Fluid Spatial Dynamics of West Nile Virus in the United States: Rapid Spread in a Permissive Host Environment

Francesca Di Giallonardo, a Jemma L. Geoghegan, a Douglas E. Docherty, b Robert G. McLean, a Michael C. Zody, c James Qu, c Xiao Yang, c Bruce W. Birren, c Christine M. Malboeuf, a Ruchi M. Newman, a Hon S. Ip, b* Edward C. Holmes b

Marie Bashir Institute for Infectious Diseases and Biosecurity, Charles Perkins Centre, School of Biological Sciences and Sydney Medical School, The University of Sydney, Sydney, NSW, Australia; a U.S. Geological Survey, National Wildlife Health Center, Madison, Wisconsin, USA; b Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA.

ABSTRACT

The introduction of West Nile virus (WNV) into North America in 1999 is a classic example of viral emergence in a new environment, with its subsequent dispersion across the continent having a major impact on local bird populations. Despite the importance of this epizootic, the pattern, dynamics, and determinants of WNV spread in its natural hosts remain uncertain. In particular, it is unclear whether the virus encountered major barriers to transmission, or spread in an unconstrained manner, and if specific viral lineages were favored over others indicative of intrinsic differences in fitness. To address these key questions in WNV evolution and ecology, we sequenced the complete genomes of approximately 300 avian isolates sampled across the United States between 2001 and 2012. Phylogenetic analysis revealed a relatively star-like tree structure, indicative of explosive viral spread in the United States, although with some replacement of viral genotypes through time. These data are striking in that viral sequences exhibit relatively limited clustering according to geographic region, particularly for those viruses sampled from birds, and no strong phylogenetic association with well-sampled avian species. The genome sequence data analyzed here also contain relatively little evidence for adaptive evolution, particularly of structural proteins, suggesting that most viral lineages are of similar fitness and that WNV is well adapted to the ecology of mosquito vectors and diverse avian hosts in the United States. In sum, the molecular evolution of WNV in North America depicts a largely unfettered expansion within a permissive host and geographic population with little evidence of major adaptive barriers.

IMPORTANCE

How viruses spread in new host and geographic environments is central to understanding the emergence and evolution of novel infectious diseases and for predicting their likely impact. The emergence of the vector-borne West Nile virus (WNV) in North America in 1999 represents a classic example of this process. Using approximately 300 new viral genomes sampled from wild birds, we show that WNV experienced an explosive spread with little geographical or host constraints within birds and relatively low levels of adaptive evolution. From its introduction into the state of New York, WNV spread across the United States, reaching California and Florida within 4 years, a migration that is clearly reflected in our genomic sequence data, and with a general absence of distinct geographical clusters of bird viruses. However, some geographically distinct viral lineages were found to circulate in mosquitoes, likely reflecting their limited long-distance movement compared to avian species.

West Nile virus (WNV) has imposed a significant disease burden on the avian population of North America since its introduction in 1999 (1). The virus infects a remarkably large number of species, with WNV-associated mortality observed in over 300 bird species spanning 30 different families, half of which are native to the United States (2, 3), although active replication may not occur in all (4). WNV has also resulted in massive population losses in several U.S. bird species, with the American crow (Corvus brachyrhynchos) experiencing the highest levels of morbidity and mortality, with disease manifested in more than 90% of infected individuals (5, 6) and a 45% decrease in population size (7). The virus is transmitted within an enzootic cycle involving mosquitoes and birds, with humans and other mammals acting as dead-end hosts (8). Although most human infections are asymptomatic, 1% of infections result in fatal neuroinvasive disease, and since 1999 more than 1,000 people have died from WNV infection in the United States (9). WNV was first reported in Uganda in 1937 and was long restricted to the Old World, with cases reported in Africa, the Middle East, and Europe (10, 11). However, the virus was not regarded as a major threat to humans, and only two large outbreaks were reported in Israel in the 1950s and in South Africa in 1974 (12). In 1996, the first major outbreak associated with severe encephalitis in humans was reported in Romania, followed by a second in Russia 2 years later (13, 14). The incidence of encephalitis has increased, with epidemics occurring regularly in Eu-
Evolution of West Nile Virus in the United States

Europe, Africa, and the Middle East (8, 15). In the United States, over 40,000 human disease cases resulting from WNV infection have been recorded, with a peak of 10,000 in 2003 (9).

