Molecular Evolution of the Hepatitis Delta Virus Antigen Gene: Recombination or Positive Selection?

Maria Anisimova, Ziheng Yang

Molecular Evolution of the Hepatitis Delta Virus Antigen Gene: Recombination or Positive Selection?

Maria Anisimova,[1,2,3] Ziheng Yang[1]

[1] Department of Biology, University College London, London WC1E 6BT, UK
[2] Center for Mathematics and Physics in the Life Sciences and Experimental Biology, University College London, London WC1E 6BT, UK
[3] LIRMM, 161 Rue Ada 34392, Montpellier France

Received: 8 April 2004 / Accepted: 29 June 2004 [*Reviewing Editor:* Dr. Nicolas Galtier]

Abstract. We present the statistical analysis of diversifying selective pressures on the hepatitis D antigen gene (HDAg). Thirty-three distinct HDAg sequences from subtypes I, II, and III were tested for positive selection using maximum likelihood methods based on models of codon substitution that allow variable selective pressures across sites. Such methods have been shown to be sufficiently accurate and successful in detecting positive selection in a variety of viral and nonviral protein-coding genes. About 11% of codon sites in HDAg were estimated to be under diversifying selection. Remarkably, most of the residues predicted to evolve under positive selection were located in the immunogenic domain and the N-terminus region with reported antigenic activity. These sites are potential targets of the host's immune response. Identification of residues mutating to escape immune recognition may help to distinguish the most virulent strains and aid vaccine design. Possible interplay between positive selection and recombination on the gene is discussed but no significant evidence for recombination was found.

Key words: Positive selection — Recombination — HDV antigen gene — Maximum likelihood — Bayesian prediction

Correspondence to: Maria Anisimova; *email:* anisimov@lirmm.fr

Introduction

Viral genes are ideal for studying mechanisms of molecular adaptation. Fast replication and high mutation rate of viral genomes are crucial for evading the response of the host's immune system. To stay ahead of the co-evolutionary game, viral populations evolve faster than their hosts. For example, in RNA viruses mutation rates are 10^6–10^7 times higher than in *E. coli* and fungi (Li 1997). However, mutating randomly or too fast would compromise the functionality of the genome. Thus, mutation rate itself is subject to natural selection (Bonhoeffer and Sniegowski 2002). Only certain mutations can make a virus unrecognizable to its host immune system but safe for the virus itself. Detecting amino acids under diversifying selective pressure might help to identify potential targets of the immune response and/or the most virulent strains contributing to vaccine design. Here we focus on the hepatitis delta virus (HDV). With only one open reading frame (ORF) in a short genome, HDV provides one of the least complicated cases for studying viral adaptation.

HDV is a *subviral satellite* of the hepatitis B virus (HBV). There is no nucleotide homology between HDV and HBV genomes. Assembly of infectious HDV particles relies on HBV surface antigen (HBsAg). The HDV virion is comprised of a short genome (a single-stranded, circular 1.68-kb RNA) encapsulated within a sphere-like envelope containing the surface protein (HBsAg) of the natural helper virus. Three RNAs are found inside HDV-infected

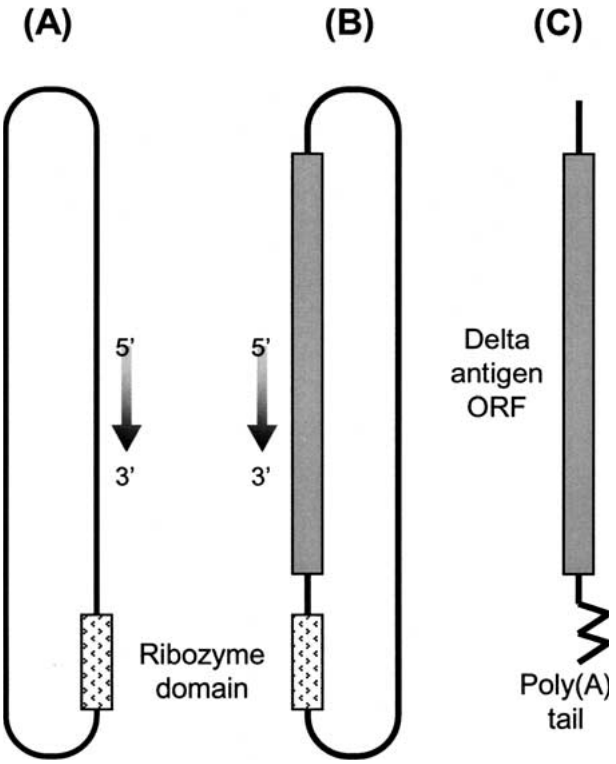


Fig. 1. Three RNAs detected in the HDV-infected cells: (**A**) genome, ~300.000 copies per cell; (**B**) antigenome, ~50,000 copies per cell; and (**C**) cytoplasmic, polyadenylated mRNA for the translation of the delta protein, ~600 per cell. Adapted from Taylor (1998).

cells (Fig. 1): (i) the HDV genome; (ii) an exact complement of the genome, known as the antigenome; and (iii) an RNA that acts as a messenger RNA for the translation of the only protein encoded by HDV—hepatitis delta antigen (HDAg). The two forms of HDAg, small (HDAg-S; 196 residues) and large (HDAg-L; 215 residues), both use the same ORF on the antigenomic strand of HDV. The small HDAg-S is synthesized early in infection and is essential for viral transcription and replication. During replication the termination codon UGA of HDAg mRNA mutates to a UGG tryptophan codon, extending the coding region by 19 amino acids. The resulting protein, HDAg-L, inhibits genome replication and assists the assembly of new viral particles.

HDV infection occurs either as a co-infection with HBV or as a superinfection of patients with chronic HBV infection. Co-infection with HDV and HBV causes more severe acute disease and higher risks of fulminant hepatitis and death compared with infection with HBV alone (e.g., Taylor 1998). However, the majority of co-infected patients completely recover. In contrast, patients with chronic HBV who become infected with HDV have not only a heavier form of chronic liver disease but also a high incidence of cirrhosis and hepatocellular carcinoma, making superinfection a very dangerous disease (e.g., Taylor 1998).

To date, three distinct genotypes of HDV have been identified based on sequence variation: Genotype I is distributed worldwide; genotype II is found mostly in Japan but also in Taiwan and Russia; and genotype III is typical to South America, where it is associated with the most severe form of HDV infection, characterized by a high mortality and a lesion in the liver called ''morula cell.'' The best protection against HDV is immunization against its natural helper virus, HBV. However, nonimmunized individuals and those already chronically infected with HBV have high risks of being infected by HDV. At the moment there is no effective therapy that directly targets HDV infection. Interferon-α treatments can be used to reduce assembly of HBV and HDV particles, but this effect is only temporary: Infection reappears once the treatment is stopped (Taylor 1998). There is currently no vaccine against HDV. Attempts to vaccinate woodchucks against HDV were not successful (e.g., Fiedler et al. 2001).

We use statistical techniques to examine the selective pressures on the HDAg gene and identify residues mutating to evade the host's immune response. The selective pressure on a protein is measured by the nonsynonymous/synonymous rate ratio ($\omega = d_N/d_S$), with $\omega < 1$, $\omega = 1$, or $\omega > 1$ indicating purifying selection, neutral evolution, or positive selection, respectively (e.g., Z. Yang and Bielawski 2000). The ω ratio on the HDAg is estimated by maximum likelihood (ML) as a parameter of an evolutionary model. The ML methods enable both testing for positive selection and identification of amino acid sites under diversifying selective pressure (Nielsen and Yang 1998; Z. Yang et al. 2000) (see Data and Methods for details). The ML models of codon evolution account for variable selection pressures across a gene by assuming a statistical distribution of the ω ratio among sites. The methods were tested by simulation (Anisimova et al. 2001, 2002, 2003) and appeared to be powerful in real data analysis. Viral genes in which positive selection has been detected using ML methods include capsid genes of foot-and-mouth virus (Fares et al. 2001; Haydon et al. 2001), the hemagglutinin gene of human influenza A (Z. Yang 2000; Z. Yang et al. 2000), the G and N genes of rabies virus (Holmes et al. 2002), and major HIV-1 genes (Nielsen and Yang 1998; W. Yang et al. 2003; Z. Yang 2001; Z. Yang et al. 2000; Zanotto et al. 1999).

