				Genome(s) phylogenet. incongruent?
				Null (H_0)		Alt. (H_1)		
				δ	P	δ	P	
Sandfly-associated	6	H_1 :EPEV basal to DENV & <i>Culex</i>	E	29.2	<0.01*	0	0.47	Yes
			prM	$2e^{-6}$	0.54	3.2	<0.01*	
		H_{01} : EPEV basal to YFV & SOKV grps	NS1-NS3 HELICc	0	0.27	21.0	<0.01*	Yes n.b. H_{01} identical to Tree7 H_0
			NS2A-NS3 DEAD	0	0.54	8.4	<0.01*	
			NS4B	$-1.3e^{-3}$	1	1.8	<0.01*	
		H_{02} : EPEV basal singly to YFV grp	3' HELICc-NS4A	$1.3e^{-5}$	0.2	14.2	<0.01*	
			C-NS5	$-3e^{-6}$	1	22.4	<0.01*	Yes
Bat-associated	7	H_1 :SOKV grp—basal MBFV	3'HELICc—NS4A	6.9	<0.01*	$5.2e^{-5}$	0.19	Yes
			E	6.3	<0.01*	8.2	<0.01*	See Tree 6 H_1 :EPEV-DENV/ <i>Culex</i>
		H_0 : SOKV grp basal to YFV grp	prM	$2e^{-6}$	0.54	13.3	<0.01*	
			NS1-NS3 HELICc	0	0.27	21.0	<0.01*	
			NS2A-NS3 DEAD	0	0.54	5.4	<0.01*	Yes
			NS4B	$-1.3e^{-3}$	0.91	1.75	<0.01*	
			C-NS5	0	0.97	18.9	<0.01*	
Insect-specific	8	H_1 : ISF paraphyly—to DENV & <i>Culex</i>	E	0	0.67	121.3	<0.01*	
			NS2A-NS3 DEAD	$1e^{-6}$	0.27	13.9	<0.01*	Yes
		H_0 : ISF monophyly	prM	4.4	<0.01*	0	0.61	
			NS1-NS3 HELICc	7.7	<0.01*	0	0.48	
			3'HELICc—NS4A	0	<0.01*	$8e^{-6}$	0.26	Yes
			NS4B	16.2	<0.01*	0	0.99	
			C-NS5	5.5	<0.01*	$1e^{-6}$	0.94	

n.b. Tree 6 H_{01} and Tree 7 H_0 are the same tree because EPEV is basal to YFV and SOKV groups and the SOKV group is basal to the YFV group. However, Tree 6 H_1 and Tree 7 H_1 are very different trees.

* = significant at 1 per cent; grp = group; grps = groups. See 'Note' in Table 2.

are competent to transmit both ZIKV and DENV in their natural settings, one species in particular viz., the anthropophilic, urban-dwelling *Aedes aegypti* is the primary transmission vector of DENV and ZIKV throughout the tropics and sub-tropics (Boyer et al. 2018). Indeed, there are many reports of mixed infections in field-isolated *Ae. aegypti* (Chaves et al. 2018), inferring epidemiological competition between these viruses in regions where they overlap geographically. Moreover, the T-cell immune response of the ZIKV hybrid genome is strongly represented by T-cell loci obtained from *Culex*-associated flaviviruses rather than those of its immediate DENV ancestor. Thus, interspecific phylogenetic incongruence provided ZIKV with a potential genetic adaptation to escape the immune barrier presented by DENV (Gaunt et al. 2020). As far as we are aware, the putative immunological selection pressure for ZIKV interspecific phylogenetic incongruence makes the observation unique. For example, interspecific recombination was reported for *Hepacivirus* (Thézé et al. 2015), but this was not linked with an identifiable selection pressure. The significance of interspecific tree incongruence extends far beyond viral genetics, evolution, and systematics. Indeed if this is a consistent

evolutionary process it could have wide-ranging potential impacts on vector or host specificity, immune response in vertebrates, virulence or disease severity, serological diagnostics, and other aspects of epidemiology.

Based on phylogenetic evidence, the *Culex*-associated flaviviruses are the direct descendants of the *Aedes*-associated viruses (Gaunt et al. 2001; Gould et al. 2001) many of which, include the ancestors of JEV, USUV, WNV, and YAOV, that share the sympatric environment of the African rainforest with the *Aedes*-associated viruses including BANV, BOUV, UGVS, and YFV. We therefore proposed that tree incongruence for ZIKV–SPOV between the DENV and *Culex* spp. MBFV was an evolutionary adaptation of ZIKV and SPOV to this ecoregion driven by selective immune pressure from T-cell epitope rich loci within the *Culex*-associated MBFV, effectively altering the immunological phenotype of ZIKV and SPOV, potentially shaping their epidemiology in terms of emerging arboviruses. In this context, SPOV was first identified outside Africa in 2016 when it was isolated from a pool of *Culex quinquefasciatus* mosquitoes in Haiti (White et al. 2018). Thus, in common with other flaviviruses, alphaviruses,

Table 4. Interspecific genomic incongruence for the RY-coded third codon nucleotide data of the MBFV using the likelihood ratio test performed using parametric bootstrapping to determine between two alternative interspecific trees for the clinically important *Aedes*- and *Culex*-associated viruses, including YFV, UGVS, JEV/USUV, WNV and ROCV. See 'Note' in Table 2.