WNV belongs to the genus Flavivirus (family Flaviviridae) of single-stranded positive-sense RNA viruses and is comprised of four lineages designated I to IV that differ by 10 to 26% in nucleotide identity (11). The virus isolated in the United States in 1999 is classified in lineage I (1). Notably, this New World WNV exhibits higher avian virulence than other lineage I viruses (16), which has been associated with a single Thr-to-Pro amino acid change at residue 249 in the viral helicase (encoded by NS3) (17). Also of importance is a Val159Ala amino acid substitution in the viral envelope (E) gene, which was first documented in 2002; viruses carrying this mutation quickly replaced those WNV strains previously circulating in the United States (18). The E-Val159Ala mutation is now fixed in the North American WNV population and has been linked to adaptation to the main Culex spp. mosquito vectors (19, 20).

Previous phylogenetic analyses have revealed two main genetic variants of WNV in the United States; an NY99 genotype, including the first U.S. strain isolated in 1999, and a more recent WN02 genotype that replaced NY99 in 2003 (21) (although other genotypes have been proposed; see below). WNV phylogenies are also characterized by relatively little spatial structure (18, 22, 23), although a relatively large clade of mosquito WNV sequences was recently documented in California (24). In addition, birds seem to harbor greater WNV diversity than mosquitoes (25, 26).

The ongoing epidemic of WNV in the United States serves as an informative example of the successful establishment of a virus in a new environment. The high incidence and mortality of WNV in the native bird population, together with human infections, have led to intensive surveillance efforts. Indeed, the combined reintroduction of the virus in a new environment. The high incidence and mortality of WNV in the native bird population, together with human infections, have led to intensive surveillance efforts. Indeed, the combined

**MATERIALS AND METHODS**

**Sample collection.** Wild birds were submitted to the USGS National Wildlife Health Center for determination of cause of death. A complete postmortem examination was performed, and when WNV was suspected, a tissue pool of kidney and spleen tissues was submitted for virological testing. Feather pulp samples were used for corvids, which are highly susceptible to WNV (28). Suspect samples were inoculated into flasks containing the Vero cell line, and the identity of WNV in cultures exhibiting cytopathic effects was determined via reverse transcription-PCR (RT-PCR) (29).

**Genome sequencing.** RNA amplification was performed as described previously (30). Illumina library construction was performed using NexteraXT (Illumina) according to the manufacturer’s protocol. Sequencing was performed on the Illumina HiSeq2500 platform, generating paired-end 101-bp reads, and de novo genome assembly was performed using the VICUNA assembly program (31).

**Comparative sequence data.** Complete coding sequences of WNV were downloaded from GenBank and combined with the sequences generated here. Sequences from humans and other nonavian species were excluded, as they represent dead-end hosts and are not involved in the natural transmission cycle. Hence, we reasoned that although small in number, the inclusion of these sequences could bias estimates of evolutionary rate. In addition, it has previously been shown that human sequences fall at multiple places across the WNV phylogeny (23). Sequence alignment was performed using MAFFT (32), and these results were inspected manually. This resulted in a total of 696 unique WNV genome sequences of 10,299 nucleotides in length (298 sequences from this study and 398 from GenBank; see Table S1 in the supplemental material). These data included viruses from 38 U.S. states, 51 bird species, and 281 sequences isolated from mosquitoes and covered a time span of 14 years, from 1999 to 2012. Prior to phylogenetic analysis, sequences were inspected for possible recombination by using the single breakpoint recombination (SBP) tool available in the program HyPhy (33, 34). No recombinants were found. A maximum likelihood (ML) tree of the full data set was estimated with RAxML (version 7.2.8) (35), employing the GTR + F nucleotide substitution model and 200 bootstrapping replicates. For Bayesian analyses (see below), the sequences were subdivided into three data sets, denoted A, B, and C, by randomly sampling (with replacement) 300 sequences for each (see Table S1). Finally, we created a data set comprising only those sequences generated here, referred to as the “this study” data set.

**Evolutionary rates.** We estimated evolutionary rates by using the Bayesian Markov chain Monte Carlo (MCMC) method implemented in the program BEAST (36). As computation time was excessive for the complete data set, we performed this analysis on WNV data subsets A, B, and C. Prior to this analysis, we used JModelTest (37) to determine the best-fit model of nucleotide substitution, which was found to be GTR + F. All subsequent Bayesian analyses therefore were run using a GTR + F nucleotide substitution model, an uncorrelated log-normal relaxed clock, and a logistic population growth prior (with the latter found to be a better fit to the data than constant population size and exponential growth demographic models). In addition, all analyses were run for at least 100 million steps, sampling every 10,000 steps to allow convergence for all parameters. The first 10% of the estimated distribution was removed as burn-in. Maximum clade credibility trees (MCC) were estimated using TreeAnnotator implemented in BEAST, with statistical support for individual nodes given as posterior probability values.