While experimental studies of the HDV genome are numerous, statistical studies of HDV have been rare, with the most recent dating as far back as 5 years ago (Krushkal and Li 1995; Wu et al. 1999). Using pairwise comparisons (Li 1993), Krushkal and Li (1995) estimated the average nonsynonymous rate

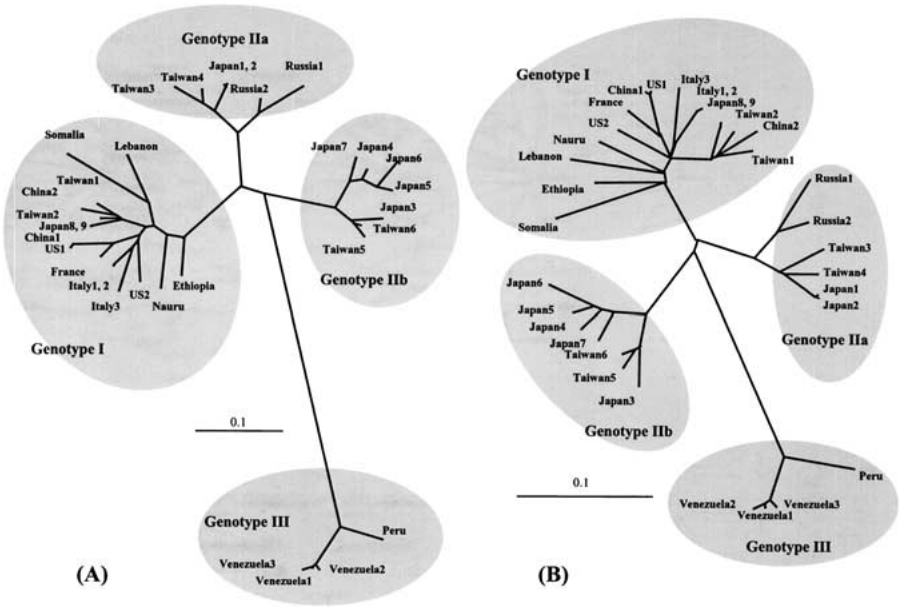


Fig. 2. Phylogenies of 33 HDAg-S strains used for testing data for positive selection: (**A**) the ML tree with ML-estimated branches; (**B**) the NJ tree with NJ-estimated branches.

on HDAg to be more than twice as high as the average synonymous rate. Later Wu et al. (1999) analyzed coding regions of HDV strains taken from two patients at different time points during the infection. They calculated the average pairwise nonsynonymous and synonymous rates (Comeron 1995) and suggested that HDAg might have been evolving under positive selection. The pairwise comparison methods do not offer proper significance testing; moreover, they assume a constant selective pressure along the sequence and cannot pinpoint locations of sites evolving by positive selection. Assuming a uniform selective pressure across sites also reduces the power of the method to detect positive selection in those genes where only a small fraction of sites are under diversifying selection, with the rest being functionally conserved.

In an attempt to find rigorous evidence of positive selection in HDV, we compiled a data set of 33 geographically dispersed hepatitis D antigene strains and examined it using likelihood ratio tests and Bayes inference. Since these approaches have been shown to be less accurate in recombinant data (Anisimova et al. 2003), we used several approaches to test for presence of recombination.

Data and Methods

HDV Data

We retrieved all available complete HDAg sequences representing all three genotypes (I, II, and III) and a variety of geographic areas. Most of the complete sequences were of genotype I, while only four sequences of genotype III were available. Multiple strains from the same patients (all from genotype I) were removed to reduce overrepresentation of genotype I. Thus, the presented study is cross-sectional, unlike the study of Wu et al. (1999), who analyzed a longitudinal sample. Thirty-three distinct strains of HDAg-S gene, 196 codons each, were finally used in the analysis: The GenBank accession numbers are AB015442, AB01543, AB015446, AB015447, AB037947–AB037949, AF018077, AF104263, AF104264, AF209859, AF309420, AJ309879, AJ309880, D01075, L22063, L22066, M28267, M58299, M58301, M58303, M58305, M58629, M84917, M92448, U19598, U25667, U81988, U81989, X04451, X63373, X77627, and X85253. The sequence alignment was first obtained by using DAMBE (Xia and Xie 2001) and then improved manually taking into consideration known genotypic differences (e.g., Casey and Gerin 1998). The alignment was submitted to the EMBL Nucleotide Sequence Database (accession number ALIGN_000712).

Phylogeny Reconstruction

A maximum likelihood (ML) tree was estimated using a new fast heuristic search algorithm implemented in PhyML (Guindon and Gascuel 2003) under the HKY85 + gamma model of nucleotide substitution (Hasegawa et al. 1985; Yang 1994). The transition/transversion rate ratio κ and the shape parameter of the gamma distribution α were estimated as free parameters. Three distance-based trees were inferred assuming HKY85 with $\kappa = 3$ (a rounded ML estimate): a neighbor-joining (NJ) tree using standard algorithm implemented in PAUP*4.0 (Swofford 2000), a BioNJ tree using an improved NJ algorithm implemented in BIONJ (Gascuel 1997), and a minimum evolution tree using a fast algorithm implemented in FastME (Desper and Gascuel 2002). All trees contained four phylogenetic clades: genotype I, subtypes A and B of genotype II, and genotype III (e.g., Fig. 2). We observed only minor discrepancies among the inferred phylogenies, all concerning closely related lineages. Subsequently, only the ML tree (the most likely) was used in the analysis (Fig. 2A); the NJ tree (the least likely of four inferred) was used to investigate how small discrepancies in the assumed topology affect results of LRTs for positive selection.

Testing for Recombination

The data were tested for recombination within the HDAg gene. The scaled recombination rate, $\rho = 2Nr$, was estimated using the composite likelihood method, where N is the effective population size and r is the recombination rate per nucleotide site per generation. First proposed by Hudson (Hudson 2001) and recently extended (McVean et al. 2002) to allow for finite-site mutation models, the composite likelihood method is based on combining the coalescent likelihoods of all pairwise comparisons of segregating sites. As it does not properly account for the nonindependence in the multiple pairwise comparisons, this method does not produce reliable confidence intervals for ρ. However, it is a fast, flexible, and well-performing alternative to the full likelihood (e.g., Fearnhead and Donnelly 2001; Kuhner et al. 2000), which is computationally intractable for large data sets. Note that the composite likelihood method assumes complete neutrality of the locus. The hypothesis of no recombination was tested using the likelihood permutation test (LPT) as in McVean et al. (2002). We also use two other permutation tests, which detect a decrease in r^2 and $|D'|$, measures of linkage disequilibrium, with an increase in the physical distance (e.g., Awadalla et al. 1999). Both the composite likelihood analysis and the three permutation tests were carried out using the LDhat package (McVean et al. 2002).

The programs PLATO and PIST were also used to test for the presence of recombination. PLATO employs a sliding window analysis to detect regions in the sequence alignment that cannot be described by the evolutionary model of the entire data set (Grassly and Holmes 1997). While usually used for recombination detection, PLATO can detect any other significant deviations from the overall evolutionary patterns (e.g., differences in selection pressure), as pointed out by Grassly and Holmes (1997). Finally, PIST (the informative-sites test) is based on the observations that (i) recombination inflates apparent rate heterogeneity among sites and (ii) the composition of polymorphic sites differs for nonrecombinant data with genuine rate heterogeneity and data with an equivalent artificial rate heterogeneity generated by recombination. PIST compares whether the proportion of two-state parsimony informative sites within all polymorphic sites is higher than expected for nonrecombinant data (Worobey 2001).

