MS. CECILIA MARIA DELFINO (Orcid ID : 0000-0001-6915-193X)

Article type : Original Paper

A comprehensive bioinformatic analysis of hepatitis D virus (HDV) full-length genomes

Running title: Bioinformatic analysis of HDV full-genomes

Cecilia María Delfino[1*], Carolina Susana Cerrudo[2*], Mirna Biglione[3], José Raúl Oubiña[1], Pablo Daniel Ghiringhelli[2§], Verónica Lidia Mathet[1§]

[1]Consejo Nacional de Investigaciones Científicas y Tecnológicas (CONICET) -Universidad de Buenos Aires (UBA). Instituto de Investigaciones en Microbiología y Parasitología Médica, (IMPAM). Ciudad Autónoma de Buenos Aires, Argentina.

[2]Laboratorio de Ingeniería Genética y Biología Celular y Molecular – Área Virosis de Insectos (LIGBCM-AVI), Instituto de Microbiología Básica y Aplicada (IMBA), Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes, Roque Sáenz Peña 352 (B1876BXD). Bernal, Provincia de Buenos Aires. Argentina.

[3]Consejo Nacional de Investigaciones Científicas y Tecnológicas Consejo Nacional de Investigaciones Científicas y Tecnológicas (CONICET) -Universidad de Buenos Aires (UBA). Instituto de Investigaciones Biomédicas en Retrovirus y SIDA (INBIRS). Ciudad Autónoma de Buenos Aires, Argentina.

[*]Both authors have contributed equally

[§] Both authors have contributed equally

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/jvh.12876

Corresponding author:

Cecilia María Delfino, Ph.D.

CONICET-UBA. Instituto de Investigaciones en Microbiología y Parasitología Médica, (IMPAM). Paraguay 2155 (C1121ABG). Ciudad Autónoma de Buenos Aires, Argentina.

Phone: 54-11-5950-9500 ext. 2175. E-mail: lic.ceciliadelfino@gmail.com

ABSTRACT

In association with hepatitis B virus (HBV), hepatitis delta virus (HDV) is a subviral agent that may promote severe acute and chronic forms of liver disease. Based on the percentage of nucleotide identity of the genome, HDV was initially classified into three genotypes. However, since 2006, the original classification has been further expanded into eight clades/genotypes. The intergenotype divergence may be as high as 35-40% over the entire RNA genome, whereas sequence heterogeneity among the isolates of a given genotype is <20%; furthermore, HDV recombinants have been clearly demonstrated. The genetic diversity of HDV is related to the geographic origin of the isolates. This study shows the first comprehensive bioinformatic analysis of the complete available set of HDV sequences, by using both nucleotide and protein phylogenies (based on an evolutionary model selection, gamma distribution estimation, tree inference, and phylogenetic distance estimation), protein composition analysis and comparison (based on the presence of invariant residues, molecular signatures, amino acid frequencies, and mono- and di-amino acid compositional distances), as well as amino acid changes in sequence evolution. Taking into account the congruent and consistent results of both nucleotide and amino acid analyses of GenBank available sequences (recorded as of January, 2017), we propose that the eight hepatitis D virus genotypes may be grouped into three large genogroups fully supported by their shared characteristics.

Keywords: HDV; bioinformatics; genotype; genogroup; nucleotide and protein phylogenies

INTRODUCTION

Hepatitis delta virus (HDV) is a subviral agent that can lead to severe acute and chronic forms of liver disease in association with hepatitis B virus (HBV) [1]. HDV is endemic in many populations with a high prevalence of HBV, ranging from 60 % among Mongolian subjects [2] to >20% in Central Africa, Romania, Pakistan, Iran, in the mountainous region of Venezuela and Colombia, as well as in the Amazon Basin in South America. In addition, 3-8% prevalence was recorded in some USA states. However, some evidence suggests that the prevalence of HDV does not proportionally correlate with that of HBV [3-5]. The hepatitis delta antigen (HDAg) is the only known functional protein of this virus and an internal component of HDV virion particles. Together with the viral RNA genome, HDAg constitutes the viral nucleocapsid [6]. The protein exhibits two distinct forms of dissimilar size: S-HDAg with 195 amino acids (24 kDa) and L-HDAg with 214 amino acids (27 kDa). These isoforms are identical, except for the 19-20 additional amino acids at the C-terminus of L-HDAg [7]. The latter event is the result of the cellular editing activity of ADAR1 on the antigenomic RNA, which mutates the amber stop codon UAG to UGG, allowing the elongation of S-HDAg by the addition of 19 or 20 amino acids in a genotype-dependent manner. Besides its structural function, HDAg also plays a very critical role in the HDV life cycle by participating in various steps of viral replication, including viral RNA synthesis (by S-HDAg) and virus assembly (by L-HDAg), among others [8-11].

