

# A proposal to sequence a number of complete genomes of uncommon strains of the Hepatitis C Virus

Carla Kuiken, Ph.D.

Principal investigator, Hepatitis C Database  
Los Alamos National Laboratory  
Los Alamos, New Mexico  
<http://hcv.lanl.gov>

## **I. Outline**

The Hepatitis C Virus (HCV) is a highly variable virus. Currently 6 genotypes are distinguished, and each genotype is further subdivided into a varying number of subtypes. These subtypes are often defined only on the basis of short sequence fragments, which are clearly distinct from other sequences in that region. As a consequence, a myriad of poorly defined subtypes now exist, and new ones are added monthly. This is leading to a confusing and inconsistent classification. Hepatitis C is an emerging infection and a burgeoning worldwide public health problem, with 170 million infected and an estimated 20 million deaths in the coming decades. Having a clear overview of the variability is important for vaccine development, treatment and epidemiology, but currently our knowledge of the HCV variant family is very limited. This proposal is intended to address that problem.

## **II. Importance of HCV as an emerging infection**

While 15-45% of new HCV infections are cleared spontaneously, an estimated 170 million people or 3% of the world population is now chronically infected (Wasley and Alter, 2000). Consequences of chronic HCV infection include chronic hepatitis, cirrhosis, and liver cancer. A recent Canadian study (Krahn et al., 2004) estimated that lifetime HCV-associated mortality is around 1 in 8, while twice as many will develop cirrhosis of the liver. Most likely this number will be higher in less developed countries. With 170 million infected worldwide, this means 20 million HCV-related deaths in the coming decades.

HCV infection is increasingly recognized as a major public health problem in the United States, accounting for 8-10,000 deaths annually, with that rate predicted to double or triple over the next 10 to 20 years (CDC, 1998). HCV is now the leading cause of liver transplantation in the United States. The estimated prevalence of HCV in the general United States population is 1.8% (Alter et al., 1999). Estimates of the prevalence among subset of the National Health and Nutrition Examination Survey III 3, to as high as 35% among veterans selected for HCV testing in a large VA healthcare system in California (Cheung, 2000). A recent estimate of the prevalence of HCV among users of Veterans Healthcare Administration (VHA) services, based on a sample of all patients in the

