Published in final edited form as:

Gene. 2007 September 15; 399(2): 152–161.

Rapid Fixation of a Distinctive Sequence Motif in the 3'Noncoding Region of the Clade of West Nile Virus Invading North America

Austin L. Hughes^{1,*}, Helen Piontkivska², and Ivo Foppa³

¹ Department of Biological Sciences, University of South Carolina, Columbia SC 29208

² Department of Biological Sciences, Kent State University, Kent OH 44242

³ Department of Epidemiology and Biostatistics, Arnold School of Public Health, University of South Carolina, Columbia SC 29208

Abstract

Phylogenetic analysis of complete genomes of West Nile virus (WNV) by a variety of methods supported the hypothesis that North American isolates of WNV constitute a monophyletic group, together with an isolate from Israel and one from Hungary. We used ancestral sequence reconstruction in order to obtain evidence for evolutionary changes that might be correlated with increased virulence in this clade (designated the N.A. clade). There was one amino acid change (I→T at residue 356 of the NS3 protein) that occurred in the ancestor of the N.A. clade and remained conserved in all N.A. clade genomes analyzed. There were four changes in the upstream portion of the 3' noncoding region (the AT-enriched region) that occurred in the ancestor of the N.A. clade and remained conserved in all N.A. clade genomes analyzed, changes predicted to alter RNA secondary structure. The AT-enriched region showed a higher rate of substitution in the branch ancestral to the N.A. clade, relative to polymorphism, than did the remainder of the non-coding regions, synonymous sites in coding regions, or nonsynonymous sites in coding regions. The high rate of occurrence of fixed nucleotide substitutions in this region suggests that positive Darwinian selection may have acted on this portion of the 3'NCR and that these fixed changes, possibly in concert with the amino acid change in NS3, may underlie phenotypic effects associated with increased virulence in North American WNV.

Keywords

Evolution of virulence; Flavivirus; natural selection; RNA helicase; West Nile virus

1. Introduction

West Nile virus (WNV), a mosquito-borne single-stranded positive-sense RNA virus classified in the genus *Flavivirus* (family *Flaviviridae*), was named for its initial discovery in the West Nile Province of Uganda (Smithburn et al. 1940). It was first observed in the Western Hemisphere in 1999, when an outbreak was reported in the New York City area (Lanciotti et al. 2002; Nash et al. 2001). Since that time, WNV has spread rapidly across North America, causing extensive mortality in wild bird species, which are its primary natural hosts (McClean 2006). The American Crow *Corvus brachyrhynchos* has suffered particularly high mortality

Corresponding author: Tel +1-803-777-9186; Fax: +1-803-777-4002; Email: austin@biol.sc.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. 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.

from WNV (Yaremych et al. 2004; McClean 2006). Experimental infections of birds of various species have provided additional evidence of a high vulnerability of the American Crow, as well as in certain other members of the crow family (Corvidae) to North American WNV [6]. High mortality was also observed in some other species such as the House Finch *Carpodacus mexicanus* and the Ring-billed Gull *Larus delawarensis* (Komar et al. 2003). High rates of mortality in avian hosts might be explainable by two not necessarily mutually exclusive factors: (1) an unusual lack of resistance to this virus on the part of certain North American bird species; and (2) unusually high virulence of WNV invading North America.

Evidence of differential vulnerability among avian species to infection with WNV provides support for the hypothesis that certain hosts are especially susceptible to WNV infection (Komar et al. 2003). On the other hand, there is also experimental evidence that North American WNV has distinctive biological properties causing greater virulence than seen in certain other WNV isolates. For example, the North American WNV has been shown to be substantially more virulent in American Crows than was a Kenyan isolate (Brault et al. 2004). One factor contributing to this difference may be the enhanced ability of North American WNV to replicate at the high body temperatures ($>43^{\circ}$) found in infected American Crows (Kinney et al. 2006). There are reports that both naturally occurring (Davis et al. 2003) and experimentally induced (Wicker et al. 2006) mutations attenuate virulence of North American WNV.

Comparing sequences of North American WNV isolates with those from the Old World may help identifying nucleotide substitutions that may account for the distinctive biological characteristics of the former (Brault et al. 2004). Phylogenetic methods provide a particularly powerful tool for such comparisons, because they make it possible to reconstruct changes ancestral to a group of related sequences. If virulence has played a role in the rapid of WNV in the new world, any sequence changes that occurred in the ancestor of North American WNV and have remained fixed as the virus has spread across North America are particularly strong candidates for causing high virulence.

Here we use phylogenetic methods to reconstruct the evolutionary relationships of complete WNV genomes from throughout the world. We use these methods to test the hypothesis that the North American WNV represent a monophyletic group or clade (i.e., the group of sequences descended from a common ancestor) within worldwide WNV and to reconstruct nucleotide sequence changes in both the coding sequence and noncoding regions of the WNV genome. We use statistical methods to test the hypothesis that distinctive features of North American WNV were fixed by positive Darwinian selection and that they thus represent adaptive characteristics of the virus.

