Journal Pre-proofs

Evolutionary dynamics of lineage 2 West Nile virus in Europe, 2004-2018: Phylogeny, selection pressure and phylogeography

Serafeim C. Chaintoutis, Anna Papa, Danai Pervanidou, Chrysostomos I. Dovas

PII: S1055-7903(19)30328-8
DOI: https://doi.org/10.1016/j.ympev.2019.106617
Reference: YMPEV 106617

To appear in: Molecular Phylogenetics and Evolution

Received Date: 30 May 2019
Revised Date: 10 September 2019
Accepted Date: 11 September 2019


This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2019 Published by Elsevier Inc.
Evolutionary dynamics of lineage 2 West Nile virus in Europe, 2004-2018: Phylogeny, selection pressure and phylogeography

Serafeim C. Chaïntoutisa, Anna Papat, Danai Pervanidouc, Chrysostomos I. Dovasa*

a Diagnostic Laboratory, School of Veterinary Medicine, Faculty of Health Sciences, Aristotle University of Thessaloniki, Greece
b National Reference Centre for Arboviruses, Department of Microbiology, Medical School, Faculty of Health Sciences, Aristotle University of Thessaloniki, Greece
c Department of Epidemiological Surveillance and Intervention, National Public Health Organization - Greece, Athens, Greece

*Correspondence at: Chrysostomos I. Dovas. E-mail: dovas@vet.auth.gr

Abstract

West Nile virus (WNV) is an arbovirus causing neuroinvasive disease to humans and equines. Since 2004, lineage 2 WNV strains have been identified in Europe and have been implicated in severe outbreaks, with that of 2018 exceeding the total number from the previous seven years. The aim of this study was to explore the evolutionary process that shapes the genetic diversity of lineage 2 WNV strains (belonging to the Central European/Hungarian subclade) and reconstruct the origin and transmission routes in Europe, and especially in the Balkans. For this purpose, a high number of whole genome sequences (WGSs) were analyzed, along with newly characterized sequences, including strains from the 2018 WNV transmission season in Greece. Maximum likelihood and Bayesian inference methods were used to perform the phylogenetic and phylodynamic analyses and phylogeographic reconstruction. The majority of the Central European/Hungarian lineage 2 strains are grouped in 2 phylogenetic subgroups (Central/South-West European and Balkan) with bush-like topology. Purifying selection shapes their evolution, however, strong evidence of positive selection was revealed in seven non-structural protein codons of NS1, NS4B, and NS5. Thirty-two amino-acid substitutions were fixed in different phylogenetic subgroups, indicating that random genetic drift is responsible for the majority of evolutionary changes. Migration, followed by subsequent local evolution is responsible for continuously evolving strains throughout Europe. In total, 10 virus transitions between discrete geographical locations, involving virus spread from Central Europe to other regions, were highly supported. Three novel, independent
introductions from Hungary and Bulgaria were responsible for the 2018 re-emergence of WNV in Northern Greece, indicating that Hungary remains an important ecological niche for the virus and has a central role for the dissemination of novel strains in the Balkans. In Northern Greece, tMRCA estimations indicated that a 1-to 2-year period of silent enzootic transmission precedes spread to dead-end hosts. Reconstruction of WNV population dynamics, from WGS data, revealed epidemic patterns characterized by 2- to 5-year oscillations in Europe. Future studies are necessary to determine the possible driving factors for these fluctuations i.e. avian herd immunity and climatic conditions affecting mosquito and bird populations. Maintaining adequate epidemiological surveillance with emphasis on obtaining WGS data, in areas at risk, is crucial for understanding the epidemiology and transmission patterns of WNV. It can further support integrated programs for risk assessment of virus circulation and dynamics, aiming to targeted prevention and response measures for veterinary and public health in Europe.

Keywords
West Nile virus, lineage 2, Central European/Hungarian subclade, phylogeny, evolutionary dynamics, phylogeography

1. Introduction

West Nile virus (WNV) (*Flavivirus* genus, *Flaviviridae* family) is a mosquito-transmitted virus. It possesses a single-stranded positive-sense RNA genome of approximately 11 kb and encodes a polyprotein of 3,434 amino-acids (Rossi et al., 2010). Following cleavage by viral and host proteases, 10 mature proteins are generated: 3 structural (C, prM and E) and 7 non-structural (NS) proteins (NS1, NS2A, NS2B, NS3, NS4A, NS4B and NS5) (Mukhopadhyay et al., 2005). Phylogenetic analyses revealed that WNV strains are grouped into eight evolutionary lineages (Fall et al., 2017), with lineages 1 and 2 being associated with disease and outbreaks (Petersen et al., 2013).

BIRDS are the main reservoir host of the virus, whereas humans and equids are considered dead-end hosts. Approximately 80% of human infections are asymptomatic; 20% present a mild febrile syndrome, while less than 1% of cases present a neuroinvasive disease (encephalitis, meningitis or acute flaccid paralysis). Elderly and immunocompromised people are at increased risk for severe disease (Lim et al., 2011). Equines may exhibit neurological symptoms, in around 10% of infections (Castillo-Olivares and Wood, 2004), although the respective ratio for the lineage 2 strain circulating in Greece during 2010
was higher (19%) (Bouzalas et al., 2016). Migratory birds are considered responsible for the introduction of the virus into new areas (Rappole and Hubálek, 2003).