	Tree no.	MBFV Tree	Combined Locus ^a	Likelihood ratio test (SOWH)						Robust Typology
				Null (H_0)		Alt. (H_1)		Alt (H_2)		
				δ	P	δ	P	δ	P	
<i>Aedes</i> -associated	1	H_1 : YFV-basal YFV grp	prM, E, C-NS5	0.0	0.37	0.774	0.014	-	-	H_0
		H_0 : YFV-WSLV/SEP monophyly	NS1-NS3 HELICc, NS2A-NS3 DEAD, 3'HELICc-NS4A, NS4B	0.0	0.99	0.0	0.994	-	-	None
	2	H_1 : UGVS-BOUV	E, 3'HELICc-NS4A,	0.0	0.34	2.82	0.036	-	-	H_0
		H_0 : UGVS-BANV	prM, NS1-NS3 HELICc, C-NS5	0.014	0.66	2.54	0.044	-	-	H_0
<i>Culex</i> -associated	3	H_1 : USUV-ALFV/MVEV monophyly (JEV-USUV paraphyly)	prM, E, NS1-NS3 HELICc, NS2A-NS3 DEAD, 3'HELICc-NS4A, NS4B	0.02	0.12	3.9	0.008	-	-	H_0
		H_0 : JEV-USUV	C-NS5	0.653	0.17	0.524	0.276	-	-	None
	4	H_1 : WNV-YAOV monophyly	prM, NS1-NS3 HELICc, NS2A-NS3 DEAD	2.67	0.046	1.52	0.01	0.0	0.24	H_2
		H_0 : WNV-YAOV paraphyly	NS5	0.17	0.048	0.0	0.996	1.11	0.14	H_1 or H_2
	5	H_1 : ROCV/ILHV-basal <i>Culex</i> I	prM, E, NS1-NS3 HELICc, NS2A-NS3 DEAD	4.2	0.032	3.32	0.16	0.0	0.24	H_2
H_0 : ROCV/ILHV-NTAV gp (<i>Culex</i> I) monophyly		3'HELICc-NS4A, NS4B, C-NS5	2.35	0.036	2.30	0.024	0.0	0.83	H_2	

^aCombined locus of wobble codons comprising all the genes listed for a given row as described in Fig. 1.

and bunyaviruses, SPOV has the potential to emerge in the New World.

Identification of interspecific hybrid genomes using tree incongruence is not unique to ZIKV and SPOV. Hybrid genomes were identified in a further 10 of 50 MBFV-related flaviviruses analysed by both parametric bootstrapping and Bayesian amino acid trees. This currently represents the widest diversity of interspecific tree incongruence amongst the positive-stranded RNA arboviruses to our knowledge. Phylogenetic incongruence that involved recent, or shallow nodes on the MBFV phylogeny represent extant (non-extinct) MBFV lineage viruses of clinical importance. They are dual-host-affiliated insect-specific flaviviruses (dISF) that are characterized. The phylogenetic incongruence identified here that is basal within the phylogenetic tree, is restricted to MBFV lineage viruses that lack dual-host affiliation, viz., sandfly-vectored MBFV lineage virus (EPEV; Tree 6), the SOKV group (bat associated; Tree 7), and the ISFV group (Tree 8).

Epidemiologically, both ZIKV and SPOV originated and co-circulate in an African sylvatic environment common to other pathogenic and clinically important MBFV including DENV, UGVS, WNV, and YFV. In context, the ZIKV and SPOV interspecific phylogenetic incongruence breakpoints can be defined as 'hotspots' because they are far from unique within the MBFV. The study was facilitated by the statistical phylogenetic power of parametric bootstrapping, which has received minimal attention in virology, and it was broadly supported by Bayesian phylogenetic models. Parametric bootstrapping is applicable because it

repeatedly produced one alternative tree hypothesis for a given member of the MBFV lineage and Bayesian analysis identified the possibility of a third phylogeny. This approach provides a template to look for further and more powerful evidence of arbovirus interspecific phylogenetic incongruence.

Viral phylogenetic incongruence could result from an evolutionary process involving recombination (reviewed by Simon-Loriere and Holmes 2011; Thézé et al. 2015) and selection or adaptation particularly through parallel or convergent evolution (reviewed by Stern et al. 2017; Geoghegan and Holmes 2018), tree-building methods (Felsenstein 1978; Huelsenbeck 1997; Alfaro et al. 2003; Drummond et al. 2005; Dávalos et al. 2012), or systematic bias in the nucleotide data (Phillips et al. 2004). Both recombination and parallel evolution have been identified as possible virulence factors through *in vivo* virus passage (Stern et al. 2017).

In this study the congruence between maximum likelihood parametric bootstrapping and Bayesian methods both for amino acid and first and second codon positions infers that interspecific phylogenetic incongruence is an evolutionary process rather than a methodological discrepancy, notably for the viruses of clinical importance and including YFV.