To determine the extent of temporal structure in the data, we performed a root-to-tip regression on the ML tree of the complete data set in Path-O-Gen (http://tree.bio.ed.ac.uk/software/pathogen/). We also performed a tip-date randomization test to evaluate the temporal signal in the data (38). For this, we ran 10 randomizations of data subset A in BEAST under the same evolutionary parameters as outlined above. A data set can be considered to possess temporal structure if none of the rate distributions from the randomizations overlap the rate distribution from the true data set (39).

**Phylogeny-trait association tests.** To determine which epidemiological variables were best associated with the data, we performed a series of phylogeny-trait association tests using the parsimony score (PS), association index (AI), and maximum clade size (MC) statistics available within the Bayesian Tip-association Significance (BaTS) program (40), utilizing the posterior distribution of trees recovered in the BEAST analysis (data sets A to C). In particular, we sought to determine whether these WNV sequence data were more structured by geographic region (including U.S. state and avian flyway), host species, or year of sampling, rather than that expected by chance alone. To this end, we divided the United States into four regions according to the U.S. Census Bureau (https://www.census.gov.geo/reference/gtc/gtc_census_divreg.html): (i) North East; Connecticut (CT), Massachusetts (MA), Rhode Island (RI), New Jersey (NJ), New York (NY), and Pennsylvania (PA); (ii) South; District of Columbia (DC), Florida (FL), Louisiana (LA), Georgia (GA), Maryland (MD), North Carolina (NC), Virginia (VA), Alabama (AL), Kentucky (KY), Mississippi (MS), Tennessee (TN), Arkansas (AR), Oklahoma (OK), and
RESULTS

Rapid dispersal of WNV in the United States. We sequenced WNV from dead birds sampled between 2001 and 2012, with the exception of 2011 (for which no samples were available). These data comprise 298 complete WNV genomes from 34 states and 43 different bird species, of which 101 viruses were from the American crow. Additionally, we downloaded full-length WNV genomes from GenBank, which were combined with the data generated here (see Table S1 in the supplemental material). The 398 sequences from GenBank comprised 281 isolates from mosquitoes and 117 from birds. These sequences were sampled from 11 geographic regions in the United States shown on the map below.

Phylogenetic analysis revealed a relatively star-like tree topology, both across the tree as a whole and within individual genotypes, indicative of the explosive expansion of WNV in the United States from its introduction in 1999 (Fig. 1). To display aspects of spatial structure more clearly, we color-coded the tree according to the U.S. region from which they were isolated: Midwest, Northeast, South, and West. Visually, there was little clustering according to geographic location, especially in the avian samples, which is compatible with rapid continental spread in the absence of major geographic barriers. However, two large sequence clusters were apparent in the western United States (Fig. 1, in green), reflecting a set of viruses isolated from mosquitoes in California (24). A more quantitative analysis of geographic structure is presented below.

These data also contained four clusters of sequences that represent the four different WNV genotypes proposed previously (11), each of which receives strong bootstrap support (at least 80%): NY99, Intermediate, WN02, and SW03 (Fig. 2). Whereas earlier studies contained only a limited number of viruses from the Intermediate genotype (18, 46), our samples added 29 Intermediate genotype viruses from multiple states, indicating that it was more widely distributed than previously realized. The Intermediate genotype differs from the NY99 genotype by 14 synonymous and 1 nonsynonymou substitution and from the WN02 genotype by 15 synonymous and 2 nonsynonymous substitutions (Table 1). In contrast, the WN02 genotype differs in only nine silent mutations and one amino acid substitution from the NY99 genotype. Notably, we found that the E-Val159Ala amino acid substitution, which is strongly associated with the WN02 genotype and might represent an adaptation to American Culex species (20), also occurs in phylogenetically distinct sequences within the Intermediate and NY99 genotypes (occurring twice in each) and was always supported by high bootstrap values (87 to 100%). Hence, this mutation has clearly evolved convergently multiple times. Also of note is the putative SW03 genotype (Fig. 2, purple), which falls within the phylogenetic diversity of the WN02 genotype, such that the latter is no longer a monophyletic group. The SW03 genotype seemingly appeared in 2002 and has predominantly circulated in the Midwest and West of the United States. The consensus sequence of this genotype differs from WN02 by 12 synonymous and 2 characteristic amino acid substitutions (NS4A-Ala85Thr and NS5-Lys314Arg) (11, 22, 47) (Table 1). As with E-Val159Ala, both the NS4A-Ala85Thr and NS5-Lys314Arg substitutions have evolved independently across the WNV phylogeny, this time in several sequence clusters in the SW02 genotype. Despite these mutational differences, the SW03 genotype is
not as clearly visible on our phylogeny as the other genotypes, because the majority of consensus mutations defining this genotype do not fall on the immediate branch leading to it (i.e., they are not synapomorphic).

