BIOINFORMATICS ORIGINAL PAPER

Vol. 21 no. 3 2005, pages 379–384
doi:10.1093/bioinformatics/bth485

The Los Alamos hepatitis C sequence database

Carla Kuiken, Karina Yusim, Laura Boykin and Russell Richardson*

HCV database, Los Alamos National Laboratory, Los Alamos, NM, USA

Received on March 16, 2004; revised on July 30, 2004; accepted on August 14, 2004
Advance Access publication September 17, 2004

ABSTRACT

Motivation: The hepatitis C virus (HCV) is a significant threat to public health worldwide. The virus is highly variable and evolves rapidly, making it an elusive target for the immune system and for vaccine and drug design. At present, some 30 000 HCV sequences have been published. A central website that provides annotated sequences and analysis tools will be helpful to HCV scientists worldwide.

Results: The HCV sequence database collects and annotates sequence data and provides them to the public via a website that contains a user-friendly search interface and a large number of sequence analysis tools, based on the model of the highly regarded Los Alamos HIV database. The HCV sequence database was officially launched in September 2003. Since then, its usage has steadily increased and is now at an average of ∼280 visits per day from distinct IP addresses.

Availability: The HCV website can be accessed via http://hcv.lanl.gov and http://hcv-db.org

Contact: hcv-info@lanl.gov

INTRODUCTION

The hepatitis C virus (HCV) has infected 4 million people in the United States and ∼170 million people worldwide. HCV infection is spontaneously resolved in ∼25% of cases (Alter *et al*., 1992; Hoofnagle, 1997), and the remaining 75% suffer from latent infection. Latent HCV infection can lead to cirrhosis and liver cancer, and is the major cause of liver transplantation in the United States. A recent Canadian study (Krahn *et al*., 2004) estimated that the lifetime HCV-associated mortality is around 1 in 8; a much larger number (an estimated 1 in 4) will develop cirrhosis of the liver. This number will most probably be higher in less developed countries. With a total of 170 million people infected worldwide, it is estimated that there will be ∼20 million HCV-related deaths in the next few decades.

The infection has been prevalent for decades and was previously called non-A, non-B hepatitis, but the virus was not discovered until 1989 (Choo *et al*., 1989). It spreads mainly through blood and blood products. Reliable tests have been available since the discovery of the virus (Vrielink *et al*., 1995), but occasional outbreaks still occur in the Western world, either because the carrier is in the very early stage of the infection when antibodies are not yet detectable or because of transmission by trace amounts of virus, e.g. via kidney dialysis equipment (CDC, 2003). The epidemic is still spreading via contaminated blood and needles worldwide.

HCV is a positive-sense RNA virus with a genome of ∼9600 bases, which encodes a single polyprotein that is cleaved into three structural proteins (Envelopes 1 and 2 and p7) and six non-structural proteins named NS2–NS5B. It has been classified as a flavivirus, the family that includes Dengue, yellow fever and the West Nile virus, with which it shares many structural features. However, the genetic distance between HCV and other flaviviruses is >50% over the entire genome (Simmonds, 1999). HCV is an extremely variable virus that forms quasispecies within the host. Six different genotypes have now been defined worldwide, subdivided into more than 80 subtypes.

Research on HCV is hampered by several technical problems that have not yet been resolved. Foremost is the lack of good animal models other than chimpanzees (the use of which is ethically problematic and very expensive), and the difficulty in culturing the virus *in vitro* (Lanford and Bigger, 2002; Grakoui *et al*., 2001). These problems make both drug development and vaccine studies difficult and expensive. At present, the only effective treatment against HCV is a long (6–12 months), expensive and highly toxic regimen of interferon and ribavirin. This regimen is effective in 40–80% of the cases depending on the HCV genotype. For unknown reasons, the efficacy against the most prevalent genotype in the United States, genotype 1, is <50% (Pawlotsky, 2003), and even lower in African-Americans (Fleckenstein, 2004).