However, MBFV members that lack a dual-host transmission cycle, notably EPEV and the SOKV group demonstrated phylogenetic incongruence between viruses in deep (ancestral) nodes with a single locus exception (ISFV), only for non-parametric and parametric bootstrapping. This suggests caution for interpreting

Table 5. The nonsynonymous to synonymous substitution rate ratio for incongruent lineages was calculated under a maximum likelihood branch-specific model. The numbers underlined simply highlight that the hypothesis tree being tested is congruent with the amino acid tree for the combined locus under investigation. See 'Note' in Table 2.

Tree no.	MBFV Tree	Combined Locus ^a	d_N/d_S (ω) ^b			Positive selection	
			Null (H_0)	Alt. (H_1)	Alt (H_2)		
Aedes-associated	1	H_1 : YFV-basal YFV grp	prM, E, C-NS5	7.9	0.25	–	H_0 only
		H_0 : YFV-WSLV/SEP monophyly	NS1-NS3 HELICc, NS2A-NS3 DEAD, 3'HELICc-NS4A, NS4B	0.03	0.008	–	No
	2	H_1 UGSV-BOUV	E, 3'HELICc-NS4A,	360.1	3.3	–	H_0 & H_1
		H_0 : UGSV-BANV	prM, NS1-NS3 HELICc, C-NS5	0.002	176.5	–	H_1 only
Culex-associated	3	H_1 : USUV-ALFV/MVEV monophyly (JEV-USUV paraphyly)	prM, E, NS1-NS3 HELICc, NS2A-NS3 DEAD, 3'HELICc-NS4A, NS4B	149.3	24.6	–	H_0 & H_1
		H_0 : JEV-USUV	C-NS5	333.0	999.0	–	H_0 & H_1
	4	H_1 : WNV-YAOV monophyly	prM, NS1-NS3 HELICc, NS2A-NS3 DEAD	16.66	4.9	42.8	H_0, H_1, H_2
		H_0 : WNV-YAOV paraphyly	NS5	0.059	51.7	10.2	H_1 & H_2
	5	H_1 : ROCV/ILHV-basal Culex I	prM, E, NS1-NS3 HELICc, NS2A-NS3 DEAD	33.8	4.1	0.01	H_0 & H_1
	H_0 : ROCV/ILHV-NTAV gp (Culex I) monophyly	3'HELICc-NS4A, NS4B, C-NS5	577.4	29.0	42.9	H_0, H_1, H_2	

^aCombined locus of all three codon positions comprising all the genes listed for a given row as described in Fig. 1. The locus will represent H_0 or H_1 for amino acid data described in Fig. 5.

^bThe background ω is 0.002–0.04, the largest value being for Tree 1 H_0 , the prM, E and NS5 locus.

phylogenetic incongruence at deep nodes based on single methods.

RY-coding of the wobble codon used in this study, albeit comparatively novel to virus phylogenetics, is a mainstream approach for overcoming compositional bias in insect evolution with a long history (Delsuc, Phillips, and Penny 2003; Breinholt and Kawahara 2013; Song et al. 2016; Yang et al. 2018). RY-coding has performed strongly in simulation studies using maximum likelihood phylogenetic analysis (Phillips et al. 2004; Ishikawa et al. 2012). In this study RY-coding was restricted to incongruent taxa that are not subject to complete saturation and is therefore a conservative application. Parametric and non-parametric bootstrapping of the RY-coded trees either inferred or robustly concluded there was a single phylogeny for all the clinically important MBFV and did not support genome-wide phylogenetic incongruence. This splits the analysis into: 'between locus incongruence' and; 'within locus incongruence'. The 'within locus incongruence' is phylogenetic incongruence between data-types predominantly amino acid trees, but also first and second codon position trees against the RY-coded wobble codon trees. Regardless of whether RY-coded wobble codons are 'neutral', it is difficult to consider interspecific recombination as the evolutionary process causing 'between locus phylogenetic incongruence' for the clinically important MBFV. Moreover, for 2 of the 5 occurrences of 'within

locus incongruence', both being SLEV group viruses, there was complete phylogenetic incongruence between data-types across their genomes, viz. Tree 4 and 5.

Positive selection analyses (ω) for the clinically important MBFV support an explanation that interspecific phylogenetic incongruence was caused by selection pressure on amino acids physically localized within a locus resulting in similar or identical mutations after two or more lineages diverged. Two specific evolutionary processes that could generate interspecific 'between and within locus' phylogenetic incongruence are parallel and convergent evolution. It is important to reiterate simulation studies on saturated neutral substitutions infer the PAML method used (Yang and Nielsen 2002), was strongly biased towards false negative ω rather than false positives (Gharib and Robinson-Rechavi 2013). However, further simulation studies on recombination and selection available through forward in time methods (Peng et al. 2015; Haller and Messer 2019; Jariani et al. 2019) would be beneficial to support the findings of Walid and Robinson-Rechavi (2013). Positive selection is a very clear substitution process that could result in the same locus producing two opposing trees.