Based on the percentage of nucleotide identity of the genome, HDV was initially classified into three genotypes, but since 2006 the original classification has been further expanded into eight clades/genotypes [12]. The intergenotype divergence may be as high as 35-40% over the entire RNA genome, whereas the sequence heterogeneity among the isolates of a given

genotype is <20% [5]. The genetic diversity of HDV is related to the geographic origin of the isolates. HDV1 is distributed worldwide but predominates in Europe and North America; HDV2 is found in Japan, Taiwan and Russia; HDV3 is exclusively detected in the Amazonian region [13-15]; HDV4 in Taiwan and Okinawa (Japan), whereas HDV5-8 predominate in Central and West Africa [4,5]. Moreover, HDV recombinants appear to play an important role both in *in vitro* experiments as well as in natural infections [16-18].

Our study shows the first comprehensive bioinformatic analysis of the complete available set of HDV sequences. Based on the congruent and consistent results of both nucleotide and amino acid analyses of GenBank available sequences (recorded as of January, 2017), we propose that eight current hepatitis D virus genotypes may be grouped into three large Genogroups.

MATERIALS AND METHODS

Sequences

All non-redundant nucleotide sequences of complete genomes obtained from GenBank database (N = 213 sequences) and all amino acid sequences (N = 379 full-length S and L protein sequences) available in the UniProt database were analyzed (recorded as of January, 2017). Different subsets of sequences were used for the particular analyses throughout the work. After a first analysis, HDV sequences ascribed to any woodchuck isolate or identified in bibliography as putative recombinants were ruled out. Names, accession numbers and individual references of nucleotide and protein sequences used are summarized in Supplementary Table 1.

Phylogeny

Phylogenetic and molecular evolutionary analyses were conducted by using the MEGA v. 6.0 software [19] and Mr. Bayes software provided by CIPRES server [20]. In addition, Protest 3.4 [21] and jModelTest 2.1 [22] software were used to select the most suitable evolutionary model. For the nucleotide and protein phylogeny, a subset of 144 complete unambiguous genomic sequences and their associated L-HDAg sequence were used throughout. This selected subset of sequences have no more than 4 undefined amino acids in sequence and each full genome sequence has its corresponding L sequence identified in the database.

Evolutionary model selection

Nucleotides. In the MEGA v. 6.0 software, Maximum likelihood (ML) [23] was used to evaluate 24 dissimilar nucleotide substitution models, with the following set of parameters: Tree to be used = Automatic (Neighbor-Joining, NJ tree), Statistical method = ML, Substitution types = Nucleotide, Gaps/Missing data = Partial deletion, Site coverage cutoff = 95%, Select codon positions = All+Noncoding and Branch swap filter = Moderate. Models tested were GTR (General Time Reversible), HKY (Hasegawa-Kishino-Yano), TN93 (Tamura-Nei), T92 (Tamura 3-parameter), K2 (Kimura 2-parameter), and JC (Jukes-Cantor). In all cases the models were evaluated on their standard form and on their variants +F (frequencies), +G (gamma distributed), +I (invariant sites), +F+G, +G+I, +G+I+F and +I+F. The best fitted model of DNA evolution was also obtained using jModelTest 2.1 [22] with the Akaike information criterion. The best nucleotide model selected based on the results of the two programs was GTR+G+I. The value of the shape parameter for the discrete Gamma Distribution was = 0.815 and the proportion of invariant sites = 0.149.

Proteins. ML [23] in the MEGA v. 6.0 software, was used to evaluate 48 dissimilar amino acid substitution models, with the following set of parameters: Tree to be used = Automatic (NJ tree), Statistical method = ML, Substitution types = Amino acid, Gaps/Missing data = Use all sites and branch swap filter = Moderate. Models tested were cpREV (General Reversible Chloroplast), Dayhoff, JTT (Jones-Taylor-Thornton), mtREV24 (General Reversible Mitochondrial), rtREV (General Reverse Transcriptase), and WAG (Whelan and Goldmann). In all cases, the models were evaluated on their own standard form and on their own variants +F (frequencies), +G (gamma distributed), +I (invariant sites), +F+G, +G+I, +G+I+F and +I+F. The best fitted model of protein evolution was also obtained using Protest 3.4 [21] software with the Akaike information criterion. The best model selected based on the results of the two programs was JTT+G+I. The value of the shape parameter for the discrete Gamma Distribution was = 0.878 and the proportion of invariant sites = 0.096.