2. Methods

Analyses were based on 81 genomes of West Nile virus (WNV); two genomes of Kunjin virus (KUNV); and four genomes of Japanese encephalitis virus (JEV; see Figure 1 for accession numbers). The WNV genome encodes a single polyprotein which is subsequently cleaved to form the 10 proteins Core, PreM, Env, NS1, NS2a, NS2b, NS3, NS4a, NS4b, and NS5 (Brinton 2002). Polyprotein sequences were aligned at the amino acid level using CLUSTAL X (Thompson et al. 1997), and the alignment was imposed on the nucleotide sequences. The 5' noncoding region (5'NCR) and the 3' noncoding region (3'NCR) were aligned separately. We used the RNAfold program from the Vienna RNA package (Mathews et al. 1999) with default parameters to predict the secondary structure of the 3'NCR. In phylogenetic analyses, any site at which the alignment postulated a gap in any sequence was excluded so that a comparable number of sites was used for each pairwise comparison.

Because synonymous sites were saturated with changes or close to saturation in the most distant comparisons, initial phylogenetic analyses were based on the amino acid sequence of the polyprotein; JEV sequences were used as an outgroup to root amino acid sequence trees. Additional phylogenetic analyses were based on nucleotide sequences of a monophyletic group of 75 WNV genomes identified from the amino acid sequence trees, rooted with an outgroup identified from the amino acid sequence trees. Synonymous sites in comparisons among these 75 sequences were far from saturation (see Table 1 below). Separate trees were constructed for the coding and non-coding regions of these 75 genomes.

A variety of different methods of phylogenetic reconstruction were used, including the following: maximum parsimony (Swofford 1999); quartet maximum likelihood (Strimmer and von Haeseler 1996); the neighbor-joining (NJ) method (Saitou and Nei 1987); and the Bayesian method (Huelsenbeck and Ronquist 2001). All methods yielded similar results, and only the results of NJ and Bayesian analyses are reported here. NJ trees of amino acid sequences were constructed on the basis of the JTT distance (Jones et al. 1992), while those of nucleotide sequence on the basis of the logdet distance (Lockhart et al. 1994). Bayesian trees of amino acid sequences were based on the JTT + Γ model, which allows for rate variation among sites, while those of nucleotide sequences were based on the GTR + I + Γ model (Rodriguez et al. 1990). All parameters were estimated from the data. Four chains were run for 2,000,000 and 1,000,000 generations for amino acid and nucleotide sequences, respectively, and trees were sampled every 100 generations. Bayesian posterior probabilities were inferred from the last 5000 sampled trees. The reliability of branching patterns in NJ trees was estimated by bootstrapping (Felsenstein 1985); 1000 bootstrap pseudo-samples were used.

Ancestral sequences were reconstructed by the maximum parsimony method using the Paup* program (Swofford 1999) based on the NJ tree. We used the MEGA 3.1 program (Kumar et al. 2004) to analyze patterns of nucleotide substitution. Because of the strong transitional bias in WNV genomes, we used the modified Nei-Gojobori method (Zhang et al. 1998) to estimate the number of synonymous nucleotide substitutions per synonymous site (d_S) and the number of nonsynonymous nucleotide substitutions per nonsynonymous site (d_N). The transition:transversion ratio (R) was estimated from 4-fold degenerate sites in the dataset (R = 7.16). To estimate the number of nucleotide substitutions per site (d) in noncoding regions, we used Kimura's (1980) 2-parameter method. Nucleotide diversity was estimated by the mean of all pairwise d values among a set of allelic sequences (Nei and Kumar 2000). Similarly, synonymous nucleotide diversity was estimated by the mean of all pairwise d_S values among a set of allelic coding sequences; and nonsynonymous nucleotide diversity was estimated by the mean of all pairwise d_N values among a set of allelic coding sequences (e.g., Hughes 2005). The standard errors of mean d_S , mean d_N , and mean d were estimated by the bootstrap method (Nei and Kumar 2000) using 1000 bootstrap pseudo-samples.

A number of authors have described different regions of the 3'NCR of flaviviruses based on such factors as nucleotide content, degree of conservation, occurrence of repeated sequence motifs, and predicted secondary structure (Proutski et al. 1997; Markoff 2003; Tajima et al. 2006; Gritsun and Gould 2006). In general, the more upstream portion of the 3' NCR has been found to be more variable than the downstream portions (Markoff 2003; Tajima et al. 2006). In the present paper, we analyzed separately the region designated the AT-enriched region by Gritsun and Gould (2006) on the basis of their alignment of a wide variety of flavivirus 3'NCR sequences. This region, occurring immediately after the stop codon of the polyprotein, consisted of 103 nucleotides in our alignment of WNV genomes.