Yellow fever (Tree 1) is a particularly interesting example of interspecific 'between loci' phylogenetic incongruence within the flaviviruses where the E and C-NS5 loci are phylogenetically basal

R_Y code wobble codon trees

Tree 3: prM, E, NS1-HELICc, NS2a-DEAD, NS4A, NS4B

Tree 5 (i) : prM, E, NS1-HELICc, NS2a-DEAD

Tree 5 (ii): NS4A, NS4B, C-NS5

Bootstraps: Tree 3, Tree 5 (i), Tree 5 (ii)

Amino acid trees

Bootstraps: posterior prob.: E, NS1-HELICc, (NS4B*)

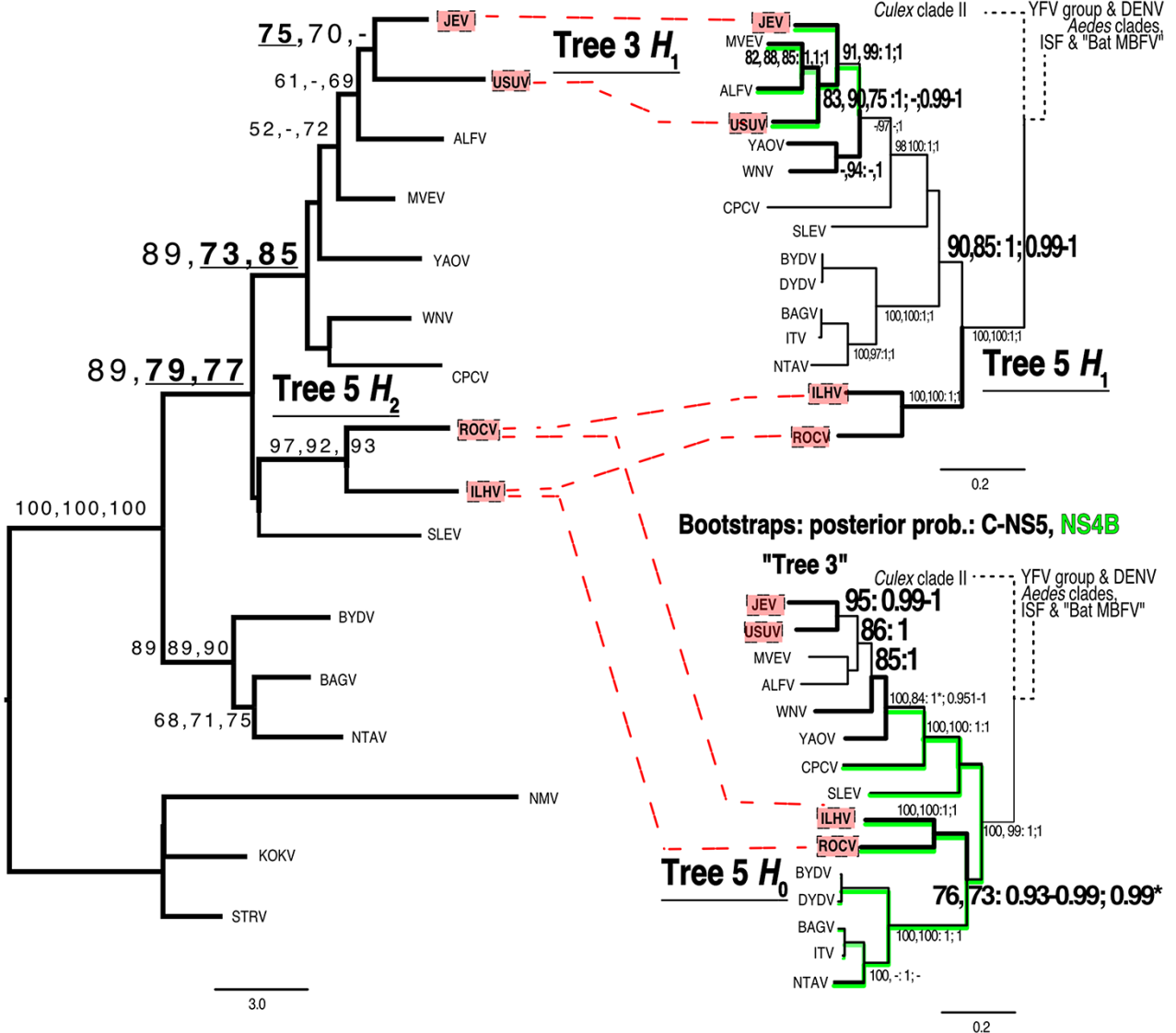


Figure 6. R_Y-coded wobble codon maximum likelihood trees from concatenated loci and their comparison to single locus amino acid maximum likelihood and Bayesian trees. Note the Bayesian MCMC would not converge for the R_Y-codon 'wobble codon data' but did achieve convergence for non-R_Y-coded, nucleotide data (Fig. 5).

within the YFV-related group, whereas all NS-proteins form sister groups with WSLV and SEPV. This would be expected to have profound immunological implications during vertebrate host infections for example by removing major CD4⁺ and CD8⁺ T-cell epitopes normally expressed by the original parent, of YFV. The implications are that WSLV and SEPV could generate T-cell cross-protective responses against YFV and this might, in part, explain the absence of YFV in New Guinea where SEPV is known to be present (Kuno 2020). This potential scenario has parallels with the ZIKV and SPOV interspecific hybrid genome/proteome associated with minimizing the cross-protective CD4⁺ and CD8⁺

T-cell immune responses resulting from previous DENV infections (Gaunt et al. 2020), but which rendered it susceptible to cross-protective responses induced by the recombinant *Culex*-associated MBFV.

