Validation of a Genotype-Independent Hepatitis C Whole Genome Sequencing Assay

Hope R Lapointe[1], Weiyan Dong[2], Winnie WY Dong[2], Don Kirkby[2], Conan Woods[2], Art FY Poon[3], Anita YM Howe[2], P Richard Harrigan[1], Chanson J Brumme[2]

**1 University of British Columbia - Division of AIDS, Vancouver, Canada, 2 BC Centre for Excellence in HIV/AIDS, Vancouver, Canada,
3 Department of Pathology and Laboratory Medicine, Western University, London, Canada**

Background

- Hepatitis C virus (HCV) infection is a major public health concern in Canada and worldwide. In 2018, the World Health Organization estimated that 71 million people globally had chronic HCV infection. Nearly 400,000 die annually from HCV infection, largely due to liver cirrhosis and hepatocellular carcinoma.
- The therapeutic approach towards hepatitis C virus (HCV) chronic infection has shifted in recent years with the advent of oral direct-acting antiviral (DAA) agents. Previously, first-line treatment options were limited to pegylated interferon-, ribavirin- and first-generation protease inhibitor-containing regimens.
- **DAA-based regimens are now recommended as first-line therapy for all patients with HCV infection.** With this highly effective repertoire of medications, overall sustained virologic response (SVR) rates are expected to exceed 95% regardless of HCV genotype, baseline HCV RNA levels, race, HIV coinfection, and hepatic fibrosis.
- However, cases of treatment failure have been associated with the **emergence of resistance-associated substitutions (RAS)**, detected in both treatment-naïve and treatment-experienced people.
- Additionally, only some DAA regimens are considered pan-genotypic and cover all **HCV genotypes**; some DAA regimens are HCV genotype- and/or subtype-specific.

Objectives

To better guide clinical decision-making, we developed and validated **a near-whole genome, HCV genotype-independent sequencing strategy** on the Illumina MiSeq next-generation sequencing platform.

Methods – Laboratory Protocol

- HCV-positive samples classified as genotypes (GT) 1-6 by external laboratories were used for this validation. Samples were collected from treatment-naïve HCV-infected individuals as well as DAA-treated persons who did not achieve sustained virologic response.
- Viral RNA was extracted on a Biomerieux easyMag RNA/DNA Extractor and underwent nested reverse-transcription-polymerase chain reaction (RT-PCR) to amplify the HCV genome in two fragments:
  - Core to NS5b codon 336 (Fig. 1).
  - NS5b codon 228 to 3' end of NS5b.
- Normalized, purified amplicon libraries were prepared by Nextera XT and sequenced on the Illumina MiSeq platform.

Methods – Data Processing

- MiSeq data were processed by an in-house pipeline (MiCall) utilizing bowtie 2 that incorporates HCV reference sequence selection and an iterative mapping process for paired-end reads:
  i. Paired-end reads are mapped to 57 HCV genotype reference sequences curated by the Los Alamos National Laboratory HCV Sequence Database.
  ii. Using the reference with the most mapped reads, the latter are collapsed into a sample-specific consensus to which reads are remapped. Consensus building and re-mapping occurs iteratively until ≥ 95% of reads map to the sample-specific consensus, or no additional reads are mapped by additional rounds of re-mapping.
  iii. Paired-end reads are merged and error correction rules are applied:
    - Bases with quality score <15 are discarded.
    - Mismatching base calls in overlapping reads are resolved by retaining the higher quality base.
  iv. Nucleotide consensus sequences are aligned to FDA reference strain sequences for downstream identification of RAS and indels.
- Sequences were compared to data obtained from a previously validated in-house assay optimized for HCV GT1 on the Illumina MiSeq. Genotypes were inferred through sequence analysis; if no sequence was generated, classification was based on the genotype provided by the sample provider.
- A minimal threshold for minority species detection was estimated from the coefficient of variation of minor species quantification as a function of mean nucleotide frequency.

