MAJOR ARTICLE

Global Origin and Transmission of Hepatitis C Virus Nonstructural Protein 3 Q80K Polymorphism

Rosemary M. McCloskey,[1] Richard H. Liang,[1] Jeffrey B. Joy,[1] Mel Krajden,[2] Julio S. G. Montaner,[1,3] P. Richard Harrigan,[1,3] and Art F. Y. Poon[1,3]

[1]BC Centre for Excellence in HIV/AIDS, [2]BC Centre for Disease Control, and [3]Department of Medicine, University of British Columbia, Vancouver, Canada

Hepatitis C virus (HCV) has a naturally occurring polymorphism, Q80K, in the nonstructural protein 3 (NS3) gene encoding the viral protease, which has been associated with reduced susceptibility to the direct-acting antiviral inhibitor simeprevir. Q80K is observed predominantly in HCV genotype 1a and seldom in other HCV genotypes; moreover, it has a markedly high prevalence in the United States. Here, we reconstruct the evolutionary history of this polymorphism to investigate why it is so highly localized in prevalence and whether it is stably transmitted between hosts. We found that the majority (96%) of HCV infections carrying Q80K were descended from a single lineage in which a Q80K substitution occurred around the 1940s in the United States, which implies that this polymorphism is likely highly transmissible. Furthermore, we identified 2 other substitutions in NS3 that may interact with Q80K and contribute to its stability. Our results imply that the current distribution and prevalence of Q80K are unlikely to change significantly in the short term.

Keywords. hepatitis C virus; Q80K; simeprevir; ancestral reconstruction; molecular phylogenetics; virus evolution; phylogeography.

The polymorphism Q80K in hepatitis C virus (HCV) nonstructural protein 3 (NS3) has been associated with a reduced response to the protease inhibitor simeprevir in combination with pegylated interferon and ribavirin, and routine screening for Q80K in patients infected with HCV genotype 1a is required as part of the Food and Drug Administration's (FDA's) adoption of simeprevir [1]. Consequently, information about the polymorphism's transmissibility and geographic distribution could help inform the future deployment of simeprevir to treat HCV genotype 1a infections worldwide. Surveys of treatment-naive individuals and baseline genotyping for clinical trials have indicated that Q80K is present in a sizeable fraction of untreated HCV genotype 1a infections, ranging from 5% to 47%, depending on geographic region [2–6]. It is likely, therefore, that the polymorphism did not arise in response to drug pressure and that its presence at baseline in HCV genotype 1a infections must be due to some other factor. It is also unknown whether Q80K arises de novo and/or is transmitted between hosts.

Here, we examined the origin, transmission, and geographic spread of Q80K globally. Using publicly available sequence data, we reconstructed the geographic and temporal origin of this mutation. Using both phylogenetic and association-based methods, epistasis was explored as a possible explanation for Q80K's prevalence and stability. Our results show that Q80K arose early in the HCV genotype 1a epidemic, appears to be transmissible, and is likely to persist in roughly its present distribution in the short term.

MATERIALS AND METHODS

Data Collection and Alignment

On 5 July 2014, we retrieved all HCV sequences in GenBank, using the search query "hepatitis + C + virus

Received 14 August 2014; accepted 15 October 2014; electronically published 10 November 2014.

Presented in part: Conference on Retroviruses and Opportunistic Infections, Boston, Massachusetts, 3–6 March 2014 (abstract 656LB); 21st International HIV Dynamics and Evolution meeting, Tucson, Arizona, 7–10 May 2014 (abstract 5).

Correspondence: Art F. Y. Poon, PhD, BC Centre for Excellence in HIV/AIDS, 680-1081 Burrard St, St. Paul's Hospital, Vancouver, BC, Canada V6Z1Y6 (apoon@cfenet.ubc.ca).

The Journal of Infectious Diseases® 2015;211:1288–95

DOI: 10.1093/infdis/jiu613

[orgn]," producing 160 071 records. Records that were not annotated with both country and year were removed, leaving 45 449 sequences. Each sequence was aligned pairwise against the HCV genotype 1a reference genome H77 (accession no. NC_004102), using MAFFT v6.850b [7]. The BioPython library [8] was used to strip out insertions relative to H77, and the alignment was clipped to codon positions 9–174 of NS3 (corresponding to nucleotides 3421–919 of H77). Sequences in which at least 50% of this region consisted of gaps or ambiguous nucleotides or where codon 80 was missing or ambiguous were removed, leaving 7594 sequences.

To select only HCV genotype 1a sequences, the relevant region of the 2012 genotype reference alignment from the Los Alamos National Laboratory HCV Database [9] was appended to the NS3 sequences. FastTree (v2.1.7) [10] with a general time-reversible model was used to build an approximate maximum likelihood tree from the nucleotide alignment. We selected the largest clade in this tree that contained all HCV genotype 1a reference sequences and no other reference sequences. Only 2656 sequences within this 1a clade were retained.