Given the diverse range of flaviviruses from arthropods, mammals, and birds, there is unlikely to be a singular selective force driving interspecific phylogenetic incongruence, with its consequent impact on flavivirus pathogenesis and epidemiology. For example, evidence of interspecific phylogenetic incongruence involving the C-NS5 protein was detected in USUV, JEV and probably in WNV and YAOV. This would strongly influence the cross-

reactive CD4⁺ response, in an infected host if it was comparable to that generated against DENV, and presumably would also alter virus replication characteristics.

Interspecific phylogenetic incongruence that resulted from interspecific recombination should not be excluded and could be facilitated by a range of ecologic mechanisms for example co-circulation and co-infections. Flavivirus infections in mosquito larvae were first reported in 1938 during YFV studies (Whitman and Antunes 1938) and also identified in DENV, SLEV, and WNV (Bara et al. 2013; Kuno 2016; Du et al. 2019). One of several possible mechanisms of co-infection is when an infective larva dies, other nearby larvae become infected through ingestion (Bara et al. 2013; Kuno 2016; Kuno et al. 2017; Du et al. 2019) and in addition there are many predators of mosquito larvae including birds. Mosquito vertical transmission is an alternative or contributory method of transmission and could play an essential role in co-infection (Burgdorfer and Varma 1967; Lequime et al. 2016; Tesh et al. 2016).

All the viruses in this study with known vectors or close association with those viruses in the clade structure, demonstrated phylogenetic incongruence and remain within the phyloepidemiological boundaries that have been defined previously (Gaunt et al. 2001; Gould et al. 2001). This fundamental observation makes sense for the necessity of co-infection of the vector and/or vertebrate, generating mixed infections in vector and/or host cells as a prerequisite for either recombination or adaptation/selection involving co-circulation.

The interspecific phylogenetic incongruence inferred here has many potential implications, not just for evolution, emergence, and biodiversity but also for their future potential impact on the development and widespread use of live-attenuated chimeric vaccines in regions of the world where co-infections between related flaviviruses are known to occur (Seligman and Gould 2004). Interspecific co-infections increase the potential for recombination and/or different mechanisms of adaptation such as convergent evolution. There is a consensus that MBFV transmission in general is not a constant parameter due to shifting environmental factors (French and Holmes 2020). For example, African YFV disperses sequentially from a sylvatic, to perisylvatic environment and then to the urban ecoregion of large towns and cities (Christopher 1960), where the domestic variant of *Ae. aegypti* is the primary transmission vector. In this context co-circulation, a prerequisite for interspecific recombination, could be temporally dynamic and environmentally sensitive. The potentially damaging global impact of human activities on the environment, including climate change, deforestation, pollution, urbanization, exploitation of fossil fuels, and widespread distribution of live-attenuated chimeric vaccines, must not be overlooked.

Supplementary data

Supplementary data is available at *Virus Evolution* online.

Acknowledgements

We thank Steve Whitbread for technical assistance with the high-performance computer (HPC) at LSHTM and EPH Faculty, LSHTM for access. In addition, we thank Google Cloud Platform (GCP) for HPC services. We wish to extend thanks to the Associate Editor.

Funding

This work was initiated under the European Commission Horizon 2020 Zika virus consortia grant program primarily through ZIKAlliance Project number (pn) 734548 and ZikaPLAN pn 734584.

Conflict of interest: The authors declare no competing interests.

