Infection, Genetics and Evolution 54 (2017) 486–490

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Short communication

Heterogeneous recombination among Hepatitis B virus genotypes

Nadine Castelhano[a,b], Natalia M. Araujo[c], Miguel Arenas[a,b,d,*]

[a] Instituto de Investigação e Inovação em Saúde (i3S), University of Porto, Porto, Portugal
[b] Institute of Molecular Pathology and Immunology of the University of Porto (IPATIMUP), Porto, Portugal
[c] Laboratory of Molecular Virology, Oswaldo Cruz Institute, FIOCRUZ, Rio de Janeiro, Brazil
[d] Department of Biochemistry, Genetics and Immunology, University of Vigo, Vigo, Spain

ARTICLE INFO

Keywords:
Hepatitis B virus
Recombination
HBV genome
HBV genotypes
HBV evolution

ABSTRACT

The rapid evolution of Hepatitis B virus (HBV) through both evolutionary forces, mutation and recombination, allows this virus to generate a large variety of adapted variants at both intra and inter-host levels. It can, for instance, generate drug resistance or the diverse viral genotypes that currently exist in the HBV epidemics. Concerning the latter, it is known that recombination played a major role in the emergence and genetic diversification of novel genotypes. In this regard, the quantification of viral recombination in each genotype can provide relevant information to devise expectations about the evolutionary trends of the epidemic. Here we measured the amount of this evolutionary force by estimating global and local recombination rates in > 4700 HBV complete genome sequences corresponding to nine (A to I) HBV genotypes. Counterintuitively, we found that genotype E presents extremely high levels of recombination, followed by genotypes B and C. On the other hand, genotype G presents the lowest level, where recombination is almost negligible. We discuss these findings in the light of known characteristics of these genotypes. Additionally, we present a phylogenetic network to depict the evolutionary history of the studied HBV genotypes. This network clearly classified all genotypes into specific groups and indicated that diverse pairs of genotypes are derived from a common ancestor (i.e., C–I, D–E and, F–H) although still the origin of this virus presented large uncertainty. Altogether we conclude that the amount of observed recombination is heterogeneous among HBV genotypes and that this heterogeneity can influence on the future expansion of the epidemic.

1. Introduction

Hepatitis B virus (HBV) infection is a major cause of chronic liver disease, including cirrhosis and hepatocellular carcinoma, affecting 257 million people worldwide (WHO, 2017). HBV is a unique enveloped double-stranded DNA (dsDNA) virus that employs an error-prone reverse transcriptase to replicate its genome. Indeed, the estimated nucleotide substitution rate is higher than that observed in most dsDNA viruses, ranging between 10^{-6} and 10^{-4} substitutions per site per year (Osiowy et al., 2006; Paraskevis et al., 2015; Zhou and Holmes, 2007). As expected, the genetic diversity of the current HBV strains results from accumulating mutation and recombination events (Simmonds and Midgley, 2005; Zhou and Holmes, 2007). At present, human HBV is classified into 10 genotypes (A to J) and a growing number of subgenotypes based on nucleotide sequence divergence of the entire genome by 8% and 4%, respectively (Araujo et al., 2011; Tong and Revill, 2016). Despite established rules, HBV subgenotyping is not straightforward and a number of misclassifications have been reported (Pourkarim et al., 2010; Schaefer et al., 2009; Shi et al., 2013). HBV genotypes have distinct geographical distributions and have been increasingly associated with differences in clinical and virological responses, such as severity of liver disease and resistance to antiviral therapies (McMahon, 2009). Novel HBV variants generated by recombination have been documented worldwide. Some of them are now endemic in certain geographical regions, indicating the potential for spreading in a wide range of human populations and developing their own epidemiology (Araujo, 2015).