While there is a limited knowledge about the immunogenicity of HCV, it is widely believed that, as in the case of HIV, both the generation of escape and resistance mutations and the high variability itself will create formidable problems for drug and vaccine design (Farci and Purcell, 2000). In HIV, there is a renewed interest in rational vaccine design, a relatively new discipline that attempts to define the optimal vaccine strain, possibly an artificially created one, that minimizes the differences from the circulating strains while maximizing the immunogenicity of the reagent (Gaschen *et al*., 2002). For hepatitis C, both drug and vaccine design are in their infancy;

*To whom correspondence should be addressed.

therefore, a database that allows researchers to study genetic variability by facilitating retrieval, alignment and analysis of all publicly available HCV sequences could play an important role in helping these efforts.

The HIV databases and website in Los Alamos were designed for HIV, but the similarities between HIV and HCV made the HIV setup very suitable for expansion to HCV. Most importantly, an infrastructure had been developed for HIV that allows fast, dynamic alignment of large numbers of sequences stored in the database (Gaschen *et al*., 2001). Tools designed for manipulations of HIV sequences, such as extracting genes, finding the coordinates of a sequence relative to the reference strain, retrieving and aligning all sequences of a given region from the database, and scanning sequences for nucleotide or protein motifs, could be easily adapted to be used for HCV. The HIV database has proven to be an important and very effective tool to assist the research on vaccine and drug development, and the HCV database is expected to provide a similar service for hepatitis C research.

THE PURPOSE AND DESIGN OF THE DATABASE

The HCV database aims to be a resource for scientists working on HCV genetics, evolution, variability, and vaccine and drug design. The database is designed and mostly operated by biologists with extensive experience in sequence analysis, assisted by an editorial board consisting of international experts in the field of HCV research.

The HCV sequences deposited in GenBank from the backbone of the database. Once a month the new sequences are downloaded and the available ancillary information is extracted from the GenBank records. This information may include country, sampling year, isolate names, genotype and subtype, host species, etc. For most of the sequences (excluding those shorter than 150 nt), relevant annotation information from the associated publications is added to the database. Sequences that do not have a genotype and subtype assigned to them are manually typed using phylogenetic analysis and BLAST searches.

At present, annotation fields in the database include:

Sequence information. Genotype, subtype, start and stop coordinates relative to the reference strain HCV-H, sampling country, sampling city, sampling date and sampling tissue.

Patient information. Health status, age, gender, ALT level, treatment and result, co-infection with HIV and hepatitis B, infection date, infection country, infection city, infection route, and infection outcome, HLA type and epidemiological relationship with other patients.

Annotation is being added continuously, both to the newly downloaded sequences and to the sequences already existing in the database. Table 1 presents an overview of the numbers of annotated fields in the database.

Table 1. Numbers of annotated sequences in the database (June 2004)

Total sequences	25 542
With genotype	21 774
With sampling country	16 961
With patient information	15 252
With sample tissue	5220
With drug treatment information	5170
With treatment response	4148
With sampling year	2871

The information in the database can be accessed via a versatile but user-friendly search interface that allows searches on some 30 different fields and lets the user automatically exclude the sequences from non-human hosts, the sequences from patent applications and the sequences that have a close epidemiological relationship (either from one patient or from a cluster of linked infections). The search results can be sorted and selected in various ways and include a graphic representation that shows, at a glance, how long each sequence is and where it is located in the genome. A graphical overview showing which regions and genotypes are included in the entire set of retrieved sequences can be generated (Fig. 1). An important feature is the ability to search by genomic region, so the user can locate all the sequences in the database that span, e.g. E1 and E2 and include or exclude sequences that are located in that region but do not cover it completely. In addition, the sequences that have been retrieved can be downloaded as an alignment. This alignment may need manual inspection and improvement, but it forms a very useful starting point. Alternatively, the sequences can be downloaded unaligned and/or translated to amino acids in any reading frame.