Testing for Positive Selection

Codon substitutions along a tree can be modeled by a Markov process that takes into account codon frequencies, transition/transversion bias, the nonsynonymous/synonymous rate ratio (ω), and branch lengths (Goldman and Yang 1994). A likelihood ratio test (LRT) for positive selection (Nielsen and Yang 1998; Z. Yang et al. 2000) compares two codon substitution models, one of which accounts for positive selection and the other of which does not. For each model, parameters and likelihood scores are estimated by ML, and twice the log-likelihood difference is compared with the χ^2 distribution with the degree of freedom equal to the difference in the number of free parameters in those models (Anisimova et al. 2001). The gene is inferred to be under positive selection if (i) ML estimates suggest that there are sites under positive selection (with $\omega > 1$), and (ii) the LRT is significant. Six models of codon substitution (Z. Yang et al. 2000) were used in the analysis: M0 (one ratio), M1 (neutral), M2 (selection), M3 (discrete), M7 (β), and M8 (β and ω). M0 (one ratio) assumes one ω ratio for all sites. Ml (neutral) has two site classes: conserved sites with $\omega_0 = 0$ and neutrally evolving sites with $\omega_1 = 1$, in proportions p_0 and p_1, respectively. Model M7 (β) assumes that the ω ratio follows the beta distribution with parameters p and q, which is approximated by 10 site classes with ω ratios between 0 and 1. Models M0 (one ratio), M1 (neutral), and M7 (β) are used as null hypotheses for comparison against M3 (discrete), M2 (selection), and M8 (β and ω), respectively. Model M3 (discrete) allows three discrete site classes with ratios ω_0, ω_1, and ω_2 taken in proportions p_0, p_1, and $p_2 = 1 - p_0 - p_1$. Model M2 (selection) adds to M1 (neutral) an extra class with the ratio ω_2 in the proportion p_2. Likewise, model M8 (β and ω) is an extension of M7 (β): It adds to M7 an extra discrete ω class to account for possible sites under positive selection. Thus, M8 has two more free parameters: the proportion p_0 of sites from the beta distribution and the ω ratio for the discrete site class in the proportion $p_1 = 1 - p_0$.

LRTs for positive selection (M0 vs. M3, M1 vs. M2, and M7 vs. M8) were performed for both ML and NJ trees, and the results for the two trees were compared. Posterior probabilities of each site falling into distinct site classes are calculated using the Bayesian approach (Nielsen and Yang 1998; Z. Yang et al. 2000). Sites with high posterior probabilities of coming from a class with $\omega > 1$ are likely to be under positive selection. The ML analysis of the codon data was conducted using the codem1 program in the PAML package (Yang 1997).

Table 1. Correlation of r^2 and $|D'|$ statistics with physical distance, d, and results of recombination tests carried out using the LDhat package

Data	$\hat{\rho}$	Corr(r^2, d)	Corr($\|D'\|$, d)	P_{LPT}	p_{r^2}	$P_{\|D'\|}$
All codon positions	10.1	3.1×10^{-5}	-2.5×10^{-4}	0.205	0.656	0.103
3rd codon positions	11.1	5.2×10^{-4}	6.1×10^{-4}	0.700	0.920	0.151
3rd codon positions after removing rare alleles with frequency < 5%	11.1	1.1×10^{-3}	2.0×10^{-3}	0.797	0.918	0.171

Note. $\hat{\rho}$ is the composite likelihood estimate of the scaled recombination rate per base; P_{LPT}, p_{r^2}, and $P_{|D'|}$ are P values of permutation tests based on composite likelihood, r^2, and $|D'|$ statistics, respectively.

Results

Test for Recombination in HDAg

Tests for recombination were performed using the full data set including all codon positions and using third codon positions only. To minimize the effect of selection on recombination detection methods, it is more appropriate to use only the third codon positions since the first and second codon positions are heavily influenced by selection. The composite likelihood estimates of scaled recombination rate ρ were around 10 for both the full data and the third codon positions. However, the method does not provide a confidence interval for the estimate, and so its significance is unclear. All three permutation tests implemented in LDhat suggested that ρ was not significantly different from 0 in either the full data set or the third codon positions ($P > 0.05$; Table 1).

Table 2. Detecting putative recombination regions with PLATO

Data	Tree	Gamma shape, α	Gamma categories	κ	Likelihood score, *l*	Anomalous regions	*Z* value[a]
3rd codon positions	**ML**	**0.716**	**4**	**3.0**	**−2146.44**	**5–48**	**3.78**
	ML	**0.716**	**10**	**3.0**	**−2144.82**	**5–48**	**4.19**
	ML	0.5[b]	4	3.0	−2149.95	None	n/a
	ML	0.716	4	4 7[b]	−2157.41	44–48	3.85
	ML	∞[b]	4	3.0	−2238.52	44–48	13.14
						8–17	9.11
						113–120	5.55
						165–175	3.56
	NJ[b]	0.716	4	3.0	−2336.50	44–48	5.57
						8–17	4.01
All codon positions	**ML**	**0.413**	**4**	**2.2**	**−5089.06**	**∼5–17**	**4.57**
	ML	**0.413**	**10**	**2.2**	**−5083.75**	**∼5–17**	**4.89**
	ML	0.413	4	4.5[b]	−5140.33	∼3–29	6.00
	ML	0.8[b]	4	2.2	−5120.01	∼5–24	7.15
						∼180–181	4.28
						∼37–38	3.93
	NJ[b]	0.413	4	2.2	−5166.28	∼5–24	4.45

Note. All analyses were performed under the HKY85 + gamma model. Rows corresponding to the best-fitting parameters are in boldface. Locations of anomalous regions refer to amino acids.

[a]Bonferroni-corrected *Z* values significant at 5%: > 3.27 for third codon positions and > 3.58 for all data.

[b]Parameters deviating from the best-fitting estimates.

McVean et al. (2002) suggested that rare variants are not informative about recombination but can distort the recombination signal, especially when there is an excess of rare mutations. For the HDAg gene, removing rare alleles (with frequency < 5%) from the third codon positions made no significant difference (Table 1). The average correlation coefficient between linkage disequilibrium statistics and distance was always very close to 0 (Table 1).

Next, the PLATO program was used to identify anomalous regions under the HKY85 + gamma model (Table 2). ML estimates of the gamma shape parameter α and the transition/transversion ratio were $\hat{\alpha} = 0.716$ and $\hat{\kappa} = 3.0$ for third codon positions and $\hat{\alpha} = 0.412$ and $\hat{\kappa} = 2.2$ for all codon positions. ML estimates of nucleotide frequencies were $f_A = 0.2618$, $f_C = 0.2790$, $f_G = 0.3503$, and $f_T = 0.1089$ for the third codon positions and $f_A = 0.3161$, $f_C = 0.2446$, $f_G = 0.3236$, and $f_T = 0.1157$ for all codon positions. For the third codon positions, PLATO detected significant deviations from the overall underlying model in the region between amino acid 5 and amino acid 48 (rows 1 and 2 in Table 2). Altering model assumptions about rate heterogeneity, transition/transversion bias, and phylogenetic relationships caused noticeable variations in PLATO results. For example, when α = 0.5 was assumed, no anomalous regions were detected, whereas assuming rate constancy over sites caused PLATO to infer four potential hot spots. Similar discrepancies were observed in PLATO results when all data were analyzed (Table 2).