3. Results

3.1. Phylogenetic Analyses

The NJ tree of WNV polyprotein sequences, rooted with JEV sequences, strongly supported the hypothesis that WNV and Kunjin virus (KUNV) do not constitute distinct monophyletic groups (Scherret et al. 2001). Several WNV sequences fell outside a cluster of KUNV and other WNV, and this pattern received highly significant (99%) bootstrap support (Figure 1). In addition, those WNV sequences falling outside KUNV were not themselves a monophyletic group; rather there were three separate lineages each supported by a highly significant internal branch (Figure 1). Bakonyi et al. (2006) showed a similar topology in a phylogeny using a smaller number of sequences. Building on earlier analyses (Lanciotti et al. 1999, 2003), Bakonyi et al. (2006) proposed five separate lineages in the WNV phylogeny as follows: (1) the WNV sequences clustering with KUNV constitute lineage 1a; (2) KUNV itself constitutes lineage 1b; (3) lineage 2 is the group including AY688948 (Israel); DQ116961 (Hungary); DQ318019 (Senegal), and DQ318020 (Central African Republic); (4) lineage 3 is represented by AY765264 (Czech Republic); and (5) lineage 4 is represented by AY277251 (Russia). The present analysis strongly supported the monophyly of each of these lineages (99% bootstrap in every case). The Bayesian analysis based on amino acid sequences showed a similar pattern, with strong support (100% posterior probability) for the same lineages (Supplementary Figure S1).

Within lineage 1a, a group of three sequences from Ethiopia (AY603654), Egypt (AF260968), and China (AY490240) formed an outgroup to all others; and this relationship received moderately strong (93%) bootstrap support (Figure 1). However, the phylogenetic analysis based on amino acid sequences did not provide clear resolution to most relationships within lineage 1a. All North American WNV were included in lineage 1a, but North American WNV did not form a monophyletic group (Figure 1). Rather, North American WNV were part of a cluster including AF481864 (Israel) and DQ118127 (Hungary); this cluster received only moderate (78%) bootstrap support (Figure 1). Relationships within the latter cluster were not well resolved (Figure 1). A sequence from Tunisia (AY268133) was the sister group to this cluster, but bootstrap support for this relationship was not strong (57%). The Bayesian analysis showed a similar pattern; in the latter, the cluster of North American WNV along with the Old World isolates AF481864 and DQ118127 was supported by 100% posterior probability (Supplementary Figure S1).

In order to provide better resolution of relationships within lineage 1a, we constructed a phylogenetic tree of nucleotide sequences from coding regions of lineage 1a members (Figure 2). The tree was rooted with the three sequences (from Ethiopia, Egypt, and China) that formed an outgroup to the other lineage 1 sequences in the tree based on amino acid sequences (Figures 1 and 2). In the nucleotide sequence tree, North American WNV clustered together, but this pattern received only 62% bootstrap support (Figure 2). However, the cluster of the North American sequences with AF481864 from Israel received 99% bootstrap support, and the cluster of DQ118127 with the former sequences received 100% bootstrap support (Figure 2). Likewise, the analysis provided 100% bootstrap support for the position of AY268133 from Tunisia as a sister group to the group of North American sequences along with AF481864 from Israel and DQ118127 from Hungary (Figure 2). The Bayesian analysis showed similar results (Supplementary Figure S2).

In order to test whether the evolutionary relationships of noncoding portions of the WNV lineage 1a genome are similar to those of coding regions, we reconstructed the phylogeny separately for the non-coding portions of the genome (combined 5'NCR and 3'NCR; Figure 3). The tree was again rooted with the three sequences (from Ethiopia, Egypt, and China) that formed an outgroup to the other lineage 1 sequences in the tree based on amino acid sequences

(Figures 1 and 3). In this tree, there was 99% support for a cluster including North American sequences along with AF481864 from Israel and DQ118127 from Hungary (Figure 3). AY268133 from Tunisia was a sister group to this cluster, although with only 58% bootstrap support (Figure 3). The Bayesian analysis showed similar results (Supplementary Figure S3).

3.2. Patterns of Nucleotide Substitution

Table 1 shows nucleotide diversity estimates in the 5'NCR, in the AT-enriched region of the 3'NCR, and in the remainder of the 3'NCR for comparisons among all lineage 1a members, and for pairwise comparisons among all N.A. clade members. Nucleotide diversity was greater in the AT-enriched region than that in the remainder of the 3'NCR in all cases, and the difference was statistically significant in the case of all lineage 1a members (Table 1). By contrast, in all cases, nucleotide diversity in the remainder of the 3'NCR (excluding the AT-enriched region) was significantly less than synonymous nucleotide diversity (mean d_S) for the same comparison (Table 1). These results provide evidence of strong purifying selection and thus of functional constraint on the 3'NCR outside of the AT-enriched region.

We used maximum parsimony to reconstruct ancestral sequences of (1) the most recent common ancestor of the clade including North American sequences along with AF481864 from Israel and DQ118127 from Hungary (henceforth referred to as the “N. A. clade”); and (2) the immediate ancestor of the latter; i.e., the most recent common ancestor of the N. A. clade and AY268133 from Tunisia. Differences between these two reconstructed sequences represent changes on the branch leading to the N.A. clade.