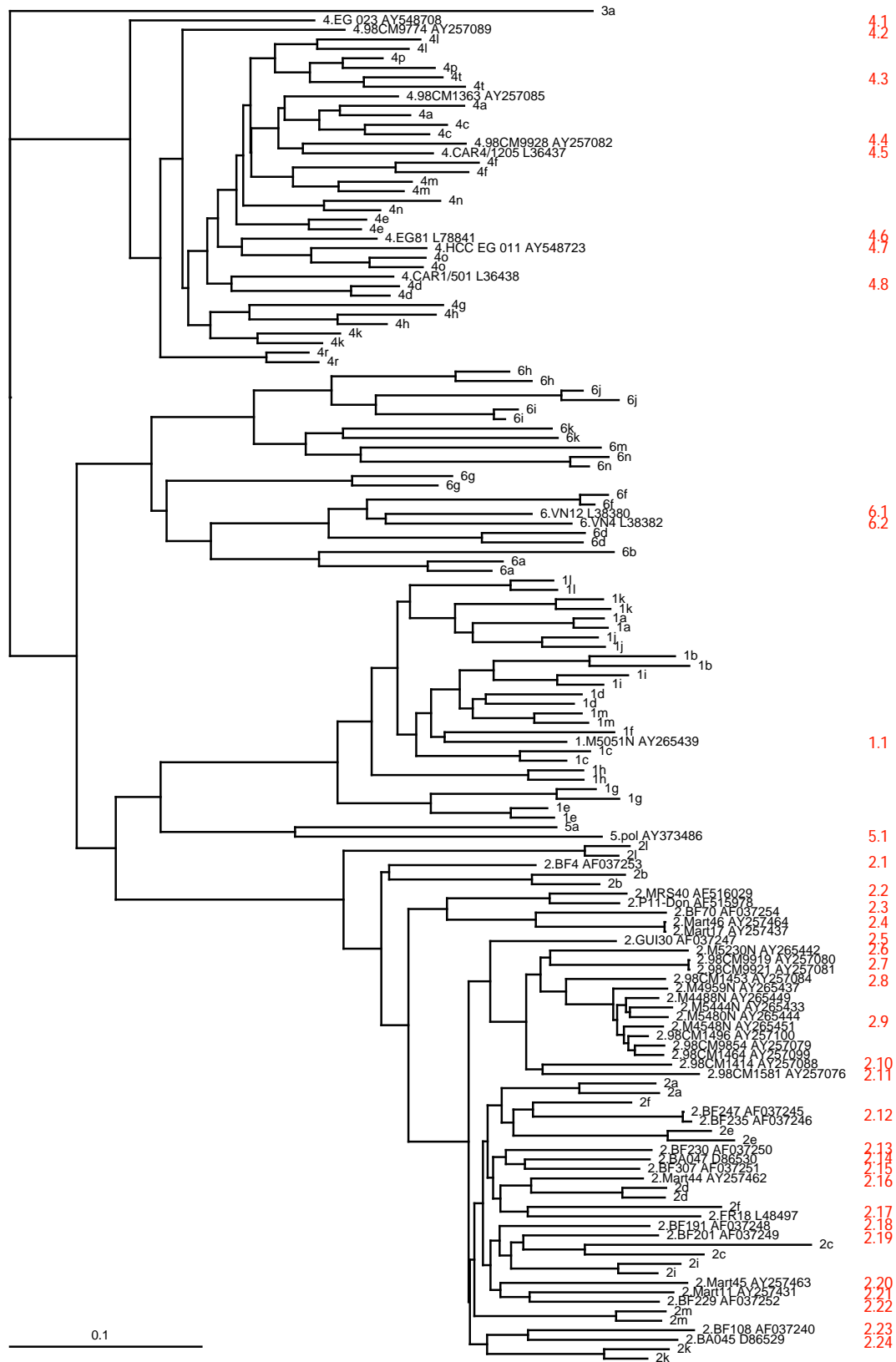


Figure 1. Phylogenetic tree of part of the NS5B region based on an alignment of 270 nt length (after gapstripping, corresponding to nt 8283-8602 of HCV-H). The tree was built using the Neighbor-Joining algorithm, based on F84 distances. One genotype 3 representative was used as an outgroup because there are no unsubtype genotype 3 variants in this region. Sequences belong to a subtype are identified only by geno/subtype, unsubtype sequences by their genotype, name, and accession number.

Northwestern US, shows that almost 12% of them were HCV positive (Sloan et al., 2004). For 2010 through 2019, computer models project \$10.7 billion in direct medical costs for HCV care in this country, and a societal cost of \$54.2 billion for the loss of 1.83 million years of life in those younger than 65 (Wong et al., 2000).

HCV has proven to be difficult to treat. The gold standard for treatment is currently a combination of pegylated interferon and ribavirin. The efficacy of this treatment has improved over the past five years, but is still not anywhere near 100%. Furthermore, there are dramatic differences between variants (see below) of HCV: genotype 1 virus, which is predominant in the United States and other Western countries, is much more resistant to treatment than other genotypes. The sustained response rate for genotype 1 infections is still below 50%, compared to close to 80% for genotypes 2 and 3 (Vrolijk et al., 2004). In addition, the response rate is influenced by ethnicity, with African Americans showing less success (Howell, Jeffers, and Hoofnagle, 2000).

### **III. Genetics of the Hepatitis C virus**

HCV is an enveloped, positive-strand RNA virus belonging to the genus of hepaciviruses in the family of flaviviruses, which also includes viruses like Dengue, West Nile, Yellow fever virus, and Japanese encephalitis virus. Its genome is around 9.6 kilobases long, and encodes one large polyprotein precursor of around 3000 amino acids. This precursor is cleaved by a combination of viral and host proteinases into 10 separate proteins, three of which are structural while the others play a regulatory role in viral replication and host cell behavior.