Table 3. Results of recombination tests carried out with PIST: Third codon positions

Tree	α	Gamma categories	κ	*P* value
ML	**0.716**	**4**	**3.0**	**< 0.07**
ML	**0.716**	**10**	**3.0**	**< 0.10**
ML	1[a]	4	3.0	< 0.11
ML	0.6[a]	4	3.0	< 0.05*
ML	0.716	4	2.0[a]	< 0.01**
ML	0.716	4	4.0[a]	< 0.32
ML	∞[a]	4	3.0	< 0.58
NJ[a]	0.716	4	3.0	< 0.08

Note. All analyses were performed under the HKY85 + gamma model. Rows corresponding to the best-fitting parameters are in boldface.

*Tests significant at 5% level.

**Tests significant at 1% level.

[a]Parameters deviating from the best-fitting estimates.

Finally, PIST program was used to test for recombination, under the HKY85 + gamma model. No significant evidence for recombination was detected at the third codon positions when the ML parameter estimates and the ML tree were assumed (rows 1 and 2 in Table 3). Increasing the number of discrete categories to approximate gamma distribution from 4 to 10 improved the fit of the model and resulted in higher *P* values, making the presence of recombination even more unlikely (row 2 in Table 3). To explore the effect of model misspecification on the performance of PIST, data were also analyzed using (i) the NJ tree or (ii) different values for α and κ.

Table 4. ML estimates, results of LRTs, and amino acids predicted to be under positive selection in the HDAg gene based on the ML tree

Null model	l_0	Alternative model	l_1	Positively selected sites
M0 (one ratio)		**M3 (discrete)**		
$\omega = 0.375$	−5541.89	$\omega_0 = 0.05$, $\omega_1 = 0.62$, $\boldsymbol{\omega_2 = 2.40}$ $p_0 = 0.56$, $p_1 = 0.30$, $\boldsymbol{p_2 = 0.14}$	−5248.17	**4S, 6S*, 9N*, 13gap, 17I*, 23 N, 24G, 28I, 38I*, 117A, 122H*, 140R, 150V***, *159G*, *173L*, **181S***, *189 V*, **192N**
M1 (neutral)		**M2 (selection)**		
$\omega_0 = 0$, $\omega_1 = 1$ $p_0 = 0.38$, $p_1 = 0.62$	−5408.53	$\omega_0 = 0$, $\omega_1 = 1$, $\boldsymbol{\omega_2 = 4.22}$ $p_0 = 0.37$, $p_1 = 0.51$, $\boldsymbol{p_2 = 0.12}$	−5338.85	**4S**, **6S***, **9N***, **13gap**, **17I***, **23 N**, *24G*, **28L**, **38I***, **117A**, **122H***, *140R*, **150V***, **181S***, **192N**
M7 (β)		**M8 (β & ω)**		
$p = 0.23$, $q = 0.41$	−5385.25	$p_0 = 0.89$, $p = 0.30$, $q = 0.79$ $\boldsymbol{p_1 = 0.11}$, $\boldsymbol{\omega = 2.58}$	−5245.88	*4S*, **6S***, **9N***, **13gap**, **17I***, *23 N*, **38I***, **117A**, **122H***, **150V***, **181S***

Note. All LRTs are significant at $p << 0.001$. ML estimates showing the presence of positive selection are in boldface. Sites inferred to be under positives election with probabilities >0.99 are indicated by an asterisk; those with probabilities >0.95 are in boldface, and those with probabilities >0.9 are in italics.

[a]ML estimates and sites inferred under model M2 are shown for the suboptimal peak. The same estimates and sites were obtained at the only optimum when M2 was restricted to have $\omega_2 \geq 1$.

Small changes to parameters α and κ caused considerable fluctuations in P values, with some tests becoming significant (e.g., rows 3–6 in Table 3). Alterations of two or more parameters resulted in a whole spectrum of P scores, from <0.001 to 0.95 (results not shown). When all codon positions were analyzed, most of the analyses gave a significant signal for recombination, suggesting that variable selective pressure and variable rates among sites might have contributed to the increased heterogeneity signal (results not shown).

In sum, neither the permutation tests nor the informative sites test (PIST) provided a significant evidence for recombination in HDAg. PLATO detected significant deviations from the overall underlying model in the N-terminus region of the HDAg, which could have been caused by either recombination or deviation in other model assumptions. We found that both PLATO and PIST are sensitive to misspecification of the evolutionary model.

Positive Selection in HDAg

Results of LRTs comparing models M0 (one ratio) vs. M3 (discrete) and M7 (β) vs. M8 (β and ω) for the ML tree are presented in Table 4. ML parameter estimates were found to be almost identical for the ML and NJ trees, with a slightly stronger signal of positive selection for the ML tree. Moreover, the ML estimates under models M3 and M8 were also in agreement. When the ML tree was assumed, model M3 suggested that about 14% of sites were under positive selection with $\omega_2 = 2.4$, while model M8 suggested about 11% of sites with $\omega = 2.58$. All LRTs were highly significant, suggesting that HDAg-S gene evolves under positive selection (Table 4). Table 4 also lists sites under positive selection predicted by the Bayes method under M3 or M8 with the posterior $P \geq 0.90$ for the ML tree. For both trees model M3 predicted approximately six or seven more sites than predicted by M8. Furthermore, when different trees were assumed, the predicted sites under positive selection were almost identical (e.g., see Fig. 3).

Model M2 (selection) was also used in the analysis. This model assumes three site classes. The first two have $\omega_0 = 0$ and $\omega_1 = 1$ fixed and account for completely constrained sites and completely neutral sites, while a third class has ω_2 estimated from the data. Two local peaks were found under M2, with log-likelihood scores $l = -5277.48$ (optimal) and $l = -5338.85$ (suboptimal) (Fig. 4). Both scores showed a significant improvement over M1 (neutral). At the optimal peak, almost half of the sites ($p_2 = 0.42$) were estimated to be under purifying selection with $\omega_2 = 0.16$, and no sites were detected to be under positive selection. Estimates at the suboptimal peak suggested the presence of positive selection with $\omega_2 = 4.22$ in proportion $p_2 = 0.12$ of sites. The results suggest the presence of both sites under purifying selection and sites under positive selection, and model M2, with only one free ω_2 to vary, can accommodate only one of the two types. Strikingly, positively selected sites detected under M2 using the suboptimal peak fully agreed with those detected under M8, with the posterior probabilities being slightly higher under M2. We thus conclude that M2 also suggests the presence of sites under positive selection. Indeed, if the selection model M2 insists on the presence of sites under positive selection with $1 \leq \omega_2 < \infty$, the local peak with $\omega_2 = 4.22$ will become the only peak for this modified model M2a, and the results will be consistent with those from model M8.

