Infection, Genetics and Evolution 44 (2016) 61–68

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

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

Heterogeneous genomic locations within NS3, NS4A and NS4B identified for genotyping and subtyping of Hepatitis C virus: A simple genome analysis approach

Sajid Mansoor [a,b], Aneela Javed [a,*], Amjad Ali [a], Atika Mansoor [b]

[a] Atta-ur-Rahman School of Applied Biosciences (ASAB), National University of Sciences & Technology, H-12, Islamabad, Pakistan
[b] Institute of Biomedical & Genetic Engineering (IBGE), G-9/1, Islamabad, Pakistan

ARTICLE INFO

Article history:
Received 7 March 2016
Received in revised form 13 June 2016
Accepted 14 June 2016
Available online 16 June 2016

Keywords:
HCV
Genotyping
Subtyping
Phylogenetic analysis
NS3
NS4A
NS4B
Genome analysis

ABSTRACT

Hepatitis C virus (HCV) displays excessive genetic heterogeneity and exists in several genotypes and subtypes. Characterizing these genotypes and subtypes becomes extremely important for diagnostic and epidemiological reasons. Present study analyzed HCV genome using a simple genome analysis approach. We combined manual sectioning of a reference genome alignment (*RA*) followed by a comprehensive comparative phylogenetic analysis. The main aim was to identify heterogeneous locations on HCV genome suitable for genotyping/subtyping. HCV reference dataset, comprising of whole genome sequences from all HCV genotypes and subtypes, was aligned into an *RA*. The *RA* was manually clipped into overlapping sections of 500 bases, each 50 bases apart. Phylogeny for each section and *RA* was estimated using neighbor-joining phylogenetic method. Clustering pattern between section phylogenies and *RA* phylogeny was compared for similarity. Sections (locations on genome) with clustering similar to whole genome were selected since it displays comparable genetic heterogeneity making these sections suitable for genotyping/subtyping. Based on this conception, we identified new genomic locations on NS3, NS4A and NS4B suitable for genotyping and subtyping. Exact genomic positions for known genotyping locations, core and NS5B were also identified. Furthermore, phylogenetic analyses at such small genomic scale provided opportunities to explore evolutionary relationships usually overlooked.

1. Introduction

Hepatitis C Virus (HCV) infection is one of the most destructive liver ailments affecting 4% of the world population. It persists with severe chronic infections in approximately 80% individuals, 5% may succumb to liver cirrhosis and hepatocellular carcinoma. HCV is therefore a strenuous global health issue with serious morbidity and mortality (Mohd Hanafiah et al., 2013). The small enveloped virus particle contains a positive-sense RNA genome and belongs to *Hepacivirus* of the family *Flaviviridae*. It evolves rapidly due to a high replication rate ($>10^{10-12}$ virions per day) and an error prone RNA-polymerase causing its extreme genetic heterogeneity (Ogata et al., 1991).

HCV exists in 7 subtypes with 67 confirmed subtypes plus some provisionally assigned subtypes (Smith et al., 2014). These genotypes and subtypes are geographically and epidemiologically diverse; subtypes 1a, 1b, 2a, 2b and 3a are global strains causing majority of infections. Genotype 4 and rare subtypes of 1 and 2 are endemic to Central Africa. The 5a strain is found only in Africa. Genotype 6 has numerous subtypes but their infections are mostly localized to South East Asia. Discovery of recombinants (like 2 k/1b) has added in to the complexion of HCV genetic variation (Nakano et al., 2012).

Identification of genotypes and subtypes is essential for HCV diagnosis and epidemiological analysis, having pertinent treatment consequences. Genotype identification is clinically relevant to determine therapy duration and for the prediction of treatment response (Zein et al., 1996). Genotypes 1 and 4 are resistant, requiring 48 week pegylated interferon-α plus ribavirin treatment but for genotype 2, 3, and 6, a treatment of 24 week is sufficient (Ghany et al., 2009; Shiffman et al., 2007; Fried et al., 2002; Poynard et al., 2009). Recently FDA approved Sofosbuvir (Sovaldi), is also prescribed according to genotype specification of the patient (Stedman, 2014; Gane et al., 2013; Jacobson et al., 2013; Lawitz et al., 2013).

