A near full-length open reading frame next generation sequencing assay for genotyping and identification of resistance-associated variants in hepatitis C virus

Pedersen, Martin Schou; Fahnøe, Ulrik; Hansen, Thomas Arn; Pedersen, Anders Gorm; Jenssen, Håvard; Bukh, Jens; Schønning, Kristian

Published in:
Journal of Clinical Virology

DOI:
10.1016/j.jcv.2018.05.012

Publication date:
2018

Document Version
Peer reviewed version

Citation for published version (APA):
Pedersen, M. S., Fahnøe, U., Hansen, T. A., Pedersen, A. G., Jenssen, H., Bukh, J., & Schønning, K. (2018). A near full-length open reading frame next generation sequencing assay for genotyping and identification of resistance-associated variants in hepatitis C virus. *Journal of Clinical Virology*, *105*, 49-56. https://doi.org/10.1016/j.jcv.2018.05.012

Title: A near full-length open reading frame next generation sequencing assay for genotyping and identification of resistance-associated variants in hepatitis C virus

Authors: M.S. Pedersen, U. Fahnøe, T.A. Hansen, A.G. Pedersen, H. Jenssen, J. Bukh, K. Schønning

PII: S1386-6532(18)30147-1
DOI: https://doi.org/10.1016/j.jcv.2018.05.012
Reference: JCV 4008

To appear in: *Journal of Clinical Virology*

Received date: 23-12-2017
Revised date: 22-5-2018
Accepted date: 26-5-2018

Please cite this article as: Pedersen MS, Fahnøe U, Hansen TA, Pedersen AG, Jenssen H, Bukh J, Schønning K.A near full-length open reading frame next generation sequencing assay for genotyping and identification of resistance-associated variants in hepatitis C virus.*Journal of Clinical Virology* (2018), https://doi.org/10.1016/j.jcv.2018.05.012

A near full-length open reading frame next generation sequencing assay for genotyping and identification of resistance-associated variants in hepatitis C virus.

Pedersen, M.S.[1,2,3], Fahnøe, U.[2], Hansen, T.A.[1], Pedersen, A.G.[4], Jenssen, H.[3], Bukh, J.[2], Schønning, K[1,5].

[1]Department of Microbiology, Copenhagen University Hospital, Hvidovre,

[2]Copenhagen Hepatitis C Program (CO-HEP), Department of Infectious Diseases and Clinical Research Centre, Copenhagen University Hospital, Hvidovre, and Department of Immunology and Microbiology, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark,

[3]Department of Science and Environment, Roskilde University, Denmark,

[4]DTU Bioinformatics, Technical University of Denmark, Denmark,

[5]Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark.

[*]Corresponding author:

Kristian Schønning

Department of Clinical Microbiology, Copenhagen University Hospital, Hvidovre.

Kettegård Alle 30, 2650 Hvidovre, Denmark.

kristian.schoenning@regionh.dk

Manuscript: 2554 words (2500 allowed).

Abstract: 249 words (250 allowed).

Highlights

- A near full-length HCV amplicon was generated using a single set of primers.
- HCV genotype 1a, 1b, 2b, 3a, 3b, 3h, 4a, 4d, 4o, 4r samples were sequenced.
- Samples with HCV RNA down to 4 Log IU/mL were sequenced.
- Method generated variation was estimated below 1%.
- The method identified dual infections and the presence of subgenomic replicons.

Abstract

Background: The current treatment options for hepatitis C virus (HCV), based on direct acting antivirals (DAA), are dependent on virus genotype and previous treatment experience. Treatment failures have been associated with detection of resistance-associated substitutions (RASs) in the DAA targets of HCV, the NS3, NS5A and NS5B proteins.

Objective: To develop a next generation sequencing based method that provides genotype and detection of HCV NS3, NS5A, and NS5B RASs without prior knowledge of sample genotype.

Study design: In total, 101 residual plasma samples from patients with HCV covering 10 different viral subtypes across 4 genotypes with viral loads of 3.84-7.61 Log IU/mL were included. All samples were de-identified and consequently prior treatment status for patients was unknown. Almost full open reading frame amplicons (~ 9 kb) were generated using RT-

PCR with a single primer set. The resulting amplicons were sequenced with high throughput sequencing and analysed using an in-house developed script for detecting RASs.