In the coding sequence, d_S along the branch leading to the N.A. clade was much greater than d_N along the same branch, and the difference was highly significant (Table 1). Similarly, synonymous nucleotide diversity (mean d_S) was significantly greater than nonsynonymous nucleotide diversity (mean d_N) in all of lineage 1a and in the N.A. clade (Table 1). These results reflect a predominance of purifying selection acting to eliminate nonsynonymous mutations in the coding region of the polyprotein. Along the branch leading to the N.A. clade, mean d was significantly greater in the AT-enriched region than in the remainder of the 3'NCR, whereas d in the remainder of the 3'NCR was significantly lower than d_S in the coding region (Table 1). Thus, in the ancestry of the N.A. clade, there was evidence of strong purifying selection on the 3'NCR outside of the AT-enriched region but not on the AT-enriched region itself.

Fixed Differences vs. Polymorphism—Evidence of rapid evolution in the ancestor of a clade can be obtained by comparing changes fixed along the branch ancestral to the clade with polymorphisms within the clade (McDonald and Kreitman 1991). According to our phylogenetic reconstruction, there were 7 nucleotide changes in the AT-enriched region along branch ancestral to the N.A. clade and none in the remainder of the non-coding regions (both 5'NCR and 3'NCR). By contrast, within the N.A. clade, there were 11 polymorphic sites in the AT-enriched region and 21 in the remainder of the non-coding regions. The difference between fixed differences and polymorphisms was highly significant ($P = 0.002$; Fisher's exact test). This difference indicates an acceleration of change in the AT-enriched region, relative to polymorphism, along the branch leading to the N.A. clade in comparison to the other non-coding regions.

The ratio of fixed differences to polymorphisms in the AT-enriched region (7:11) was also significantly different from that (78: 467) at synonymous sites in the coding region ($P = 0.011$; Fisher's exact test). Similarly, the ratio of fixed differences to polymorphisms in the AT-enriched region was also significantly different from that (4:111) at nonsynonymous sites in the coding region ($P < 0.001$; Fisher's exact test). These results indicate that nucleotide changes along the branch leading to the N.A. clade occurred more rapidly, relative to polymorphism,

in the AT-enriched region than at either synonymous or nonsynonymous sites in the coding region.

Of the seven fixed differences in the AT-enriched region that occurred on the branch ancestral to the N.A. clade, four were both unique to the N.A. clade and conserved throughout the N.A. clade. At each of these sites, the substitution that occurred in the ancestor of the N.A. clade did not occur elsewhere in the phylogeny, and the site was not polymorphic within the N.A. clade. These four substitutions were U→A at position 7 of the 3'NCR; A→G at position 46 of the 3' NCR; A→G at position 70 of the 3'NCR; and A→G at position 99 of the 3'NCR. The proportion of unique conserved changes (4 of 103 or 4.3%) was significantly higher in the AT-enriched region than in the remainder of the non-coding region, where there were no unique conserved changes in 364 aligned sites ($P = 0.002$; Fisher's exact test). The proportion of unique conserved changes in the AT-enriched region was also significantly different from that at synonymous sites in the coding region (37 at 2932 sites or 1.3%; $P = 0.048$; Fisher's exact test). There was only one amino acid change, AUC (I) → ACC (T) at position 356 of the NS3 protein, that was fixed along the branch ancestral to the N.A. clade and was conserved throughout the N.A. clade.

3.3. RNA Secondary Structure Prediction

Secondary structure prediction of the 3'NCR by the RNAfold program revealed substantial differences between the N.A. clade and other members of lineage 1a, in spite of a relatively small number of nucleotide differences (Figure 4). For high-resolution images of the secondary structures shown in Figure 4 see Supplementary Figure S4. For example, AF260968 (from Egypt; Figure 4A) and DQ164194 (from the USA; Figure 4B) had nucleotide differences at only 21 of 631 nucleotide sites in the 3'NCR (3.3%); but, at the level of secondary structure, 212 sites (33.6%) differed as regards to whether they formed part of stems or loops. When the RNAfold secondary structure prediction was applied to the reconstructed common ancestor of the N.A. clade (Figure 4C) and to the latter's immediate ancestor (Figure 4D), the difference in structure was even more striking than that between extant sequences. There were only 7 nucleotide differences between these two sequences (1.1%), including the four unique changes fixed in ancestor of the N.A. clade and conserved throughout the N.A. clade. Yet 90 sites (14.7%) differed as regard to whether they formed stems or loops. One feature shared by extant members of the N.A. clade (Figure 4B) and by the predicted ancestor of the N.A. clade (Figure 4C) was a stem-loop structure in which the loop included nucleotide 46, one of the positions with a unique substitution conserved in the N.A. clade. This stem-loop structure was absent from lineage 1a members outside of the N.A. clade (Figure 4A) and from the predicted immediate ancestor of the N.A. clade ancestor (Figure 4D).

In order to test the hypothesis that the four unique differences conserved in the N.A. clade were primarily responsible for the structural differences between the ancestor of the N.A. clade and the latter's immediate ancestor, we analyzed an artificial sequence which differed from the latter only at those four positions (Figure 4E). (Note that this sequence does not correspond to an inferred ancestor but was constructed only to test the effect of these four changes on secondary structure.) The overall structure of the artificial sequence was predicted to be very similar to that of the ancestor of the N.A. clade (compare Figures 4C and 4E) but rather different from the sequence lacking those four changes (compare supplementary Figures 4D and 4E). In fact, as a result of these four changes (representing 0.6% of all sites), 84 sites (13.3%) differed as regards to whether they formed part of stems or loops. Thus the small number of unique nucleotide substitutions conserved in the N.A. clade were predicted to have a large impact on the secondary structure of the 3'NCR. When the effects of these four changes were analyzed separately, they did not each have an equally large effect on the predicted secondary structure. For example the stem-loop structure highlighted in Figures 4B and 4C was predicted to occur

in the event of the A→G substitution at position 70 alone, but not when any of the other four substitutions occurred in the absence of the others (not shown).