Six clades or genotypes are generally distinguished in HCV (figure 1)(Simmonds, 1999). To some extent the genotypes are correlated with geography and risk group. Genotype 1 is mostly found in Japan, Europe and the US; however, among European drug users clade 3 is more common. In Egypt, which has an exceptionally high prevalence, almost all infections are caused by clade 4. Clade 6 is mostly found in South-east Asia, and all clades except 6 are represented in sub-Saharan Africa.

Each clade is further divided into subtypes. Subtypes are not as readily distinguished as genotypes are, especially on the basis of shorter sequences, which may partly explain why they are not as rigorously defined. The subtype classification is important for several reasons. First, they can be crucial in detection of recombination. Only one circulating recombinant has so far been found for HCV (Kalinina et al., 2002), but more are widely expected; the ability to accurately classify subtypes is indispensable for identifying within-genotype recombination. For design and testing of prophylactic and protective vaccines against HCV, the ability to distinguish and test for potential differences between subtypes will also be important (Koff, 2003). And although so far the genotype differences in treatment efficacy have not been seen at the subtype level, the subtype distinctions could bear on treatment of the infection (Howard, 2002).

Sixty-five genotype/subtype combinations are currently distinguished, while almost a thousand genotyped sequence fragments have not yet been classified into subtypes. It is expected that new genotypes, probably with associated subtypes, will be added to the repertoire, and there are credible reports of a new genotype that is waiting for confirmation by complete-genome sequencing (pers. comm. Erwin Sablon, Innogenetics, Belgium). While at least one complete genome is available for each genotype, at the subtype level the problems are clear: of the 65 genotype/subtype combinations, only 18 are represented by at least one complete genome, and 10 by the desired number of 2 independent complete genomes.

#### **IV. Sorting out the classification of Hepatitis C**

The classification of the Hepatitis C virus is still complex, despite an earlier effort to create some order (Robertson et al., 1998), and at the subtype level the confusion is growing with the number of sequence fragments. One important reason is the fact that for many geno/subtypes, only small fragments have been sequenced. This means that misclassifications can very easily arise. To illustrate this, suppose investigator A has identified a set of novel core gene sequences, which he has named as 2e. Later, Investigator B identifies a set of novel NS5B sequences; seeing that 2e is "taken", he names his new sequences as 2f. 2e and 2f could very well represent the same variant, but this cannot possibly be recognized in the absence of complete genome sequences to link the fragments together. Problems will arise once complete genomes are sequenced, as variants in many earlier publications will then need to be re-classified, and afterwards will be very hard to trace back. This could cause confusion for many years to come. A similar situation occurred in HIV research, where an initial classification based on only partial sequences turned out to be erroneous when complete genome sequences became available. The revised classification has been available for over 10 years, but the confusion still occasionally resurfaces. By carefully annotating these name changes, the HCV database can play a role in minimizing this problem.

Another problematic group consists of sequences that have been genotyped but not yet subtyped. A histogram of the genomic regions in this group is shown in figure 2. Unfortunately, many of these are located in the 5' UTR, which is a region that rarely offers sufficient resolution to distinguish different subtypes, and sometimes not even enough to reliably classify genotypes. These 5' UTR sequences will not be analyzed further. Figure 1 identifies the NS5B sequences that have been genotyped but not subtyped; these most likely represent at least 36 unique subtypes (shown with their full name and accession number, numbered in red), and therefore are prime candidates for complete genome sequencing. It is likely that at least some of these will turn out to be representatives of some of the other geno/subtypes for which this region has not yet been sequenced, so the number of genomes that need to be generated to resolve this part of the puzzle will probably be less than  $2 \times 36$ . The number of genomes needed to completely resolve the current HCV classification will probably not increase above 200. This does not include new variants that will likely be found in the future.