The NY99 genotype entered the Northeast United States region in 1999 (Fig. 2). Around 2001, the Intermediate and WN02 genotypes emerged, dispersing to the South and Midwest (Intermediate) and the West (WN02) regions. These two descendant genotypes therefore appear to have generally spread in different geographical regions compared to NY99, as there is only a little overlap in genotype sampling per U.S. state, suggesting that they arose by allopatric separation. In 2003, WN02 replaced both the NY99 and Intermediate genotypes throughout the United States. Interestingly, the SW03 and WN02 genotypes were isolated in the same states during the same time period, such that they cocirculated (Fig. 2), although since 2009 the SW03 genotype has only been isolated in California.

**Limited evidence for adaptive evolution in WNV.** To help determine whether the spread of WNV across the United States since 1999 was associated with host adaptation in a new environment, we estimated the ratio of nonsynonymous to synonymous substitutions per site ($dN/dS$) for different genes and branches. The $dN/dS$ ratio for the complete coding sequence of the full WNV data set (696 sequences) was 0.083, while the values for the external and internal branches were 0.086 and 0.076, respectively (Table 2). These ratios are suggestive of a slight excess of transient deleterious mutations in WNV evolution, as observed previously (48). In most cases, individual genes exhibited a pattern similar to that in the whole genome. The strongest examples of excess deleterious mutations were observed in the membrane and envelope genes (external and internal $dN/dS$ values of 1.5 and 1.4, respectively). In contrast, three genes, NS1, NS4A, and NS4B, exhibited a slight excess of nonsynonymous mutations on internal versus external branches, although the ratios observed (~0.9) were perhaps more suggestive of a general neutral evolution. The capsid gene exhibited the greatest excess of nonsynonymous substitutions on internal branches (ratio of external/internal $dN/dS$, 0.7). Although this could indicate the past occurrence of positive selection on some amino acid sites and is supported by the observation that the capsid gene also had the highest overall $dN/dS$ ratio (0.315), no positively selected sites were identified in this gene when using site-specific bioinformatic approaches (see below).
strikingly, one of the lowest $dN/dS$ values in the data set (0.056) was observed for the envelope gene, indicating that it has been subject to the strongest purifying selection.

we further estimated the site-specific selection pressures acting on WNV, reflected in values for $dN$ and $dS$. Only two amino acid sites, NS2A-119 and NS4A-135, were found to harbor significant evidence for positive selection across all three methods used (SLAC, FUBAR, and MEME) (Fig. 3, red bars). Site NS4B-240 was also found to be subject to positive selection according to the FUBAR analysis, while site NS2A-52 was similarly identified as selected by MEME. Strikingly, the only evidence for positive selection at amino acid sites that delineate the different WNV genotypes was observed at NS5-314, where a Lys-to-Arg substitution distinguishes SW03 (Table 1), although this was marginal in all cases ($P = 0.048$ in MEME; $P = 0.086$ in SLAC; posterior probability of 0.947 in FUBAR). Notably, the E-Val159Ala substitution exhibited no signal of positive selection under either method, and we found no evidence for positive selection on any individual lineage on the WNV phylogeny. Although there was evidence for positive selection ($P = 0.05$) at an additional 17 amino acid sites according to the MEME analysis, these involved either mutations at tips of the tree alone (10/17 cases) or in very small clusters ($n = 3$ sequences), so that their biological significance is difficult to determine and false positives cannot be discounted. Finally, we determined that the mean pairwise identity across the WNV genome was 99.6% per site, with 69.6% of nucleotide sites identical in all sequences (Fig. 3). Interestingly, and in contrast to what is observed in most RNA viruses, greater genetic diversity was apparent in the nonstructural genes, particularly NS3 and NS4, than the structural genes.