References

- Aaskov, J. et al. (2007) 'Multiple Recombinant Dengue Type 1 Viruses in an Isolate from a Dengue Patient'. *Journal of General Virology*, 88: 3334–40.
- Alfaro, M. E. et al. (2003) 'Bayes or Bootstrap? A Simulation Study Comparing the Performance of Bayesian Markov Chain Monte Carlo Sampling and Bootstrapping in Assessing Phylogenetic Confidence', *Molecular Biology and Evolution*, 20: 255–66.
- Alkan, C. et al. (2015) 'Ecuador Paraiso Escondido Virus, a New Flavivirus Isolated from New World Sand Flies in Ecuador, Is the First Representative of a Novel Clade in the Genus Flavivirus', *Journal of Virology*, 89: 11773–85.
- Baidaliuk, A. et al. (2020) 'Novel Genome Sequences of Cell-fusing Agent Virus Allow Comparison of Virus Phylogeny with the Genetic Structure of *Aedes Aegypti* Populations', *Virus Evolution*, 6: veaa018.
- Bara, J. J. et al. (2013) 'Susceptibility of Larval *Aedes Aegypti* and *Aedes Albopictus* (Diptera: Culicidae) to Dengue Virus', *Journal of Medical Entomology*, 50: 179–84.
- Blok, J. et al. (1992) 'Comparison of a Dengue-2 Virus and Its Candidate Vaccine Derivative: Sequence Relationships with the Flaviviruses and Other Viruses', *Virology*, 187: 573–90.
- Boyer, S. et al. (2018) 'An Overview of Mosquito Vectors of Zika Virus', *Microbes and Infection*, 20: 646–60.
- Breinholt, J., and Kawahara, A. (2013) 'Phylotranscriptomics: Saturated Third Codon Positions Radically Influence the Estimation of Trees Based on Next-gen Data', *Genome Biology and Evolution*, 5: 2082–92.
- Burgdorfer, W., and Varma, M. G. (1967) 'Trans-stadial and Transovarial Development of Disease Agents in Arthropods', *Annual Review of Entomology*, 12: 347–76.
- Carney, J. et al. (2012) 'Recombination and Positive Selection Identified in Complete Genome Sequences of Japanese Encephalitis Virus', *Archives of Virology*, 157: 75–83.
- Chaves, B. et al. (2018) 'Coinfection with Zika Virus (ZIKV) and Dengue Virus Results in Preferential ZIKV Transmission by Vector Bite to Vertebrate Host', *The Journal of Infectious Diseases*, 218: 563–71.
- Christopher, S. R. (1960) in (ed.) *Aedes Aegypti (L.): The Yellow Fever Mosquito: Its Life History, Bionomics and Structure*, pp. 57–60. Cambridge University Press: Cambridge.
- Church, S. H. et al. (2015) 'Automation and Evaluation of the SOWH Test with SOWHAT', *Systematic Biology*, 64: 1048–58.
- Cook, S. et al. (2012) 'Molecular Evolution of the Insect-specific Flaviviruses', *Journal of General Virology*, 93: 223–34.
- Cusick, M. F. et al. (2011) 'Hepatitis C Virus Induces Regulatory T Cells by Naturally Occurring Viral Variants to Suppress T Cell Responses', *Clinical & Developmental Immunology*, 2011: 806061.
- Dávalos, L. et al. (2012) 'Understanding Phylogenetic Incongruence: Lessons from Phyllostomid Bats', *Biological Reviews*, 87: 991–1024.
- Delsuc, F., Phillips, M., and Penny, D. (2003) 'Comment on "Hexapod Origins: Monophyletic or Paraphyletic?"' *Science*, 301: 5639.

