Infection, Genetics and Evolution 43 (2016) 329–337

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

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

Research paper

Deep sequencing increases hepatitis C virus phylogenetic cluster detection compared to Sanger sequencing

Vincent Montoya [a,b], Andrea Olmstead [a,b], Patrick Tang [c], Darrel Cook [a], Naveed Janjua [a,b], Jason Grebely [d], Brendan Jacka [d], Art F.Y. Poon [e], Mel Krajden [a,b,*]

[a] BC Centre for Disease Control, Vancouver, BC, Canada
[b] Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, BC, Canada
[c] Sidra Medical and Research Center, Doha, Qatar
[d] The Kirby Institute, UNSW Australia, Sydney, NSW, Australia
[e] BC Centre for Excellence in HIV/AIDS, St Paul's Hospital, Vancouver, BC, Canada

ARTICLE INFO

Article history:
Received 23 March 2016
Received in revised form 3 June 2016
Accepted 4 June 2016
Available online 6 June 2016

Keywords:
Deep sequencing
Hepatitis C virus
Quasispecies
Phylogenetics
Transmission

ABSTRACT

Effective surveillance and treatment strategies are required to control the hepatitis C virus (HCV) epidemic. Phylogenetic analyses are powerful tools for reconstructing the evolutionary history of viral outbreaks and identifying transmission clusters. These studies often rely on Sanger sequencing which typically generates a single consensus sequence for each infected individual. For rapidly mutating viruses such as HCV, consensus sequencing underestimates the complexity of the viral quasispecies population and could therefore generate different phylogenetic tree topologies. Although deep sequencing provides a more detailed quasispecies characterization, in-depth phylogenetic analyses are challenging due to dataset complexity and computational limitations. Here, we apply deep sequencing to a characterized population to assess its ability to identify phylogenetic clusters compared with consensus Sanger sequencing. For deep sequencing, a sample specific threshold determined by the 50th percentile of the patristic distance distribution for all variants within each individual was used to identify clusters. Among seven patristic distance thresholds tested for the Sanger sequence phylogeny ranging from 0.005–0.06, a threshold of 0.03 was found to provide the maximum balance between positive agreement (samples in a cluster) and negative agreement (samples not in a cluster) relative to the deep sequencing dataset. From 77 HCV seroconverters, 10 individuals were identified in phylogenetic clusters using both methods. Deep sequencing analysis identified an additional 4 individuals and excluded 8 other individuals relative to Sanger sequencing. The application of this deep sequencing approach could be a more effective tool to understand onward HCV transmission dynamics compared with Sanger sequencing, since the incorporation of minority sequence variants improves the discrimination of phylogenetically linked clusters.

1. Introduction

Approximately 80 million people worldwide are infected with hepatitis C virus (HCV) (Gower et al., 2014). About 75% of those infected develop chronic infection and 25% clear infection spontaneously (Grebely et al., 2014). In most high-income countries, people who inject drugs (PWID) account for >80% of new infections and are key drivers of onward HCV transmission (Shepard et al., 2005; Martin et al., 2013). With the approval of highly effective anti-HCV therapies, the opportunity to eliminate HCV is becoming feasible, but due to limited financial resources treatments are often prioritized to individuals with advanced liver disease. Controlling onward HCV transmission will also require targeting curative treatments to people with a high risk of onward HCV transmission (Grebely and Dore, 2014; Hellard et al., 2014).

Sequencing and phylogenetic analyses are powerful tools that can be used to understand population level transmission dynamics (Pybus et al., 2005; Magiorkinis et al., 2009; Jacka et al., 2014; Poon et al., 2014; Olmstead et al., 2015; Jacka et al., 2016). Sequences from individuals that cluster tightly together on a phylogenetic tree typically represent individuals who are part of a shared transmission network (Hué and Clewley, 2004; Hue et al., 2005; Gifford et al., 2007). Phylogenetic cluster analyses can support outbreak investigations in clinical settings and can characterize larger transmission networks among high-risk groups (i.e. PWID) (Lewis et al., 2008; Jacka et al., 2014). Tracking clusters over time may be used to identify populations who may benefit from 'Treatment as Prevention' or other harm reduction strategies and to evaluate the effectiveness of these efforts for reducing onward HCV transmission (Grebely and Dore, 2014; Hellard et al., 2014).