Results: The method successfully amplified and sequenced 94% (95/101) of samples with an average coverage of 14,035; four of six failed samples were genotype 4a. Samples analysed twice yielded reproducible nucleotide frequencies across all sites. RASs were detected in 21/95 (22%) samples at a 15% threshold. The method identified one patient infected with two genotype 2b variants, and the presence of subgenomic deletion variants in 8 (8.5%) of 95 successfully sequenced samples.

Conclusions: The presented method may provide identification of HCV genotype, RASs detection, and detect multiple HCV infection without prior knowledge of sample genotype.

Keywords: RAS, NGS, genotyping, subgenome, RT-PCR, replicon.

1. Background

Hepatitis C virus (HCV) is a single-stranded positive-sense RNA virus belonging to the *Flaviviridae* family with a genome of ~ 9,600 nucleotides. The genome encodes three structural proteins (core, E1, E2) and seven non-structural proteins (p7, NS2, NS3, NS4A, NS4B, NS5A, NS5B) [1,2]. There are 7 recognized genotypes with up to 35% nucleotide divergence and more than 80 accepted subtypes [3,4]. Treatment with direct acting antivirals (DAAs) targeting the NS3 protease, NS5A and the NS5B polymerase has increased efficacy of antiviral treatment [5]. Treatment duration and choice of DAAs is dependent on HCV genotype and prior patient treatment experience. Treatment failures have been associated with resistance-associated substitutions (RASs) in NS3, NS5A and NS5B [5–7].

An ideal method for detecting RASs should cover current DAA targets NS3, NS5A, and NS5B for all genotypes, have a high analytical sensitivity, and be capable of detecting RASs in minority virus populations [8]. Amplification and sequencing of parts of the NS3 and NS5A has been used to detect RASs across multiple genotypes with high analytical sensitivity [9,10].

Alternatively, all RAS sites in HCV may be obtained within one amplicon with subtype-specific primers covering the entire ORF [11,12], subtype-specific primers with nested PCR generating a near full-length amplicon [13], or pan-genotypic primers generating a near full-length amplicon for multiple genotypes [14]. The advantage of the latter method is that prior knowledge of sample subtype is not required for amplification. Other methods have omitted the initial HCV specific amplification and used template-independent amplification [15], direct RNA sequencing [16], in combination with ribosomal RNA depletion [17], or HCV probe based target enrichment [18], but may display varying analytical sensitivity.

We adopted a strategy of generating a near full-length amplicon using pan-genotypic primers and subsequent next generation sequencing. Compared to the similar method

published by Trémeaux *et al.* [14], all steps have been optimized to increase sensitivity. The procedure provides in-depth RASs analysis, near full-length ORF consensus sequence and genotype simultaneously across multiple HCV genotypes, and it may therefore be used to replace standard HCV genotyping assays for primary characterization of specimens from chronically HCV infected individuals.

2. Objectives

Our aim was to develop a method that provides genotype and detection of resistance-associated substitutions in NS3, NS5A, and NS5B of HCV without prior knowledge of sample genotype.

3. Study design

3.1. HCV Specimens

The cell culture derived HCV RNA sample TNcc contains a full-length subtype 1a genome [19]. Plasmids containing HCV strain specific whole genome sequences from isolates TN [19], J6/JFH-1 [20], DBN3acc [7], SA13 [21,22] and KH6a [21,23] have been described previously.

Residual plasma from samples sent for routine genotyping were de-identified and stored at -20°C prior to amplification. The treatment history of the patients was consequently unknown. A total of 101 samples from patients with HCV were included between January 4, 2016 and January 9, 2017 at the Department of Clinical Microbiology, Copenhagen University Hospital, Hvidovre, Copenhagen and included 10 different subtypes across 4 genotypes. Samples containing HCV subtypes other than 1a and 3a were preferentially included and samples with HCV subtypes 1a and 3a were selected to achieve a broad range of viral loads. Viral concentrations were 3.84-7.61 Log IU/mL (mean 5.99 Log IU/mL).