TOOLS PROVIDED ON THE WEBSITE

The website also provides pre-made alignments of HCV complete genomes, genes and proteins that have been manually corrected and cleaned, i.e. closely related sequences have been removed. Subsets of these, the reference alignments, contain three to four representatives of each available genotype and subtype, and can be used as the background sequences in various analyses. Graphical overviews of the total number and synonymous/non-synonymous variation of all genes and proteins (using sliding window analysis) are available. In addition, an alignment of flavivirus complete genomes is provided along with some results of the divergence analysis of these viruses.

The Geography tool can be used to plot frequencies of the different genotypes stored in the database as a function of their geographical origin (Fig. 2). This tool can be very useful to get a general idea of which genotypes have been found in which countries, as well as the density of sampling in

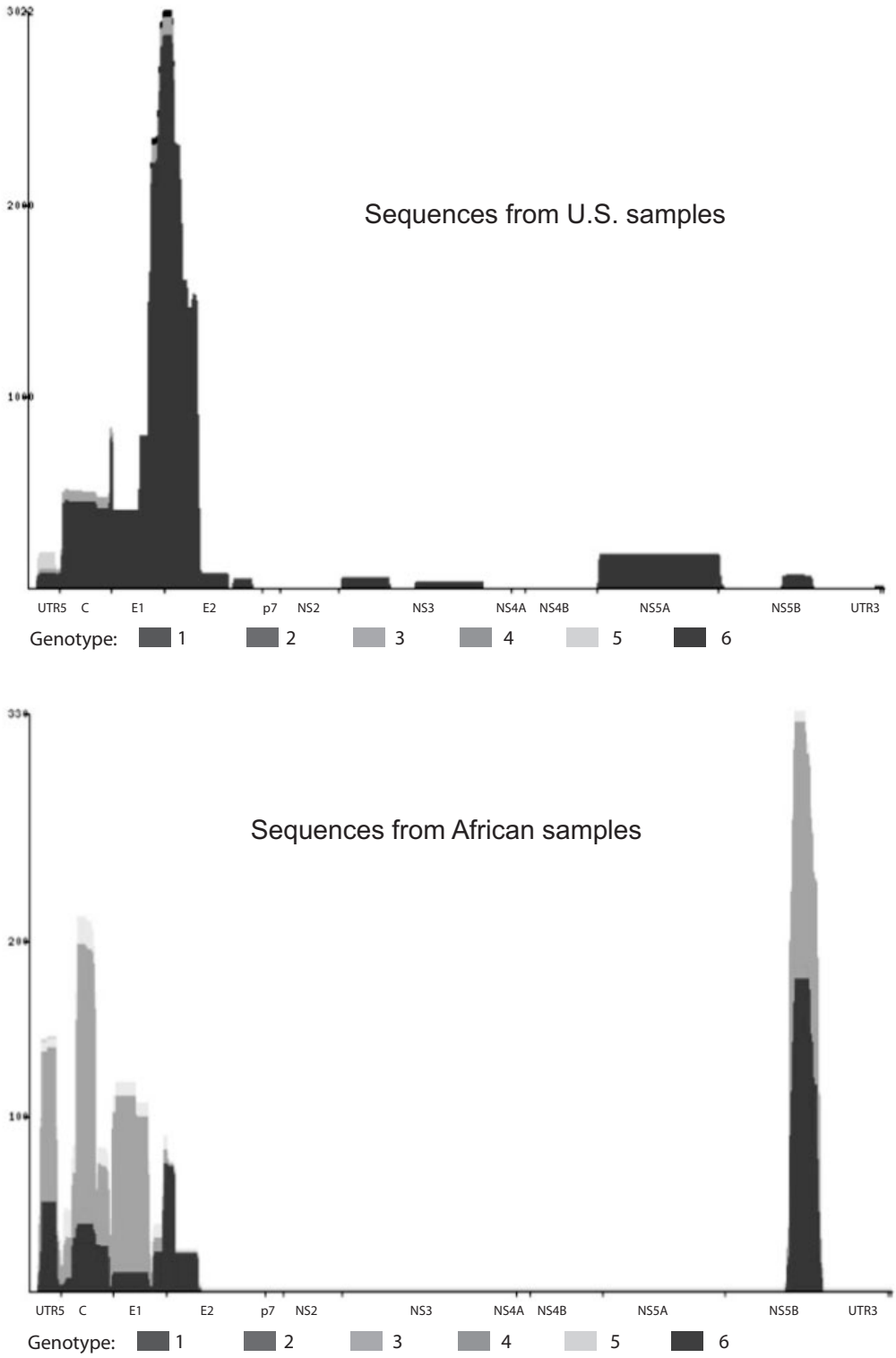

Fig. 1. Histograms showing the number, genotype and genomic region for two sets of sequences: those sampled in Africa and those sampled in the United States. Note that the scales of the vertical axes are