We restricted these data to only 1 sequence per person. In the absence of consistent annotation in GenBank, we deduced this information from sequence variation. A maximum likelihood tree was built from the alignment by use of the same methods described above. All clusters of tip nodes in this tree with pairwise distances of <0.04 expected substitutions per site were reduced to a single tip by retaining only the node most distant from the root, and their associated sequences were removed from the alignment. This cutoff was determined from the distribution of pairwise distances in the tree (Supplementary Figure 1), which exhibited 3 modes, including one near zero, that we interpreted as representing intrahost variation. It is possible that members of transmission clusters sampled very near to the time of transmission would also have been excluded by this step, resulting in a slightly more conservative data set but not biasing the results. This left 677 sequences.

We added 47 sequences from individuals in the Vancouver Injection Drug Users Study cohort [11], which had been processed with the same procedure, for a total of 724 sequences. Nearly all of these sequences were sampled during 1996–1997, while the majority of the GenBank sequences were sampled after 2000. Preliminary experiments indicated that these less recently sampled Canadian sequences may have been biasing the results. In particular, phylogeographic analyses of subsampled alignments almost always placed the root of HCV genotype 1a in Canada. Therefore, we constructed 2 different data sets: one including every available sequence, and one including only those sampled since the year 2000. Additionally, because Q80K occurs frequently in genotypes 5 and 6 [6, 12], we retrieved all NS3 sequences in GenBank of these 2 genotypes (n = 5 and 50, respectively), using GenBank annotation for genotype classification.

Phylogenetic Inference and Ancestral Reconstruction

Using the alignment with codon 80 removed, we again built a maximum likelihood tree. A modified v1.4.0 of the RootToTip utility, originally distributed as part of the Path-O-Gen application within BEAST [13], was used to root the tree under the molecular clock assumption and to produce a linear regression of root-to-tip distances against sampling times. The chronos function within the R package ape [14] (modified to accept tips at varying dates) was used to fix the tips of the tree to their year of sampling and rescale the internal branches accordingly. Because the penalized likelihood method used by chronos to estimate divergence times does not produce confidence intervals (CIs), we estimated the timing of internal nodes by a linear regression of tree distances against sampling times [15].

To reconstruct the ancestral sequence at each internal node, we used the MG94 × REV codon substitution model [16], which combines a general time-reversible model with differing rates for synonymous and nonsynonymous codon substitutions. HyPhy v2.2 [17] was used to jointly optimize model parameters and ancestral states by maximum likelihood. All reported substitutions are relative to the inferred ancestral sequence of HCV genotype 1a (Supplementary Data). In particular, we reconstructed both S174 and A91 in this sequence, as opposed to residues N174 and S91, which are present in the H77 HCV genotype 1a reference sequence. Vallet et al also report these residues in the consensus sequence of a group of HCV genotype 1a–infected individuals [5].

By using this phylogeny and set of ancestral sequences, it is straightforward to infer which mutations occurred along a branch in the tree by comparing the observed or reconstructed sequences at each end of the branch. We determined the branches where the Q80K substitution arose or reverted, as well as the S174N and A91S/T substitutions that we identified as of potential interest (see the "Site Interactions" subsection). To compare Q80K's evolutionary history with that of other mutations associated with HCV drug resistance, we also determined which branches gave rise to any mutation in a published list [18]. We were only able to investigate 5 of 13 listed positions. At 7 positions, the listed mutations (F43S, A165F/N/S/T/V, V158I, and I170A/T, Q80R, I132V, and D168A/E/G/H/N/T/V/Y) were each present in <10 sequences. We could not investigate position 175 because of low coverage in public data.

Site Interactions

The Fisher exact test was used to test the statistical association of each mutation with Q80K in observed sequences. This method is less specific than phylogenetically informed tests for interaction and may produce spurious correlations [19]. However, it is useful for identifying large numbers of potential interactions for further investigation. There were 18 sequences with an amino acid other than K/Q at position 80 (10 had L, 5 had N, 2 had R, and 1 had M), which were not included in this test. Only

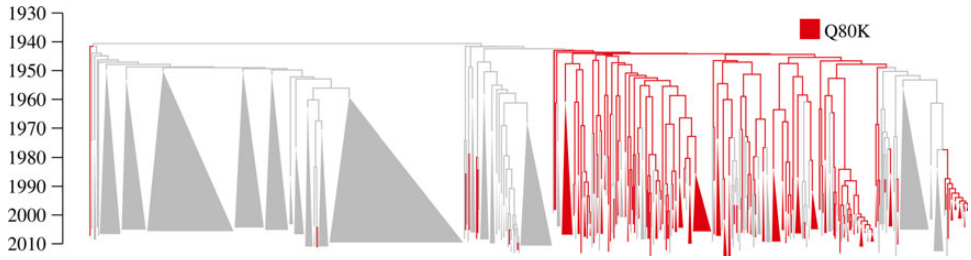

Figure 1. Global phylogeny of hepatitis C virus showing lineages possessing the Q80K polymorphism in nonstructural protein 3 (NS3). Almost all circulating strains with this polymorphism cluster into a clade descending from a single substitution event near the root of the tree. The phylogenetic tree was constructed from nucleotide sequences using codon positions 9–174 of NS3.

mutations that occurred in at least 5 Q80Q and 5 Q80K variant sequences were examined.