WNV circulation in Europe has been reported through serological studies since the 1950s. The first major outbreak in Europe was recorded in 1996 in Romania (Tsai et al., 1998). Since then, cases have been recorded in Southern, Eastern and Western European countries. Prior to 2004, only lineage 1 strains were detected in animals and humans in Europe (Zeller and Schuffenecker, 2004). Lineage 2 was first detected in Europe in 2004, when it was isolated from the brain of a goshawk (*Accipiter gentiles*) in Hungary (Bakonyi et al., 2006). The first large outbreak in Europe caused by WNV lineage 2 occurred in 2010 in Greece (Papa et al., 2010). The strain (Nea Santa-Greece-2010) belonged to the Central European/Hungarian subclade of lineage 2 (Papa et al., 2011). Cases and outbreaks caused by these strains occurred in several European countries over Central Europe and eastern Mediterranean region, including Austria, Greece, Hungary, Serbia and Italy (Hernandez-Triana et al., 2014). Another lineage 2 strain was recorded in 2007 in southern Russia comprising the Eastern European/Russian subclade (Platonov et al., 2008). The same strain caused large outbreaks also in Romania (Dinu et al., 2015); in 2018 it was detected in a clinical case in Eastern Greece, although the rest of the cases in Greece were caused by the Central European/Hungarian subclade (Papa et al., 2019).

Since 2010 when WNV emerged in Greece and caused a large outbreak (262 reported cases with 17% fatality) (Danis et al., 2011), up to 2018, cases occurred every year, except 2015 and 2016 (Papa, 2017). In 2018, an unusually early onset and a large number of WNV infections were recorded in European countries, including Greece, Italy, Hungary and Romania (Haussig et al., 2018). The total number of reported autochthonous infections in 2018 (\(N = 2,083\)) exceeded the total number from the previous seven years (\(N = 1,832\)) (European Centre for Disease Prevention and Control, 2018). The first cases were notified by Greece in week 26 (25 June-1 July). In total, 316 human cases and 50 deaths were recorded in Greece (National Public Health Organization-Greece, 2018). Several outbreaks in equines were also reported through the Animal Disease Notification System.

Understanding the epidemiology and transmission patterns of WNV in Europe is crucial for further assessing the risk of WNV circulation and activity in the next transmission seasons. The aim of the present study was to reconstruct the origin and transmission routes of lineage 2 WNV strains of the Central European/Hungarian subclade in Europe, and especially in the Balkans. For this purpose, whole genome sequences (WGSs) available in GenBank were analyzed, along with newly characterized sequences from Greece, including those from the 2018 WNV transmission season. Additionally, the forces that drive the evolution of the lineage 2 WNV strains in Europe were investigated, through implementation of evolutionary time-scale and selection pressure analyses.
2. Materials and methods

2.1. Viral sequences

The present study was performed using newly characterized WGSs from WNV lineage 2 strains detected in Greece during 2010-2012 (N = 6) and 2018 (N = 7). The viral strains obtained in 2018 originated from human (N = 3), equine (N = 1), avian (N = 1), mosquito (N = 1), and canine (N = 1) samples. All Greek samples originated from hosts in Central Macedonia region (specifically, from the regional units of Thessaloniki and Pella), except one which was obtained from a dog residing in Thrace region (specifically, from the regional unit of Rhodope). Viral strains from 2010-2012 originated from mosquito (N = 3) and avian (N = 3) hosts, sampled within the framework of domestic bird and mosquito-based WNV surveillance surveys which were performed by the School of Veterinary Medicine, Aristotle University of Thessaloniki, in Central Macedonia, Greece (Chaintoutis et al., 2014, 2013, Chaskopoulou et al., 2013, 2011). The WGS of these strains was determined using a WNV lineage 2-specific PCR-based next generation sequencing protocol (Chaintoutis et al., 2019). The obtained sequences were deposited in the European Nucleotide Archive (ENA) database, under accession numbers ERZ805097-ERZ805098, ERZ805102-ERZ805111 and ERZ805097.

Fifteen previously characterized WGSs from Greek WNV strains were included in the dataset (Barzon et al., 2015, 2013; Papa et al., 2011). WGSs from lineage 2 WNV strains from other European countries available in GenBank were also added. Sequences belonging to the Eastern European/Russian subclade were not included, since they are phylogenetically distant to the Central European strains, representing a possibly separate introduction of WNV lineage 2 strain in Europe (Ravagnan et al., 2015).

All selected sequences were imported in MEGA 6 (Tamura et al., 2013) and were aligned using MUSCLE (Edgar, 2004). Only the complete coding region (ORF) was subjected to analyses. Recombination can adversely affect the accuracy of phylogenetic reconstruction and may result in higher rates of false-positives in selection pressure analysis (Anisimova et al., 2003; Posada and Crandall, 2002). Consequently, the data were initially tested for the identification of possible recombinant sequences using the Genetic Algorithm for Recombination detection (GARD) (Kosakovsky Pond et al., 2006), implemented on the Datamonkey 2.0 adaptive evolution web server (Weaver et al., 2018).

TempEst v1.5.1 was used in order to assess the presence of temporal signal in the investigated heterochronous sequences, and identify those with discrepancies between their sampling date and genetic divergence (Rambaut et al., 2016). After checking the molecular clock hypothesis, 2 sequences whose
sampling date was not consistent with their genetic divergence and phylogenetic position, as revealed by the root-to-tip regression analysis were excluded. These sequences belonged to isolates “Berliner” and “Tammy” (GenBank acc. no. KP780839 and KP780840, respectively), obtained from brains of kea (Nestor notabilis) with chronic infections (~3 and > 6 years, respectively). This resulted in independent evolution of these viral strains in the brains of the infected birds, and to their high divergence compared to field strains from other vertebrate hosts (acute infections) (Bakonyi et al., 2016).