**Evolutionary dynamics of WNV in the USA.** We utilized a Bayesian coalescent method to reveal the dynamics of WNV spread across the United States. Prior to this analysis, we used a root-to-tip analysis of the ML tree to determine whether these
sequence data possessed sufficient temporal signal to accurately infer evolutionary rates. This revealed a strongly significant ($P < 0.0001$) correlation between time of sampling and genetic divergence (Fig. 4A). Similarly, a strong clock-like structure was revealed by use of a tip-date randomization test, in which none of the rate distributions from the randomizations overlapped with the true tip-date rate (Fig. 4B).

Our Bayesian MCMC analysis revealed that the population dynamics of WNV in the United States best fit a model of logistic population growth, reflecting an initial period of exponential population growth, corresponding to the spread of the NY99 genotype, followed by a lower growth rate during the time period in which the WN02 genotype dominated the viral population (21). The mean nucleotide substitution rate for each data subset was as follows (mean [95% highest posterior density, or HPD, interval]): A, $4.47 \times 10^{-4}$ ($[4.08 \text{ to } 4.89] \times 10^{-4}$); B, $4.27 \times 10^{-4}$ ($[3.86 \text{ to } 4.61] \times 10^{-4}$); C, $4.16 \times 10^{-4}$ ($[3.81 \text{ to } 4.51] \times 10^{-4}$). Similarly, the times to the most recent common ancestor were estimated: A, 1998.0 (1997.1 to 1998.5); B, 1997.8 (1997.1 to 1998.5); C, 1997.9 (1997.2 to 1998.6), indicating that they diverged shortly after the first appearance of WNV in the United States. Finally, the emergence of SW03 was estimated to be around 2001 for all three data sets (range of 95% HPD intervals, 2000.5 to 2002.1). As an internal control, we repeated the BEAST analysis with sequences isolated in this study (birds) or from mosquitoes only, which resulted in very similar rate estimates (range of $[3.50 \text{ to } 4.65] \times 10^{-4}$ and $[4.34 \text{ to } 5.12] \times 10^{-4}$ substitutions/site/year, respectively), suggesting that they are robust.

**Clustering by host species and sampling location.** We performed a series of Bayesian phylogeny-trait association tests to reveal the extent of clustering by host species, geographic location, and year of sampling. Although this analysis is conservative, because the null hypothesis of overall random clustering by the trait in question is nearly always rejected, it does provide a useful way to assess the relative strength of clustering of different traits. This analysis revealed a significant phylogenetic association with mosquitoes in all three data subsets (i.e., mosquito sequences are more closely clustered together than expected by chance alone) (Table 3), although most of these involve small clusters of sequences. In addition, given that there is often focused sampling of mosquito WNVs, exemplified by a recent study in California

![FIG 3](https://example.com/figure3.png)

**FIG 3** Genetic variability across the WNV genome. The graph shows the value of $dN - dS$ (left y axis; estimated by the SLAC method) for each codon across the WNV coding sequence (displayed in the top panel). Negative values for $dN - dS$ are shown in gray, and positive values are shown in black. Sites that show consistent evidence for positive selection based on the SLAC, FUBAR, and MEME methods are marked in red. The black line on top (right x axis) indicates the average nucleotide identity estimated as the average value across 100 nucleotides, using a sliding window of 5-nucleotide steps.
and set of host species, we found relatively little geographic sampling bias in the analysis of clustering by host (see Discussion). In marked contrast, no such strong clustering was observed in birds, although only the American crow, American white pelican, and the blue jay had a sufficient sample size for individual analyses. Only the American crow data set for this study \( (P = 0.01) \) and the American white pelican in data subset C \( (P = 0.005) \) exhibited statistically significant clustering. Hence, overall, these data paint a picture of extensive population mixing between different bird species.

To better assess the impact of geographical structure, we divided the WNV sequences into four regions according to the U.S. Census Bureau: North East, Midwest, South, and West (Table 3). The Midwest did not show significant clustering in any data set, whereas significant values were observed in the North East and West regions in all four data sets. The clustering according to U.S. state only exhibited significant values in all available data sets for Connecticut and for California in the three data subsets (Table 3). Importantly, 154 of the 160 sequences from Connecticut and 101 of the 106 sequences from California were isolated from mosquitoes (see above), again revealing that the geography and host species are confounding variables in this case. Indeed, it was recently shown that mosquito sequences sampled between 2003 and 2011 in California form two large clusters, indicative of strongly local evolution (24).