Tree inference

Nucleotides. An initial phylogenetic tree was inferred in MEGA software by using the NJ method with the following set of parameters: Bootstrap with 500 replicates [24], Gaps/Missing data = Pairwise deletion, Model = Nucleotide (T92+G model) [25], patterns among sites = Same (Homogeneous), rates among sites = Different (Gamma Distributed) and gamma parameter = 0.815. Bayesian inference analyses were performed under the GTR+G+I model (Rates = invgamma, lset Nucmodel = 4by4, lset Nst = 6, lset Nbetacat = 5, prset pinvarpr = fixed(0.149), prset shapepr = fixed(0.815)). MCMC chain length was 10,000,000 generations (nruns = 2, nchains = 4, temp = 0.200, ngen = 10,000,000) with trees sampled every 1,000 generations (samplefreq = 1,000); the first 2,500 trees were discarded as burn-in (burninfrac = 0.250). Bayesian 50% majority rule consensus trees as inferred and posterior probabilities more than 0.50 are given for appropriate clades.

Proteins. The original phylogeny tree was inferred in MEGA software by using the NJ method with the following set of parameters: Bootstrap with 500 replicates [24], Gaps/Missing data = Pairwise deletion, Model = Amino (JTT+G model) [26], patterns among sites = Same (Homogeneous), rates among sites = Different (Gamma Distributed) and gamma parameter = 0.878. The tree was also inferred by Bayesian analyses using Mr. Bayes software with the following parameters: JTT+G+I (lset Rates = invgamma, lset Nucmodel = protein, lset Nst = 6, lset Nbetacat = 5, prset pinvarpr = fixed(0.096), prset aamodelpr = fixed(jones), prset shapepr = fixed(0.878)). MCMC chain length was 20,000,000 generations (nruns = 2, nchains = 4, temp = 0.200, ngen = 10,000,000) with trees sampled every 1,000 generations (samplefreq = 1,000); the first 2,500 trees were discarded as burn-in (burninfrac = 0.250). Bayesian 50% majority rule consensus trees as inferred are given.

Estimation of phylogenetic distances

Nucleotides. Evolutionary distances were calculated pairwise by using MEGA v. 6.0 software with the following set of parameters: Variance estimation method = Bootstrap, Number of Bootstrap replications = 500, Substitutions type = Nucleotide, Model = Kimura 2-parameters, Substitutions to include = d: Transitions + Transversions, Rates among sites = gamma distributed, Gamma parameter = 0.815, Pattern among lineages = Same (Homogeneus) and Gaps/Missing data = Pairwise deletion.

Proteins. Evolutionary distances were calculated pairwise by using the MEGA 5 software with the following set of parameters: Variance estimation method = Bootstrap, Number of Bootstrap replications = 500, Substitutions type = Amino acid, Model = JTT+G, Rates among sites = gamma distributed, Gamma parameter = 0.878, Pattern among lineages = Same (Homogeneus) and Gaps/Missing data = Pairwise deletion.

Protein similarity analyses

A selected set of 281 non-redundant full-length S and L protein sequences was used for these analyses. Sequence identity and similarity studies were performed by an all-against-all strategy and by carrying out pairwise alignments with ClustalW2 software [27] by using default parameters and Gonnet matrices for similarities [28]. Identity was calculated as the percentage of identical residues in the pairwise alignment, and similarity was calculated as the sum of identities plus similarities (strong plus weak) and expressed as a percentage. Relative similarity profiles were calculated by using the ClustalX [29] consensus symbol (*****, **:**, **.** and blank space) as the input sequence, in an overlapping window-based strategy. Arbitrary values of +1 for identical (*****), +0.5 for strong (**:**) and +0.25 for weak (**.**) conservative changes, and −0.5 for non-identical (blank spaces) residues were used. The sum of assigned values for each residue in each window (11 amino acids) was divided by the window width and allotted to the central position to generate the plots.

Protein composition analyses and comparison

Invariant residues

For each group (Group 1, 2 and 3), multiple aligned S-HDAg and L-HDAg amino acid sequences were separately processed in two parts to determine the invariant residues: firstly, the overlapped region and secondly, the C-terminal Large antigen domain. For each aligned region, only residues with a frequency of 1.0 were represented in the pseudoconsensus sequence; the remaining residues were named with an “x”.

Molecular signatures

For this particular analysis all full-length S and L protein sequences (N = 379) were used. Blocks of specific sequences were determined based on protein multiple alignments of S and L proteins of each group. Molecular signatures are written employing the usual syntactic rules (PROSITE-like) [30].

Amino acid frequencies

Amino acid frequencies were determined by using an *ad hoc* program (Ghiringhelli, unpublished). Results were tabulated and mean frequencies as well as standard deviation for each group were calculated.

Mono- and di-amino acid compositional distances

Mono- and di-amino acid compositional distances were calculated in a similar way to that used to calculate the distance of dinucleotide frequencies [31]. Equation 1 and equation 2 were used to mono- and di-amino acid compositional distances calculation, respectively.

(1)

$$\delta_{aa} = \left[\frac{\sum_{x=1}^{20}\left(Q_{aa_x} - R_{aa_x}\right)^2}{20}\right] * 10^5$$

Where δ_{aa} is the amino acid compositional distance, Q_{aa_x} is the amino acid frequency [1, 2, …, 20] in the query species and R_{aa_x} is the mean amino acid frequency [1, 2, …, 20] in the reference group used.