The final dataset was comprised by 83 sequences across 9 countries, including 28 sequences from Greece. Pairwise distances among the sequences were calculated using MEGA 6.

2.2. Phylogenetic analysis

Maximum likelihood (ML) phylogenetic analysis was performed through the IQ-TREE software (Nguyen et al., 2015). Prior the analysis, ModelFinder (Kalyaanamoorthy et al., 2017), implemented in IQ-TREE was used to select the best-fitted nucleotide substitution model for the dataset. The strain “goshawk-Hungary/04” (GenBank acc. no. DQ116961) was used as outgroup. The ML estimate tree was subsequently visualized in FigTree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/). Support was evaluated using the Shimodaira-Hasegawa approximate likelihood ratio test (SH-aLRT) (Guindon et al., 2010) and the ultrafast bootstrap approximation (UFBoot) (Minh et al., 2013) with 1,000 replicates each.

2.3. Discrete phylogeography and phylodynamics

The evolutionary rate and demographic history of European lineage 2 WNV strains, including the time to the most recent common ancestor (tMRCA) were estimated using Bayesian inference. The Bayesian Markov Chain Monte Carlo (MCMC) algorithm was run through BEAST v1.10.4 software (Suchard et al., 2018). Analysis parameters and traits, including sampling location and date, as well as substitution and clock models were set in BEAUti graphical user interface. Specifically, a discrete phylogeographic diffusion approach between countries was implemented. Greek sequences from Central Macedonia and Thrace regions were named as subgroups GrM and GrTh, respectively. Sampling location was kept for all sequences, except for strain “Belgium/2017/Antwerpen” (GenBank acc. no. MH021189), which corresponds to a sequence from a traveler who returned to Belgium from Hungary. The sampling time-point was precisely retrieved for the majority of the sequences, while in 33 sequences of which only the sampling year was known, the uncertainty of the sampling date was accommodated by estimating tip dates uniformly from a 3-month time-interval precision spanning calendar weeks 29 to 42. Marginal
likelihood estimator values, obtained through path sampling (PS) and stepping-stone (SS) sampling analysis between prior and posterior, were assessed to compare and select the molecular clock and nucleotide substitution models (Baele et al., 2012). Based on the result of this analysis, the uncorrelated lognormal relaxed clock (ULRC) was applied (Drummond et al., 2006) which was preferred over the strict clock model. The Shapiro-Rambaut-Drummond (SRD06) codon model was used (Shapiro et al., 2006) which was preferred over the GTR+Γ4+I and T93+Γ4+I nucleotide substitution models. The temporal viral population dynamics was reconstructed using the Bayesian non-parametric coalescent-based Skygrid model (Gill et al., 2013). Graphical representation of the virus effective population size estimate (Ne) through time, as well as the calculation of the nucleotide substitution rate were performed through Tracer v1.7 (Rambaut et al., 2018). Statistical uncertainty for all parameters was expressed in 95% highest posterior density (HPD) intervals.

The patterns of WNV transmission between geographical locations were investigated using a discrete state continuous-time Markov chain (CTMC) model. In this model, transitions were estimated between pairs of locations, assuming an asymmetric non-reversible transition model, which employs a Bayesian stochastic search variable (BSSV). SpreeAD3 v0.9.6 (Bielejec et al., 2016) was used for the visualization of the geographical dispersal of the viral strains in Europe through time, as well as for the computation of the Bayes factors (BF) associated with each migration route, in order to identify support for non-zero transmission routes. The directional transition rates between discrete locations (countries) were considered significant when the respective BF was > 6 (Lemey et al., 2009). Markov jumps procedure was also used to reconstruct the history of transitions (Minin and Suchard, 2008).

MCMC analyses were run for $2.5 \times 10^8$ generations, with subsampling every $10^4$ iterations. Analysis of the generated samples was performed in Tracer v1.7 (Rambaut et al., 2018). Results were acceptable if the effective sample size (EES) was > 200 and MCMC convergence and mixing were adequate. Using the TreeAnnotator software (BEAST package), the trees sampled during the MCMC analysis were summarized in a maximum clade credibility (MCC) tree, after discarding the first 10% of the samples as burn-in. The resulting MCC tree was subsequently visualized using FigTree v1.4.4. Posterior probabilities (PP) were used as estimates of the statistical supports of the tree nodes.