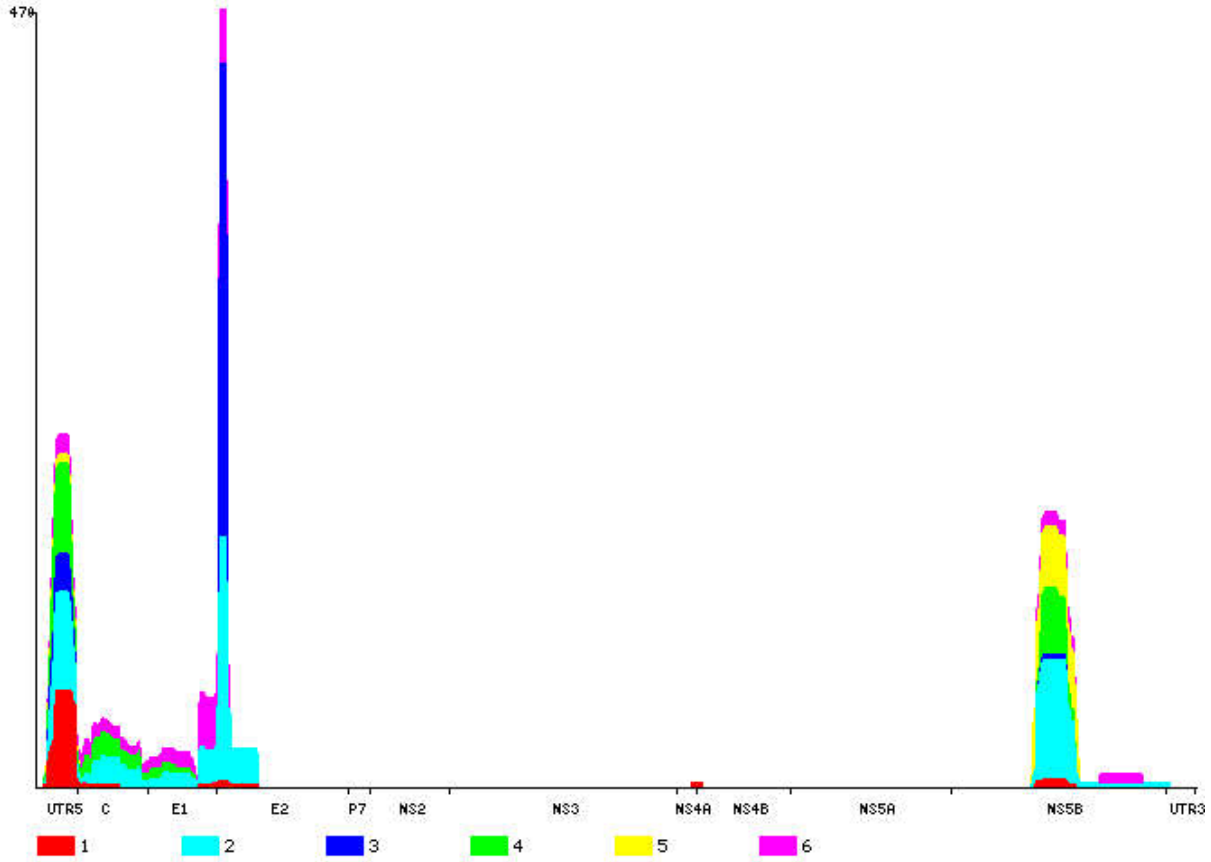


Figure 2. Histogram showing the distribution of sequences with a genotype but no subtype over the genome. The NS5B subregion used for the tree in Figure 1 contains a concentration of these variants; the other region is the 5' UTR, which is not very well suited for geno/subtyping because it is too conserved.

The only foolproof way to resolve the HCV classification issue is to generate at least 2 complete genome sequences for each genotype and subtype. Judging by the present state and the slow increase in the number of complete genomes of uncommon variants, it cannot realistically be expected that the scientific community will sort the nomenclature problem out on its own. Instead, as is happening now, people with interesting samples will continue to sequence small regions, and quickly name and publish them. Sequencing complete genomes of the known subtypes will have to be done as a community service, which is the reason to submit this proposal.