(2)

$$\delta_{\mathrm{diaa}} = \left[\frac{\sum_{\mathrm{x}=1}^{400}\left(\mathrm{Q}_{\mathrm{diaa}_{\mathrm{x}}} - \mathrm{R}_{\mathrm{diaa}_{\mathrm{x}}}\right)^2}{400}\right] * 10^6$$

Where δ_{diaa} is the di-amino acid compositional distance, Q_{diaa_x} is the di-amino frequency [1, 2, …, 400] in the query species and R_{diaa_x} is the mean di-amino acid frequency id [1, 2, …, 400] in the reference group used.

RESULTS

Phylogeny

The phylogenetic analyses of full-length HDV sequences available from GenBank as of January 2017 (Suppl. Table 1) was performed using two different methods, NJ (MEGA) and Bayesian inference analyses (Mr. Bayes, CIPRES). The tree topology obtained with the two methods was similar, both at full-genome nucleotide and L-HDAg protein levels (Fig. 1A and 1B). This analysis shows the existence of three major clades, of which the most distant seems to be the one corresponding to genotype 3 (with 99/100 node consistency for NJ and Bayesian inference, respectively). The clade containing all genotype 1 sequences (86 sequences) has the same node consistency values as genotype 3 (37 sequences). Finally, the last clade link together the sequences (21 sequences) corresponding to the remaining genotypes (HDV-2, HDV-4, HDV-5, HDV-6, HDV-7, and HDV-8) with values of 93/100 (NJ/Bayesian inference) for nucleotides and 88/100 (NJ/Bayesian inference) for amino acids. These results suggest that the sequences of the HDV2, and HDV-4 to HDV-8 genotypes are closely related and may share additional characteristics. For this reason, additional characterizations at the

amino acid level were carried out comparing the three large groups; called Group 1, 2 and 3 (G1, G2 and G3).

The evolutionary distance analyses for nucleotide and amino acid sequences were calculated for intragroup (G1, G2 and G3) and intergroup (G1 *vs* G2, G1 *vs* G3 and G2 *vs* G3) relationships and plotted as box plots. No significant differences were recorded when intragroup sequences were tested. In contrast, a significant distance was measured when intergroups were analyzed, G1 *vs* G3 and G2 *vs* G3 exhibiting the highest values (Fig. 2A and 2B).

Protein sequence similarity

As shown in Fig. 3A, the S and L overlapped amino acid sequences from group 3 exhibit the highest conservation for a given position throughout the protein, as compared with groups 1 and 2, remarkably between amino acids 126 and 156, as well as between amino acids 170 and 190. The former region encompasses the arginine-rich motif (ARM2) within the RNA binding domain; the latter overlaps with the proline and glycine-rich domain, which is associated with the packaging signal.

Protein sequence similarity was analyzed both within each group (G1, G2 and G3), as well as between groups (G1 *vs* G2, G1 *vs* G3 and G2 *vs* G3), showing a higher similarity between G1 and G2 (Fig. 3B).

Compositional studies

Typically, group 1 members show S and L proteins with constant lengths (195 and 214 amino acids, respectively). In sharp contrast, such lengths for groups 2 and 3 are variable: the S protein encompasses 193-195 amino acids for G2 and 190-194 for G3, while the L protein spans 208-214 for G2, and 210 and 214 residues for G3.

Intragroup amino acid composition is more homogeneous, either within G1 or within G3, as compared with G2. The intergroup analysis is somehow influenced by some bias inherent to each group. For example, G1 exhibits as an average a lysine excess (39% higher than G2, and 16% higher than G3). On the other hand, G2 exhibits an arginine excess (28% higher than G1, and 14% higher than G3). However, despite such differences, the theoretical average net charge keeps a constant value: +12.5. Even more strikingly, the bias in amino acid composition promotes dissimilar values in the average hydrophobicity profile inherent to each group: G1 = 1.66, G2 = 1.82 and G3 = 1.73 (Fig. 4A).

A measure of the differences between pairs of protein sequences (from different viral isolates) are the mono- and di-amino acid compositional distances. Usually, the relationship among viral isolates is established only by sequence comparison. However, the comparison of mono- and di-amino acid compositional distances provides another measure of similarity that can be used. Mono- and di-amino acid compositional distances analysis clearly shows the high similarity of proteins in intragroup comparisons (Figs. 4B and 4C).

Protein sequence structure and conservation

The analysis of invariant residues of S-HDAg and L-HDAg showed 30 fully invariant amino acids, some of them placed either in the nuclear localization signal or in the arginine-rich motif (Fig. 5A). Of these 30 invariant amino acids, 6 corresponded to G (20.0 %), 5 to R (16.7 %), 3 either to E or to L (10.0 % each), 2 either to P, K, M, or to F (6.7 % each) and 1 either to D, N, V, A or to Q (3.3 % each).