* Corresponding author at: Clinical Prevention Services, BC Centre for Disease Control, 655 W 12th Ave, Vancouver, BC 45Z4R4, Canada.
E-mail address: Mel.Krajden@bccdc.ca (M. Krajden).

http://dx.doi.org/10.1016/j.meegid.2016.06.015
1567-1348/

To date, most viral phylogenetic studies have used conventional PCR-based Sanger sequencing (consensus sequencing) to identify transmission clusters (Holmes and Grenfell, 2009; Grabowski and Redd, 2014). Consensus sequencing generates the dominant viral sequences present in an infected individual at a single point in time. As a result, these sequences fail to fully characterize diverse intra-host HCV populations, which evolve under viral and host dependent selective pressures that differ spatially (between individuals) and temporally (during transmission vs. established infections) (Wang et al., 2010; Bull et al., 2011; Preciado et al., 2014). Following a transmission event, the ability to reconstruct phylogenetic relationships between individuals decreases over time as the sequences in the donor and recipient viral populations evolve (Escobar-Gutierrez et al., 2012; Cruz-Rivera et al., 2013). Tracking transmissions is further complicated when the founder virus (transmitted variant) is a minority variant within either the donor or recipient at the time of sampling (Saito et al., 2004; Liu et al., 2006; D'Arienzo et al., 2013; Preciado et al., 2014; Li et al., 2015) (Fig. 1) because non-clonal Sanger sequencing fails to detect variants that are less than ~20% of the viral population (Palmer et al., 2005; Delobel et al., 2007). Exemplifying this restriction, a recent experimental analysis of viral genomic adaptations to new host environments showed that the majority of positively selected mutations were present at levels below this percent threshold (Bordería et al., 2015). False identification of transmission links can also occur when insufficient sequence and epidemiological data are available (Resik et al., 2007; Abecasis et al., 2011; Vandamme and Pybus, 2013).

Deep sequencing allows in-depth characterization of HCV intra-host sequence diversity and has been applied to differentiate acute from chronic HCV infections (Montoya et al., 2015) and to identify genetic bottlenecks during HCV transmission (Wang et al., 2010; Bull et al., 2011). However, few studies have combined large deep sequencing datasets with comprehensive phylogenetic analyses to examine the relationships between infected individuals (Wang et al., 2010; Bull et al., 2011; Le et al., 2015; Sato et al., 2015). Enhanced resolution of the viral quasispecies within an infected individual could facilitate identification of otherwise hidden transmission clusters and improve the ability to discriminate between related and unrelated transmissions. Here, we describe the application of a NS5B deep sequencing and phylogenetic approach to characterize and identify clusters in a population of HCV seroconverters (Montoya et al., 2015). This method, adapted from a consensus Sanger sequencing approach described by Prosperi et al. (Prosperi et al., 2011), identifies clusters using a dynamic patristic distance threshold based upon each individual's intra-host viral population. Our findings suggest that deep sequencing applied to characterize HCV quasispecies can improve phylogenetic cluster identification compared to consensus Sanger sequencing.