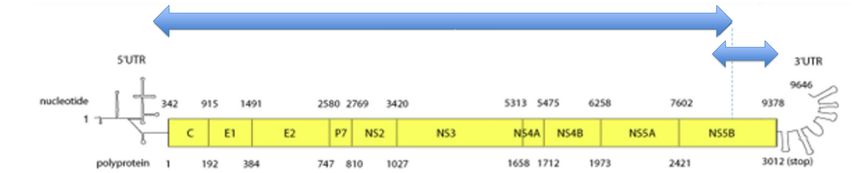


Figure 1. HCV genome map depicting coverage of a genotype-independent, near-whole genome sequencing method wherein the genome is amplified in two fragments then sequenced in parallel.

Results

Accuracy

The genotype-independent HCV method showed >99.85% nucleotide concordance with the HCV GT1-optimized next-generation sequencing method in NS3, NS5a, and NS5b using a 20% mixture-calling threshold.

RAS quantification demonstrated a linear relationship between both genotyping methods for NS3 and NS5a. No systematic bias in amplifying specific RAS was observed.

- Linear regression could not be performed reliably for NS5b, but substitutions at positions 159, 282 and 320 were conserved.

Sensitivity

Roughly 90% sequencing success rates (>100-fold NGS read coverage across NS3, NS5a and NS5b (codon 1-336)), was observed for most genotypes in samples with HCV RNA >5 $\log_{10}$ IU/mL (Table 1).

Table 1: Sequencing success rates NS3, NS5a and NS5b genes using a pan-genotype assay for GT1-6.

Genotype	Genotype Subtypes	Samples Attempted	Samples Sequenced (%)		
			NS3	NS5a	NS5b
1	1a	78	72 (92%)	72 (92%)	72 (92%)
	1b	11	11 (100%)	11 (100%)	11 (100%)
	1e	3	2 (67%)	2 (67%)	2 (67%)
2	2b	2	2 (100%)	2 (100%)	2 (100%)
3	3a	20	17 (100%)	17 (100%)	17 (100%)
4	4a, 4n	7	5 (71%)	5 (71%)	2 (29%)
5	5a	4	3 (75%)	3 (75%)	3 (75%)
6	6a, 6e, 6h, 6k, 6t	21	20 (95%)	20 (95%)	17 (85%)

The lower limit of detection was determined to be at >5 $\log_{10}$ IU/mL (Table 2).

- Drop in NS5b success is due to decreased coverage, primarily in proximity to the amplification primer site.

Table 2. Sequencing success rates in GT3-6 samples with different viral loads using a pan-genotype HCV assay.

$\log_{10}$ HCV RNA (IU/mL)	NS3	NS5A	NS5B
>7	2/2 (100%)	2/2 (100%)	2/2 (100%)
6.6 - 7.0	1/1 (100%)	1/1 (100%)	1/1 (100%)
6.1 - 6.5	4/4 (100%)	4/4 (100%)	4/4 (100%)
5.6 - 6.0	10/10 (100%)	10/10 (100%)	10/10 (100%)
5.1 - 5.5	20/23 (87%)	20/23 (87%)	15/23 (65%)
≤5	7/13 (7/54%)	7/13 (54%)	7/13 (54%)

Minority Species Detection

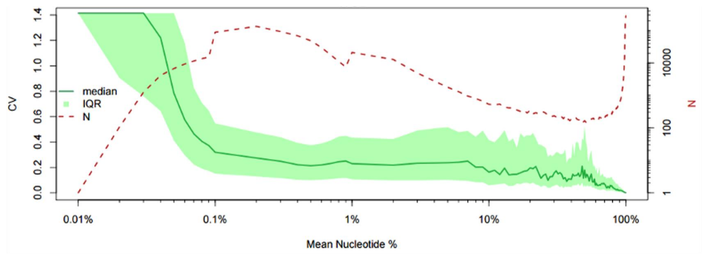


Figure 2. Median coefficient of variation (CV) of minor species with interquartile range (IQR) at different mean **nucleotide** frequencies (defined as the mean prevalence of a substitution detected across two replicates) from roughly 100 samples. N is defined as the total number of nucleotide substitutions detected at a particular mean substitution frequency.

An absolute lower limit of 0.2% for reproducible minor species detection was estimated but warrants a more conservative threshold given MiSeq error rates (0.5%) and limitations of PCR (Figure 2).