4. Discussion

Phylogenetic analyses by various methods supported the five lineages of West Nile virus (WNV) proposed by Bakonyi et al. (2006), one of which corresponds to Kunjin virus (KUNV). Thus, WNV as currently defined, is paraphyletic. As shown by previous analysis of a smaller number of sequences, one WNV genome from Israel (AF481864) and one from Hungary (DQ118127) form a clade with sequences from North America. For convenience, we designated this clade as the “N.A. clade,” while recognizing that this clade originated outside North America and that it includes members outside North America. Bakonyi et al. (2006) have discussed scenarios whereby this clade may have arisen in central Europe and have been carried from there both to Israel and to New York.

If North American WNV is characterized by unusual virulence, we would expect to see nucleotide sequence changes responsible for that virulence to have been fixed in the N.A. clade. Overall, the branch leading to the N.A. clade was characterized by purifying selection on the coding region, but one unique amino acid sequence change occurred that remained conserved in the N.A. clade. This change (I→T at residue 356 of the NS3 protein) may have functional importance. The NS3 protein has RNA helicase activity, and the three-dimensional structure of the homologous helicase from another flavivirus, Yellow fever virus (YFV), has been determined (Wu et al. 2005). By analogy with YFV, residue 356 is located near the beginning of a β -strand that includes the initial two residues of a sequence motif (Motif IV) believed to play a role in translocating RNA through the helicase (Wu et al. 2005).

Because the amino acid replacement at residue 356 of NS3 involved only a single codon, it was not possible to determine by available statistical methods whether this change occurred as a result of positive Darwinian selection. On the other hand, there was evidence of rapid evolution in a portion of the 3′ non-coding region (3′NCR) on the branch leading to the N.A. clade. In the AT-enriched region (Gritsun and Gould 2006), four unique nucleotide changes occurred that remained conserved throughout the N.A. clade. Moreover, the AT-enriched region showed a higher rate of substitution in the branch ancestral to the N.A. clade, relative to polymorphism, than did the other non-coding regions, synonymous sites in coding regions, or nonsynonymous sites in coding regions. This in turn suggests the possibility that positive Darwinian selection may have acted on the AT-enriched region in the ancestor of the N.A. clade.

The fact that substitutions along the branch ancestral to the N.A. clade were much more prevalent in the AT-enriched region than expected by the pattern of polymorphism within the N.A. clade is strong evidence against the hypothesis that the observed pattern could be explained by a high mutation rate alone. The same observation argues against the hypothesis that purifying selection on the AT-enriched region is unusually relaxed in the N.A. clade. The comparison between polymorphism and divergence should not be used uncritically as a test for positive selection because the fixation of slightly deleterious alleles during a bottleneck might lead to a pattern indistinguishable from that associated with positive selection (Ohta 1993; Eyre-Walker 2002). However, most slightly deleterious mutations will occur at nonsynonymous mutations in coding regions (Hughes et al. 2003; Hughes 2007); but, in the case of the N.A. clade, the ratio of nonsynonymous divergence to polymorphism (4:110) was much lower than that of synonymous divergence to polymorphism (78:468). Thus, the coding region provided no suggestion of the fixation of deleterious mutations during a bottleneck in the ancestry of the N.A. clade, suggesting that such a process is unlikely to have occurred in the AT-enriched region either.

In addition to the evidence of an accelerated rate of nucleotide substitution in the AT-enriched region, the same region showed evidence of an unusually high proportion of unique conserved changes in the branch ancestral to the N.A. clade; in other words, changes that occurred nowhere else in the phylogeny and remained conserved throughout the N.A. clade. The high rate of conservation within the N.A. clade of changes in this AT-enriched region is consistent with the hypothesis that these substitutions play an adaptive role in the biology of the viruses in this clade. Further support for this hypothesis derives from the observation that these positions were unchanged in the N.A. clade in spite of the high variability of the AT-enriched region in lineage 1a WNV outside the N.A. clade (Table 1).

Several studies have emphasized the low genetic variation among WNV isolates in North America and evidence that a dominant variant has emerged as WNV has spread across the continent (Bertolotti et al. 2007; Davis et al. 2002, 2003, 2004; Ebel et al. 2004). Ebel et al. (2004) reported a particularly low frequency of nucleotide variants in the 3'NCR. The latter observation is consistent with the hypothesis that nucleotide changes in the 3'NCR in the ancestor of the N.A. clade of WNV have played a role in the rapid spread of this virus in the New World.