2.4. Selection pressure analysis and identification of fixed amino-acid substitutions

For selection pressure analysis, 10 nucleotide sequence datasets were constructed, corresponding to each of the 10 WNV proteins. In these datasets, all homologous sequences from lineage 2 strains available in GenBank as of January 2019 were added. For each of the 10 aligned sequence datasets,
selection pressure acting on the sequences was estimated as a ratio of non-synonymous ($d_N$) to synonymous ($d_S$) nucleotide substitutions ($d_N/d_S$ or $\omega$ ratio) per site. Based on the $\omega$ values, codons were characterized as neutral ($\omega = 1$), under positive selection ($\omega > 1$), or under purifying selection ($\omega < 1$) (Wong et al., 2004). For this purpose, the following methods available on the Datamonkey 2.0 adaptive evolution web server (Weaver et al., 2018) were used: i) the single-likelihood ancestor-counting (SLAC) method, ii) the fixed-effects likelihood (FEL) method (Kosakovsky Pond and Frost, 2005), iii) the fast, unconstrained Bayesian approximation (FUBAR) method (Murrell et al., 2013), as well as iv) the mixed-effects model of evolution (MEME) method, which was implemented to identify instances of both episodic and pervasive positive selection (Murrell et al., 2012). Significance was considered at $P < 0.1$ or PP > 0.8 for FUBAR. Furthermore, likelihood ratio tests (LRTs) were performed, to compare three sets of nested site-specific models, namely M3 (discrete) vs. M0 (one-ratio), M2a (positive selection) vs. M1a (nearly neutral) and M8 ($\beta$ & $\omega$) vs. M7 ($\beta$), via the CODEML algorithm (Yang, 2007), implemented in EasyCodeML v1.21 (Gao et al., 2019). When LRTs yielded significant results, the Bayes Empirical Bayes (BEB) approach was used to identify codons under positive selection (Yang et al., 2005), i.e. positions with $\omega > 1$ and PP > 0.95.

For the identification of amino-acid substitutions fixed on specific subgroups of the ML tree, the amino-acid sequences generated by the polyprotein genomic sequence dataset were mapped to the phylogeny. Substitutions observed in single sequences, and not in clades, were not taken into consideration.

3. Results

3.1. Phylogenetic analysis and phylogeography

ML and Bayesian inference analyses revealed that, all WNV lineage 2 strains detected between 2004-2018 in Central and Southeastern European countries were segregated into a monophyletic group, associated with a single introduction event in Hungary. The mean tMRCA calculated for the European lineage 2 dataset was 16.19 years, i.e. in August 2002 (Table 1). Following its entrance in Hungary, the virus spread to other countries, most importantly, westwards to Austria, and southwards to Greece. Both trees (ML and MCC) are similar, characterized by the existence of 2 main subgroups with bush-like topology. More specifically, strains detected in Austria, Italy, Czech Republic and Germany are grouped within the first major subgroup (Central/South-West European subgroup, node “H”, Fig. 1) whereas those from Balkan countries were segregated within the second subgroup (Balkan subgroup, node “R”, Fig. 1).
Greek WNV strains do not comprise a monophyletic group. The strains obtained in Northern Greece during the 2018 transmission season were associated with novel, independent introductions, which were not associated with the “Nea Santa-Greece-2010” subgroup. The mean tMRCA for the strains detected in Greece during 2010-2013 was estimated at June 2008, whereas the MRCA of a descendant GrTh subgroup spreading in the Thrace region of Greece was estimated to emerge at September 2011 (Table 1, Fig. 2). Additionally, the mean tMRCA for the newly introduced strains which were obtained and sequenced from Northern Greece within the framework of the present study in 2018, could only be calculated for the subgroup of GrM strains (node “T”, Fig. 1) and was estimated at May 2017 (Table 1). Estimation of the introduction date was not feasible regarding the other two strains that were associated with novel introductions in Greece (acc. no. ERZ805102 in Central Macedonia region and ERZ805104 in Thrace region), due to the limited number of available sequences.

In total, 10 transitions between discrete geographical locations were highly (definitive, very strongly and strongly) supported (Fig. 3). Visualization of the migration routes revealed that the spread from Hungary to Central Macedonia region of Greece occurred in 3 independent transitions. Regarding the 2018 strain detected in a dog residing in Thrace region of Greece (acc. no. ERZ805104), it was revealed that the transition history of the virus in that region involves its spread from Hungary through Serbia and, then, through Bulgaria (Fig. 3).

3.2. Evolutionary rate and population dynamics

The mean evolutionary rate for the European lineage 2 WNV strains was $3.7 \times 10^{-4}$ nucleotide substitutions/site/year (95% HPD interval: $3 \times 10^{-4}$-$4.41 \times 10^{-4}$ s/s/y). The average pairwise distance between all analyzed polyprotein genomic sequences was 53.8 nucleotides (range: 3-87). Analysis using the non-parametric Bayesian Skygrid coalescent model indicated that, after lineage 2 WNV introduction in Europe in 2002, an initial increase in the virus effective population size was observed up to 2008, followed by exponential increases at 2- to 5-year intervals (Fig. 4).

3.3. Fixed amino-acid substitutions and selection pressure analysis

In total, 32 fixed amino-acid substitutions were revealed in supported phylogenetic subgroups of the European lineage 2 strains. Out of these, 1 was observed at NS1, 2 at NS2B and NS3, 4 NS2A and NS4B, 5 at NS1 and 7 at E and NS5 (Fig. 1). Moreover, except for the goshawk/Hungary 2004 strain (GenBank acc. no. DQ116961) the whole subclade is characterized by 3 substitutions ($S_{378}P$, $G_{666}E$ and
A_{2554}T). Similarly, the Balkan subgroup (Node “R”, Fig. 1) is characterized by 3 amino-acid substitutions (V_{1493}I, H_{1754}P and T_{2322}A), which can be regarded as phylogenetic signatures of all viral strains grouped within it. In codons 449, 2287, 2386 and 2731, the observed amino-acid substitution occurred in different, non-related subgroups. Variation in codon 449 (i.e. position E_{159}) was observed, as 3 different substitutions (I→T/M and T→A) were observed in different supported phylogenetic subgroups (Fig. 1).