In an intergroup (global) multiple alignment, each group exhibits 2 specific deletions of amino acids, only one of which is shared with one of the two remaining groups. Within groups 2 and 3 there are few sequences exhibiting a small number of additional deletions. In the last 10 amino acids corresponding to the C-terminal of L-HDAg (packaging signal) is

located one of the regions showing the highest intragroup conservation as well as the maximal intergroup diversity.

Based on a deep analysis of global and individual (group-specific) multiple alignments, we observed two blocks of residues which constitute group-specific subregions (including fully conserved and variable positions). We proposed them as group-specific molecular signatures (Fig. 5B). Both could be integrated in a sequence pattern that will permit to quickly assign new sequences obtained by massive parallel sequencing methods to the different groups (Fig. 5C).

DISCUSSION

Hepatitis D occurs worldwide, since around 5% of hepatitis B surface antigen (HBsAg) carriers in the world are infected by HDV as well [32]. Initially, based on the coding region of HDAg, HDV was classified into 3 large genotypes (1, 2, and 3) [33, 34], with different geographic distribution. Up to the present time, the HDV has been classified into eight clades/genotypes: HDV-1, HDV-2 (initially genotype IIa), HDV-3, HDV-4 (initially genotype IIb), HDV-5, -6, -7 and -8 [5, 12]. Over the years the HDV genotyping was carried out by different methodologies: hybridization, restriction fragment length polymorphism (RFLP) or sequence analysis corresponding to particular regions of the genome [35]. More recent studies of novel genomes included full length genomic sequencing and genotype comparison. However, when performing a phylogenetic analysis other authors have selected only some of the genomes from databases as reference sequences to compare them with the newly obtained genome [36]. Additionally, the reference sequences taken into account vary according to the authors.

In this study, phylogenetic trees were performed by using two methods: Neighbor-Joining (MEGA) and Bayesian inference (Mr. Bayes, CIPRES server) over non-redundant complete genome sequences. To our knowledge, this is the first time that phylogenetic analysis of all full-genome (deposited in the GenBank -until January, 2017) was performed in such comprehensive manner. These analyses, regarding both full-length genome sequences and L-HDAg, rendered congruent results allowing us to propose the grouping of the eight current genotypes into three large groups (G1, G2 and G3; Fig. 1).

Within the proposed Groups 1 and 3, all sequences belong to the current homonymous genotypes 1 and 3. In contrast, the proposed Group 2 encompasses HDV sequences (previously) ascribed to genotypes 2 and 4-8. The three phylogenetic analyses shows similar support values for NJ and Bayesian tree inferences (i.e., G1: 99 and 100; G2: 93 and 100, G3: 99 and 100; for NJ and Bayesian posterior probabilities, respectively), hence, all values support the proposed three Groups (Fig. 1). Further genetic distance analyses performed by using the K 2-p model for nucleotides and JTT+G model for proteins, supported the existence of the above mentioned three groups (Fig. 2). Moreover, the two amino acid deletions observed within each group appeared to be specific for each proposed group, being present in all sequences of such group. Interestingly, in all cases, only one of such variants is shared with one of the two remaining groups (Fig. 5A). Additional deletions existing in few members of groups 2 and 3 might be due either to sequencing misreadings or to a recent evolutionary divergence event.

The highest amino acid conservation was recorded in group 3, as compared with groups 1 or 2. It encompassed two regions between residues 126 and 156, as well as between amino acids 170 and 190 throughout the S and L overlapped amino acid sequences. The former region includes the arginine-rich motif (ARM2) within the RNA binding domain; the latter overlaps with the proline and glycine rich domain that is next to the packaging signal. The intragroup

(G1, G2 and G3) and intergroup (G1 *vs* G2, G1 *vs* G3 and G2 *vs* G3) protein sequence similarity demonstrates a higher similarity between G1 and G2 (Fig. 3), and supports the hypothesis of a common ancestry.

When the amino acid composition of each group was analyzed, the frequencies of each residue were calculated. Importantly, amino acids with a significant bias contribution were clearly observed: G, K, L and S for G1; D, H, R and T for G2 and A, P and Q for G3. In this regard, it is worth mentioning a higher frequency of R within G2, as compared with that observed with the remaining G1 and G3. The intra- and inter-group analysis of mono- and di-amino acid distances shows a greater internal homogeneity within and between G1 and G3, as compared with G2. Nevertheless, the distribution of mono- and di-amino acid minimal distances in intra- and inter-group analyses supports the hypothesis of three main groups or clades (Fig. 4). Additionally, it could be observed that several isolates, previously identified as belonging to either genotype 2, 4, 5, 6, 7 or 8, were reported from the same geographical areas (Asia and Africa), reinforcing the notion that they could be grouped within the same group (G2).