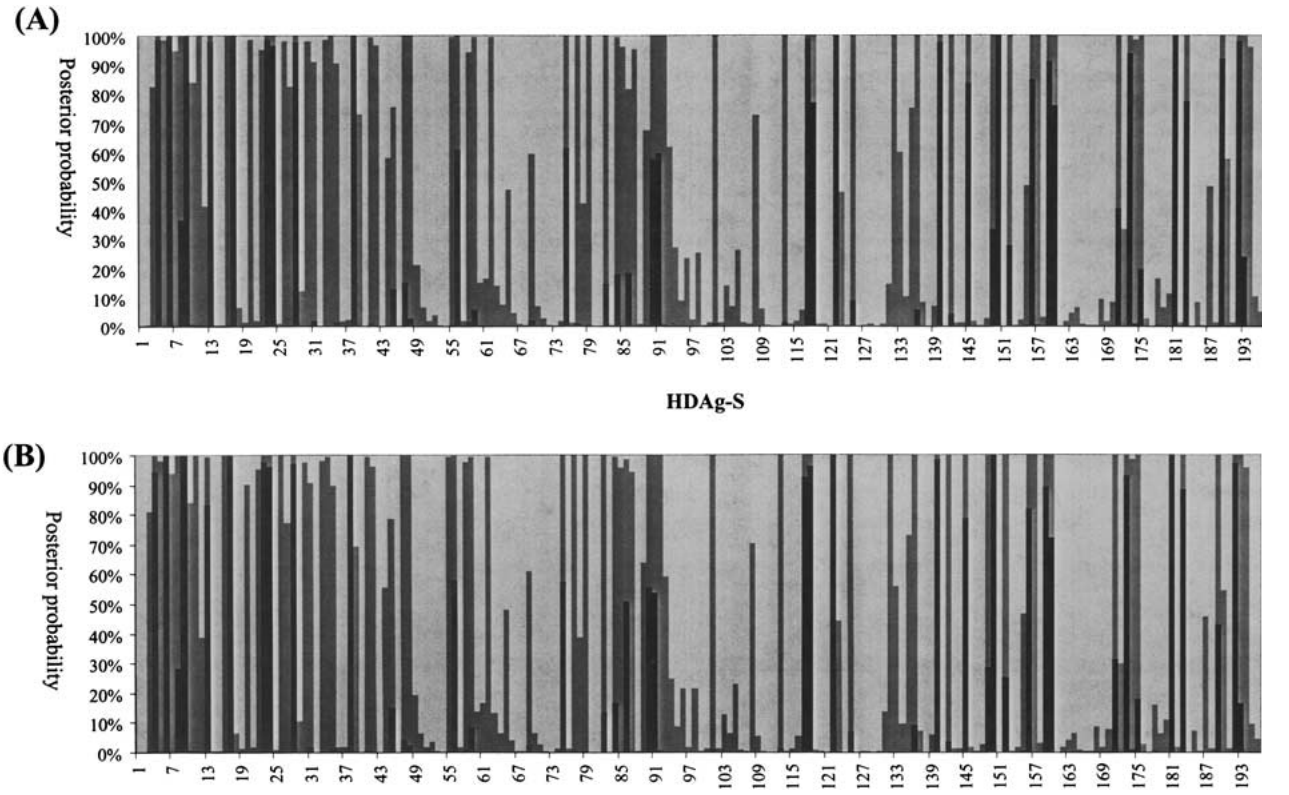
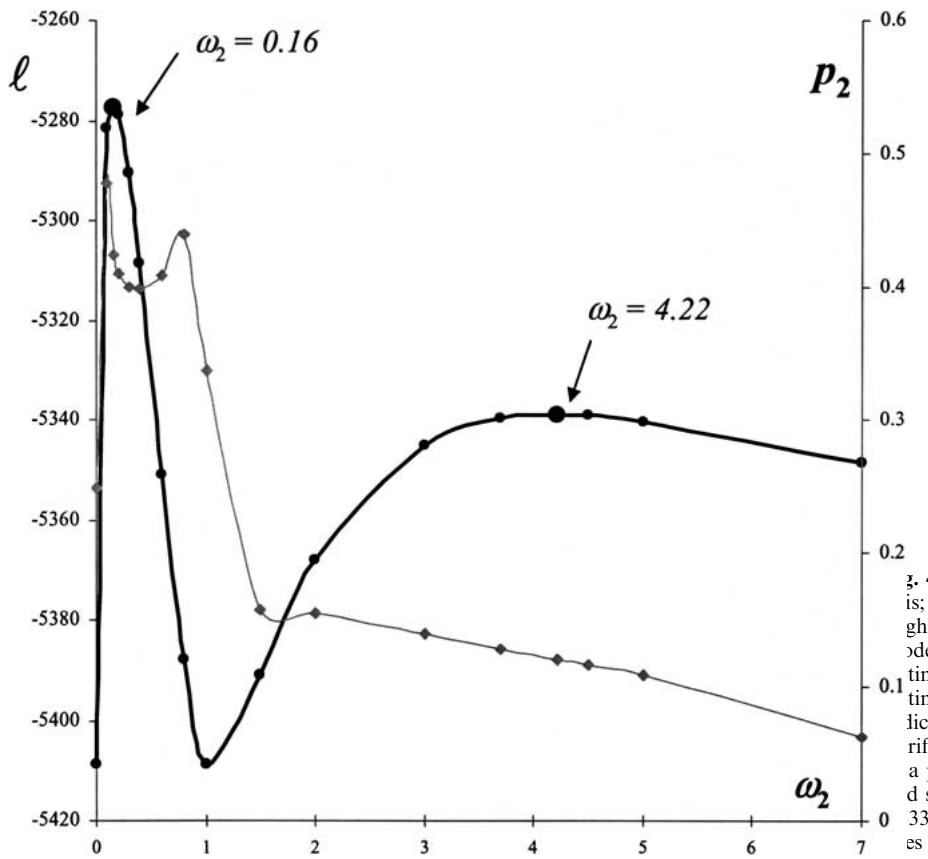


Fig. 3. Posterior probabilities for sites in HDAg-S to evolve under positive (black) or purifying (gray) selection or under relaxed functional constraints (light gray) in discrete model M3: (**A**) assuming the ML tree; (**B**) assuming the NJ tree.



g. 4. Log-likelihood surface (left is; black line) and parameter p_2 ght axis; gray line) under selection odel M2 illustrating two local tima along the parameter ω_2: timal peak with $l = -5277.48$ licating the presence of sites under rifying selection with $\omega_2 = 0.16$ a proportion $p_2 = 0.42$ of sites; d suboptimal peak with $l =$ 338.85, suggesting the presence of es under positive selection with $\omega_2 = 4.22$ in a proportion $p_2 = 0.12$ of sites.