3.2 RNA extraction

RNA extraction was performed with the ZR Viral RNA kit (Zymo Research, Irvine, CA, USA) as previously described [17] with the modification that 300 µL sample was centrifuged at 1,500g for 10 minutes to pellet cell debris before 200 µL supernatant was used as input.

3.3. RT-PCR of near full-length open reading frame

Reverse transcription was performed with 12 µL template and 1 µL RNasin Plus RNase Inhibitor (Promega Corporation, Madison, WI, USA) with Maxima RT minus H reverse transcriptase (Thermo Fischer Scientific, Waltham, MA, USA) with 2.5 µM RT-primer oligo dA20 [24] at 40°C for 2 hours. Reactions were inactivated at 85°C for 5 minutes and placed on ice. Amplification was done in 4 replicates using Q5 Hot Start High-Fidelity DNA Polymerase (New England BioLabs, Ipswich, MA, USA) with High GC enhancer in a 25 µL reaction volume with 2.5 µL cDNA, 0.5 µM each of forward primer 5UTR_F_298 (Position in H77 298-315: AGGGTGCTTGCGAGTGCC) and reverse primer NA2.9304.v2 (Position in H77 9284-9304: CGGGCAYGHGACASGCTGTGA; modified from [25]) in a two-step PCR with initial denaturation at 98°C for 30 seconds, 35 cycles with 10 seconds denaturation at 98°C, and 8.5 minutes annealing/extension at 72°C, with a final extension step at 72°C for 8 minutes. The PCR products were purified using DNA Clean and Concentration Zymo Research, Irvine, CA, USA) and eluted in 25 µL elution buffer. Concentrations were measured with Qubit dsDNA HS assay kit (Thermo Fischer Scientific, Waltham, MA, USA).

3.4. Library preparation and sequencing

Library preparation was performed using NEBNext Ultra II DNA Library Prep kit for Illumina (New England BioLabs, Ipswich, MA, USA) with fragmentation of 25 ng PCR product with NEBNext dsDNA Fragmentase (New England BioLabs, Ipswich, MA, USA) at 37°C for 17.5 minutes before inhibition with 0.5 M EDTA (Sigma-Aldrich, ST Louis, MO, USA). NEBNext Multiplex Oligos (New England BioLabs, Ipswich, MA, USA) were added in 7 PCR cycles. Size selection was performed with Agencourt AMPure XP Beads (Beckman Coulter, Brea, CA, USA) to obtain fragments of 300 bp. Resulting libraries were evaluated with Bioanalyzer High Sensitivity DNA kit (Agilent Technologies, Santa Clare, CA, USA), normalized with NEBNext Library Quant kit for Illumina (New England BioLabs, Ipswich, MA, USA) before pooling. Batches of 24 samples were loaded to a MiSeq Reagent kit v2 (300 cycles) (Illumina, San Diego, CA, USA).

3.5. PCR for characterization of subgenomic replicons

A sample containing subgenomic replicons were additionally amplified using alternative reverse primer NS2_R_3166 (Position in H77 3149-3166: GGCCCTCACAAAGTATGG) in a three-step PCR with initial denaturation at 98°C for 30 seconds, 35 cycles with 10 seconds denaturation at 98°C, 10 seconds annealing at 65°C and 1.5 minutes extension at 72°C, with a final extension step at 72 °C for 2 minutes.

3.6. Identification of RASs

A pipeline with software packages was developed to analyse the NGS data. Reads were excluded or trimmed with Sickle [26] to obtain reads with Phred quality scores ≥30. Passing reads were assembled *de novo* with IVA [27], mapped against the generated consensus sequence using BWA [28], and SNP-calling was done using LoFreq SNP-caller [29] with a cut-off set at 1%. RAS analysis was performed with the author's own software

with a local, modified database extracted from Geno2Pheno [30]. Samples failing de novo assembly with IVA were handled by an iterative mapping strategy with BWA and subtype references [31] before SNP-calling.