Finally, we tested whether phylogenetic pattern was structured by avian flyway, although it is obviously difficult to distinguish these effects from those of geographic region in general. Accordingly, we grouped the sequences according to which of four gen-  

Finally, we tested whether phylogenetic pattern was structured by avian flyway, although it is obviously difficult to distinguish these effects from those of geographic region in general. Accordingly, we grouped the sequences according to which of four gen-

DISCUSSION

The U.S. epizootic occurrence of WNV is noteworthy for two reasons: the virus spread across the contiguous United States within 4 years, and it was able to infect a wide range of birds and mosquitoes, such that it has a broad host range, with very high mortality in some avian species (49). Previous studies of the molecular evolution of WNV in North America have largely utilized sequences from mosquitoes, been restricted to local areas, relied on partial genomes \( (18, 26, 46, 47, 50) \), or mainly focused on those viruses sampled from humans within the context of WNV in birds \( (23) \). Hence, the current study is notable in that we obtained \( \sim 300 \) complete genome sequences of WNV from birds across the United States. With these data in hand, we were able to reveal that the evolution of WNV in the United States is characterized by relatively little phylogenetic structure (such as distinct clusters of sequences separated by long branches), suggesting that the virus was able to spread in a rapid and generally unconstrained manner. This might, in part, reflect the patterns and dynamics of bird migration across the United States that clearly facilitate rapid virus dispersal. Although we have documented some evidence that phylogenetic structure is shaped by avian flyway, a pattern that has also been observed for avian influenza virus \( (51) \), it is difficult to distinguish the impact of avian flyway from other geographic effects on these data. It is also likely that long distance viral movements, rather than simple homoge-  

DISCUSSION

The U.S. epizootic occurrence of WNV is noteworthy for two reasons: the virus spread across the contiguous United States within 4 years, and it was able to infect a wide range of birds and mosquitoes, such that it has a broad host range, with very high mortality in some avian species (49).

Previous studies of the molecular evolution of WNV in North America have largely utilized sequences from mosquitoes, been restricted to local areas, relied on partial genomes \( (18, 26, 46, 47, 50) \), or mainly focused on those viruses sampled from humans within the context of WNV in birds \( (23) \). Hence, the current study is notable in that we obtained \( \sim 300 \) complete genome sequences of WNV from birds across the United States. With these data in hand, we were able to reveal that the evolution of WNV in the United States is characterized by relatively little phylogenetic structure (such as distinct clusters of sequences separated by long branches), suggesting that the virus was able to spread in a rapid and generally unconstrained manner. This might, in part, reflect the patterns and dynamics of bird migration across the United States that clearly facilitate rapid virus dispersal. Although we have documented some evidence that phylogenetic structure is shaped by avian flyway, a pattern that has also been observed for avian influenza virus \( (51) \), it is difficult to distinguish the impact of avian flyway from other geographic effects on these data. It is also likely that long distance viral movements, rather than simple homoge-

DISCUSSION

The U.S. epizootic occurrence of WNV is noteworthy for two reasons: the virus spread across the contiguous United States within 4 years, and it was able to infect a wide range of birds and mosquitoes, such that it has a broad host range, with very high mortality in some avian species (49).

Previous studies of the molecular evolution of WNV in North America have largely utilized sequences from mosquitoes, been restricted to local areas, relied on partial genomes \( (18, 26, 46, 47, 50) \), or mainly focused on those viruses sampled from humans within the context of WNV in birds \( (23) \). Hence, the current study is notable in that we obtained \( \sim 300 \) complete genome sequences of WNV from birds across the United States. With these data in hand, we were able to reveal that the evolution of WNV in the United States is characterized by relatively little phylogenetic structure (such as distinct clusters of sequences separated by long branches), suggesting that the virus was able to spread in a rapid and generally unconstrained manner. This might, in part, reflect the patterns and dynamics of bird migration across the United States that clearly facilitate rapid virus dispersal. Although we have documented some evidence that phylogenetic structure is shaped by avian flyway, a pattern that has also been observed for avian influenza virus \( (51) \), it is difficult to distinguish the impact of avian flyway from other geographic effects on these data. It is also likely that long distance viral movements, rather than simple homoge-  

DISCUSSION

The U.S. epizootic occurrence of WNV is noteworthy for two reasons: the virus spread across the contiguous United States within 4 years, and it was able to infect a wide range of birds and mosquitoes, such that it has a broad host range, with very high mortality in some avian species (49).