To incorporate phylogenetic information, we used the methods of Poon et al [19] to search for networks of coevolving amino acids, using a local implementation of the Spidermonkey tool [20]. This tool adjusts for phylogenetic confounding by mapping substitutions to the tree, which are phylogenetically independent observations. Markov chain Monte Carlo methods are used to search the space of Bayesian networks describing relationships among sites. Each site is modeled as a node in the network, with 2 possible states: evolving or static. Given a phylogeny relating the sampled sequences, each branch is treated as a separate observation of the network: a site is evolving if a nonsynonymous substitution happens along that branch at that site; otherwise, the site is static.

For computational tractability, nodes in the network were constrained to be conditionally dependent on a maximum of 2 other nodes. Two replicate Markov chains were run for 10^7 steps each. The first half of the chain was discarded as burn in, and 1000 samples were taken uniformly along the remainder. Convergence was assessed with the Gelman-Rubin diagnostic [21], as implemented in the coda library for R [22, 23]. The potential scale reduction factor was 1.02 (upper limit of the 95% CI, 1.09), consistent with convergence between the 2 chains.

The locations of identified sites on the NS3 protein were visualized with PyMOL [24], using PDB entry 4B73 [25].

Phylogeography

To investigate the geographic history of the Q80K variant, we modeled each sequence's country of origin as a discrete character state, which evolves along the phylogenetic tree [26]. Two different models were considered: a 1-parameter equal rates model, wherein all migration events are equiprobable, and a 2-parameter continent-aware model, in which different rates were assigned to intercontinent and intracontinent migrations. Numerical instability, likely arising from the large size of the phylogeny and limited data (a single state) at each node, prevented fitting more parameter-rich models. Model were fit using the ace function within the ape package for *R* [14, 23]. The 2 models were compared by a likelihood ratio test.

RESULTS

Single Origin of Q80K Substitution

We investigated the origin and stability of the Q80K polymorphism in 677 HCV genotype 1a NS3 sequences obtained from GenBank and the Vancouver Injection Drug Users Study cohort of injection drug users in Vancouver, Canada [11]. The sequences originated from 13 countries, sampled between 1977 and 2014. A maximum likelihood phylogeny relating these data was partitioned into 2 large clades (Figure 1), as reported previously [27, 28]. Root-to-tip regression [15] on the maximum likelihood tree dated the most recent common ancestor of the HCV genotype 1a clade to 1938 (95% CI, 1925–1948), roughly consistent with previous estimates (see Discussion). When sequences before 2000 were excluded, the most recent common ancestor was dated to 1943 (95% CI, 1920–1956). For these data sets, the nucleotide substitution rates (estimated from slopes of regression) were 1.4×10^{-3} and 1.6×10^{-3} substitutions per site per year, respectively.

Our analysis indicated that a single Q80K substitution occurred along a branch near to the root of the HCV genotype 1a clade, ancestral to almost all analyzed sequences (206 [96%]) with the Q80K polymorphism (Figure 1). This result was robust to the exclusion of pre-2000 sequences. Henceforth, we refer to the entire subtree rooted at this branch as the "Q80K clade." The Q80K clade comprised 342 sequences, including the aforementioned 206 sequences carrying the Q80K polymorphism. A total of 136 sequences carried an amino acid other than K at position 80, owing to 19 reversions at internal nodes, 12 reversions at tips, and 6 substitutions at tips to a different amino acid (3 to N, 1 to L, and 1 to M). The most recent common ancestor of the Q80K clade was estimated to have existed in 1940, by chronos, or in 1955 (95% CI, 1945–1963), by root-to-tip regression [15].

We found no similar phylogenetic history for any other previously defined NS3 resistance mutation. For the 5 resistance

Table 1. Magnitude and Significance of Associations of Other Hepatitis C Virus Nonstructural Protein 3 Substitutions With Q80K

Substitution	Odds Ratio (95% CI)	*P* Value
A91S/T	6.2 (4.3–8.9)	$<10^{-10}$
S174N	3.8 (2.6–5.6)	$<10^{-10}$
V29A	7.6 (4.1–14.8)	$<10^{-10}$
P67S	0.3 (.1–.5)	1.0×10^{-7}
T98A	6.2 (2.4–18.0)	1.7×10^{-5}

Only substitutions significantly associated with Q80K, without accounting for common ancestry, are shown. Odds ratios and *P* values were obtained with the Fisher exact test.

Abbreviation: CI, confidence interval.

positions covered by at least 10 sequence variants in our data, multiple substitutions were mapped to mostly either terminal or preterminal branches that did not form a distinct cluster (Supplementary Figure 2).