The application of the selection pressure analysis methods revealed that the majority of the lineage 2 WNV polyprotein codons were subjected to strong purifying selection. Specifically, 2,361 codons (68.8% out of the total) were under purifying selection as evidenced by the FUBAR method. The E protein of lineage 2 WNV strains was the structural protein under the strongest purifying selection (79.2% of the sites). Similarly, NS5 and NS4A were the NS proteins under the strongest purifying selection (83.3% and 77.9% of the sites, respectively) (Fig. 5).

Strong evidence of positive selection was revealed in 7 NS protein sites of the lineage 2 strain polyprotein (Fig. 5, Table 2). Specifically, 2 of these sites were recognized by two methods (MEME or M8 of the CODEML algorithm and FUBAR); 4 were confirmed by 3 methods (FEL or M8, MEME and FUBAR, and one was confirmed by all methods used, except for SLAC, which did not reveal any site under positive selection. The positively selected NS1 sites included amino-acid substitutions fixed in subgroups comprising by Italian sequences (nodes “G” and “E”, respectively, Fig. 1). NS4B sites were associated with amino-acid substitutions which were observed in 6 Greek sequences, and most importantly, in vertebrate hosts. Out of them, only NS4B_{11} received a highly significant PP value of 0.96 in CODEML analysis (M8), compared to positions 14 and 20 of NS4B (Table 2). Regarding the 2 positively selected sites of NS5, these were associated with substitutions fixed in phylogenetic subgroups (nodes “H” and “B”). Additionally, NS5_{368} was also associated with a substitution with sporadic occurrence (Table 2). Furthermore, 19 sites were putatively subjected to positive selection as evidenced by only 1 method (MEME, FUBAR, or M8 of the CODEML algorithm, in 15, 3 and 1 sites, respectively). These sites were associated with amino-acid substitutions observed in non-European lineage 2 WNV strains, or sporadically, in 1-2 European viral strains.

4. Discussion

Since 2004, when WNV lineage 2 was detected for the first time in Europe (Bakonyi et al., 2006), different strains have spread to numerous countries of Central and Southeastern Europe, including Greece, and comprising a major public and veterinary health issue. Previous studies have examined the spread of WNV lineage 2 followed its introduction into Europe (Zehender et al., 2017; Ziegler et al.,
The present study focuses on the evolutionary processes (phylogeny, selection pressure and phylogeography) of the Central European/Hungarian lineage 2 WNV strains, with emphasis on the descendant Southeastern European strains. Our Bayesian analyses incorporated more realistic nucleotide substitution and coalescent models and utilized larger and more temporally and geographically diverse datasets, including proper approximation of sampling dates. Greek strains detected during the most recent seasonal outbreak (2018), as well as during 2010-2012 were also sequenced and included in the analysis. The aforementioned, permitted a more precise reconstruction of the transmission history of the WNV lineage 2 strains in Europe, and particularly in the Balkans, revealing novel introductions from Hungary.

All strains detected between 2004 and 2018 in Central and Southeastern Europe comprise a monophyletic group. The clade is associated with a single introduction event in Hungary, probably via migratory birds (Bakonyi et al., 2006). More specific, our tMRCA estimate suggests that the ancestral strain of this clade was introduced into Europe as a single event in August 2002, i.e. approximately 2 years prior it was first detected in Hungary (August 2004), presenting similarities to a tMRCA (March 2002) estimated previously (Zehender et al., 2017). Following its adaptation in Hungarian ecological niches, the virus spread to other European countries, like Austria (April 2006) and Italy (June 2008). Spread to Southeastern Europe occurred after the emergence of the Balkan strains MRCA, which was located in Hungary (tMRCA estimated at December 2006). Regarding Greece, it was shown that the mean tMRCA for the strains detected in 2010-2013 was estimated at June 2008, approximately 1.9 years before the first detection of WNV in human cases in Greece (Papa et al., 2010), whereas, for Thrace region of Greece, the mean tMRCA was estimated at September 2011, approximately 10 months prior to the occurrence of human cases in this area. Similarly, the tMRCA for the novel WNV introduction in Central Macedonia, Greece in 2018, was estimated at May 2017, approximately 1.1 years prior to the occurrence of human cases. This data support that after the first introduction of the virus in a geographical area, a 1 to 2-year period of silent enzootic transmission precedes virus spreading in humans. Zehender et al., (2017) also suggested that WNV circulates among reservoirs and vectors for several years before causing animal and human outbreaks, and so surveillance systems should be maintained in high risk areas even in the absence of human cases (Zehender et al., 2017).

Reconstruction of WNV population dynamics using the Skygrid model indicated that, upon introduction in Europe, the $N_e$ increased up to 2008. A maintenance phase followed, where the $N_e$ fluctuated, revealing epidemic patterns characterized by 2- to 5-year oscillations. Coalescence-based models have been used to depict complex epidemic patterns of flaviviral infections from genetic data (Allicock et al., 2012), and in infectious diseases the coalescence rate has been related to the pathogen transmission rate (Frost and Volz, 2010). In our case, the different epidemic oscillations observed could
be driven by different factors such as avian herd immunity and climatic conditions affecting mosquito and bird populations, in combination with re-introductions through migration. This pattern of virus transmission is evident in a well-studied area, the Central Macedonia region of Greece. Our phylogeographic reconstruction shows that, after the epidemic of 2010, strains belonging to the “Nea Santa-Greece-2010” subgroup continued to overwinter and circulate in the region until 2013. At the end of the 2010 transmission period, high seroprevalence (up to 68%) was recorded in pigeons residing within Central Macedonia region (Chaintoutis et al., 2014), indicating high degree of avian herd immunity. Chicken serological surveillance indicated continual decrease of virus circulation until 2014 (Chaintoutis et al., 2016; S. C. Chaintoutis et al., 2015). Human cases also decreased in the area and were not reported after 2013, until their reappearance 5 years later due to novel strain introductions from Hungary.