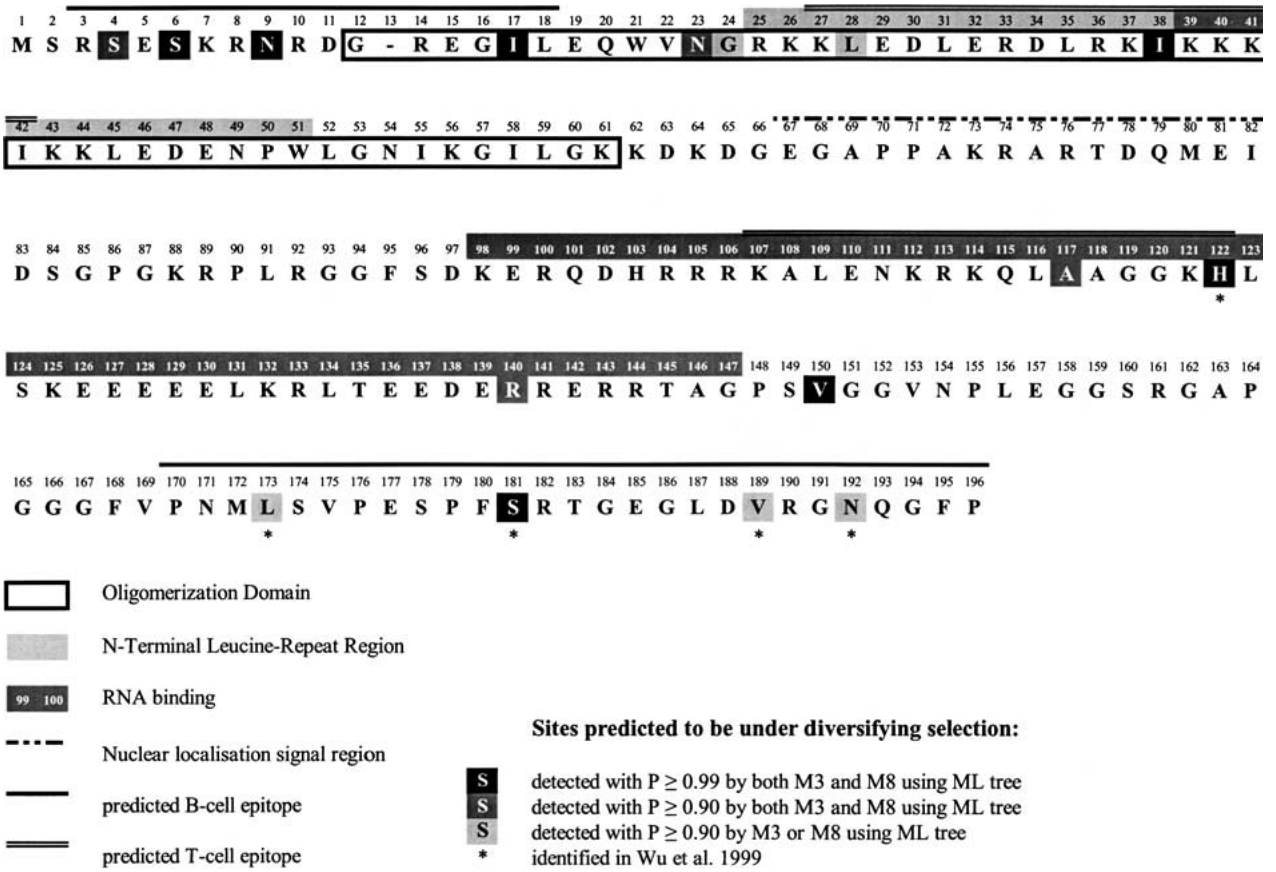


Fig. 5. Positively selected sites shown against domains of the HDAg gene: oligomerization domain (Rozzelle et al. 1995); leucine repeat region (Cheng et al. 1998); RNA-binding domain (Lai et al. 1993); predicted B- and T-cell epitopes (Nisini et al. 1997; Wang et al. 1990). The GenBank accession number of the reference sequence is AF104263.

Finally, we verified our results using the LRT suggested by Swanson et al. (2003), where the null model is M8 with a constraint $\omega = 1$ and the alternative is M8 with $\omega \geq 1$. This test was also significant; moreover, the analysis under alternative model converged on an identical likelihood peak, resulting in the prediction of the same sites with posterior probabilities identical to those calculated under the standard M8 model.

Figure 5 illustrates the locations of predicted sites under positive selection in relation to functional domains identified in experimental studies (the complete crystal structure of HDAg protein is not available). The distribution of predicted sites along the sequence is clearly not random: half of them are located in the N-terminus and the others are in the C-terminus. The middle region of the gene (residues 39 to 116) appears to be mostly conserved, presumably due to functional constraints (Fig. 3). Most of the sites were found in the putative B- and T-cell epitopes. This pattern supports the finding of Wu et al. (1999), who used visual inspection of the alignment to suggest that some amino acid residues in HDAg within predicted T- and B-cell epitopes might be under selective pressure to evade immune response (Wang et al. 1990). Remarkably, almost all sites that exhibited an accelerated rate of amino acid substitutions in the longitudinal study by Wu et al. (1999) were predicted to be under positive selection in this cross-sectional study. This study identified 11 more sites under positive selection than was suggested by Wu et al. (1999).

Discussion

Evidence of Recombination in the HDAg-S Gene

Wu et al. (1999) reported a number of putative recombination regions identified by PLATO in a longitudinal sample of HDAg strains. However, the authors used all three codon positions, and therefore the performance of PLATO was compromised by strong selection acting on the first and second positions. Our analysis of all codon positions using PIST and PLATO supports this suggestion (see Results). Moreover, all identified by Wu et al. (1999) putative recombination events in HDAg are single-amino acid changes (see Fig. 4 of Wu et al. 1999), which could have been caused by other evolutionary forces. Our results also clearly demonstrate that misspecification

of the evolutionary model (e.g., failure to account for rate heterogeneity among sites) can radically affect the performance of PLATO (Table 2).

In this (cross-sectional) study, no convincing evidence for recombination in the HDAg gene was found. None of the permutation tests produced significant results for either the third codon positions or the full data. According to simulations (McVean et al. 2002), for the HDAg gene with Watterson estimate of the scaled mutation rate $\hat{\theta}_W = 0.174$, the power of the LPT to detect recombination should be sufficiently high (about 95%). The power of other permutation tests is lower on average by 10–20% (McVean et al. 2002). While PLATO suggested the presence of an anomalous region between residue 5 and residue 48, without a more thorough analysis of the detected regions one cannot conclude whether it is a result of recombination or other underlying evolutionary processes (Grassly and Holmes 1997). Significant deviations from the overall underlying model in certain regions may reflect not recombination but variable selective forces in a protein or variation in other parameters of the evolutionary model. Visual examination of the alignment did not reveal obvious recombined segments but rather single-amino acid mutations, as in Wu et al. (1999). It is worth noting that the detected anomalous region coincides with the N-terminal leucine repeat region with reported antigenic activity (Cheng et al. 1998). When PIST was used to analyze third codon positions it could not reject the null hypothesis of no recombination.

We conclude that recombination, if present in the sample of 33 HDAg-S strains analyzed here, is unlikely to be frequent as to affect the inference of positive selection, judging by the recent simulation study (Anisimova et al. 2003).

Diversifying Selection in HDAg

The LRTs comparing models M0 vs. M3 and M7 vs. M8 provided significant evidence for positive selection acting on the HDAg-S gene. Note that the methods used here compare synonymous and nonsynonymous rates, and can only detect recurrent diversifying positive selection affecting particular sites. Episodic or directional positive selection usually does not significantly elevate the nonsynonymous rate and will not be detected by such methods.

Eighteen codon sites were inferred to be under diversifying positive selection by the Bayesian method. Eleven of those sites were inferred with posterior probabilities $>90\%$ by both models M3 and M8, with seven "most likely" sites and four "very likely" sites (Fig. 5). The remaining 7 of the 18 sites were inferred with a probability of $>90\%$ by at least one of models M3 and M8 and are "possibly" under positive selection. To verify whether such criteria are stringent enough we referred to simulation results (Anisimova et al. 2002). In the HDAg-S data set there is $\sim$0.10 nucleotide substitution per codon per branch or $S \approx 6.57$ changes along the tree. This is equivalent to $d_N = 1.8$ nonsynonymous changes per nonsynonymous site and $d_s = 3.4$ synonymous changes per synonymous site along the tree, and represents a near-optimum level of sequence divergence. In simulations, similar levels of sequence divergence produced very reliable results with sites predicted under M3 at a posterior $P > 0.6$ being at least 90% likely to be truly under positive selection. Model M8 significantly outperforms M3 with posterior probabilities being even more conservative (Anisimova et al. 2002). Under model M2, two local peaks exist in the likelihood surface. Our analysis of the model suggests that the presence of local peaks is due to the design of the model, and the estimates at the suboptimal peak are biologically sensible and in full agreement with models M3 and M8. M2 also suggests the presence of sites under positive selection. In computer simulations, Bayes prediction under model M2 was shown to be very accurate (Anisimova et al. 2002).

Most sites found to be under positive selection in this study were in the N-terminus, where antigenic activity was previously reported (Rozzelle et al. 1995), and in immunogenic domains of the C-terminus (Nisini et al. 1997; Wang et al. 1990). Experimental studies suggest that HDAg-L antigenic domains include 41% of the HDAg molecule (Wang et al. 1990). Surprisingly, some of experimentally predicted epitopes are within functionally conserved regions. For example, Rozzelle et al. (1995) reported considerable antigenicity within the N-terminal oligomerization domain that formed an antiparallel coiled-coil (four-helical bundle). Coiled-coils are responsible for the function of many RNA-binding proteins and are required for the entry of influenza and HIV viral particles into their target cells (Oakley and Hollenbeck 2001). Likewise, Nisini et al. (1997) found epitopes within functional domains such as the coiled-coil (aa 27–42), nuclear localization signal region (aa 67–82), and RNA-binding domain (aa 107–122). In this study, residue 140R was found to be under positive selection, however, it also belongs to the arginine-rich motif responsible for RNA-binding activity (Lee et al. 1993). Further, five residues under positive selection (sites 17I, 23N, 24G, 28L, 38I; Fig. 5) were also found in the oligomerization domain (Fig. 6). Most of them correspond to the *a* and *d* positions of the heptad repeat. Zuccola et al. (1998) point out that dimers of HDAg are stabilized by hydrophobic interactions other than those of the heptad repeat. Residues 28 and 38 seem to be interacting when monomers dimerize (Fig. 6). We find no noticeable correlation between the amino acids at