Epistatic Interactions

By tabulating co-occurring mutations without adjusting for common ancestry, we identified 6 substitutions that were significantly associated with Q80K in observed sequences (Table 1). Of these, we selected the 2 most significant, A91S/T and S174N, for further investigation. Both of these substitutions were more common in our data than Q80K. Of the 723 sequences with coverage at position 91, 229 (32%) showed either A91S or A91T (147 had S, and 82 had T), and 198 (27%) carried both Q80K and A91S/T. Similarly, 361 of 691 (52%) displayed S174N, and 157 (23%) carried both Q80K and S174N. Of 690 sequences with coverage at both positions, 88 (13%) carried all 3 mutations.

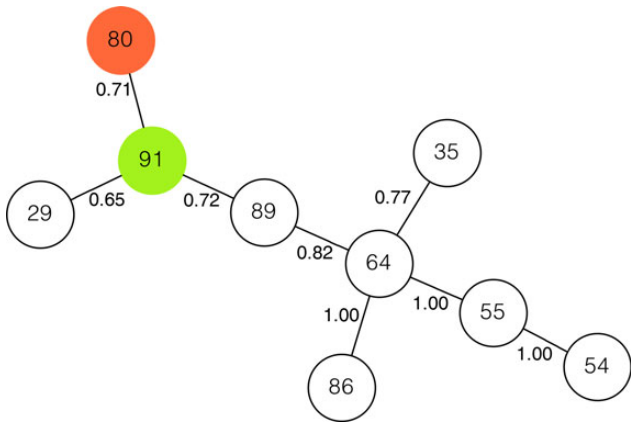

Figure 2. Partial Bayesian graphical network of potential interactions between sites on hepatitis C virus nonstructural protein 3 (NS3). Node labels are amino acid position in NS3, and edge labels are posterior probabilities of interaction between adjoining nodes. Only edges with a posterior probability of ≥95% are shown.

In the phylogenetically corrected Bayesian network (Figure 2), positions 80 and 91 were associated with a probability of 0.71 and were part of a larger interaction network involving multiple residues. No direct interactions involving position 174 were found, likely because of the low statistical power of the test when few substitutions involving this position occurred in the phylogeny. We note that this network was very sensitive to where mutations were mapped to the tree as a result of uncertainty in phylogenetic reconstruction.

Of the 5 genotype 5 sequences we were able to obtain, 4 carried a K at position 80, of which 3 carried S174N and 1 carried A91S/T. In genotype 6, 17 of 50 sequences (34%) contained Q80K. All 17 of these also carried S174N, but none carried A91S/T.

Reconstruction by maximum likelihood indicated that the A91S/T substitution occurred in several places in the Q80K clade, most notably along a branch ancestral to about half of the sequences (163 [48%]) in the clade (Figure 3*A*). This branch succeeded the root of the Q80K clade by about 3 years. On the other hand, the S174N substitution succeeded Q80K in the phylogeny by only 1 year and defined a clade of 341 sequences (Figure 3*B*). We refer to this clade as the "S174N clade." Although the existence of these clades was robust, the exact placement of the branches varied with the addition of more data or the use of different methods. During experimentation, the root of the S174N clade was sometimes placed deeper in the tree than the root of the Q80K clade. Likewise, it was unclear whether the A91S/T clade was the result of 2 mutation events that happened independently (one from A to T, and the other from A to S) or sequentially (from A to T, and then from T to S).

We investigated the possibility that the putative interaction between these sites is related to their proximity on the NS3 protein structure, as was previously postulated to explain the frequent coemergence of resistance mutations at other positions [29, 30]. Residues 80 and 174 lie immediately beside each other, with a mean interatomic distance of 6.6Å (Figure 4), and are bound by a polar contact. On the other hand, residues 80 and 91 are separated by 23.7Å, with a fold roughly composed of residues 48 to 54 lying between them.

Geographic History of Q80K

We performed a maximum likelihood reconstruction on the geographic location of each ancestral node, using 2 models. Using a likelihood ratio test, we found that the 2-parameter continent-aware model, with parameters for intercontinent and intracontinent transition rates, was supported over the equal-rates model ($P < 10^{-5}$). Under this model, the probabilities that root branches of the Q80K, A91S/T, and S174N clades were located in the United States were >98%, whether or not sequences sampled before 2000 were excluded. However, it is likely that this result is influenced by sampling bias, as the vast majority (229) of the sequences originated in the United States (with