Phylogenetic analysis was done using the assembled consensus sequences obtained. Sequences were aligned to reference subtypes [31] using MAFFT [32]. Phylogenetic trees were constructed using PhyML [33].

3.7. Structural genome variant analysis

All samples were screened for recombination with all available options in RDP4 [34]. Samples were investigated for structural variants with LUMPY with standard settings [35].

4. Results

4.1. Analytical sensitivity and inclusivity of long range RT-PCR

Analytical sensitivity was evaluated using low passage cell culture derived HCV TNcc RNA [19] for which all primers have a perfect match. This provides an estimate of maximally obtainable analytical sensitivity. Two 10-fold dilution series were prepared in parallel and tested in the RT-PCR in two replicates. All four reactions were amplified consistently down to the 10^{-4} dilutions (Figure 1). HCV RNA in the two 10^{-4} dilutions were 3.87 Log IU/mL and 3.88 Log IU/mL, and provided a crude estimate of lower limit of amplification.

Samples with genotype 5 and 6 were not available as clinical samples. Instead plasmids SA13 and HK6a with 5a and 6a specific sequences were successfully amplified to evaluate inclusivity of these genotypes.

4.2. Assessment of assay generated nucleotide variation

The library preparation was evaluated with direct input of plasmids and amplicons of the plasmids TN, J6/JFH-1, DBN3acc, SA13, HK6a. All but one of the samples had less

than 1% variation per position across the approximately 9 kb amplicon. The PCR product of DBN3acc had 4 consecutive bases at 2.7% present as minor variant above the 1% threshold. The sequenced plasmid only showed 0.3% variation at these positions. Sequencing the same libraries again on different flow cells on different days reproduced the original results indicating that variation had been introduced during PCR amplification.

4.3. Reproducibility of the procedure

Three clinical samples (1, 5, and 29) were extracted, amplified and sequenced twice in different runs. Samples 1, 5, and 29 had 0, 16, and 3 discrepancies at the consensus level across nucleotide positions, 9007, 9007, and 9004 nucleotides, respectively, yielding an inter-assay reproducibility of base calling at the consensus level of 100%, 99.8% and 99.9%. The SNPs variant frequencies were generally close between the duplicates, as seen in Figure 2. At a 5% duplicate SNP frequency threshold, samples 1, 5, and 29 had 54, 247, and 197 variant sites. Sample 1 had one site with 8% difference between the duplicates. Samples 5 and 29 contained 4 and 3 sites, respectively, with more than 20% difference in frequency between the duplicates.

4.4. Validation on clinical samples

Of the 101 samples included, 94% (95/101) were successfully amplified and sequenced. The samples had the following genotype distribution 1a: 26/27, 1b: 14/15, 2b: 10/10, 3a: 27/27, 3b: 3/3, 3h: 1/1, 4a: 2/6, 4d: 10/10, 4o: 1/1, 4r: 1/1 (see supplementary Table S1). Viral load ranged between 3.84-7.61 Log IU/mL (Figure 3). Successfully amplified samples had a mean coverage of 14,035 (277-57,732), and reading frames were open and encoded a single polyprotein. Recombinant sequences were not detected using RDP4.

The samples failing amplification were not strictly correlated to viral load (Figure 3). Of the 6 samples failing amplification, 4 were of subtype 4a with a broad range of viral loads.

Two of these, sample 84 and 88, were amplified with subtype specific full-length open reading frame primers and both showed a mismatch in the reverse primer. A genotype 1a sample failing amplification, number 17, was also amplified with subtype specific primers and also had a mismatch in the reverse primer.

Reads from sample 46 assembled into two subtype 2b near full-length genomes 10% divergent at consensus level. Reads mapping to the two genomes were represented at a 3:2 ratio in favour of the genome designated 46.1.