Previous studies of the molecular evolution of WNV in North America have largely utilized sequences from mosquitoes, been restricted to local areas, relied on partial genomes \( (18, 26, 46, 47, 50) \), or mainly focused on those viruses sampled from humans within the context of WNV in birds \( (23) \). Hence, the current study is notable in that we obtained \( \sim 300 \) complete genome sequences of WNV from birds across the United States. With these data in hand, we were able to reveal that the evolution of WNV in the United States is characterized by relatively little phylogenetic structure (such as distinct clusters of sequences separated by long branches), suggesting that the virus was able to spread in a rapid and generally unconstrained manner. This might, in part, reflect the patterns and dynamics of bird migration across the United States that clearly facilitate rapid virus dispersal. Although we have documented some evidence that phylogenetic structure is shaped by avian flyway, a pattern that has also been observed for avian influenza virus \( (51) \), it is difficult to distinguish the impact of avian flyway from other geographic effects on these data. It is also likely that long distance viral movements, rather than simple homoge-
Indeed, our dN/dS analysis failed to detect positive selection on E-Val195Ala, even though this mutation has been experimentally demonstrated to increase fitness in mosquitoes (19). In this context, it is also notable that we observed a far higher proportion of nonsynonymous mutations on internal (compared to external) branches than a previous study (48), indicative of weaker purifying selection in our data. Why these estimates are so different is unclear, although the earlier study was based on global WNV sequences from multiple lineages, whereas we have only considered viruses from a single epizootic occurrence. It is therefore possible that fewer mutations were deleterious in our study set, such that they reached higher frequencies in the population, representing transient polymorphisms rather than true fixation events.

Also of note was our observation that the highest dN/dS ratios (but no positively selected sites) were associated with the capsid protein and one of the lowest ratios was for the envelope gene; not
only is the latter a structural gene, but also it contains the putative 
Val159Ala fitness determinant. Despite the limitations of bioin-
formatics-based analyses of natural selection, at the very least this 
observation suggests that other mutations in the envelope gene are 
strongly deleterious, acting to reduce dNdS values. In turn, it may 
also be that the capsid is able to tolerate more nonsynonymous 
mutations without reducing the viral fitness. Similarly, it is sur-
prising that we consistently observed more positive selection in 
nonstructural than structural genes, indicating that immune-
driven selection is unlikely to be the cause of any fitness increases.

Previous estimates of the rate of evolutionary change in WNV
in the United States vary from $2.3 \times 10^{-3}$ to $8.2 \times 10^{-3}$ substitu-
tions/site/year (21–26, 46, 60). The evolutionary rate estimated in 
our study, (4.08 to 4.32) $\times 10^{-4}$ substitutions/site/year, is within 
the range of those estimated previously and close to that deter-
mined for SLEV (4.10 $\times 10^{-4}$ substitutions/site/year) (61). Simi-
lar evolutionary rates have been reported in three other arthro-
pod-borne viruses associated with human disease: dengue virus 
(62), Japanese encephalitis virus (63), and the emerging Chikun-
gunya virus (CHIKV). In the case of CHIKV, the highest rates 
(sometimes in excess of $1 \times 10^{-3}$ substitutions/site/year) have 
been observed in the lineage associated with the recent human 
epidemic that began in the Indian Ocean (64, 65). Although this 
higher rate may simply represent a strong time dependency, such 
that rates are elevated in the short term due to the presence of 
transient deleterious mutations and depressed in the long term 
due to site saturation (64, 66), it is possible that evolutionary rates 
have increased due to more transmission cycles per unit time and/or stronger immune selection associated with epidemic 
spread. Indeed, it is interesting that some estimates of the rate of 
WNV evolution that have included human sequences (23) are 
higher than those reported here, in which WNVs from humans 
were excluded.