The mean evolutionary rate for the European WNV lineage 2 strains (based on the analysis of the coding region only) was estimated at 3.7×10^{-4} s/s/y, slightly lower compared to that reported in an Italian study (5.18×10^{-4} s/s/y) which analyzed European sequences dated up to 2015 (Zehender et al., 2017). However, this can be attributed to the different dataset analyzed and the applied nucleotide substitution and coalescent models. Furthermore, this estimated mean rate is slightly higher compared to the evolutionary rate reported for WNV lineage 2 in a study wherein strains from different continents were included (2.73×10^{-4} s/s/y), whereas it was identical to the mean evolutionary rate estimated for all WNV lineages (3.74×10^{-4} s/s/y) (McMullen et al., 2013). The majority of the codons in the lineage 2 WNV strains were found to be under strong purifying selection, which appears to constrain the changes at protein level and justifies their low evolutionary rate. This fact, along with the observed small overall mean distance between the analyzed polyprotein genomic sequences of the European lineage 2 strains (53.8 nucleotides), underlines the importance of obtaining and analyzing full polyprotein genomic sequences, in order to correctly estimate their phylogenetic relationships and phylogeography.

In total, 142 amino-acid sites were polymorphic in our polyprotein dataset. However, only 32 sites possessed amino-acid substitutions which were fixed in phylogenetic subgroups of the European strains. The majority of the observed substitutions where probably generated as a result of random genetic drift, since a limited number was found to be subjected to positive selection pressure and associated to virus adaptation during its entrance in an ecological niche. The phylogenetic and phylogeographic reconstruction revealed that, the majority of the Central European/Hungarian lineage 2 strains are grouped in 2 phylogenetic subgroups (Central/South-West European and Balkan) with bush-like topology. This suggests that after viral introduction in each of these geographical regions, the virus has undergone population expansion permitting the fixation of neutral substitutions, and in situ evolution due to the pressure exerted by environmental factors, including the hosts to which the virus had to adapt. Similar
conclusions were obtained by the investigation of the evolutionary dynamics of lineage 1 WNV in the USA (Añez et al., 2013). In Europe, the continuous dispersal of WNV resulted in the emergence of novel strains, which were responsible for outbreaks in defined geographical locations. In those areas, while adapting to the local environmental conditions, the viral strains accumulated genetic changes and several fixed amino-acid substitutions were observed in the phylogenetic subgroups. Following these changes, the virus strains either perpetuated in the area and spread to other locations, or were eliminated if their sustained transmission was not feasible (Pesko and Ebel, 2012). In the latter case, extinction of strains with positively selected codons is also possible due to ecological factors responsible for seasonal virus population bottlenecks, e.g. virus overwintering through a limited population of infected mosquitoes, as well as the high degree of avian herd immunity after a period of high virus transmission.

The aforementioned local evolution events are also supported by the selection pressure analysis results. Through our analyses, no evidence of positive selection in structural proteins was found. However, seven positively selected codons, which could potentially impact viral fitness, were recognized in 3 NS proteins (NS1<sub>35</sub>, NS1<sub>146</sub>, NS4B<sub>11</sub>, NS4B<sub>14</sub>, NS4B<sub>20</sub>, NS5<sub>298</sub> and NS5<sub>638</sub>). Interestingly, these positively selected sites differ from those recognized in lineage 1 WNV strains circulating in the US (Añez et al., 2013). The observed differences can be attributed to the different ecological factors associated with the studied regions, such as different mosquito, bird and mammal host species, the different virus lineage, the unique geoclimatic factors of each area, etc. Both positively selected NS1 sites observed are associated with amino-acid substitutions that characterize two separate subgroups wherein Italian sequences segregate. Flaviviral NS1 is required for efficient viral genome replication, and is also involved in immune evasion. Regarding WNV, it has been shown that this protein interacts with innate immunity components, such as the toll-like receptors (Morrison and Scholle, 2014). Two positively selected sites of NS4B (NS4B<sub>11</sub>, NS4B<sub>20</sub>) are associated with amino-acid substitutions observed in Greek sequences, and most importantly, in vertebrate hosts. NS4B<sub>14</sub> was identified by FUBAR, as well as by M8 of the CODEML algorithm, although not being under strongly positive selection. This site is also of interest, since it is associated with the S<sub>2287</sub>G amino-acid substitution that characterizes the whole Balkan subgroup. The aforementioned sites are located close to the N-terminus of the protein. The N-terminal region of the flaviviral NS4B (amino-acids 1-125) is associated with blocking the type I interferon (IFN-α/β) signaling, thus inhibiting IFN response (Munoz-Jordan et al., 2005). This protein has also a role in the formation of replication complex and virus pathogenesis (neurovirulence and/or neuroinvasiveness). The functional significance of these amino-acid substitutions has not been determined. However, the fact that this NS4B region comprises a preferential target for positive selection in several flaviviruses, suggests an important role in viral adaptation (Sironi et al., 2016). NS5 functions as the viral RNA-dependent RNA polymerase (RdRp), which is subjected to a diverse range of biochemical conditions, i.e.
within avian, mammalian and mosquito hosts. It has been shown that even subtle point mutations in RdRp away from its active site can have a significant biological effect on WNV fitness, which is also host-dependent (Van Slyke et al., 2012). The two positively selected NS5 sites revealed herein are associated with amino-acid substitutions fixed within the Central/South-West European subgroup. These sites should also be studied in detail to elucidate their possible role for WNV replication in different hosts. Consequently, all 6 positively selected sites should be prioritized in future pathogenesis studies, through construction of infectious clones.