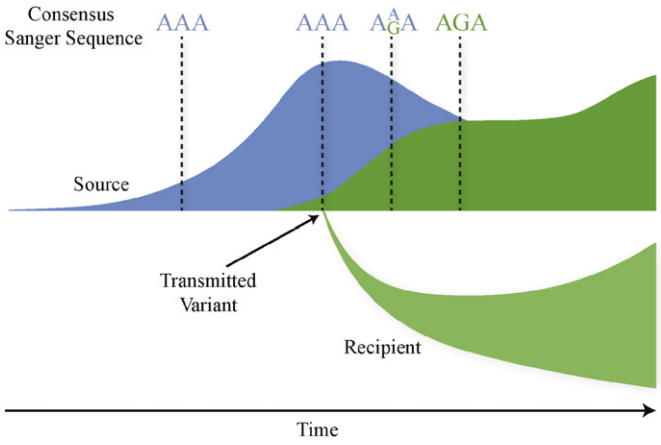

Fig. 1. A hypothetical HCV transmission scenario between individuals with increasing viral diversification over time. The ability of consensus Sanger sequencing to detect viral variants transmitted from a source to a recipient depends on the concentration of the variant at the time of sampling and on the discriminatory power of the technique. In this example, the major variant, AAA, but not the transmitted minor variant, AGA, is detectable by consensus sequencing of the source in the first two time points. The concentration of the AGA variant increases over time in the source, which enables its detection by consensus sequencing in the third and fourth time points. The likelihood that a given variant is transmitted depends both on its concentration and its transmission/replication fitness relative to other variants in the population.

2. Methods

2.1. Study population and design

Approximately 95% of all HCV serological screening in British Columbia, Canada, is performed at the BC Centre for Disease Control Public Health Laboratory. Automated linkage of new anti-HCV positive tests with historical test results permits identification of seroconverters, i.e., individuals who previously tested anti-HCV negative and on subsequent testing were anti-HCV positive. In total, 18% (374 of 2089) of individuals newly identified as anti-HCV positive in 2011, were identified as seroconverters.

As previously described, 94 samples were selected from 77 confirmed seroconverters to maximize the number of well-defined acute and chronic HCV infections as well as the availability of sequential samples from individuals (Montoya et al., 2015). Sixteen individuals had more than one sample: three with more than one sample within the acute timeframe, three with samples that spanned both the acute and chronic timeframes, and 10 with more than one sample within the chronic timeframe.

The estimated infection duration was defined as the time between the collection date of the sequenced sample and the midpoint between an individual's last negative and first positive anti-HCV tests. Acute HCV infection was defined as an estimated duration of less than six months and chronic infection as greater than six months. No identifying or risk factor information was available for the individuals in this study. Epidemiological data was limited to age, sex, HIV status and residential health authority. In this study, individual samples were assigned alphanumeric identifiers. The number represented the individual subject whereas the letter (e.g., A, B, C) was used to identify sequential samples (e.g., 1A, 1B, 1C).

2.2. NS5B amplicons

HCV RNA was extracted from serum using the MagMAX™-96 Viral RNA Isolation kit (Life Technologies, Carlsbad, CA, USA). Complimentary DNA (cDNA) was reverse transcribed using the SuperScript® VILO™ cDNA Synthesis Kit (Life Technologies) according to the manufacturer's protocol. For each sample, a 389 bp fragment corresponding to nucleotides 8250–8638 of the NS5B region of the HCV H77 genome (Los Alamos HCV Sequence Database (hcv.lanl.gov/content/index)) was amplified as previously described (Montoya et al., 2015). Equimolar concentrations of two dsDNA gBLOCK oligo controls (Integrated DNA Technologies, Coralville, Iowa), which correspond to HCV isolates H77 (Genotype 1a) and NZL1 (accession D17763, genotype 3a), were also amplified (Montoya et al., 2015).

2.3. Sanger sequencing

To obtain consensus Sanger sequences from each sample, NS5B amplicons were sequenced using the BigDye Terminator Cycle Sequencing technology (Life Technologies). Cycle sequencing products were ethanol precipitated and denatured using Hi-Di Formamide (Life Technologies) (McGovern et al., 2010). Sanger sequencing was performed using the ABI Prism 3730xl DNA Sequencer (Life Technologies).