- Douady, C. J. et al. (2003) 'Comparison of Bayesian and Maximum Likelihood Bootstrap Measures of Phylogenetic Reliability', *Molecular Biology and Evolution*, 20: 248–54.
- Drummond, A. J. et al. (2005) 'Bayesian Coalescent Inference of past Population Dynamics from Molecular Sequences', *Molecular Biology and Evolution*, 22: 1185–92.
- et al. (2006) 'Relaxed Phylogenetics and Dating with Confidence', *PLoS Biology*, 4: e88.
- Du, S. et al. (2019) 'Aedes Mosquitoes Acquire and Transmit Zika Virus by Breeding in Contaminated Aquatic Environments', *Nature Communications*, 10: 1324.
- Faye, O. et al. (2014) 'Molecular Evolution of Zika Virus during Its Emergence in the 20(th) Century', *PLoS Neglected Tropical Diseases*, 8: e2636.
- Felsenstein, J. (1978) 'Cases in Which Parsimony or Compatibility Methods Will Be Positively Misleading', *Systematic Biology*, 27: 401–10.
- Forslund, K. et al. (2004) 'VisRD-visual Recombination Detection', *Bioinformatics*, 20: 3654–5.
- French, R. K., and Holmes, E. C. (2020) 'An Ecosystems Perspective on Virus Evolution and Emergence', *Trends in Microbiology*, 28: 165–75.
- Gaunt, M. W. et al. (2001) 'Phylogenetic Relationships of Flaviviruses Correlate with Their Epidemiology, Disease Association and Biogeography', *Journal of General Virology*, 82: 1867–76.
- et al. (2020) 'Recombination of B- and T-cell Epitope-rich Loci from Aedes- and Culex-borne Flaviviruses Shapes Zika Virus Epidemiology', *Antiviral Research*, 174: 104676.
- Geoghegan, J. L., and Holmes, E. C. (2018) 'The Phylogenomics of Evolving Virus Virulence', *Nature Reviews: Genetics*, 19: 756–69.
- Gharib, W. H., and Robinson-Rechavi, M. (2013) 'The Branch-Site Test of Positive Selection Is Surprisingly Robust but Lacks Power under Synonymous Substitution Saturation and Variation in GC', *Molecular Biology and Evolution*, 30: 1675–86.
- Goldman, N. (1993) 'Statistical Tests of Models of DNA Substitution', *Journal of Molecular Evolution*, 36: 182–98.
- et al. (2000) 'Likelihood-based Tests of Topologies in Phylogenetics', *Systematic Biology*, 49: 652–70.
- Gould, E. A. et al. (2001) 'Evolution, Epidemiology, and Dispersal of Flaviviruses Revealed by Molecular Phylogenies', *Advances in Virus Research*, 57: 71–103.
- Grard, G. et al. (2010) 'Genomics and Evolution of Aedes-borne Flaviviruses', *Journal of General Virology*, 91: 87–94.
- Haller, B. C., and Messer, P. W. (2019) 'SLiM 3: Forward Genetic Simulations beyond the Wright–Fisher Model', *Molecular Biology and Evolution*, 36: 632–7.
- Hillis, D. M. et al. (1996) 'Applications of Molecular Systematics: The State of the Field and a Look to the Future', in Hillis, D. M., Moritz, C., and Mable, B. K. (eds) *Molecular Systematics*, pp. 515–43. Sunderland, MA: Sinauer.
- Holmes, E. C. et al. (1999) 'Phylogenetic Evidence for Recombination in Dengue Virus', *Molecular Biology and Evolution*, 16: 405–9.
- Huelsenbeck, J. P. (1997) 'Is the Felsenstein Zone a Fly Trap?' *Systematic Biology*, 46: 69–74.
- Huelsenbeck, J. P., and Bull, J. J. (1996) 'A Likelihood Ratio Test to Detect Conflicting Phylogenetic Signal', *Systematic Biology*, 45: 92–8.
- Huerta-Cepas, J. et al. (2016) 'ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data', *Molecular Biology and Evolution*, 33: 1635–8.
- Huhtamo, E. et al. (2009) 'Characterization of a Novel Flavivirus from Mosquitoes in Northern Europe that Is Related to Mosquito-borne Flaviviruses of the Tropics', *Journal of Virology*, 83: 9532–40.
- et al. (2014) 'Novel Flaviviruses from Mosquitoes: Mosquito-specific Evolutionary Lineages within the Phylogenetic Group of Mosquito-borne Flaviviruses', *Virology*, 464–465: 320–9.
- Huson, D. H., and Scornavacca, C. (2012) 'Dendroscope 3: An Interactive Tool for Rooted Phylogenetic Trees and Networks', *Systematic Biology*, 61: 1061–7.
- Ishikawa, S. A. et al. (2012) 'RY-coding and Non-homogeneous Models Can Ameliorate the Maximum-likelihood Inferences from Nucleotide Sequence Data with Parallel Compositional Heterogeneity', *Evolutionary Bioinformatics*, 8: 357.
- Jariani, A. et al. (2019) 'SANTA-SIM: Simulating Viral Sequence Evolution Dynamics under Selection and Recombination', *Virus Evolution*, 5: 1.
- Jones, D. T., and Kandathil, S. M. (2018) 'High Precision in Protein Contact Prediction Using Fully Convolutional Neural Networks and Minimal Sequence Features', *Bioinformatics*, 34: 3308–15.
- Kalinina, O. et al. (2002) 'A Natural Intergenotypic Recombinant of Hepatitis C Virus Identified in St. Petersburg', *Journal of Virology*, 76: 4034–43.
- Kolodziejek, J. et al. (2013) 'Barkedji Virus, a Novel Mosquito-borne Flavivirus Identified in Culex Perexiguus Mosquitoes, Israel, 2011', *Journal of General Virology*, 94: 2449–57.
- Kuno, G. et al. (1998) 'Phylogeny of the Genus Flavivirus', *Journal of Virology*, 72: 73–83.
- et al. (2009) 'Correlations of Phylogenetic Relation with Host Range, Length of ORF or Genes, Organisation of Conserved Sequences in the 3' Noncoding Region, and Viral Classification among the Members of the Genus Flavivirus', in Feng, Z., and Long, M. (eds) *Viral Genomes: Diversity, Properties and Parameters*, pp. 1–33. Hauppauge, NY: Nova Science Publishers.
- (2016) 'The Boundaries of Arboviruses: Complexities Revealed in Their Host Ranges, Virus-host Interactions and Evolutionary Relationships', in Vasilakis, N., and Gubler, D. J. (eds) *Arboviruses: Molecular Biology, Evolution and Control*, pp. 219–68. Poole, UK: Caister Academic Press.
- et al. (2017) 'Vertebrate Reservoirs of Arboviruses: Myth, Synonym of Amplifier, or Reality', *Viruses*, 9: 185.
- (2020) 'The Absence of Yellow Fever in Asia: History, Hypotheses, Vector Dispersal, Possibility of YF in Asia, and Other Enigmas', *Viruses*, 12: 1349.
- Le, S. Q., and Gascuel, O. (2008) 'An Improved General Amino Acid Replacement Matrix', *Molecular Biology and Evolution*, 25: 1307–20.
- Lemey, P. et al. (2009) 'Identifying Recombinants in Human and Primate Immunodeficiency Virus Sequence Alignments Using Quartet Scanning', *BMC Bioinformatics*, 10: 126.
- Lequime, S. et al. (2016) 'Determinants of Arbovirus Vertical Transmission in Mosquitoes', *PLOS Pathogens*, 12: e1005548.
- Moureaux, G. et al. (2015) 'New Insights into Flavivirus Evolution, Taxonomy and Biogeographic History, Extended by Analysis of Canonical and Alternative Coding Sequences', *PLoS One*, 10: e0117849.
- Olsen, G. J. (2013) 'Parsimony', in Maloy, S., and Hughes, K. (eds) *Brenner's Encyclopedia of Genetics*, 2nd edn. pp. 229–32. Reference Module in Life Science, London, UK: Academic Press.
- Peng, B. et al. (2015) 'Genetic Data Simulators and Their Applications: An Overview', *Genetic Epidemiology*, 39: 2.
- Phillips, M. J. et al. (2004) 'Genome-scale Phylogeny and the Detection of Systematic Biases', *Molecular Biology and Evolution*, 21: 1455–8.
- Rambaut, A. et al. (2018) 'Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7', *Systematic Biology*, 67: 901–4.
- Rambaut, A., and Grassly, N. C. (1997) 'Seq-Gen: An Application for the Monte Carlo Simulation of DNA Sequence Evolution