RNA secondary structure prediction suggested that the 3'NCR has a very different secondary structure in the N.A. clade from that seen in other members of lineage 1a and that the four unique substitutions conserved in the N.A. clade brought about substantial differences in secondary structure. Changes in the secondary structure of this portion of the 3'NCR may influence viral replication, as is known to be true of the 3' terminal 100 nucleotides of the genome of flaviviruses (Yu and Markoff 2005). In JEV, the NS3 helicase associates with NS5, which contains the viral RNA dependent polymerase (Chen et al. 1997). Thus, an intriguing hypothesis is that the nucleotide changes in the 3'NCR have co-evolved with the amino acid change at residue 356 of NS3, and that together these substitutions affect the interactions of the helicase and the RNA polymerase with their RNA substrate. The possible phenotypic effects of the fixed changes in the 3'NCR of the N.A. clade, as well as that of the fixed amino acid change in NS3, remain to be addressed experimentally; for example, by site-directed mutagenesis.

Virulence of WNV, as of any pathogen, is likely to involve a complex pattern of interaction between virus and host. The fact that different avian species show differences in susceptibility to WNV highlights the importance of host interactions (Komar et al. 2003). If the fixed nucleotide changes in the 3'NCR of the N.A. clade play a role in increased virulence, this need not imply that all hosts species will be equally affected by this virulence. Furthermore, outside of the N.A. clade, there may be other virulent lineages of WNV in which different mutational patterns underlie enhanced virulence.

Moreover, the unusual pattern of virulence seen in WNV infection of North American birds may also depend on unique genetic characteristics of the host species that render them especially vulnerable. The hypothesis that species especially vulnerable to WNV are particularly numerous in the Nearctic avifauna might explain observations suggesting lesser susceptibility of wild birds in the Neotropics, even though WNV infecting tropical America presumably originated in North America (Komar and Clark 2006). One possible factor in this difference in susceptibility between Nearctic and Neotropical birds may be the reduced genetic diversity of the former, evidently an effect of glaciation on effective population sizes of North American birds (Hughes and Hughes 2007). For example, reduced genetic diversity at host immune defense loci such as those of the major histocompatibility complex (Hughes 1991; Hughes and Nei 1988) might be a factor in increased vulnerability to infection on the part of the Nearctic avifauna.

The clade of West Nile virus that has invaded North America was found to have a small number of unique fixed nucleotide substitutions. Four nucleotide substitutions in the 3' non-coding region (3'NCR) evolved rapidly in the ancestor of this clade, a pattern of accelerated evolution suggestive of positive selection. These four nucleotide changes are predicted to have a major impact on the secondary structure of the 3'NCR, which in turn may influence viral virulence; for instance, by an effect on viral replication. Our analyses suggest experiments that may help to determine the molecular basis of the high virulence that characterizes WNV infecting certain avian hosts in North America.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This research was supported by grant GM43940 from the National Institutes of Health to A.L.H. and by the Research Council of Kent State University.

References

- Bakonyi T, Ivanics E, Erdélyi K, Ursu K, Ferenczi E, Weissenböck H, Nowotny N. Lineage 1 and 2 strains of encephalitic West Nile virus, Central Europe. *Emerg Infect Dis* 2006;12:618–623. [PubMed: 16704810]
- Bertolotti L, Kitron U, Goldberg TL. Diversity and evolution of West Nile virus in Illinois and The United States, 2002–2005. *Virology* 2006;360:143–149. [PubMed: 17113619]
- Braut AC, Langevin SA, Bowen RA, Panella NA, Biggerstaff BJ, Miller BR, Nicholas K. Differential virulence of West Nile strains for American Crows. *Emerg Infect Dis* 2004;10:2161–2168. [PubMed: 15663854]
- Brinton MA. The molecular biology of West Nile virus: a new invader of the Western Hemisphere. *Annu Rev Microbiol* 2002;56:371–402. [PubMed: 12142476]
- Chen CJ, Kuo MD, Chen LJ, Hsu SL, Wang YM, Lin JH. RNA-protein interactions: involvement of NS3, NS5, and 3' noncoding regions of Japanese encephalitis virus genomic RNA. *J Virol* 1997;71:3466–3473. [PubMed: 9094618]
- Davis CT, Beasley DW, Guzman H, Raj P, D'Anton M, Novak RJ, Unnasch TR, Tesh RB, Barrett AD. Genetic variation among temporally and geographically distinct West Nile virus isolates, United States, 2001, 2002. *Emerg Infect Dis* 2003;9:1423–1429. [PubMed: 14718086]
- Davis CT, Beasley DW, Guzman H, Siirin M, Parsons RE, Tesh RB, Barrett AD. Emergence of attenuated West Nile virus variants in Texas, 2003. *Virology* 2004;330:342–350. [PubMed: 15527859]
- Davis CT, Ebel GD, Lanciotti RS, Braut AC, Guzman H, Siirin M, Lambert A, Parsons RE, Beasley DW, Novak RJ, Elizondo-Quiroga D, Green EN, Young DS, Stark LM, Drebot MA, Artsob H, Tesh RB, Kramer LD, Barrett AD. Phylogenetic analysis of North American West Nile virus isolates, 2001–2004: evidence for the emergence of a dominant genotype. *Virology* 2005;342:252–265. [PubMed: 16137736]
- Ebel GD, Carricaburu J, Young D, Bernard KA, Kramer LD. Genetic and phenotypic variation of West Nile Virus in New York, 2000–2003. *Am J Trop Med Hyg* 2004;71:493–500. [PubMed: 15516648]
- Eyre-Walker A. Changing effective population size and the McDonald-Kreitman test. *Genetics* 2002;162:2017–2024. [PubMed: 12524367]
- Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 1985;39:783–791.
- Gritsun TS, Gould EA. Direct repeats in the 3' untranslated regions of mosquito-borne flaviviruses: possible implications for virus transmission. *J Gen Virol* 2006;87:3297–3305. [PubMed: 17030864]
- Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 2001;17:754–755. [PubMed: 11524383]
- Hughes AL. MHC polymorphism and the design of captive breeding programs. *Conservation Biol* 1991;5:249–251.