Interestingly, the selection pressure analysis results revealed that the NS3\textsubscript{249} site was not under positive selection. This site possesses the H\textsubscript{1754}P amino-acid substitution, which can be considered a phylogenetic signature of the whole Balkan subgroup. The presence of proline in this position has been associated with increased virulence of WNV lineage 1a strains in American crows (Brault et al., 2007), although the exact mechanism by which this substitution leads to increased virulence is yet unknown. A recent pathogenesis study, which was conducted in an avian model, indicated that the presence of proline at this position is neither sufficient nor necessary to confer pathogenicity to lineage 2 WNV strains in birds (Dridi et al., 2015). However, the investigation of the pathogenicity of strains possessing this substitution in mammalian models may lead to different results. Thus, further research is required, via reverse genetics methods, to fully determine the role of this amino-acid substitution in the virulence of lineage 2 strains in dead-end hosts.

Molecular and evolutionary data on the circulating virus strains are also of importance for vaccine design, as well as for the long-term assessment of their ability to confer immunity, i.e. to induce antibodies capable of neutralizing newly emerging viral strains. The flaviviral E protein is the main target of the host’s neutralizing immune responses (Roehrig, 2003). Regarding WNV, an inactivated vaccine for equines, based on WNV lineage 1 antigens, was able to induce cross-protective immunity under field conditions against encephalitis due to infections by lineage 2 strains in Greece (Serafeim C. Chaintoutis et al., 2015). Despite the genetic drift observed in the lineage 2 strains of the Central European/Hungarian subelade, the ability of this vaccine to provide long-term protection of equines against all European lineage 2 strains is evidenced, since their E protein is subjected to strong purifying selection and no positively selected sites were observed.

In conclusion, Central European/Hungarian lineage 2 WNV, since its single event introduction in Europe, estimated in 2002, continues to evolve. However, most new amino-acid substitutions are not fixed, possibly due to bottleneck effects possibly driven by the drastic reduction of mosquito populations during the winter and avian herd immunity. The virus is subjected to strong purifying selection and random genetic drift is responsible for the great majority of evolutionary changes. Hungary remains an
important ecological niche for the virus, contributing to gene flow through bird migration. This is followed by subsequent local evolution of strains due to natural selection in affected areas. In this regard, different amino-acid residues of NS1, NS4B and NS5 were found to be subjected to positive selection and some of them were fixed in various subgroups. Substitutions occurring independently in different subclades are also of special importance and should be further studied, in order to determine their effect on the epidemiology and pathogenesis. In Europe, maintaining adequate enhanced epidemiological surveillance and laboratory capacity, in areas at risk, is crucial. Integrated surveillance programs should emphasise on WGS data, with additional sampling among the locations and over time, in order to obtain more detailed patterns for better understanding WNV epidemiology and transmission aiming to further support the risk assessment for the virus circulation and dynamic, and guide targeted prevention and response measures for veterinary and public health.

Acknowledgements

The study was partially supported by EU Horizon 2020 research and innovation program COMPARE (grant agreement no. 643476) and by the Greek Scientific Veterinary Agreement program, funded by Zoetis Inc. The laboratory diagnosis of the human WNV cases performed in the National Reference Centre for Arboviruses was financially supported by the National Public Health Organization-Greece (Ministry of Health). Serafeim C. Chaintoutis was supported by the Onassis Foundation.

References


**Fig. 1.** Maximum likelihood estimate tree of the 83 European lineage 2 WNV strain polyprotein genomic sequences. Taxon labels include GenBank or ENA acc. no., sampling location, sampling date, strain designation and host. Greek sequences are labeled in boldface. Au: Austria; Bul: Bulgaria; Cz: Czech Republic; Ger: Germany; GrM: Greece, Central Macedonia region; GrTh: Greece, Thrace region; Hun: Hungary; It: Italy; Ser: Serbia; Sl: Slovenia. The scale bar indicates nucleotide substitutions per codon site. Supported clades are indicated by Shimodaira-Hasegawa approximate likelihood ratio test (SH-aLRT) and ultrafast bootstrap approximation (UFBoot) support values of \( \geq 80\% \) and \( \geq 95\% \), respectively. The letters in circles correspond to specific subclades of the tree and respective fixed amino-acid substitutions. Fixed substitutions marked with asterisks represent positions with multiple occurrences in different non-related subclades. Positively selected codons confirmed by 2 or 3 methods are shown in boldface. Positions of the substitutions are numbered based on the polyprotein codons (positions 1 to 3434).

**Fig. 2.** Time-calibrated Bayesian maximum clade credibility (MCC) tree inferred for the phylogeographic history of European lineage 2 WNV strains, using 83 polyprotein genomic sequences. Taxon labels include GenBank or ENA acc. no., sampling location, sampling date, strain designation and host. Greek sequences are labeled in green. Terminal branches are colored according to the geographical location of the virus strain at the tip; Au: Austria; Bul: Bulgaria; Cz: Czech Republic; Ger: Germany; GrM: Greece, Central Macedonia region; GrTh: Greece, Thrace region; Hun: Hungary; It: Italy; Ser: Serbia; Sl: Slovenia. Internal branches are colored according to the most probable location of origin of the corresponding ancestral node, as inferred by Bayesian discrete phylogeography. Ancestral nodes represent mean age, and horizontal bars indicate 95% HPD (highest posterior density) interval for the age of each node in the MCC tree. Supported clades are indicated by posterior probabilities (PP) of > 0.9.