Similarly, genotype remains one of the most important predictor of sustained virological response (SVR) in patients with chronic HCV. Response rate for patients with genotype 1 or 4 is around 50% while 70% or so achieve SVR in case of genotype 2 or 3 following pegylated interferon-α plus ribavirin treatment (Richter, 2002). Alternatively, Sofosbuvir gives greater SVR achievement (almost 90%) for all genotypes under different treatment durations (Schiff, 2015). There are few studies indicating relevance of subtype identification with

Abbreviations: bs, bootstrap; GTR, general time reversible; HCV, Hepatitis C virus; MCL, maximum composite likelihood; ML, maximum likelihood; N-J, neighbor-joining; *RA*, reference alignment; SVR, sustained virological response.

* Corresponding author.
E-mail address: javedaneela@googlemail.com (A. Javed).

http://dx.doi.org/10.1016/j.meegid.2016.06.031
1567-1348/

treatment outcome. Difficult-to-treat subtype 1a is associated with a lower response to therapy than subtypes 1b, 4a, and 4d (Legrand-Abravanel et al., 2009). Other subtypes may well present differential therapy responses too, however limited data is available on this.

All this connote the significance of correct genotype (and in some cases subtype) identification for the appropriate management of HCV. Blind treatment recommendations may not only produce resistant viral strains, but may well be a cause of severe social and physical distress for the patient, both grave health concerns worldwide.

A number of genotyping techniques are currently in use. Most commercially available methods rely on the amplification of highly conserved 5′-untranslated region (5′-UTR) and/or core from clinical isolates. Type-specific assays, such as restriction fragment length polymorphism analysis (Buoro et al., 1999), line probe reverse hybridization (Stuyver et al., 1993, 1995) or sequence analysis (Simmonds et al., 1994; Weck, 2005) is then performed for genotyping. Commercial genotyping method can identify genotypes with great accuracy however none is recommended for subtyping (Stuyver et al., 1996; Chinchai et al., 2003; Chen and Weck, 2002; Chevaliez and Pawlotsky, 2007; Nolte et al., 2003; Cai et al., 2013).

Direct sequencing and phylogeny inference in the core and/or NS5B region; therefore continue to be the golden standard for identification of genotypes particularly subtypes. This method accurately identifies subtypes and provides evolutionary information that is used to establish an epidemiological picture of circulating variants (Smith et al., 2014; Pybus et al., 2009). Although sequencing only in core or NS5B is considered sufficient for subtype identification, recombinants can only be typed by whole genome sequencing (Kalinina et al., 2002; Nakano et al., 2012; Kurbanov et al., 2008a, 2008b; Moreau et al., 2006; Morel et al., 2010; Tallo et al., 2007). It can be contemplated that multiple sequencing sites distinctly apart may eliminate the need for whole genome sequencing, in case of recombinants.

Furthermore locations on the genome for genotyping may mutate overtime making it difficult to amplify known genomic locations. For that reason, updated knowledge of genomic conservation and variation for the various genotypes and subtypes becomes vital. We argued that genomic locations heterogeneous enough to clearly distinguish between genotypes and subtypes but sufficiently conserved (for easy amplification) are suitable for genotyping. Secondly, we argued that such heterogeneous genomic locations, suitable for genotyping/subtyping, can be identified by determining locations on the genome that exhibit genetic heterogeneity similar to the whole genome. Because these locations best depict the genetic variation HCV displays among the various genotypes and subtypes.

Based on these premise, we designed the present study to identify heterogeneous locations on the HCV genome for genotyping and subtyping. We adopted a simple genome analysis approach; a reference alignment from all genotype and subtype genomes was generated, proceeded by manual sectioning and a comprehensive comparative phylogenetic analysis. The approach not only pointed out exact locations appropriate for genotyping/subtyping on the core and NS5B but also identified new ones.