**V. The current proposal**

This proposal entails a collaborative effort to sequence complete genomes of the HCV variants that are now known and have not been completely sequenced yet. A quick inventory among scientists working in this field shows that the nomenclature and classification problem is widely regarded as significant; some letters of support from renowned researchers are appended to this document. Samples are scattered around the world, but all groups that were approached thus far to ask for sample donations agreed in

principle to contribute some. The appendix contains letters of support from four groups (Dr. Donald Murphy, Public Health Laboratory in Quebec, Canada; Dr Jean Ndjomou of Indiana University School of Medicine; Dr. Charles Rice of Rockefeller University, New York; and Dr. Peter Simmonds of Edinburgh University, UK) who are willing to donate samples or viral RNA, and a letter from Dr Ling Lu (Dept. of Gastro-enterology, Kansas U. Medical School) who has offered to assist this effort by generating full-length cDNA for sequencing. Details for further sample donations are being worked out with several other groups.

subtype	1	2	3	4	5	6	genotype
a	2		1		1		1
b	2		1				1
c		1	2	2			
d	2	2	2	2			1
e	2	2	2	2			
f	2	2	2	2			2
g	2	2	2	2			1
h	2	2	2	2			1
i	2	2	2				2
j	2	1					2
k	2	2	1	2			1
l	2	2		2			2
m	2			2			2
n				2			2
o				2			2
p				2			2
q				2			2
r				2			2
s				2			2
t				2			2
Total	20	18	15	33	0	18	104

not yet assigned  
 defined by 2 complete genomes

Table 1. The number of complete genome sequences that are needed to fully characterize all defined geno/subtypes in the database.

Currently, there are firm commitments to supply cDNA of around 45 HCV genomes; The total number of complete genomes that need to be sequenced is 102, based on the existence of fragments that are subtyped but not sufficiently defined. These include two genomes each for 47 variants for which no complete genomes exist, and one additional genome each for another 8 for which one complete genome is available (see table 1). As work progresses, intermediate sequence analysis might show that different variants are actually the same, and some of the sequences that have presently been assigned a genotype but no subtype might fall into place. Other only partially classified variants might be found to require complete genome sequencing.

The analysis of the complete genome sequences, and the reclassifications that could be necessary, can be done in Los Alamos. Both the analysis capacity and the expertise are available there, as well as the means to rapidly communicate the results to the research community. It seems likely that several publications will result from this effort, and as a courtesy to the collaborators and contributors to this project, I would like to request permission to embargo the sequences until the manuscripts have been accepted for publication. All sequences will be made available to the public both via GenBank and via the Los Alamos HCV database immediately after that. The manuscripts will be submitted as quickly as possible, probably within 3-6 months after the sequences become available for analysis.

*Acknowledgements* This proposal has benefited greatly from discussions with many colleagues, most importantly Drs. Peter Simmonds and Charles Rice. The HCV database is funded by NIAID/DMID through an interagency agreement with the Department of Energy.

### *References*

- Alter, M. J., Kruszon-Moran, D., Nainan, O. V., McQuillan, G. M., Gao, F., Moyer, L. A., Kaslow, R. A., and Margolis, H. S. (1999). The prevalence of hepatitis C virus infection in the United States, 1988 through 1994. *N Engl J Med* **341**(8), 556-62.
- CDC (1998). Recommendations for prevention and control of hepatitis C virus (HCV) infection and HCV-related chronic disease. Centers for Disease Control and Prevention. *MMWR Recomm Rep* **47**(RR-19), 1-39.
- Cheung, R. C. (2000). Epidemiology of hepatitis C virus infection in American veterans. *Am J Gastroenterol* **95**(3), 740-7.
- Howard, C. R. (2002). Hepatitis C virus: clades and properties. *J Gastroenterol Hepatol* **17 Suppl**, S468-70.
- Howell, C., Jeffers, L., and Hoofnagle, J. H. (2000). Hepatitis C in African Americans: summary of a workshop. *Gastroenterology* **119**(5), 1385-96.
- Kalinina, O., Norder, H., Mukomolov, S., and Magnus, L. O. (2002). A natural intergenotypic recombinant of hepatitis C virus identified in St. Petersburg. *J Virol* **76**(8), 4034-43.
- Koff, R. S. (2003). Hepatitis vaccines: recent advances. *Int J Parasitol* **33**(5-6), 517-23.
- Krahn, M., Wong, J. B., Heathcote, J., Scully, L., and Seeff, L. (2004). Estimating the prognosis of hepatitis C patients infected by transfusion in Canada between 1986 and 1990. *Med Decis Making* **24**(1), 20-9.
- Robertson, B., Myers, G., Howard, C., Brettin, T., Bukh, J., Gaschen, B., Gojobori, T., Maertens, G., Mizokami, M., Nainan, O., Netesov, S., Nishioka, K., Shin i, T., Simmonds, P., Smith, D., Stuyver, L., and Weiner, A. (1998). Classification, nomenclature, and database development for hepatitis C virus (HCV) and related viruses: proposals for standardization. International Committee on Virus Taxonomy. *Arch Virol* **143**(12), 2493-503.