The relevance of recombination in virus evolution [establishing new chances for viruses to escape from selective pressures caused by the immune system or antiviral therapies (Arenas, 2015; Fraser, 2005; Levy et al., 2004)] leads to a need for not only detecting its presence (i.e., by identifying the distribution of recombination hotspots), but also for quantifying its extent (Perez-Losada et al., 2015). In this concern, recombination rate is often used to appraise the frequency of recombination events in the studied data (McVean et al., 2004). Multiple studies have estimated this parameter to investigate the evolution of diverse viruses such as HIV-1 or HSV-1 (Arenas et al., 2016; Bowden et al., 2004; Lopes et al., 2014). However, to our knowledge, the

* Corresponding author at: Department of Biochemistry, Genetics and Immunology, University of Vigo, Vigo 36310, Spain.
E-mail addresses: nadines@ipatimup.pt (N. Castelhano), nmaraujo@ioc.fiocruz.br (N.M. Araujo), marenas@uvigo.es (M. Arenas).

http://dx.doi.org/10.1016/j.meegid.2017.08.015
Received 5 May 2017; Received in revised form 13 August 2017; Accepted 17 August 2017
Available online 18 August 2017
1567-1348/

Table 1
Studied HBV data and estimated nucleotide diversity, GC content and global recombination rate. For each genotype, the table indicates the number of studied genomes (sample size), the observed nucleotide diversity (π), GC content, the score of the recombination test (Φ_w) (see also Fig. S2) and the global recombination rate (ρ) (see also Fig. 1). The presented values for ρ correspond to the average of values estimated with two independent MCMC runs.

Genotype	Sample size	Nucleotide diversity[a] (π)	GC content[a,b]	PHI (Φ_w)	Recombination rate[c] (ρ)
A	775	2.89	48.60 (1.68)	9.99×10^{-1}	8.47 (7.94 – 9.00)
B	1009	2.83	48.96 (1.68)	9.92×10^{-1}	3191.54 (761.78 – 5634.78)
C	1519	3.16	48.90 (1.68)	9.58×10^{-1}	821.52 (498.92 – 1595.15)
D	853	2.91	47.96 (1.68)	1.00	215.70 (172.50 – 259.92)
E	227	1.79	48.23 (1.70)	1.00	8234.01 (3067.80 – 18,372.35)
F	226	3.95	48.86 (1.66)	6.17×10^{-1}	16.67 (13.63 – 20.78)
G	38	1.53	47.98 (1.70)	1.70×10^{-5}	3.30×10^{-2} (3.30×10^{-2} – 3.30×10^{-2})
H	26	1.31	49.33 (1.70)	5.99×10^{-1}	33.02 (27.38 – 40.10)
I	60	2.80	48.42 (1.68)	1.00	27.27 (21.97 – 32.53)

[a] Nucleotide diversity and GC content are presented in percentage.
[b] 95% CI (confidence interval) among sites is shown in parenthesis.
[c] 95% HPDI (highest posterior density interval) is shown in parenthesis.

quantification of recombination (i.e., by estimated recombination rates) was not investigated so far in HBV. Here we estimated and discussed global (entire genome) and local (variation along the genome) recombination rates for the different HBV genotypes. We also inferred and discussed the evolutionary history of HBV genotypes with a recombination-based phylogenetic network. We conclude that recombination is heterogeneous among HBV genotypes and that the future expansion of the HBV epidemic could be highly influenced by this heterogeneity.

2. Materials and methods

2.1. Genomic data and nucleotide diversity

We downloaded all HBV full-length genome sequences available in the HBVdb database (Hayer et al., 2013) (Table 1). These sequences correspond to nine HBV genotypes (from A to I, J was excluded due to the lack of information with only one sequence in the database). Next, we obtained a multiple sequence alignment (MSA) of all sequences with *MAFFT* (Katoh and Standley, 2013) and trimmed poorly aligned regions (gaps in > 95% of sequences) with *trimAl* (Capella-Gutierrez et al., 2009). The resulting MSA presented an overall coverage of 99.4% (its variation along the genome is shown in Fig. S1; Supplementary material). Finally, we calculated nucleotide diversity (π) (Nei and Li, 1979) and GC content for each genotype.