different. From the figure it is evident that there are many more sequences from the United States; that they are mostly of genotype 1 whereas African sequences are of all genotypes; and that the best region for comparative analysis now (because most background sequences would be available) in Africa would be NS5B or Core/E1, because most background information is available. For the United States it would be E1/E2.

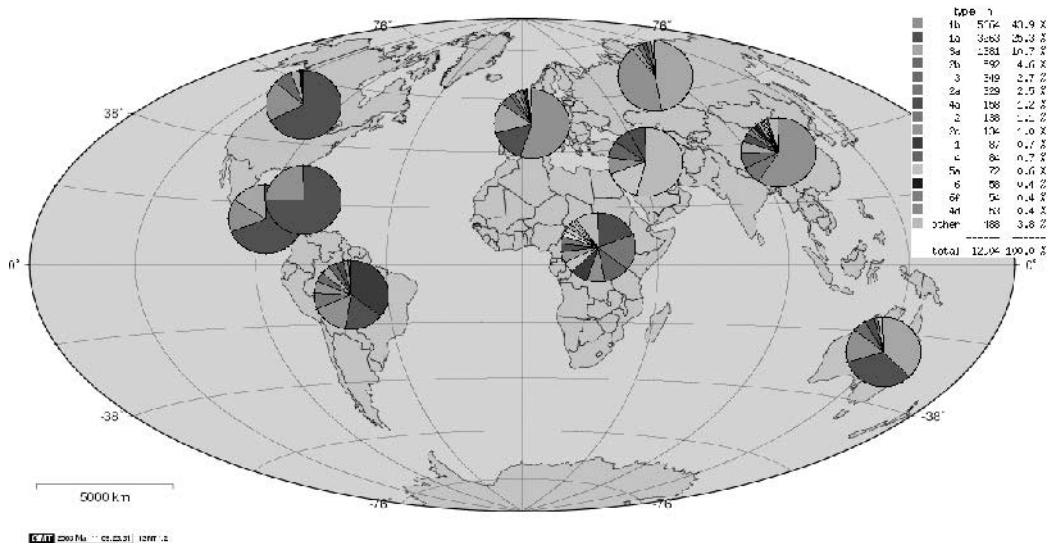

Fig. 2. World map showing the geographical distribution of sequences with different genotypes in the HCV database. The map was generated using the Geography tool, available on the HCV website.

different regions of the world. However, its results need to be interpreted with care: it is very easy to overlook the sampling biases that can distort the frequencies of the sequences in the database relative to those in the population.

The HCV sequence website offers a number of tools for common types of data analysis. Interfaces are provided to several programs that were originally written for use by the HIV database:

Syn-Nonsyn, a program that calculates, analyzes and builds trees from the numbers of silent (ds) and non-silent (dn) mutations in a codon alignment (Korber, 1997).