- Hughes AL. Evidence for abundant slightly deleterious polymorphisms in bacterial populations. *Genetics* 2005;169:533–538. [PubMed: 15545641]
- Hughes AL. Micro-scale signature of purifying selection in Marburg virus genomes. *Gene* 2007;392:266–272. [PubMed: 17306473]
- Hughes AL, Hughes MA. Coding sequence polymorphism in avian mitochondrial genomes reflects population histories. *Mol Ecol* 2007;16:1369–1376. [PubMed: 17391262]
- Hughes AL, Nei M. Pattern of nucleotide substitution at MHC class I loci reveals overdominant selection. *Nature* 1988;335:167–170. [PubMed: 3412472]
- Hughes AL, Packer B, Welch R, Bergen AW, Chanock SJ, Yeager M. Widespread purifying selection at polymorphic sites in human protein-coding loci. *Proc Natl Acad Sci USA* 2003;100:15754–15757. [PubMed: 14660790]
- Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* 1992;8:275–282. [PubMed: 1633570]
- Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* 1980;16:111–120. [PubMed: 7463489]
- Kinney RM, Huang CYH, Whiteman MC, Bowen RA, Langevin SA, Miller BR, Brault AC. Avian virulence and thermostable replication of the North American strain of *West Nile virus*. *J Gen Virol* 2006;87:3611–3622. [PubMed: 17098976]
- Komar N, Clark GG. West Nile virus activity in Latin America and the Caribbean. *Rev Panam Salud Publica* 2006;19:112–117. [PubMed: 16551385]
- Komar N, Langevin S, Hinten S, Nemeth N, Edwards E, Hettler D, Davis B, Bowen R, Bunning M. Experimental infection of North American birds with the New York 1999 strain of West Nile virus. *Emerg Infect Dis* 2003;9:311–322. [PubMed: 12643825]
- Kumar S, Tamura K, Nei M. MEGA3: integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinf* 2004;5:150–163.
- Lanciotti RS, et al. Origin of the West Nile virus responsible for an outbreak of encephalitis in the Northeastern United States. *Science* 1999;286:2333–2337. [PubMed: 10600742]
- Lanciotti RS, Ebel GD, Deubel V, Kerst AJ, Murri S, Meyer R, Bowen M, McKinney N, Morrill WE, Crabtree MB, Kramer LD, Roehrig JT. Complete genome sequences and phylogenetic analysis of West Nile virus strains isolated from the United States, Europe, and the Middle East. *Virology* 2002;298:96–105. [PubMed: 12093177]
- Lockhart PJ, Steel MA, Hendy MD, Penny D. Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol Biol Evol* 1994;11:605–612.
- McClean R. West Nile virus in North American birds. *Ornithol Monogr* 2006;60:44–64.
- Markoff L. 5'- and 3'-noncoding regions in flavivirus RNA. *Adv Vires Res* 2003;59:177–228.
- Matthews DH, Sabina J, Zuker M, Turner DH. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* 1999;288:911–940. [PubMed: 10329189]
- McDonald JH, Kreitman M. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* 1991;351:114–116.
- Nash D, Mostashari F, Fine A, Miller J, O'Leary D, Murray K, Huang A, Rosenberg A, Greenberg A, Sherman M, Wong S, Layton M. The outbreak of West Nile virus infection in the New York City area in 1999. *New England J Med* 2001;344:1807–1814. [PubMed: 11407341]
- Nei, M.; Kumar, S. Molecular evolution and phylogenetics. New York: Oxford University Press; 2000.
- Ohta T. Amino acid substitution at the *Adh* locus of *Drosophila* is facilitated by small population size. *Proc Natl Acad Sci USA* 1993;90:4548–4551. [PubMed: 8506297]
- Proutski V, Gaunt MW, Gould EA, Holmes EC. Secondary structure of the 3'-untranslated region of yellow fever virus: implications for virulence, attenuation and vaccine development. *J Gen Virol* 1997;78:1543–1549. [PubMed: 9225027]
- Rodriguez F, Oliver JF, Marí A, Medina JR. The general stochastic model of nucleotide substitution. *J Theor Biol* 1990;142:485–501. [PubMed: 2338834]
- Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 1987;4:406–425. [PubMed: 3447015]