- The assay demonstrated minimal variation in its ability to reliably detect nucleotide substitutions when present in roughly 0.1-10% of sequence reads. However, nucleotide detection becomes highly variable for rare substitutions <0.1% prevalence.

Precision and Reproducibility

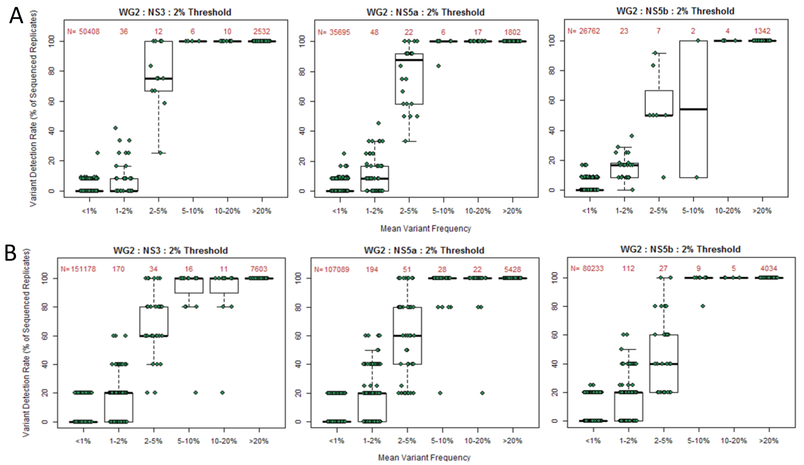


Figure 3. Precision and reproducibility were estimated from **amino acid** substitutions. Substitutions observed at a frequency ≥2% in a given replicate were considered "detected." Variant detection rate (defined as the % of replicates per sample, in which a substitution was detected) is categorized by expected substitution frequency (defined as the mean prevalence of a substitution across all replicates of a sample). Replicate testing of each sample began from the same RNA extract; all steps beginning from the RT-PCR were repeated. Numbers indicated in red represent the total number of substitutions in each bin. (A) Precision of detecting amino acid substitutions (intra-assay variability) was determined through replicate testing (n=12) of four samples (3 GT1a, 1 GT1b) processed on a single separate MiSeq run. (B) Reproducibility of detecting amino acid substitutions (inter-assay variability) was determined through replicate testing (n=5) of 12 samples (10 GT1a, 1 GT1b, 1 GT3) processed on separate days on five separate MiSeq runs.

- Precision analyses demonstrated that substitutions above 5% mean prevalence were detected in nearly 100% of replicates (Fig. 3A).
  - >80% repeatability was observed in roughly half of the substitutions with a mean prevalence between 2 and 5%.
  - Lower precision was observed below 2% mean prevalence (the mixture-calling cutoff).
  - Substitutions present below 1% prevalence were correctly identified as conserved positions, with only a few outliers.

Almost 100% reproducibility was achieved at detecting amino acid substitutions above 5% prevalence, when using a 2% mixture-calling threshold (Figure 3B).

- 75% reproducibility was observed in roughly half of the substitutions with a mean prevalence of 2-5%.
- Reproducibility decreased –genotype further once below the mixture-calling cutoff; substitutions with a mean prevalence between 1 and 2% ranged from 0 and 75%.
- Conserved amino acids (<1% mean prevalence) were also identified correctly, with few outliers.

Specificity

No evidence of cross-contamination, interfering substances (e.g. hemoglobin) or off-target amplification of HIV or hepatitis B virus from coinfected samples was observed in this study.

Conclusions

- This study highlights the performance of a whole genome NGS assay and bioinformatic pipeline for genotype-independent HCV genotyping and RAS detection. The assay protocols and software are freely available; protocols can be obtained by contacting Dr. Chanson Brumme (cbrumme@cfenet.ubc.ca), and software is available at https://github.com/cfe-lab/MiCall.
- This HCV genotyping method demonstrated similar performance to a validated GT1 assay and can now be extended to other HCV genotypes.
- This method has been implemented clinically and has been used to deliver ~3000 resistance test reports to physicians across Canada.

CROI Conference on Retroviruses and Opportunistic Infections

Genome BritishColumbia · BRITISH COLUMBIA CENTRE for EXCELLENCE in HIV/AIDS · UBC · Providence HEALTH CARE How you want to be treated.