Glycosite, which tallies and plots N-linked glycosylation sites (Zhang *et al*., 2004). The Entropy program calculates and plots the Shannon entropy (a measure of variability) for each position in an alignment (Korber *et al*., 1994). It can also compare the entropy of all positions in two alignments, and perform a permutation-based statistical significance test to find positions with different variability.

Bill Bruno provided the code for FindModel, which is similar to Posada and Crandall's Modeltest script (Posada and Crandall, 1998), but uses Ziheng Yang's PAML (Yang, 1997) as a back end. FindModel analyses your sequence alignment to determine which evolutionary model fits it best; you can then use this model to build a better tree.

The website also has interfaces to several public domain programs:

Treemaker generates Neighbor-joining trees. It is a user-friendly interface to the DNAdist/Neighbor/Drawtree suite from the PHYLIP package (Felsenstein, 1984). The interface works around several common problems: it gapstrips the alignments before feeding them into DNAdist; it preserves the sequence names that are longer than 10 characters and removes negative numbers in the calculated distance matrix that can cause the programs to crash. The output provided includes PNG and PDF files of the tree figure, a downloadable Postscript code and the original PHYLIP treefile and outfile.

BLAST (Altschul *et al*., 1997) searches the HCV database for nucleotide sequences that are most similar to the user's query sequence. Optionally, the search can be limited to sequences that have a valid genotype.

PCOORD (Higgins, 1992) offers a principal coordinate analysis, a data-reduction technique similar to the principal components analysis, to identify co-varying positions in groups of sequences.

In addition to these programs, a large number of tools are available for manipulating or describing sequences:

Gene Cutter is an HCV-specific tool that finds defined genes in a nucleotide input sequence, generates the appropriate amino acid translation and is able to codon-align a set of sequences, which can then be downloaded.

Consensus, a versatile tool to create consensus sequences of groups of sequences that can be modified by a large number of parameters.

Sequence Locator is a program that finds the coordinates of an input sequence relative to the reference strain HCV-H (Fig 3). This program can be used as a means to standardize primer and epitope numbering and shows the user the location of an HCV sequence fragment. It provides the amino acid translation in the correct frame if a nucleotide sequence was submitted, and aligns the amino acid sequence against

LOCATION from start of HCVH genome 245 -> 721 (shown as red bar in map)

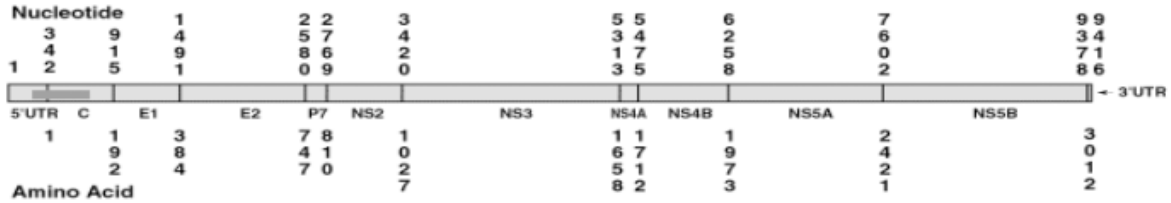

Table of genomic regions touched by query sequence		
CDS	**Nucleotide** position relative to CDS start in HCVH	Position relative to query sequence start
5UTR	245 -> 341	1 -> 97
Core	1 -> 380	98 -> 477
MSTNPKPQRKTKRNTNRRPQDVKFPGGGQIVGGVYLLPRRGPRLGVRATRKTSERSQPRG RRQPIPKARRPEGRTWAQPGYPWPLYGNEGCGWAGWLLSPRGSRPSWGPTDPRRRSRNLG KVIDTLX		