- Simmonds, P. (1999). Viral heterogeneity of the hepatitis C virus. *J Hepatol* **31 Suppl 1**, 54-60.
- Sloan, K. L., Straits-Troster, K. A., Dominitz, J. A., and Kivlahan, D. R. (2004). Hepatitis C tested prevalence and comorbidities among veterans in the US Northwest. *J Clin Gastroenterol* **38**(3), 279-84.
- Vrolijk, J. M., de Knecht, R. J., Veldt, B. J., Orlent, H., and Schalm, S. W. (2004). The treatment of hepatitis C: history, presence and future. *Neth J Med* **62**(3), 76-82.
- Wasley, A., and Alter, M. J. (2000). Epidemiology of hepatitis C: geographic differences and temporal trends. *Semin Liver Dis* **20**(1), 1-16.
- Wong, J. B., McQuillan, G. M., McHutchison, J. G., and Poynard, T. (2000). Estimating future hepatitis C morbidity, mortality, and costs in the United States. *Am J Public Health* **90**(10), 1562-9.

## Appendix: Letters of support

Dr. Peter Simmonds, Professor of Virology, University of Edinburgh, United Kingdom

Dr. Ling Lu & Dr. Curt Hagedorn, U. Kansas Medical Center, Kansas City, KS

Dr. Donald Murphy, Institut national de santé publique du Québec, Quebec, Canada

Dr. Charles M. Rice, The Rockefeller University, New York, NY

Dr. Jean Ndjomou, Indiana University School of Medicine, Indianapolis, IN

23<sup>rd</sup> July, 2004

Carla Kuiken, Ph.D.  
Principal investigator, Hepatitis C Database  
Los Alamos National Laboratory  
Los Alamos, New Mexico  
<http://hcv.lanl.gov>

Re: *A proposal to sequence a number of complete genomes of uncommon strains of the Hepatitis C Virus*

Dear Carla,

This letter confirms my enthusiasm for the planned project to obtain a more complete set of complete genome sequences of HCV for classification and genetic analysis purposes. I also confirm that I will be able to supply at least 25 samples of plasma containing novel subtypes of genotypes 3, 4 and 6 for the project.

I have been involved in the classification of HCV for several years, and I regard the project as an extremely valuable further step in documenting the variability of HCV, investigating the occurrence and frequency of virus recombination, and providing an internationally recognized sequence resource for future studies of virus variability and genotype differences.

Yours sincerely

Peter Simmonds, BM, PhD, MRCPATH

Professor of Virology, University of Edinburgh  
Centre for Infectious Diseases  
University of Edinburgh  
Summerhall  
Edinburgh, EH9 1AJ  
UK

Tel.: +44 131 650 7927  
Fax.: +44 131 650 6511

Dear Carla,

We are pleased to collaborate with you to establish complete genome sequences of hepatitis C virus (HCV). Recently, we have refined a sensitive long RT-PCR technique which can be used to efficiently amplify HCV genomic fragments of up to five kb in length from microliter quantities of serum samples. This technique has been used to obtain the complete genomic sequences of four novel HCV variants from serum samples that were < 300 \*l in volume. This technique should provide a major advantage in studies aimed at further development of an HCV sequence database. We fully agree that sequencing many more HCV genomes is necessary to prevent further confusion in the field of HCV genetics and that this work should be a high priority.