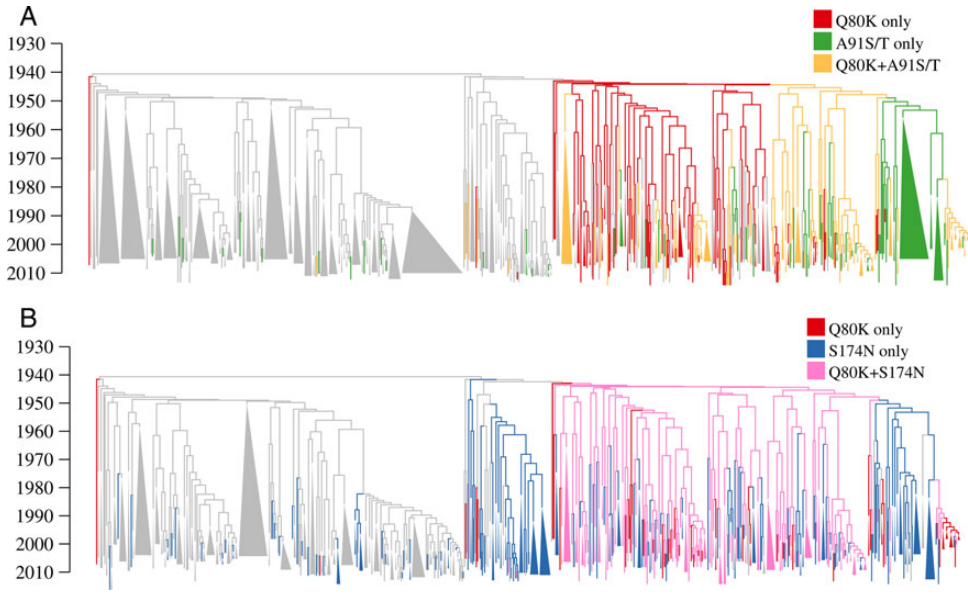

Figure 3. Global phylogeny of hepatitis C virus showing lineages possessing the A91S/T (*A*) and S174N (*B*) polymorphisms in nonstructural protein 3 (NS3). Lineages possessing the Q80K polymorphism are also indicated, showing the substantial overlap between clades representing Q80K, A91S/T, and S174N variant sequences. The phylogenetic tree was constructed from nucleotide sequences, using codon positions 9–174 of NS3.

38 from Germany, 25 from Canada, 24 from Italy, and 26 from elsewhere). The estimated intracontinent migration rate was higher than the intercontinent rate (0.6 vs 0.1 migrations/year). Estimated geographic locations for each internal node, aggregated by continent, are shown in Figure 5.

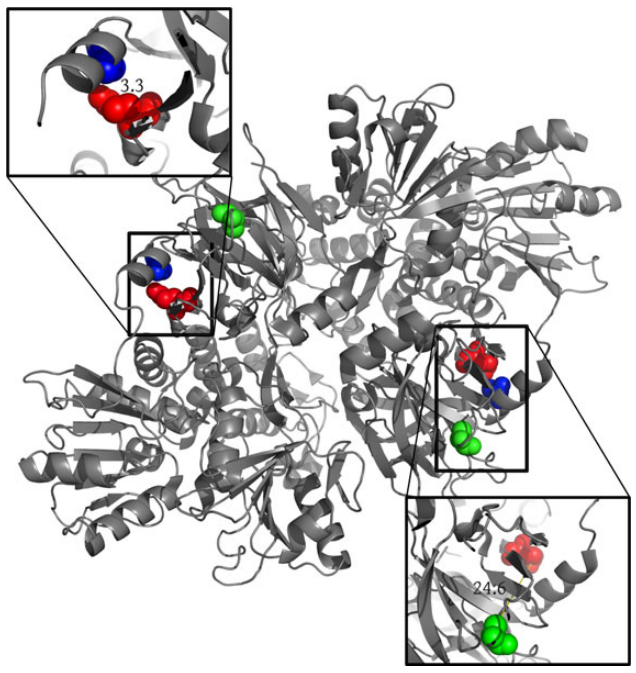

Figure 4. Protein dimer structure of hepatitis C virus nonstructural protein 3, highlighting positions 80 (red), 174 (blue), and 91 (green). Phylogenetic evidence supports epistatic interaction between these sites. Residues 80 and 174 are extremely proximal, but residues 80 and 91 are separated by a fold composed of roughly residues 48–54. Labels are interatomic distances between arbitrarily chosen atoms from each residue.

DISCUSSION

The Q80K polymorphism has been previously associated with reduced susceptibility to the HCV protease inhibitor simeprevir in combination with pegylated interferon and ribavirin, so baseline Q80K screening prior to initiating treatment is now mandated by the FDA [1]. Although primarily found in subtype 1a, the polymorphism has also been observed in HCV genotypes 5 and 6 [4]. Previous reports of the variant's prevalence within HCV genotype 1a vary by geographic region, from <10% in Sweden [2] up to 47% in the United States [30]. Our identification of a single substitution event ancestral to the vast majority of circulating Q80K lineages and our placement of this event in the 1940s United States helps explain this geographic heterogeneity. The reconstructed ancestral substitution event occurred when the HCV genotype 1a epidemic in the United States was still in its infancy and when travel between countries was infrequent enough to prevent many migration events. This was well before the exponential growth phase of the US HCV genotype 1a subtype epidemic, thought to have started in the 1960s [31].