Remarkably, 30 invariant amino acids are observed throughout the alignment from all three group-derived pseudo-consensus sequences. These invariant residues could be related to crucial structural and/or functional features. It is interesting to note that there are only 5 further invariant residues shared by G1 and G2, 23 shared by G2 and G3 and other 12 residues shared by G1 and G3 (Fig. 5A). The aforementioned results –together with the fact that amino acid deletions appear to be specific for each proposed group- provide further support to the currently proposed 3 main branches within the phylogenetic tree. Although the above mentioned small number of invariant residues shared by G1 and G2, would appear to suggest unrelated branches, the reader should take into account that an invariant amino acid is shown in a given group pseudo-consensus sequence only when their whole set (100%)

exhibits the same residue at such position. In addition, we determined group-specific molecular signatures (Fig. 5B) that could help to easily classify newly obtained HDV sequences in any of the three groups. Testing assays employing the prosite-like sequence patterns (Fig. 5C) against a global L-HDAg database correctly identify members of each group.

Previous studies suggested that HDV genotype 2 (according to the classic nomenclature) is associated with a less aggressive course of the disease, as compared with genotype 1 [37], a fact attributed to the lower packaging efficiency of G2 with respect to G1 [38]. Moreover, a variation in packaging efficiency has been observed among different isolates of the same genotype, reflecting the critical role of isoform L- length (208-214 amino acids for G2, which contrasts sharply with the invariable 214 amino acids of G1) in HDAg interactions with HBsAg [11]. Furthermore, it has been reported that both the hydrophobic nature of the L-HDAg C-terminal domain provided by C211-farnesylation, as well as the number of hydrophobic residues of the packaging signal (which differs among HDV genotypes; Fig. 5A) enhance HDV interactions with surface proteins of HBV and thus, packaging efficiency. Remarkably, we demonstrated that the last 10 amino acids located at the C-terminal of L-HDAg exhibit the highest intragroup conservation, as well as the highest intergroup diversity. Such region overlaps with the 19 amino acids of L-HDAg, corresponding to the packaging signal responsible for the interaction between HBsAg and HDAg.

In further support of the proposed HDV classification into three nucleotide groups, amino acid changes regarding the sequence evolution were observed, confirming the existence of three monophyletic groups. This might suggest that G3 evolved earlier together with a common ancestor of the latter two groups (G1 and G2).

Therefore, based on congruent and consistent results of both nucleotide and amino acid analyses, we propose that genotypes of hepatitis D virus could be reorganized into 3 main

groups or genogroups -each of them exhibiting a specific molecular signature at the amino acid level-, which still remain to be explored regarding their correlation with virological and/or clinical features.

CMD, M.B., J.R.O., C.S.C. and V.L.M. are researchers from CONICET. P.D.G. is Director of LIGBCM-AVI and IMBA (UNQ).

ACKNOWLEDGEMENTS AND DISCLOSURES

This study was funded by Agencia Nacional de Promoción Científica y Tecnológica (ANPCyT PICT 2340), Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET PIP-00086), Universidad de Buenos Aires (UBACyT 20020130100891BA) and Universidad Nacional de Quilmes (Argentina). The authors are deeply indebted to María Victoria Illas for her excellent work to improve the readability of the manuscript.

REFERENCES

1- Wedemeyer H. Hepatitis D revival. Liver Int. 2001; 31 Suppl 1, 140-144.

2- Chen X, Oidovsambuu O, Liu P et al. A novel quantitative microarray antibody capture (Q-MAC) assay identifies an extremely high HDV prevalence amongst HBV infected Mongolians. Hepatology 2016 Nov 23 [Epub ahead of print].

3- Wedemeyer H., Manns MP. Epidemiology, pathogenesis and management of hepatitis D: update and challenges ahead. Nat Rev Gastroenterol Hepatol 2010; 7(1): 31-40.

4- Alvarado-Mora MV, Locarnini S., Rizzetto M., Pinho JR. An update on HDV: virology, pathogenesis and treatment. Antivir Ther 2013; 18(3 Pt B): 541-548.

5- Lempp FA, Ni Y, Urban S. Hepatitis delta virus: insights into a peculiar pathogen and novel treatment options. Nat Rev Gastroenterol Hepatol 2016; 13(10):580-589.

6- Bonino F, Hoyer B, Shih JW, Rizzetto M, Purcell RH, Gerin JL. Delta hepatitis agent: Structural and antigenic properties of the delta-associated particle. Infect Immun 1984; 43:1000–1005.

7- Lai MM. The molecular biology of hepatitis delta virus. Annu Rev Biochem 1995; 64:259-286.