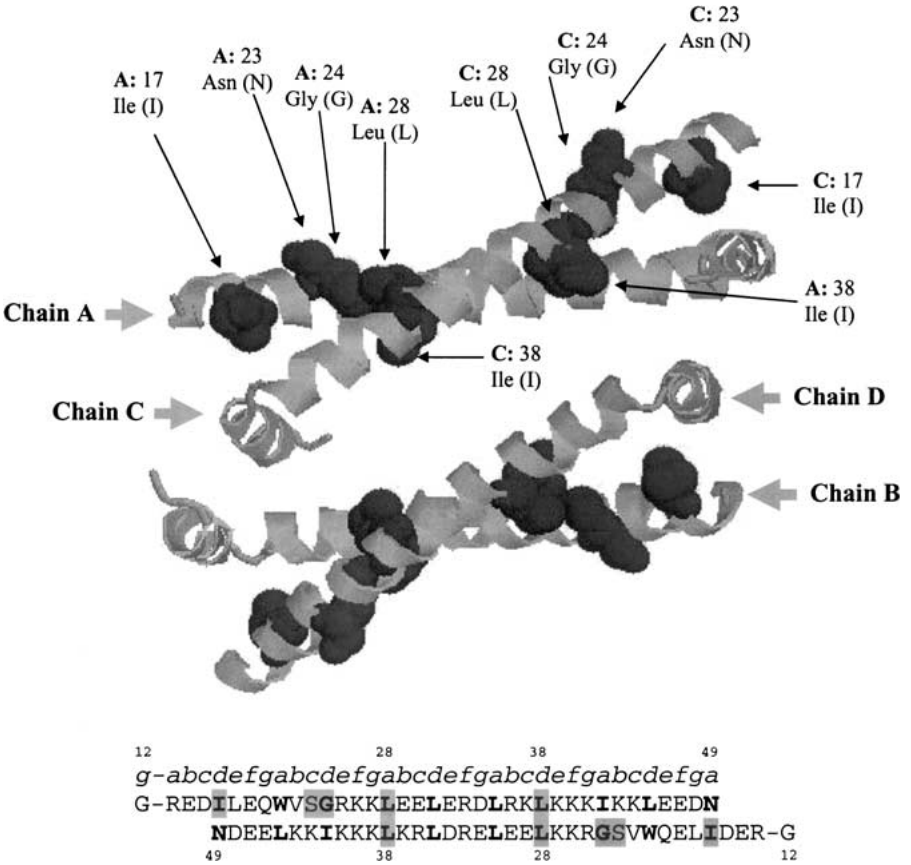


Fig. 6. Top: The crystal structure of the oligomerization domain of HDAg (Zuccola et al. 1998; PDB accession No. 1A92). Sites inferred to be under positive selection are shown in black. Letters in boldface refer to chain A, B, C, or D. The oligomerization domain of HDAg contains a leucine-zipper-like region that allows two monomers to wrap around each other in an antiparallel coiled-coil, forming dimers, tetramers (as in the picture), or higher-ordered structures. Each monomer is a long N-terminal helix interrupted by a sharp bend at site 50Pro Bottom: The amino acid sequence (PDB No. 1A92) of the long helix formed from residues 12–49 (including the gap at site 13 to accommodate genotype IIa) displayed in the antiparallel orientation. The letters above the sequence (*abcdefg*) represent the heptad repeat, where residues *a* and *d* tend to be hydrophobic (shown in boldface). Sites inferred to be under positive selection are highlighted gray.

these positions. In genotypes I and IIa most of the residues at position 28 are hydrophobic, whereas in genotypes IIb and III all residues at site 28 are hydrophilic. In all genotypes, residues found at position 38 were neutral, and most of them were hydrophobic, with rare exceptions. While the majority of leucine zippers have Val and Leu in positions *a* and *d*, respectively, deviations from these amino acids lead to variations in structure (e.g., Harbury et al. 1993). Radical amino acid changes in positions *a* and *d* throughout genotypes might reflect variations in the structure of the oligomerization domain and, as a consequence, affect the fitness of a strain or a genotype. The effect of positive selection on residues in the coiled-coil formation merits further investigation.

The presence of some sites under diversifying selective pressure in functionally conserved regions might not be so surprising. Due to structural constraints, conserved domains evolve more slowly, becoming easier targets for the host immune system, as there is less "room" for the virus to evolve escape mutations. This can lead to strong selection pressure on the virus for changes within the conserved regions. Studies in the HIV-1 *env* gene (Ross and Rodrigo 2002; Yamaguchi-Kabata and Gojobori 2000) reported similar findings: Some sites responsible for CD4 and CCR5 binding within the conserved regions C2–C4 were suggested to be under diversifying selective pressure. This has been supported by experimental studies (Thali et al. 1992).

Positive Selection or Recombination?

It is interesting that the anomalous region detected by PLATO contains positive selection sites inferred within the N-terminus. Since the oligomerization domain is functionally conserved, in this region recombination involving segments of sequences might be more unlikely, as it would be more damaging than diversifying pressure on a handful of isolated sites. Recombination tests could not reject the hypothesis of

no recombination. Moreover, the ML phytogeny reconstructed for the nucleotide sequences in the region between codon 5 and codon 48 did not support any major topological incongruencies with the ML or NJ trees (results not shown). Positive selection, and not recombination, thus appears to be responsible for diversification of sites within the N-terminus. It is the conglomeration of sites with an accelerated nonsynonymous rate within the N-terminus that was detected by PLATO: Naturally, in this region positive selection caused significant deviations from the overall evolutionary model that did not take variable selection pressure into account. In general, definite conclusions about the presence of recombination cannot be drawn on the basis of a single method (Posada and Crandall 2001). In the same way that the LRTs for positive selection can produce false positives for recombinant data, methods for detecting recombination can confuse recombination and heterogeneous selective pressure among sites or other nonhomogeneous evolution among sites or across lineages (Paraskevis et al. 2003; Worobey et al. 2002; unpublished simulations of D. Paraskevis, personal communication).

The HDAg appears to employ an extremely compact but very flexible mechanism whereby even in functionally important domains certain sites have accumulated an excess of nonsynonymous mutations, presumably to avoid antibody recognition without harming the important functions of the protein. This mechanism seems to be very efficient considering the severity of the disease, its frequent progression to chronic liver infection, and the absence of satisfactory therapies. Larger and more equally distributed samples representing all genotypes (especially genotype 3) are required for a better understanding of HDV evolutionary dynamics and for uncovering the molecular differences causing genotype III to have such a high fatality rate. Toward this goal it might be informative to compare the strength of selection acting on each genotype individually.

It is also interesting to investigate the protein interactions during double infections by HBV and HDV. The assembly mechanism of HBV is very inefficient: About 99% of the assembled particles do not contain HBV DNA and so are nonvirulent. One hypothesis is that HDV evolved to make use of otherwise futile HBV particles (Taylor 1998). It is unclear whether HDV infection makes the HBV more efficient or whether most virulent strains of HBV are more amenable to HDV replication. This could be investigated by comparing evolutionary patterns in samples of HBV strains from patients infected with only HBV and patients co- or superinfected with HDV.

Acknowledgments. We thank Joe Bielawski and Wa Yang for discussions and two anonymous referees for comments on the manuscript. This work was supported by a Medical Research Council (U.K.) studentship to M.A. and grants from the Biotechnology and Biological Sciences Research Council (U.K.) and the Human Frontier Sciences Programme (EU) to Z.Y. At the final stage of this project M.A. was supported by French ACIs and IMPBIO.