2.2. Recombination analyses

For each genotype we performed two quantitative measures of recombination, (*i*) estimation of recombination rates at both global and local levels, and (*ii*) identification of recombination with a robust recombination test. We finally inferred an implicit phylogenetic network to explore the evolutionary history of HBV genotypes. We applied straightforward and well-established methods for investigating recombination in viruses (Perez-Losada et al., 2015).

We estimated global and local population recombination rates (ρ) with *RDP 4* (Martin et al., 2015), which incorporates a sophisticated version of the MCMC approach implemented in *LDHat* (McVean et al., 2004). We performed two independent runs per alignment, each based on 10,000,000 iterations. Convergence between runs was assessed as the overlapping of the estimated recombination rates within the 95% highest posterior density interval (HPDI). Additionally, as a sanity check of the estimates obtained from RDP, we analyzed genetic signatures of recombination with the statistical test PHI (Φ_w, pairwise homoplasy index), implemented in *PHIPack* (Bruen et al., 2006). We selected this specific test due to its robustness with respect to other recombination tests and because it can be applied to large datasets (Bruen et al., 2006).

In order to explore the evolutionary history of all the studied HBV genotypes we inferred a phylogenetic network [note that recombination can bias phylogenetic tree inferences (Arenas, 2013; Arenas and Posada, 2010; Posada and Crandall, 2002; Schierup and Hein, 2000)] with the well-established program *SplitsTree 4.1* (Huson and Bryant, 2006). Unfortunately, this program crashes (out of memory) with large datasets (higher than 1000 HBV genomes) and consequently, for this analyses we considered a maximum of 100 genomes per genotype that is a sample large enough to infer the phylogeny of 9 genotypes. The procedure to perform this analysis follows below. First, we identified the best-fitting model of nucleotide substitution for the genome sequences of all genotypes that are analyzed in subsequent steps. This task was performed with *jModelTest 2.0* (Darriba et al., 2012) under the Bayesian Information Criterion (BIC), as recommended by Luo et al. (2010), to select the model SYM + G + I. Next, this substitution model was applied to infer phylogenetic networks under the Neighbor-Net method (Bryant and Moulton, 2004), considering a bootstrap based on 100 and 1000 replicates, implemented in *SplitsTree 4.1*.

3. Results and discussion

We found that recombination is highly heterogeneous among HBV genotypes. The estimated global recombination rate was extremely high in genotype E ($\rho > 8000$), followed by genotypes B and C ($\rho > 3000$ and 800, respectively). Next, genotype D presented an intermediate recombination rate ($\rho > 200$). Genotypes A, F, G, H and I presented lower recombination rates ($\rho < 50$), where genotype G indicated almost negligible recombination (Table 1 and Fig. 1). We obtained similar findings with the recombination test PHI (Table 1), actually results from these analytical methods presented a high correlation (Fig. S2; Supplementary material).

Concerning nucleotide diversity, genotypes F and C presented the highest levels of diversity whilst genotypes G and H showed the lowest levels (Table 1). Since recombination could be biased by low genetic diversity (Posada and Crandall, 2001), we explored a possible correlation between these parameters. We found lack of correlation between recombination rate and nucleotide diversity (Fig. S3; Supplementary material) suggesting that our recombination analyses are not influenced by the level of genetic diversity present in the data. The GC content did not significantly vary across HBV genotypes (Table 1 and Fig. S4; Supplementary material), which indicates lack of interrelationship between recombination and GC content.

One could expect high levels of recombination in genotypes A, B and C since most of known intergenotypes are derived from them (Araujo, 2015; Shi et al., 2012; Yang et al., 2006). However, these genotypes can present a higher number of descendant recombinant forms because they are the oldest genotypes (Paraskevis et al., 2015) and not necessarily because of a larger recombination rate. Indeed, analyzing