8- Lai M. Chapter 4: hepatitis delta virus: Biochemical Properties and Functional Roles in HDV Replication. In: Handa H, Yamaguchi Y, editors. Medical intelligence unit hepatitis delta virus. New York, USA: Springer Science+Business Media, Inc 2006: 38–51.

9- Huang WH, Chen CW, Wu HL, Chen PJ. Post-translational modification of delta antigen of hepatitis D virus. Curr Top Microbiol Immunol 2006; 307:91-112.

10- Shih HH, Jeng KS, Syu WJ et al.. Hepatitis B surface antigen levels and sequences of natural hepatitis B virus variants influence the assembly and secretion of hepatitis d virus. J Virol 2008; 82(5), 2250–2264.

11- Greco-Stewart V, Pelchat M. Interaction of Host Cellular Proteins with Components of the Hepatitis Delta Virus. Viruses 2010; 2: 189-212.

12- Le Gal F, Gault E, Ripault MP, Serpaggi J, Trinchet JC, Gordien E, Dény P. Eighth major clade for hepatitis delta virus. Emerg Infect Dis 2006; 12(9): 1447-1450.

13- Duarte MC, Cardona N, Poblete F et al. A comparative epidemiological study of hepatitis B and hepatitis D virus infections in Yanomami and Piaroa Amerindians of Amazonas State, Venezuela. Trop Med Int Health 201; 15(8): 924-933.

14- Alvarado-Mora MV, Pinho JR. Epidemiological update of hepatitis B, C and delta in Latin America. Antivir. Ther. 2013; 18(3 Pt B):429-433

15- Cicero MF, Pena NM, Santana LC et al. Is Hepatitis Delta infections important in Brazil? BMC Infect Dis 2016; 16:525.

16- Lin CC, Yang ZW, Iang SB, Chao M. Reduced genetic distance and high replication levels increase the RNA recombination rate of hepatitis delta virus. Virus Res 2015; 195: 79-85.

17- Lin CC, Lee CC, Lin SH et al. RNA recombination in Hepatitis delta virus: Identification of a novel naturally occurring recombinant. J Microbiol Immunol Infect 2015; S1684-1182(15)00910-X.

18- Sy BT, Nguyen HM, Toan NL et al. Identification of a natural intergenotypic recombinant hepatitis delta virus genotype 1 and 2 in Vietnamese HBsAg-positive patients. J Viral Hepat. 2015; 22(1): 55-63.

19- Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol 2013; 30: 2725-2729.

20- Miller MA, Pfeiffer W, Schwartz T. "Creating the CIPRES Science Gateway for inference of large phylogenetic trees" in Proceedings of the Gateway Computing Environments Workshop (GCE), 14 Nov. 2010, New Orleans, LA: pp 1 - 8.

21- Darriba D, Taboada GL, Doallo R, Posada D. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 2011; 27:1164-1165.

22- Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. Nature Methods 2012; 9(8): 772.

23- Nei M, Kumar S. Molecular Evolution and Phylogenetics. Oxford University Press, New York. Publications Ltd, 2000.

24- Saitou N, Nei M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol Biol Evol 1987; 4:406-425.

25- Tamura K. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G + C-content biases. Mol. Biol Evol 1992; 9: 678-687.

26- Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 1992; 8(3): 275-282.

27- Larkin MA, Blackshields G, Brown NP et al. Clustal W and Clustal X version 2.0. Bioinformatics 2007; 23: 2947-2948.

28- Benner SA, Cohen MA, Gonnet GH. Amino acid substitution during functionally constrained divergent evolution of protein sequences. Protein Engineering 1994; 7:1323-1332.

29- Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 1997; 25: 4876-4882.

30- Bairoch A. PROSITE: a dictionary of sites and patterns in proteins. Nucleic Acids Res 1991; 19 Supplement: 2241-2245.

31- Karlin S, Campbell AM, Mrázek J. Comparative DNA analysis across diverse genomes. Annu Rev Genet 1998; 32: 185–225

32- Rizzetto M. Hepatitis D Virus: Introduction and Epidemiology. Cold Spring Harb Perspect Med 2015; Jul 1;5(7): a021576.

33- Zhang YY, Tsega E, Hansson BG. Phylogenetic analysis of hepatitis D virus indicating a new genotype I subgroup among African isolates. J Clin Microbiol. 1996; 34: 3023-3030.

34- Casey JL, Polson AG, Bass BL, Gerin JL. Molecular biology of HDV: analysis of RNA editing and genotype variation. In: Rizzetto M, Purcell RH, Gerin JL, eds. Viral Hepatitis and Liver Disease. Turin, Italy: Edizioni Minerva Medica 1997: 290-294.

35- Le Gal F, Badur S, Hawajri NA et al. Current hepatitis delta virus type 1 (HDV1) infections in central and eastern Turkey indicate a wide genetic diversity that is probably linked to different HDV1 origins. Arch Virol 2012; 57(4):647-659.