Alignment of the query sequence to the reference sequence:

```
Query GACTGCTAGC CGAGTAGTGT TGGGTCGCGA AAGGCCTTGT GGTACTGCCT 50
      :::::::::: :::::::::: :::::::::: :::::::::: ::::::::::
HCV-H GACTGCTAGC CGAGTAGTGT TGGGTCGCGA AAGGCCTTGT GGTACTGCCT 294
Query GATAGGGTGC TTGCGAGTGC CCCGGGAGGT CTCGTAGACC GTGCACCATG 100
      :::::::::: :::::::::: :::::::::: :::::::::: ::::::::::
HCV-H GATAGGGTGC TTGCGAGTGC CCCGGGAGGT CTCGTAGACC GTGCACCATG 344
```

Fig. 3. Output of the Sequence Locator tool, which finds the location of a sequence fragment of any length. The graphic representation shows where the user's input sequence is located; the table beneath it shows the amino acid translation and coordinates of all the genes included in the fragment.

the HCV-H nucleotide sequence if the input is an amino acid sequence. The sequence locator can also be used for reverse-complement sequences.

PeptGen has been designed to help immunologists rationally design overlapping peptide sets to probe the immune response, taking into account the forbidden N- and C-terminal amino acids and the desired peptide length.

Motifscan currently allows the user to input sequences such as an HLA-anchor residue motif and finds its occurrence each time either in the HCV-H reference strain protein or in a user specified protein. Motifscan is currently being modified and improved and the updated version will appear on the HCV website in the near future (Yusim *et al*., 2004).

Primalign automatically aligns a primer or sequence fragment to the HCV complete genome alignment. The interface returns the coordinates (HCV-H numbering) and an alignment of the fragment to all the sequences in the whole genome alignment.

Epilign searches HCV-H proteins for the best match to an amino acid string, e.g. an epitope, and generates an alignment from our main alignment so that researchers can rapidly assess the level of variation in an epitope of interest.

Seq-convert is a robust tool for rendering sequences and alignments in different formats; the user specifies both input and output format.

OmniRead attempts to automatically recognize input sequences in any valid sequence format and outputs them in any one of 23 other formats.

Seqpublish formats user alignments suitable for publication: identical columns are replaced by dashes and the sequences are printed in blocks of user-determined length. This program can also generate consensus sequences.

FUTURE ENHANCEMENTS

The HCV immunology database was made public in the summer of 2004; this database provides annotation and background information about HCV immunological epitopes, along with a set of analysis tools. This database will be described more extensively in a separate publication.

Several new interfaces are being developed currently. Distplot will help users calculate, analyze and generate graphs of matrices of pairwise distances. It will take a sequence alignment and (when applicable) grouping information, calculate

various distance measures, provide summary statistics and significance tests, and plot the distances relative to each other or to one sequence or group (e.g. the first sample in a time series). Other new tools that are planned for release in 2005 are, Synchalign, TreeMaker and Diagnostics. Synchalign will automatically merge two alignments based on a common reference sequence. TreeMaker will be expanded with a module for branch length calculation. Diagnostics will graphically display a large number of sequence or alignment.

ACKNOWLEDGEMENTS

We thank Bette Korber and our colleagues at the HIV databases for their generous help and support. We also gratefully acknowledge the help of the members of the Editorial Board, especially Dr Peter Balfe, Columbia University, for his tireless testing and many suggestions for improvement. The HCV database is funded by the Division of Microbiology and Infectious Diseases of the National Institute of Allergies and Infectious Diseases (NIAID; LAUR: 04–2085).

REFERENCES

Alter,M.J., Margolis,H.S., Krawczynski,K., Judson,F.N., Mares,A., Alexander,W.J., Hu,P.Y., Miller,J.K., Gerber,M.A., Sampliner,R.E. *et al*. (1992) The natural history of community-acquired hepatitis C in the United States. The Sentinel Counties Chronic non-A, non-B Hepatitis Study Team. *N. Engl. J. Med.*, **327**, 1899–1905.

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

CDC (2003) Transmission of hepatitis B and C viruses in outpatient settings—New York, Oklahoma, and Nebraska, 2000–2002. *MMWR Morb. Mortal. Wkly. Rep.*, **52**, 901–906.

