**TEXT, SUPPLEMENTAL DIGITAL CONTENT 1**

**Methods**

**Ultra-deep sequencing**

Amplification and sequencing primers were designed to capture conserved regions of the hepatitis B virus (HBV) genome associated with variable levels of evolutionary pressure due to, for example, the presence or absence of an overlapping open reading frame or immune epitope (Table 1, Supplemental Digital Content 2; [1,2]). Extracted DNA from each clinical sample was quantified using the RealStar HBV PCR kit 1.0 (Altona Diagnostics Canada, Toronto, ON, Canada). Approximately equivalent quantities (IU/mL) of clinical HBV DNA and plasmid control HBV DNA, quantified using a conversion of 5 copies/IU, were amplified after diluting the extracted and plasmid DNA. First stage amplification was performed using 5 μL of template DNA, HBV-specific primers and HotStarTaq polymerase (Qiagen Inc., Toronto, ON, Canada). First stage PCR consisted of 30 cycles of standard thermocycling conditions which included an annealing temperature of 55°C for the core and surface primer sets and 50°C for the polymerase primer set. Illumina adapter overhang nucleotide sequences were added to the second stage HBV‐specific primer sequences (Table 1, Supplemental Digital Content 2; Illumina 16S Metagenomic Sequencing Library Preparation) and 2 μL of first stage amplification product was amplified using Roche Expand High Fidelity Polymerase (Sigma-Aldrich Canada, Oakville, ON, Canada) with 25 cycles of standard thermocycling conditions including an annealing temperature of 55°C for all primer sets. Amplicons were purified (PCRClean DX beads, Aline Biosciences, Woburn, MA, USA) and assessed by fluorometric detection (PicoGreen™ dsDNA Assay Kit, Thermo Fisher Scientific, Mississauga, ON, Canada). Approximately 10 nM purified adaptor-tagged amplicon from the clinical specimens and plasmid control were sequenced by synthesis using the Illumina MiSeq system (Reagent Nano kit v2, Illumina Canada, Victoria, BC, Canada).

**Sequence analysis**

Raw sequence reads were processed using bioinformatics tools within the Galaxy platform (https://usegalaxy.org/) in order to merge and filter paired-end reads to generate clusters of homologous sequences grouped using a threshold of 99% similarity, represented by a single haplotype sequence. The average number of reads was 6,184 for the surface antigen region, 5,079 for the core region and 5,347 for the polymerase region, with a range of 4,289 to 6,674 reads for all regions. Representative sequences were aligned using MAFFT (https://mafft.cbrc.jp/alignment/server/index.html), trimmed to remove primer sequences, and filtered to exclude sequences having non-in-frame insertions or deletions (approximately 7% sequences on average). The number of representative sequences (haplotypes) per sample is shown in Table 2, Supplemental Digital Content 2. All haplotypes, including those having a frequency of 1, were considered during analysis.

**Experimental error estimation**

Experimental error was determined by amplification and processing of a genotype A full genome plasmid (pEco63 plasmid ATCC 31518; GenBank accession AY128092) in parallel with clinical specimens at an approximately equivalent IU/mL quantity, to account for differences in viral load. The amplicon sequence generated for each sub-genome region of the plasmid by first stage PCR was cycle sequenced using an ABI 3730XL DNA analyzer (Applied Biosystems, USA) and the sequence used as the reference to map plasmid template Illumina MiSeq reads. Nucleotide statistics at each site were used to determine the mean and standard deviation (SD) single-nucleotide polymorphism (SNP) rate, or inherent error, within the experiment. The mean SNP rate per base was 0.195% (±0.135% SD). Base call variation among the amplicon sequence, including the highest SNP peak rate among all nucleotides, was considered when determining the final error rate (mean plus 3 SD), such that all SNP peaks fell below the final error rate which equalled 0.6%.

**Quasispecies complexity analysis**

The complexity of intra-host quasispecies was examined by calculating measures of incidence, such as the number of haplotypes, polymorphic sites and unique mutations, as well as measures of abundance, such as the maximum mutation frequency (Mfm), the population nucleotide diversity (π) and Simpson index (HSi), as described by Gregori, et al [3] . The mutation rate is defined as the incident frequency of mutation during genome replication [4]. Mfm is expressed as the mean substitutions per nucleotide for a set of quasispecies genomes in comparison to a determined consensus sequence for that set and therefore is a measurement of the proportion of nucleotides that are mutations within the set of sequences. The population nucleotide diversity (π) measures the expected fraction of nucleotide differences between any two genomes of a quasispecies set [3], thus π captures differences between genomes in the population, rather than differences between genomes and the consensus sequence. HSi is defined as the probability that two genomes from a viral quasispecies population selected at random belong to the same haplotype and thus is a measure of diversity within a population. The calculated value ranges from 0 (representing infinite diversity) to 1 (representing complete homogeneity with no diversity). Input data for calculating complexity measures consisted of MAFFT aligned, trimmed and filtered representative haplotype sequences. Incidence or entity level measures distinguish the frequency of haplotypes within a viral quasispecies, while Mfm and π, similar to biodiversity quantification measures used in ecology, are functional measures of viral quasispecies diversity (observed differences between haplotypes including the frequency of differences) and the Simpson index constitutes a measure of richness and evenness among haplotypes [3] . Calculated incidence and abundance values are directly compared, while higher HSi values imply a less diverse quasispecies (i.e., an HSi of 1.0 indicates a completely homogenous population).

**References**

1. Cowie BC. Is there an optimal genetic target for molecular analysis of hepatitis B virus transmission? J Clin Microbiol 2006;44:3051.

2. van de Klundert MA, Cremer J, Kootstra NA, Boot HJ, Zaaijer HL. Comparison of the hepatitis B virus core, surface and polymerase gene substitution rates in chronically infected patients. J Viral Hepat 2012;19:e34-40.

3. Gregori J, Perales C, Rodriguez-Frias F, Esteban JI, Quer J, Domingo E. Viral quasispecies complexity measures. Virology 2016;493:227-37.

4. Domingo E, Sheldon J, Perales C. Viral quasispecies evolution. Microbiol Mol Biol Rev 2012;76:159-216.

5. Stuyver L, De Gendt S, Van Geyt C, et al. A new genotype of hepatitis B virus: Complete genome and phylogenetic relatedness. J Gen Virol. 2000;81:67-74.

6. Osiowy C, Giles E. Evaluation of the INNO-LiPA HBV genotyping assay for determination of hepatitis B virus genotype. J Clin Microbiol. 2003;41:5473-5477.

7. Lindh M, Furuta Y, Ljunggren KK, et al. Detection of hepatitis B virus precore TAG mutant by an amplification-created restriction site method. J Infect Dis. 1995;171:194-197.

8. Grandjacques C, Pradat P, Stuyver L, et al. Rapid detection of genotypes and mutations in the pre-core promoter and the pre-core region of hepatitis B virus genome: correlation with viral persistence and disease severity. J Hepatol. 2000;33:430-439.

9. Chaudhuri V, Tayal R, Nayak B, et al. Occult hepatitis B virus infection in chronic liver disease: full-length genome and analysis of mutant surface promoter. Gastroenterology. 2004;127:1356-1371.

10. Illumina Inc. 16S Metagenomic Sequencing Library Preparation (15044223 B). Available at: <https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf>. Published November, 2013. Accessed October, 2016.