Based on these rates, we estimated the time to the most recent 
common ancestor of WNV in the United States to be between 1997 
and 1998. Although this is obviously before the first documented 
appearance of WNV in 1999, this difference between molecular clock 
and epidemiological estimates may simply reflect a lack of precision 
in sampling dates, as we only have information on the year of isola-
tion and not month or day. An alternative, but less likely, scenario is 
that more than one WNV lineage entered the United States prior to 
its detection in 1999. We similarly estimated the time of the most 
recent common ancestor of the Intermediate and WN02 genotypes 
to be around 1999, indicating that these two genotypes emerged 
shortly after the initial appearance of the NY99 genotype, with SW03 
emerging around 2001. Our Bayesian coalescent analysis also re-
vealed that WNV exhibits a pattern of logistic population pattern in 
the United States, with a rapid increase in viral diversity during the 
initial viral invasion followed by a more constant population size after 
the virus was established nationwide, suggesting that the virus is now 
effectively endemic in the U.S. environment.

Also of note was that our phylogeographic analysis revealed rela-
tively little geographic clustering in most cases. Hence, the virus 
seemingly experienced few geographic constraints as it spread across 
the United States. In contrast, a recent study of WNV phylodynamics 
in California proposed that resident birds are responsible for virus 
overwintering, resulting in strong geographic clustering in this case 
(24). Although plausible, only mosquito sequences were included in 
that analysis, such that the phylogeographic patterns exhibited by 
avian WNV in California is uncertain. Despite this ease of geographic 
movement, our phylogeographic analysis did reveal some (relatively 
weak) significant geographic clustering, with those viruses sampled in 
the North East and West regions of the United States the most dis-
tinct, although this is confounded by the strong association with mos-
quito sampling in Connecticut and California. In this context, it is 
important to note that our sampling is inevitably biased, likely having 
missed those bird species responsible for the majority of cross-coun-
try virus transmission. Indeed, WNV does not cause disease or death 
in all infected birds (4), such that asymptomatic infected animals 
will necessarily be missed in our sampling process that relied on dead 
birds. It has been proposed that the American robin (Turdus migrato-
rius) is the main host for WNV in North America, particularly as 
they experience low levels of mortality (4), are prevalent across 
North America, are preferential hosts for mosquitoes (67), and use migratory flyways, although the exact routes are uncertain (www.
allaboutbirds.org/guide/American_Robin/lifehistory). Unfortunately, 
we have only two sequences from the American robin in our data set, so that we cannot test its impact on WNV evolution and 
epidemiology, although this is clearly a priority for future phylogeographic studies.

Finally, although birds are the main host for WNV, it was striking 
that we only found a significant association between phylogenetic 
clustering and host species for mosquitoes and not for any of the bird 
species for which sufficient data were available. Although the cluster-
ing in the mosquito viruses likely reflects a geographical effect, as 
most of the mosquito sequences sampled here were from Connecti-
cut and California (whereas the avian WNV samples were obtained 
throughout the United States), the lack of clustering by bird species 
reveals the extensive mixing that occurs among bird species, such that 
there are few host and virological barriers to productive infection. 
WNV is evidently a highly successful generalist avian virus.

ACKNOWLEDGMENTS

This project was funded in part with federal funds from the National 
Institute of Allergy and Infectious Diseases, National Institutes of 
Health, Department of Health and Human Services, under contract 
HHSN272200900018C to the Broad Institute’s Genomic Sequencing 
Center for Infectious Diseases. F.D.G. is supported by the Swiss National 
Science Foundation grant number PZ2HP3_151594, and E.C.H. is sup-
ported by an NHMRC Australia Fellowship.

We thank Katy Griffin, Renee Long, and Tina Egstad for their dedic-
ation and expert technical assistance.

The use of trade, product, or firm names is for descriptive purposes 
only and does not imply endorsement by the U.S. Government.

FUNDING INFORMATION

HHS | NIH | National Institute of Allergy and Infectious Diseases (NIAID) 
provided funding to Michael C. Zody, James Qu, Xiao Yang, Bruce Birren, 
Christine M. Malboeuf, and Ruchi M. Newman under grant number 
HHSN272200900018C. Swiss National Science Foundation (SNSF) pro-
vided funding to Francesca Di Giallonardo under grant number 
PZ2HP3_151594, National Health and Medical Research Council 
(NHMRC) provided funding to Edward C. Holmes under grant number 
AF30.

REFERENCES

1. Lanciotti RS, Roehrig JT, Deubel V, Smith J, Parker M, Steele K, Crise 
1999. Origin of the West Nile virus responsible for an outbreak of enceph-
doi.org/10.1126/science.286.5448.2333.


15. Kilpatrick AM, Meola MA, Moody RM, Kramer LD. 2008. Temperature, viral genetics, and the transmission of West Nile virus by Culex pipi-