**Fig. 3.** Visualization of the significant migration paths of lineage 2 WNV strains in Europe overlaid on a geographical map. The bullets represent discrete geographical locations, Au: Austria; Bul: Bulgaria; Cz: Czech Republic; Ger: Germany; GrM: Greece, Central Macedonia region; GrTh: Greece, Thrace region; Hun: Hungary; It: Italy; Ser: Serbia; Sl: Slovenia. Arrows between geographical locations represent branches in the MCC tree, along which the relevant virus transmission occurs. Well supported paths [Bayes factor (BF) > 6] between locations are shown. The arrows indicate the directionality of the
migration and the intensity of their color represents the Bayes factor estimate. Interpretation of BF values: > 100 definitive support; 30-100 very strongly support; 10-30 strongly support.

**Fig. 4.** Population dynamics of the European lineage 2 WNV strains, estimated using the non-parametric Bayesian Skygrid coalescent model on the dataset of the 83 polyprotein genomic sequences analyzed. The Bayesian Skygrid plot represents temporal variation in the virus effective population size ($N_e$) through time. The blue line represents the median $N_e$ estimate and the shaded area corresponds to the 95% HPD intervals.

**Fig. 5.** Schematic representation of selection pressure analysis results on each protein of the lineage 2 WNV strains. Different color intensity was used for proteins, based on the percentage of their sites under purifying selection, as evidenced by the FUBAR method. The numbers of unique sequences in each dataset are indicated above ($N$). Arrows depict sites under positive selection, as evidenced by 2 (green), 3 (blue) or 4 (red) methods.
Table 1. Estimation of time to the most recent common ancestor (tMRCA) and virus detection for lineage 2 WNV strains in Europe and Greece.

<table>
<thead>
<tr>
<th>Clade / date of first detection / host</th>
<th>Estimated date of virus incursion</th>
<th>tMRCA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean 95% HPD interval</td>
<td>Mean 95% HPD interval</td>
</tr>
<tr>
<td>Greek, Central Macedonia / 2010.6 (August) / Human</td>
<td>2008.49 (June) 2007.18-2009.61</td>
<td>10.3 9.18-11.61</td>
</tr>
<tr>
<td>Greek, Thrace / 2012.56 (July) / Human</td>
<td>2011.76 (September) 2011.04-2012.41</td>
<td>7.03 6.38-7.75</td>
</tr>
<tr>
<td>Greek, Central Macedonia / 2018.5 (July) / Human</td>
<td>2017.36 (May) 2016.6-2018.08</td>
<td>1.43 0.71-2.19</td>
</tr>
</tbody>
</table>

HPD: highest posterior density

Table 2. Codons under positive selection by two or three selection pressure analysis methods.

<table>
<thead>
<tr>
<th>Codon</th>
<th>Protein and amino-acid position</th>
<th>Associated amino-acid substitution(s)</th>
<th>FEL</th>
<th>MEME</th>
<th>FUBAR</th>
<th>M8</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>P</td>
<td>PP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>826</td>
<td>NS1_35</td>
<td>Y→H</td>
<td>-</td>
<td>0.02</td>
<td>0.84</td>
<td>-</td>
</tr>
<tr>
<td>937</td>
<td>NS1_146</td>
<td>A→V</td>
<td>0.03</td>
<td>0.07</td>
<td>0.97</td>
<td>-</td>
</tr>
<tr>
<td>2284</td>
<td>NS4B_11</td>
<td>N→S/D/E</td>
<td>0.02</td>
<td>0.03</td>
<td>0.99</td>
<td>0.96</td>
</tr>
<tr>
<td>2287</td>
<td>NS4B_34</td>
<td>S→G</td>
<td>-</td>
<td>-</td>
<td>0.90</td>
<td>0.88</td>
</tr>
<tr>
<td>2293</td>
<td>NS4B_20</td>
<td>K→R</td>
<td>-</td>
<td>0.07</td>
<td>0.96</td>
<td>0.81</td>
</tr>
<tr>
<td>2827</td>
<td>NS5_298</td>
<td>A→T</td>
<td>0.06</td>
<td>0.04</td>
<td>0.96</td>
<td>-</td>
</tr>
<tr>
<td>3167</td>
<td>NS5_638</td>
<td>E→K/G</td>
<td>0.07</td>
<td>0.00</td>
<td>0.96</td>
<td>-</td>
</tr>
</tbody>
</table>

FEL: fixed-effects likelihood, MEME: mixed-effects model of evolution, FUBAR: fast, unconstrained Bayesian approximation, M8: Model 8 (β & ω) of the CODEML algorithm. Significance: P < 0.1 or posterior probability (PP) > 0.8 (FUBAR) or > 0.95 (M8).
Highlights

- The origin and evolution of lineage 2 WNV strains in Europe were investigated.
- Hungary has an important role for the dissemination of WNV strains in Europe.
- The Greek strains of 2018 were associated with 3 novel incursions from Hungary.
- Strong purifying selection and 7 positively selected codons were recognized.
- Random genetic drift is responsible for the majority of evolutionary changes.