- along Phylogenetic Trees', *Computer Applications in the Biosciences: CABIOS*, 13: 235–8.
- Sackman, A. M. et al. (2017) 'Mutation-driven Parallel Evolution during Viral Adaptation', *Molecular Biology and Evolution*, 34: 3243–53.
- Seligman, S. J., and Gould, E. A. (2004) 'Live Flavivirus Vaccines: Reasons for Caution', *The Lancet*, 363: 2073–5.
- Simon-Loriere, E., and Holmes, E. C. (2011) 'Why Do RNA Viruses Recombine?' *Nature Reviews. Microbiology*, 9: 617–26.
- Song, F. et al. (2016) 'Capturing the Phylogeny of Holometabola with Mitochondrial Genome Data and Bayesian Site-heterogeneous Mixture Models', *Genome Biology and Evolution*, 8: 1411–26.
- Stamatakis, A. (2014) 'RAxML Version 8: A Tool for Phylogenetic Analysis and Post-analysis of Large Phylogenies', *Bioinformatics*, 30: 1312–3.
- Stern, A. et al. (2017) 'The Evolutionary Pathway to Virulence of an RNA Virus', *Cell*, 169: 35–46.
- Suchard, M. A. et al. (2018) 'Bayesian Phylogenetic and Phylodynamic Data Integration Using BEAST 1.10', *Virus Evolution*, 4: vey016.
- Susko, E. (2014) 'Tests for Two Trees Using Likelihood Methods', *Molecular Biology and Evolution*, 31: 1029–39.
- Tesh, R. B. et al. (2016) 'Role of Vertical Transmission in Mosquito-borne Arbovirus Maintenance and Evolution', in Vasilakis, N., and Gubler, D. J. (eds) *Arboviruses: Molecular Biology, Evolution and Control*, pp. 191–217. Norfolk, UK: Caister Academic Press.
- Thézé, J. et al. (2015) 'Evolutionary and Phylogenetic Analysis of the Hepaciviruses and Pegiviruses', *Genome Biology and Evolution*, 7: 2996–3008.
- Twiddy, S. S., and Holmes, E. C. (2003) 'The Extent of Homologous Recombination in Members of the Genus *Flavivirus*', *Journal of General Virology*, 84: 429–40.
- Wang, R. et al. (2020) 'T Cell Immunity Rather than Antibody Mediates Cross-protection against Zika Virus Infection Conferred by a Live Attenuated Japanese Encephalitis SA14-14-2 Vaccine', *Applied Microbiology and Biotechnology*, 104: 6779–89.
- Weaver, S. C., and Vasilakis, N. (2009) 'Molecular Evolution of Dengue Viruses: Contributions of Phylogenetics to Understanding the History and Epidemiology of the Preeminent Arboviral Disease', *Infection, Genetics and Evolution*, 9: 523–40.
- White, S. K. et al. (2018) 'Spondweni Virus in Field-Caught *Culex Quinquefasciatus* Mosquitoes, Haiti, 2016', *Emerging Infectious Diseases*, 24: 1765–7.
- Whitman, L., and Antunes, P. C. A. (1938) 'Studies on *Aedes Aegypti* Infected in the Larval Stage with the Virus of Yellow Fever', *Experimental Biology and Medicine*, 37: 664–6.
- Wolf, Y. I. et al. (2018) 'Origins and Evolution of the Global RNA Virome', *mBio*, 9: 6.
- Worobey, M. et al. (1999) 'Widespread Intra-serotype Recombination in Natural Populations of Dengue Virus', *Proceedings of the National Academy of Sciences*, 96: 7352–7.
- Worobey, M., and Holmes, E. C. (1999) 'Evolutionary Aspects of Recombination in RNA Viruses', *Journal of General Virology*, 80: 2535–43.
- Yang, H. et al. (2018) 'Compositional and Mutational Rate Heterogeneity in Mitochondrial Genomes and Its Effect on the Phylogenetic Inferences of Cimicomorpha (Hemiptera: Heteroptera)', *BMC Genomics*, 19: 1.
- Yang, Z. (2007) 'PAML 4: Phylogenetic Analysis by Maximum Likelihood', *Molecular Biology and Evolution*, 24: 1586–91.
- Yang, Z., and Nielsen, R. (2002) 'Codon-substitution Models for Detecting Molecular Adaptation at Individual Sites along Specific Lineages', *Molecular Biology and Evolution*, 19: 908–17.