References

Anisimova MJ, Bielawski JP, Yang Z (2001) Accuracy and power of the likelihood ratio test to detect adaptive molecular evolution. Mol Biol Evol 18:1585–1592

Anisimova M, Bielawski JP, Yang Z (2002) Accuracy and power of Bayes prediction of amino acid sites under positive selection. Mol Biol Evol 19:950–958

Anisimova M, Nielsen R, Yang Z (2003) Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics 164:1229–1236

Awadalla P, Eyre-Walker A, Smith JM (1999) Linkage disequilibrium and recombination in hominid mitochondrial DNA. Science 286:2524–2525

Bonhoeffer S, Sniegowski P (2002) Virus evolution: The importance of being erroneous. Nature 420:367–369

Casey JL, Gerin JL (1998) Genotype-specific complementation of hepatitis delta virus RNA replication by hepatitis delta antigen. J Virol 72:2806–2814

Cheng JW, Lin IJ, Lou YC, Pai MT, Wu HN (1998) Local helix content and RNA-binding activity of the N-terminal leucine-repeat region of hepatitis delta antigen. J Biomol NMR 12:183–188

Comeron JM (1995) A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. J Mol Evol, 41:1152–1159

Desper R, Gascuel O (2002) Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. J Comput Biol 9:687–705

Fares MA, Moya A, Escarmis C, Baranowski E, Domingo E, Barrio E (2001) Evidence for positive selection in the capsid protein-coding region of the foot-and-mouth disease virus (FMDV) subjected to experimental passage regimens. Mol Biol Evol 18:10–21

Fearnhead P, Donnell P (2001) Estimating recombination rates from population genetic data. Genetics 159:1299–1318

Fiedler M, Lu M, Siegel F, Whipple J, Roggendorf M (2001) Immunization of woodchucks (*Marmota monax*) with hepatitis delta virus DNA vaccine. Vaccine 19:4618–4626

Gascuel O (1997) BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol 14:685–695

Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11:725–736

Grassly NC, Holmes EC (1997) A likelihood method for the detection of selection and recombination using nucleotide sequences. Mol Biol Evol 14:239–247

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52:696–704

Harbury PB, Zhang T, Kim PS, Alber T (1993) A switch between two-, three-, and four-stranded coiled coils in GCN4 leucine zipper mutants. Science 262:1401–1407

Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160–174

Haydon DT, Bastos AD, Samuel AR (2001) Evidence for positive selection in foot-and-mouth disease virus capsid genes from field isolates. Genetics 157:7–15

Holmes EC, Woelk CH, Kassis R, Bourhy H (2002) Genetic constraints and the adaptive evolution of rabies virus in nature. Virology 292:247–257

Hudson RR (2001) Two-locus sampling distributions and their application. Genetics 159:1805–1817

Krushkal J, Li WH (1995) Substitution rates in hepatitis delta virus. J Mol Evol 41:721–726

Kuhner MK, Yamato J, Felsenstein J (2000) Maximum likelihood estimation of recombination rates from population data. Genetics 156:1393–1401

Lai MM, Xia YP, Hwang SB, Lee CZ (1993) Functional domains of hepatitis delta antigen. Prog Clin Biol Res 382:21–27

Lee CZ, Lin JH, Chao M, McKnight K, Lai MM (1993) RNA-binding activity of hepatitis delta antigen involves two arginine-rich motifs and is required for hepatitis delta virus RNA replication. J Virol 67:2221–2227

Li WH (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol Evol 36:96–99

Li WH (1997) *Molecular Evolution*. Sinauer Associates, Sunderland, MA

McVeam G, Awadalla P, Fearnhead P (2002) A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics 160:1231–1241

Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929–936

Nisini RM, Paroli M, Accapezzato D, Bonino F, Rosina F, Santantanio T, Sallusto F, Amoroso A, Houghton M, Barnaba V (1997) Human CD4+ T-cell response to hepatitis delta virus: Identification of multiple epitopes and characterization of T-helper cytokine profiles. J Virol 71:2241–2251

Oakley MG, Hollenbeck JJ (2001) The design of antiparallel coiled coils. Curr Opin Struct Biol 11:450–457

Paraskevis D, Lemey P, Salemi M, Suchard M, Van De PY, Vandamme AM (2003) Analysis of the evolutionary relationships of HIV-1 and SIV cpz sequences using bayesian inference: Implications for the origin of HIV-1. Mol Biol Evol 20:1986–1996

Posada D, Crandall KA (2001) Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proc Natl Acad Sci USA 98:13757–13762

Ross HA, Rodrigo AG (2002) Immune-mediated positive selection drives human immunodeficiency virus type 1 molecular variation and predicts disease duration. J Virol 76:11715–11720

Rozzelle JE, Wang JG, Wagner DS, Erickson BW, Lemon SM (1995) Self-association of a synthetic peptide from the N terminus of the hepatitis delta virus protein into an immunoreactive alpha-helical multimer. Proc Natl Acad Sci USA 92:382–386

Swanson WJ, Nielsen R, Yang Q (2003) Pervasive adaptive evolution in mammalian fertilization proteins. Mol Biol Evol 20:18–20

Swofford DL (2000) PAUP*: Phylogenetic analysis using arsimony (*and other methods), version 4.0b 10 for Unix Sinauer, Sunderland, MA

Taylor JM (1998) Hepatitis delta. In: Mahy BWJ, Collier L (eds) *Topley and Wilson's microbiology and microbal infections*. Oxford University Press, New York

Thali MC, Furman C, Ho DD, Robinson J, Tilley S, Pinter A, Sodroski J (1992) Discontinuous, conserved neutralization epitopes overlapping the CD4-binding region of human immunodeficiency virus type 1 gp 120 envelope glycoprotein. J Virol 66:5635–5641

Wang JG, Jansen RW, Brown EA, Lemon SM (1990) Immunogenic domains of hepatitis delta virus antigen: Peptide mapping of epitopes recognized by human and woodchuck antibodies. J Virol 64:1108–1116

Worobey M (2001) A novel approach to detecting and measuring recombination: New insights into evolution in viruses, bacteria, and mitochondria. Mol Biol Evol 18:1425–1434

Worobey M, Rambaut A, Pybus OG, Robertson DL (2002) Questioning the evidence for genetic recombination in the 1918 "Spanish flu" virus. Science 296:211a

Wu JC, Chiang TY, Shiue WK, Wang SY, Sheen IJ, Huang YH, Syu WJ (1999) Recombination of hepatitis D virus RNA sequences and its implications. Mol Biol Evol 16: 1622–1632

Xia X, Xie Z (2001) DAMBE: Data analysis in molecular biology and evolution. J Hered 92:371–373

Yamaguchi-Kabata Y, Gojobori T (2000) Reevaluation of amino acid variability of the human immunodeficiency virus type 1 gp 120 envelope glycoprotein and prediction of new discontinuous epitopes. J Virol 74:4335–4350

Yang W, Bielawski JP, Yang Z (2003) Widespread adaptive evolution in the human immunodeficiency virus type I genome. J Mol Evol 57:212–221

Yang Z (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. J Mol Evol 39:306–314

Yang Z (1997) PAML: A program package for phylogenetic analysis by maximum likelihood. Cabios 13:555–556

Yang Z (2000) Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus. J Mol Evol 51:423–432

Yang Z (2001) Maximum likelihood analysis of adaptive evolution in HIV-1 gp120 env gene. Pacif Symp Biocomput pp 226–237

Yang Z, Bielawski JP (2000) Statistical methods for detecting molecular adaptation. TREE 15:496–503

Yang Z, Nielsen R, Goldman N, Pedersen A-M (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431–449

Zanotto PM, Kallas EG, de Souza RF, Holmes EC (1999) Genealogical evidence for positive selection in the *nef* gene of HIV-1. Genetics 153:1077–1089

Zuccola HJ, Rozzelle JE, Lemon SM, Erickson BW, Hogle JM (1998) Structural basis of the oligomerization of hepatitis delta antigen. Structure 6:821–830