The global HCV genotype 1a phylogeny we reconstructed for HCV was partitioned into 2 distinct clades, as previously reported [27, 28]. In particular, De Luca et al performed a similar analysis, using a smaller data set combining GenBank records and samples from Italy, and came to similar conclusions

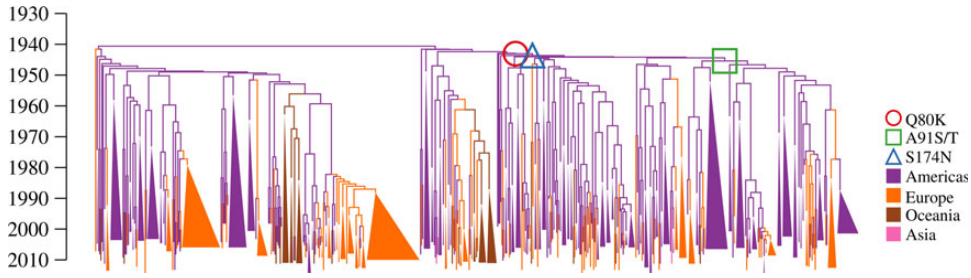

Figure 5. Inferred geographic locations of ancestral branches on global hepatitis C virus genotype 1a phylogeny, grouped by continent. The earliest branch along which a Q80K substitution in nonstructural protein 3 (NS3) occurred is indicated by a circle; this branch is ancestral to the majority of circulating Q80K strains. Similar ancestral branches for the A91S/T and S174N polymorphisms are indicated by a square and triangle, respectively. Phylogenetic tree was constructed from nucleotide sequences, using codon positions 9–174 of NS3.

regarding the global distribution of Q80K [28], although they did not infer a single origin of Q80K. We dated the most recent common ancestor of HCV genotype 1a to 1938, roughly consistent with previous estimates of 1900 [32], 1931 (95% CI, 1906–1957) [33], and 1954 (95% CI, 1929–1972) [28]. These studies analyzed different regions of the HCV genome (NS3 [28], NS5B [33], and E1 [32]); in particular, E1 is characterized by a higher rate of evolution than the other 2 [34]. We note that 2 groups [28, 33] used relaxed clock models, in which rates of evolution could change over time, whereas a third group [32] used a more rudimentary strict clock method. Similarly, our methods would have been less robust to variation in clock rates; chronos does not incorporate phylogenetic information when rescaling internal branches, while root-to-tip regression may underestimate the rate of molecular evolution in HCV [34]. Furthermore, our estimate of the nucleotide substitution rate (1.4×10^{-3} substitutions per base pair per year) is consistent with previous estimates (1.9×10^{-3} [34] and 0.9×10^{-3} [35] substitutions per base pair per year).

Our analyses based on this phylogeny suggested 2 substitutions, S174N and A91S/T, which may interact with Q80K. Of note, although we focused on the HCV genotype 1a subtype, every subtype 6 NS3 sequence in GenBank possessing the Q80K polymorphism also possessed S174N. The phylogenetic evidence from HCV genotype 1a, as well as the independent confirmation from genotype 6, indicates an interaction between positions 80 and 174 of NS3. A91S/T was not observed in genotype 6 and is more separated from Q80K on the NS3 protein structure, but it was more strongly associated with Q80K by tests both informed and naive of common ancestry. Position 174 was identified as being under positive selection, in the analysis by De Luca et al [28]. However, to our knowledge, our study is the first to report the potential interaction between these substitutions and Q80K. Previous work has identified various substitutions as co-occurring with Q80K in individual patients in the course of drug treatment, primarily at positions 36, 155, and 168 [29, 30, 36–40], but we were unable to identify these associations on a global scale. We emphasize that S174N and A91S/T are relative to the inferred ancestral sequence of HCV genotype 1a, which is imputed and not known with certainty. If the true ancestral sequence carried N174 instead of S174, it would imply a different evolutionary history, namely, that a substitution N174S occurred at nearly the same time as Q80K but in a different part of the tree.

The relatively great age of the Q80K polymorphism indicates that it did not arise as a response to drug selection. Its long phylogenetic history is in sharp contrast with those of other known resistance mutations, which are seen only on branches near the tips of the global phylogeny. This makes it clear that Q80K is not a resistance mutation per se, but rather a preexisting polymorphism that (coincidentally) confers reduced drug susceptibility. In that case, there are 2 possible scenarios to explain the large number of circulating Q80K variants: either the polymorphism is not transmissible and has frequently arisen de novo, or it is transmissible and the circulating lineages carrying it all descend from a common ancestor, which acquired the mutation naturally. The heterogeneous geographic distribution of the polymorphism, as well as the phylogenetic evidence presented here, strongly suggest the latter scenario, although no definitive conclusions can be drawn regarding transmissibility in the absence of transmission cohorts. The stability and high prevalence of Q80K in the 1a genotype of HCV suggest at first glance that it may confer a fitness advantage in the absence of treatment, perhaps due to escape from cellular immune pressure. However, if that were the case, it would almost certainly have arisen more frequently, given the high mutation rate of the virus (10^{-4}–10^{-3} substitutions per base pair per year [34, 35, 41]) and the inferred age of the epidemic [33, 41, 42]. In light of this, it is more plausible that the Q80K polymorphism was established as the result of a more rare event—≥2 contemporaneous substitutions that interacted to compensate for any loss of fitness incurred from either substitution alone. However, we do not rule

out the possibility that Q80K initially arose as an escape mutation from an immune phenotype that was simply uncommon in the general population.