- Scherret JH, Poidinger M, MacKenzie JS, Broom AK, Deubel V, Lipkin WI, Briese T, Gould EA, Hall RA. The relationships between West Nile and Kunjin viruses. *Emerg Infect Dis* 2001;7:697–705. [PubMed: 11585535]
- Smithburn KC, Hughes TP, Burke AW, Paul JH. A neurotropic virus isolated from the blood of a native Ugandan. *Am J Trop Med* 1940;20:471–492.
- Strimmer K, von Haeseler A. Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. *Mol Biol Evol* 1996;13:964–969.
- Swofford, DL. PAUP*: Phylogenetic Analysis Using Parsimony (*and other methods). Sinauer; Sunderland MA: 1999.
- Tajima S, Nukui Y, Ito M, Takasaki T, Kurane I. Nineteen nucleotides in the variable region of 3' non-translated region are dispensable for the replication of dengue type 1 virus in vitro. *Virus Research* 2006;116:38–44. [PubMed: 16253374]
- Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Diggins DG. The CLUSTAL X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* 1997;25:4876–4882. [PubMed: 9396791]
- Wicker JA, Whiteman MC, Beasley DWC, Davis CT, Zhang S, Schneider BS, Higgs S, Kinney RM, Barrett ADT. A single amino acid substitution in the central portion of the West Nile virus NS4B protein confers a highly attenuated phenotype in mice. *Virology* 2006;349:245–253. [PubMed: 16624366]
- Wu J, Bera AK, Kuhn RJ, Smith JL. Structure of the flavivirus helicase: implications for catalytic activity, protein interactions, and proteolytic processing. *J Virol* 2005;79:10268–10277. [PubMed: 16051820]
- Yaremych SA, Warner RE, Mankin PC, Brawn JD, Raim A, Novak R. West Nile virus and high death rate in American crows. *Emerg Infect Dis* 2004;10:709–711. [PubMed: 15200865]
- Yu L, Markoff L. The topology of bulges in the long stem of the flavivirus 3' stem-loop is a major determinant of RNA replication competence. *J Virol* 2005;79:2309–2324. [PubMed: 15681432]
- Zhang J, Rosenberg HF, Nei M. Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc Natl Acad Sci USA* 1998;95:3708–3713. [PubMed: 9520431]

Abbreviations

3'NCR	3' non-coding region
5'NCR	5'non-coding region
d_N	the number of nonsynonymous nucleotide substitutions per nonsynonymous site
d_S	the number of synonymous nucleotide substitutions per synonymous site
JEV	Japanese encephalitis virus
KUNV	Kunjin virus
WNV	West Nile virus
YFV	yellow fever virus

Figure 2.

NJ tree of coding nucleotide sequences of WNV lineage 1a (based on 10286 aligned sites). Branches with less than 50% bootstrap support are condensed, and topology only is shown. Following Figure 1, AY603654 (Ethiopia), AF260968 (Egypt), and AY490240 (China) were used as an outgroup to root the tree.

Figure 3.

NJ tree of noncoding nucleotide sequences of WNV lineage 1a (based on 467 aligned sites). Branches with less than 50% bootstrap support are condensed, and topology only is shown. Following Figure 1, AY603654 (Ethiopia), AF260968 (Egypt), and AY490240 (China) were used as an outgroup to root the tree.

Figure 4.

Predicted secondary structures of WNV 3' NCRs: (A) AF260968 (a representative member of lineage 1a outside the N.A. clade); (B) DQ164195 (a representative member of the N.A. clade); (C) common ancestor of the N.A. clade; (D) immediate ancestor of the common ancestor of the N.A. clade (ancestor of the N.A. clade and its nearest sister group); (E) sequence differing from that in (D) by only the four unique nucleotide substitutions conserved in the N.A. clade. The *arrow* (B and C) indicates a stem-loop structure in which position 46 of the 3'NCR contributes to the loop.

Table 1
Nucleotide substitutions in coding and noncoding regions of WNV.

Region (no. sites)	Distance	Nucleotide diversity ($\pm$ S.E) for lineage 1a, excluding N.A. clade (N= 20)	Nucleotide diversity ($\pm$ S.E) for N.A. clade (N = 55)	Changes along branch to N.A. clade ($\pm$ S.E.)
Coding (3420 codons)	d_S	0.1005 ± 0.0022^c	0.0110 ± 0.0008^c	0.0271 ± 0.0031^c
	d_N	0.0022 ± 0.0004^{cd}	0.0008 ± 0.0001^{ad}	0.0005 ± 0.0003^{ad}
5'NCR (54)	d	0.0055 ± 0.0043^a	0.0075 ± 0.0032	0.0000 ± 0.0000
AT-enriched (103)	d	0.0756 ± 0.0183^b	0.0091 ± 0.0048	0.0807 ± 0.0317^a
Remainder 3' NCR (310)	d	0.0189 ± 0.0043	0.0021 ± 0.0006	0.0000 ± 0.0000

Tests of the hypothesis that a value equals d in the 3'NCR (remainder):

^aP < 0.05;

^bP < 0.01;

^cP < 0.001.

Tests of the hypothesis that $d_N = d_S$:

^dP < 0.001.