36- Shirvani-Dastgerdi E, Amini-Bavil-Olyaee S, Alavian SM, Trautwein C, Tacke F. Comprehensive analysis of mutations in the hepatitis delta virus genome based on full-length sequencing in a nationwide cohort study and evolutionary pattern during disease progression. Clin Microbiol Infect 2015, 21(5): 510.e11-23.

37- Su CW, Huang YH, Huo TI et al. Genotypes and viremia of hepatitis B and D viruses are associated with outcomes of chronic hepatitis D patients. Gastroenterology 2006. 130(6): 1625-1635.

38- Hsu SC, Syu WJ, Sheen IJ, Liu HT, Jeng KS, Wu JC. Varied assembly and RNA editing efficiencies between genotypes I and II hepatitis D virus and their implications. Hepatology 2002; 35(3): 665-672.

SUPPLEMENTARY MATERIAL

Supplementary Table 1: Names, accession numbers and individual references of nucleotide and protein sequences.

FIGURE LEGENDS

Figure 1. Phylogenetic inferences of HDV. Phylogenetic trees of hepatitis delta virus (HDV) inferred from the 144 nucleotide sequences of HDV complete genomes (**A**) and 144 amino acid sequences of L-HDAg encoded within (**B**). In both cases, trees were reconstructed using Neighbor-Joining (NJ, MEGA) and Bayesian (Mr. Bayes, CIPRES server) inference methods. The resulting trees are shown; bootstrap values and posterior probability of the conserved nodes between the different analyses were added on the NJ-tree. In the figure the trees are drawn to scale in a radial form, with branch lengths in the same units as those of the evolutionary distances used to infer them (NJ, MEGA). Branch support measures are

indicated as a percentage at selected internodes, depicting NJ bootstrap values / posterior probabilities of Bayesian inference.

Figure 2. Evolutionary distance analyses. Intragroup (G1, G2 and G3) and intergroup (G1 *vs* G2, G1 *vs* G3 and G2 *vs* G3) relationships were plotted as box plots. The boundary of the boxes closest to zero indicates the 25th percentile, the line within the box marks the median, and the boundary of the box furthest from zero indicates the 75th percentile. Error bars above and below the box indicate the 90th and 10th percentiles, respectively. Circles indicate outlying points. Grey dashed lines indicate putative cut offs. Colour for each group is similar to that used in Figure 2. Colour for intergroup combines the basic colours of both compared groups. **A. Nucleic acids.** Genetic distances between pair of genomes were calculated using the K 2-p model. **B. Proteins.** Genetic distances between pair of proteins were calculated using the JTT+G model.

Figure 3. A. Relative similarity plots. The graph shows intragroup sequence conservation. Black, red and blue lines correspond to G1, G2 and G3, respectively. Sequence conservation of shared portions of S-HDAg and L-HDAg is represented by a continuous line. C-terminal sequence of the L-HDAg is represented by a dashed line. Several protein domains are indicated above the graph. **B. Protein sequence conservation.** Intragroup (G1, G2 and G3) and intergroup (G1 vs G2, G1 vs G3 and G2 vs G3) protein sequence similarity were analysed and drawn as box plots. Colour for each group is similar to that used in Figure 2. Colour for intergroup combines the basic colours of both compared groups.

Figure 4. Compositional studies. A. Intragroup amino acid composition. Frequencies were calculated and drawn as box plots. Residues with a significative composition bias are indicated with a red oval in the corresponding graph. **B. Intra- and intergroup amino acid compositional distances.** Relationships were calculated as indicated in the Material and Methods section and represented as box plots. **C. Intra- and intergroup di-amino acid compositional distances.** Relationships were calculated as indicated in the Material and Methods section and represented as box plots. In **B** and **C** graphs, colour for each group are similar to those used in Figure 2, colour for intergroup combines the basic colours of both compared groups.

Figure 5. A. Invariant residue analyses. For each group, invariant residues of S-HDAg and L-HDAg shared sequences are indicated in red characters (before vertical dashed line) and invariant residues of C-terminal L-HDAg portion are indicated in orange characters (after vertical dashed line). Group pseudo-consensus sequences of invariant residues are piled up in an alignment. Deletion symbols (blue Greek character) were introduced to best represent the alignment, and fully conserved invariant residues are indicated by a green dot below the column. Regions proposed as molecular signatures are boxed. **B. Molecular signatures.** For each group, typical molecular signatures are shown following syntactic rules (prosite-like). Signature A corresponds to the shared portions of S and L proteins and comprise 11 residues for each group, whereas Signature B corresponds to the C-terminal region of L protein and comprises 18 residues for G1 and G2, and 19 residues for G3. **C. L-HDAg sequence patterns.** For each group, a specific L-HDAg sequence pattern is shown following the prosite-like syntactic rules.