In conclusion, Q80K is a highly stable and, likely, a transmissible polymorphism that persists in a large fraction of HCV genotype 1a infections in the United States. However, given that the stability depends on epistatic interactions and that the HCV epidemic is no longer growing exponentially, it is unlikely that the mutation will arise de novo in another population or geographic area. In effect, the phylogenetic evidence suggests that the current prevalence and distribution of Q80K is unlikely to undergo a significant change in the future.

Supplementary Data

Supplementary materials are available at *The Journal of Infectious Diseases* online (http://jid.oxfordjournals.org). Supplementary materials consist of data provided by the author that are published to benefit the reader. The posted materials are not copyedited. The contents of all supplementary data are the sole responsibility of the authors. Questions or messages regarding errors should be addressed to the author.

Notes

Financial support. This work was supported by the Canadian Institutes of Health Research (CIHR; grant HOP-111406), Genome Canada (Genomics and Personalized Health), the Michael Smith Foundation for Health Research/St. Paul's Hospital Foundation–Providence Health Care Research Institute Career Investigator Program (scholar award to A. F. Y. P.), the Canadian HIV Vaccine Initiative for Vaccine Discovery and Social Research (CIHR new investigator award to A. F. Y. P.), and CIHR/GlaxoSmithKline (research chair in clinical virology to P. R. H.).

Potential conflicts of interest. The BC Centre for Excellence in HIV/AIDS has received funding from Janssen Pharmaceuticals (manufacturers of simeprevir). M. K. certifies no potential conflicts of interest.