2. Methodology

The methodology was devised to identify heterogeneous HCV genomic locations suitable for genotyping and subtyping. Briefly, a reference sequence data set comprising of whole genome sequences of all HCV genotypes and subtypes was selected and aligned. Phylogeny was estimated for the data set alignment which was then manually clipped into overlapping sections, each 500 bases long moving in steps of 50 bases. Finally, phylogeny was estimated for each section to identify sections with clustering similar to the whole genome phylogeny at >90% bootstrap confidence. These sections were then put together on the genome to determine the exact genomic location. This led to the identification of genomic locations suitable for genotyping and subtyping (Fig. 1). Detailed methodology is given below:

2.1. Selection of reference sequence data set

Full length genome sequences for all genotypes and subtypes, as described by Smith et al. (2014) were selected as the reference sequence data set and downloaded from the Los Alamos HCV database (Table S1). A total of 69 whole genome sequences (one for each HCV subtype plus an additional sequence for subtypes 1a and 3a) were selected as the reference sequence data set, including H77 isolate full length genome as numbering reference (Table S1). The sequences were selected because they best encompassed HCV genetic diversity and its classification into genotypes and subtypes as described by Smith et al. (2014). Multiple sequences for each subtype were avoided to keep the final analysis simple and also because reference sequences used here were very carefully selected in their report by Smith et al. (2014), relating to all the newly confirmed and previous HCV subtypes. Subtype 2e was excluded from the final alignment and subsequent phylogenetic analysis because it produced undesired gaps within *RA*.

2.2. Phylogenetic analysis of 'RA'

The sequences were aligned using MUSCLE program (8 iterations) along with full length genome sequence of HCV reference isolate H77 (AF009606) (Kuiken et al., 2006). The alignment was marked as the *reference alignment* (*RA*) for further mentions. Maximum Likelihood (ML) and Neighbor-joining (N-J) phylogeny was then estimated for *RA* using 10 different evolutionary distance models (Table S2) to create a phylogenetic comparison between both methods. The intention behind this comparison was not of mathematical nature; rather it was to choose the most appropriate method for subsequent phylogeny constructions. A comprehensive algorithmic comparison has been provided elsewhere (Kuhner and Felsenstein, 1994).

General time reversible (GTR) was the most suitable distance model for ML, as determined by the *MEGA* model selection tool. ML generates superior evolutionary relationships but is time inept, whereas N-J

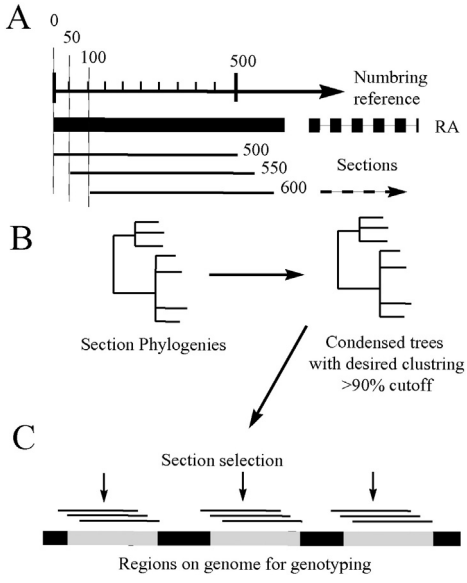

Fig. 1. Overview of the genome analysis strategy: Reference sequence data set comprising of whole genome sequences of all HCV subtypes was selected and aligned (*RA*). Phylogeny was estimated for *RA* and the alignment was then sectioned into overlapping sections, each 500 bases long moving in steps of 50 bases along the numbering reference (A). Phylogeny was estimated for each section. Sections with clustering similar to the whole genome phylogeny at >90% bootstrap confidence using condensed trees were selected (B). These sections were then put together on the genome to locate their exact genomic location (C). This led to the identification of genomic regions suitable for genotyping.