Choo,Q.L., Kuo,G., Weiner,A.J., Overby,L.R., Bradley,D.W. and Houghton,M. (1989) Isolation of a cDNA clone derived from a blood-borne non-A, non-B viral hepatitis genome. *Science*, **244**, 359–362.

Farci,P. and Purcell,R.H. (2000) Clinical significance of hepatitis C virus genotypes and quasispecies. *Semin. Liver Dis.*, **20**, 103–126.

Felsenstein,J. (1984) PHYLIP: Phylogeny Inference Package, V3.5. University of Washington, Seatlle, WA.

Fleckenstein,J. (2004) Chronic hepatitis C in African Americans and other minority groups. *Curr. Gastroenterol. Rep.*, **6**, 66–70.

Gaschen,B., Kuiken,C., Korber,B. and Foley,B. (2001) Retrieval and on-the-fly alignment of sequence fragments from the HIV database. *Bioinformatics*, **17**, 415–418.

Gaschen,B., Taylor,J., Yusim,K., Foley,B., Gao,F., Lang,D., Novitsky,V., Haynes,B., Hahn,B.H., Bhattacharya,T. and Korber,B. (2002) Diversity considerations in HIV-1 vaccine selection. *Science*, **296**, 2354–2360.

Grakoui,A., Hanson,H.L. and Rice,C.M. (2001) Bad time for Bonzo? Experimental models of hepatitis C virus infection, replication, and pathogenesis. *Hepatology*, **33**, 489–495.

Higgins,D.G. (1992) Sequence ordinations: a multivariate analysis approach to analysing large sequence data sets. *Comput. Appl. Biosci.*, **8**, 15–22.

Hoofnagle,J.H. (1997) Hepatitis C: the clinical spectrum of disease. *Hepatology*, **26**(Suppl. 1), 15S–20S.

Korber,B. (1997) Signature and sequence variation analysis. In Rodrigo,A.G. and Learn,G.H. (eds.), *Computational Analysis of HIV Molecular Sequences*. Kluwer Academic Publishers, Dordrecht, The Netherlands.

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**, 20–29.

Lanford,R.E. and Bigger,C. (2002) Advances in model systems for hepatitis C virus research. *Virology*, **293**, 1–9.

Pawlotsky,J.M. (2003) Mechanisms of antiviral treatment efficacy and failure in chronic hepatitis C. *Antiviral Res.*, **59**, 1–11.

Posada,D. and Crandall,K.A. (1998) MODELTEST: testing the model of DNA substitution. *Bioinformatics*, **14**, 817–818.

Simmonds,P. (1999) Viral heterogeneity of the hepatitis C virus. *J. Hepatol.*, **31**(Suppl 1), 54–60.

Vrielink,H., Zaaijer,H.L., Reesink,H.W., van der Poel,C.L., Cuypers,H.T. and Lelie,P.N. (1995) Sensitivity and specificity of three third-generation anti-hepatitis C virus ELISAs. *Vox Sang.*, **69**, 14–17.

Yusim,K., Szinger,J.J., Honeyborne,I., Calef,C., Goulder,P.J.R. and Korber,B.T.M. (2004) Enhanced Motif Scan: A Tool to Scan for HLA Anchor Residues in Proteins. In Korber,B., Brander,C., Haynes,B., Koup,R., Kuiken,C., Moore,J.P., Walker,B.D. and Watkins,D.I. (eds), HIV Molecular Immunology. Los Alamos National Laboratory, Theoretical Biology and Biophysics, Los Alamos, NM.

Zhang,M., Gaschen,B., Blay,W., Foley,B., Haigwood,N., Kuiken,C. and Korber,B. (2004) Tracking global patterns of N-linked glycosylation site variation in highly variable viral glycoproteins: HIV, SIV, and HCV envelopes, and influenza hemagglutinin. *Glycobiology*, **14**, 1229–1246.