All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

References

1. Lin MV, Chung R. Recent FDA approval of sofosbuvir and simeprevir. Implications for current HCV treatment. Clin Liver Dis **2014**; 3:65–8.
2. Palanisamy N, Danielsson A, Kokkula C, et al. Implications of baseline polymorphisms for potential resistance to NS3 protease inhibitors in hepatitis C virus genotypes 1a, 2b and 3a. Antiviral Res **2013**; 99:12–7.
3. Paolucci S, Fiorina L, Piralla A, et al. Naturally occurring mutations to HCV protease inhibitors in treatment-naïve patients. Virol J **2012**; 9:245.
4. Leggewie M, Sreenu VB, Abdelrahman T, et al. Natural NS3 resistance polymorphisms occur frequently prior to treatment in HIV-positive patients with acute hepatitis C. AIDS **2013**; 27:2485–8.
5. Vallet S, Gouriou S, Nousbaum J-B, Legrand-Quillien M-C, Goudeau A, Picard B. Genetic heterogeneity of the NS3 protease gene in hepatitis C virus genotype 1 from untreated infected patients. J Med Virol **2005**; 75:528–37.
6. Cento V, Mirabelli C, Salpini R, et al. HCV genotypes are differently prone to the development of resistance to linear and macrocyclic protease inhibitors. PLoS One **2012**; 7:e39652.
7. Katoh K, Toh H. Recent developments in the MAFFT multiple sequence alignment program. Brief Bioinform **2008**; 9:286–98.
8. Cock PJ, Antao T, Chang JT, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics **2009**; 25:1422–3.
9. Kuiken C, Yusim K, Boykin L, Richardson R. The Los Alamos hepatitis C sequence database. Bioinformatics **2005**; 21:379–84.
10. Price MN, Dehal PS, Arkin AP. FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS One **2010**; 5:e9490.
11. Strathdee SA, Patrick DM, Currie SL, et al. Needle exchange is not enough: lessons from the Vancouver injecting drug use study. AIDS **1997**; 11:F59–65.
12. Vallet S, Viron F, Henquell C, et al. NS3 protease polymorphism and natural resistance to protease inhibitors in French patients infected with HCV genotypes 1–5. Antivir Ther **2011**; 16:1093.
13. Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol **2007**; 7:214.
14. Paradis E, Claude J, Strimmer K. APE: analyses of phylogenetics and evolution in R language. Bioinformatics **2004**; 20:289–90.
15. Drummond AJ, Pybus OG, Rambaut A. Inference of evolutionary rates from molecular sequences. Adv Parasitol **2003**; 54:331–58.
16. Pond SK, Muse SV. Site-to-site variation of synonymous substitution rates. Mol Biol Evol **2005**; 22:2375–85.
17. Nielsen R. Statistical methods in molecular evolution. Vol. 6. Heidelberg: Springer, **2005**.
18. Mani N. Clinically relevant HCV drug resistance mutations. Ann Forum Collab HIV Res **2012**; 14:1–8.
19. Poon AF, Lewis FI, Pond SLK, Frost SD. An evolutionary-network model reveals stratified interactions in the V3 loop of the HIV-1 envelope. PLoS Comput Biol **2007**; 3:e231.
20. Poon AF, Lewis FI, Frost SD, Pond SLK. Spidermonkey: rapid detection of co-evolving sites using Bayesian graphical models. Bioinformatics **2008**; 24:1949–50.
21. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. Stat Sci **1992**; 17:457–72.
22. Plummer M, Best N, Cowles K, Vines K. CODA: Convergence diagnosis and output analysis for MCMC. R News **2006**; 6:7–11.
23. R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, **2005**.
24. DeLano WL. The PyMOL molecular graphics system. San Carlos, CA: DeLano Scientific, **2002**.
25. Saalau-Bethell SM, Woodhead AJ, Chessari G, et al. Discovery of an allosteric mechanism for the regulation of HCV NS3 protein function. Nat Chem Biol **2012**; 8:920–5.
26. Slatkin M, Maddison WP. A cladistic measure of gene flow inferred from the phylogenies of alleles. Genetics **1989**; 123:603–13.
27. Pickett BE, Striker R, Lefkowitz EJ. Evidence for separation of HCV subtype 1a into two distinct clades. J Viral Hepat **2011**; 18:608–18.
28. De Luca A, Di Giambenedetto S, Prosperi M, et al. Two distinct HCV genotype 1a clades: geographical distribution and association with natural resistance mutations to HCV NS3/4A inhibitors. Antivir Ther **2013**:A47.
29. Lenz O, Verbinnen T, Lin T-I, et al. In vitro resistance profile of the hepatitis C virus NS3/4A protease inhibitor TMC435. Antimicrob Agents Chemother **2010**; 54:1878–87.
30. Bae A, Sun S-C, Qi X, et al. Susceptibility of treatment-naive hepatitis C virus (HCV) clinical isolates to HCV protease inhibitors. Antimicrob Agents Chemother **2010**; 54:5288–97.
31. Tanaka Y, Kurbanov F, Mano S, et al. Molecular tracing of the global hepatitis C virus epidemic predicts regional patterns of hepatocellular carcinoma mortality. Gastroenterology **2006**; 130:703–14.
32. Pybus OG, Charleston MA, Gupta S, Rambaut A, Holmes EC, Harvey PH. The epidemic behavior of the hepatitis C virus. Science **2001**; 292:2323–5.
33. Magiorkinis G, Magiorkinis E, Paraskevis D, et al. The global spread of hepatitis C virus 1a and 1b: a phylodynamic and phylogeographic analysis. PLoS Med **2009**; 6:e1000198.
34. Ogata N, Alter HJ, Miller RH, Purcell RH. Nucleotide sequence and mutation rate of the H strain of hepatitis C virus. Proc Natl Acad Sci U S A **1991**; 88:3392–6.

35. Abe K, Inchauspe G, Fujisawa K. Genomic characterization and mutation rate of hepatitis C virus isolated from a patient who contracted hepatitis during an epidemic of non-A, non-B hepatitis in Japan. J Gen Virol **1992**; 73:2725–9.
36. Lenz O, Vijgen L, Berke JM, et al. Virologic response and characterisation of HCV genotype 2–6 in patients receiving TMC435 monotherapy (study TMC435-C202). J Hepatol **2013**; 58:445–51.
37. Zeuzem S, Berg T, Gane E, et al. Simeprevir increases rate of sustained virologic response among treatment-experienced patients with HCV genotype-1 infection: a phase IIb trial. Gastroenterology **2014**; 146:430–41.
38. Lenz O, de Bruijne J, Vijgen L, et al. Efficacy of re-treatment with TMC435 as combination therapy in hepatitis C virus–infected patients following TMC435 monotherapy. Gastroenterology **2012**; 143:1176–8.
39. Forns X, Lawitz E, Zeuzem S, et al. Simeprevir with peginterferon and ribavirin leads to high rates of SVR in patients with HCV genotype 1 who relapsed after previous therapy: a phase 3 trial. Gastroenterology **2014**; 146:1669–79.
40. Fried MW, Buti M, Dore GJ, et al. Once-daily simeprevir (TMC435) with pegylated interferon and ribavirin in treatment-naïve genotype 1 hepatitis C: The randomized PILLAR study. Hepatology **2013**; 58:1918–29.
41. Pybus OG, Barnes E, Taggart R, et al. Genetic history of hepatitis C virus in East Asia. J Virol **2009**; 83:1071–82.
42. Pybus O, Drummond A, Nakano T, Robertson B, Rambaut A. The epidemiology and iatrogenic transmission of hepatitis C virus in Egypt: a Bayesian coalescent approach. Mol Biol Evol **2003**; 20:381–7.