Sincerely,

Ling Lu, Ph.D.

Division of Gastroenterology/Hepatology  
University of Kansas Medical Center  
3901 Rainbow Boulevard  
4035 Delp, Mail Stop 1023  
Kansas City, KS 66160  
E-mail: llu@kumc.edu  
Phone: 913-588-3433  
Fax: 913-588-3975

Curt H. Hagedorn, MD.

Division of Gastroenterology/Hepatology  
University of Kansas Medical Center  
3901 Rainbow Boulevard  
4035 Delp, Mail Stop 1023  
Kansas City, KS 66160  
E-mail: Chagedorn@kumc.edu  
Phone: 913-588-0105  
Fax: 913-588-3975

Dear Carla,

It will be my pleasure to collaborate with you to establish complete genome sequences of hepatitis C virus. I have an important collection of HCV variants and I will be glad to donate a minimum of 10 RNA samples for this project. Sequencing of complete HCV genomes will strengthen and reduce the confusion that persists over the HCV genotype nomenclature.

Sincerely,

Donald Murphy, Ph.D.  
Institut national de santé publique du Québec  
Laboratoire de santé publique du Québec  
20045 chemin Sainte-Marie  
Sainte-Anne-de-Bellevue (Québec)  
CANADA H9X 3R5  
Telephone: (514) 457-2070 ext. 266  
Fax: (514) 457-6346

Dear Carla,

I am delighted that you are taking the initiative to determine more HCV genome sequences. There is a great need to have a set of high quality full-length genome sequences for each of the major HCV genotypes and subtypes. This will enhance our ability to do more sophisticated covariant analyses and inform experiments aimed at dissecting the function of HCV RNA elements and proteins. I wholeheartedly support this effort and am willing to assist in helping you obtain samples. We have a large collection of isolates from the New York City area and David Ho has collected numerous samples from different parts of China. In addition, I am sure that we can contact HCV investigators in other parts of the world to obtain rarer isolates. I have found the HCV medical and research community to be extremely helpful in this regard and I am sure they will happily join this effort. Given the powerful and inexpensive sequencing technologies now available it is time that we get this done for the field. I wish you success with the proposal and again, thanks for taking on this most important task. Everyone in HCV research will be very grateful to have this done. It will be a terrific resource.

With warm regards,

Charlie

Charles M. Rice, Ph.D.  
Maurice R. and Corinne P. Greenberg Professor  
Head, Laboratory of Virology and Infectious Disease  
The Rockefeller University

Scientific & Executive Director, Center for the Study of Hepatitis C  
The Rockefeller University  
New York-Presbyterian Hospital  
Weill Medical College of Cornell University

The Rockefeller University  
1230 York Avenue  
New York, NY 10021

Phone: 212-327-7046  
FAX: 212-327-7048  
ricec@rockefeller.edu

Dear Dr. Kuiken:

I would like to apologize for the delay in replying to your mail. Thank you for contacting me for the purpose stated in your mail.

I appreciate very much the initiative you have taken and believe that upon completion the work will be of great importance in the field of HCV molecular epidemiology in particular and HCV in general. I am aware that subtype classification based on the available data and criteria does still lack consistency and accuracy. Therefore, redefinition and review of the system is an imperative. But before that, representative data is a prerequisite to achieve the goal.

I would like to let you know that I have no problem to collaborate on the project. For the reasons that I mentioned above, I believe the outcome of the project would be of great significance. There is no problem in principle for me to provide you with the samples you needed. The only concern that I would like to let know is that I left Germany in 2003 and I am currently in the US. But the samples are still kept there at - 80 C. So if your research proposal is successful, we will just arrange to get the samples here. My contact information is listed below and whenever you might need some information please let me know.

Sincerely,

Jean Ndjomou, Ph.D.  
Postdoctoral Fellow  
Department of Microbiology and Immunology  
Indiana University School of Medicine  
950 W. Walnut St. R2 302  
Indianapolis, IN 46202