Sample 43 was amplified with a shorter amplification product in addition to the expected ~ 9 kb PCR band (Figure 4A). The NGS coverage plot showed a drop in coverage between position 843 (core, H77 numbering) and 3004 (NS2, H77 numbering) (Figure 4B). Amplification with primers spanning the coverage drop yielded 4 bands (2 bands ~ 2700 bp and 2 bands ~ 700 bp; Figure 4C). Sanger Sequencing of the excised gel bands demonstrated the presence of the coverage drop sequence in the long fragments but not the short fragments. The software LUMPY identified paired end reads mapping to both sides of the gap (Figure 4B). Besides the near full-length genome, 3 different shorter variants were identified all with open reading frames.

Review of coverage plots identified 8/95 (8.4%) samples (8, 25, 40, 43, 45, 85, 91, and 101) with coverage drops from the core to NS2 region. Retrospective investigation of their gel images confirmed the smaller amplification product. The structural variants identified constitute subgenomic deletion variants as previously described [36–38].

4.5. Phylogeny of the samples

Phylogeny of the near full-length genomes showed conformity to results from routine genotyping [39,40]. Sample pairs 2 and 3, 36 and 38, 49 and 52, 50 and 51, 74 and 75, 90 and 96, 94 and 95, and 97 and 98 were highly related (Figure 5) with only 0, 2, 2, 0, 15, 8,

89, and 7 SNP's between them at consensus level, respectively. Thus, these samples may have originated from the same individual or epidemiologically related individuals.

4.6. Identification of RASs

The RAS pipeline identified all sites with nucleotide polymorphisms reported to convey resistance to any DAA, both polymorphisms naturally occurring as the most frequent within a HCV genotype (genotype intrinsic RAS) and polymorphisms occurring infrequently within a genotype (genotype extrinsic RAS). Genotype intrinsic RASs constituted the majority of RASs (247/275 (89.8%) at a 15% threshold, Table 1). 28 genotype extrinsic RASs were identified (in 21 samples) at a 15% threshold; this increased to 32 (in 23 samples) and 36 (in 27 samples) at a 5% and 1% thresholds, respectively. Only the three RASs Q30K (NS5A), L159F (NS5B) and S556G (NS5B) were detected by lowering the threshold. As seen in Table 1, the majority of extrinsic RASs were within genotype 1 (Q80K (NS3) in 3 samples, Y93H (NS5A) in 4 samples and C316Y (NS5B) in 3 samples). Five samples contained 12 of the extrinsic RASs at a 15% threshold. Sample 9 had Q30R (NS5A), L31M (NS5A) and L159F (NS5B), and the two former were combined in 99% of the reads. Sample 10 had L28M (NS5A) and Q30H (NS5A) linked in 96% of the reads. Sample 35 had NS5B substitutions L159F and S556G. Sample 39 had NS5B substitutions L159F, C316N and S556G. Sample 44 had NS5A substitutions Q30K and L31M linked in 99% of the reads.

5. Discussion

The aim of the present study was to develop a sequence based assay capable of identifying HCV genotype and the presence of RASs in clinical samples without prior characterization. Such an assay should be pan-genotypic, analytically sensitive, and capable to reproducibly identify RASs present in minority HCV populations. We successfully

sequenced clinical samples containing HCV genotype 1a, 1b, 2b, 3a, 3b, 3h, 4d, 4o, and 4r, however, the method failed to amplify 4/6 genotype 4a samples. In the study, rarer genotypes were prioritized to specifically test inclusivity. In our practise subtypes 1a, 1b, 2b, and 3a constitute 90% of genotyped samples, and genotype 4a constitutes 3%, although prevalence may be considerably higher in other regions [41,42].

The method successfully amplified clinical samples down to 3.84 Log IU/mL. Only 4% of samples for routine genotyping was below 4 Log IU/mL in a recent study from our institution [43], and none was below this level in a study with 140 DAA treatment naïve patients [9]. The overall success rate of the method presented here is comparable to individual amplification of parts of NS3 and NS5A [9,10], and compares favourably with a previously reported almost full length amplification approach that obtained an overall amplification and sequencing success rate of 58% (19/33) [14].

Three samples were analysed twice and the reproducibility across all sequenced sites were >99.5% at the consensus level. The frequency of extrinsic RASs only increased modestly by lowering the threshold for SNP calling and, interestingly, 7 of 8 RASs additionally identified were in NS5B. A recent analysis of data from clinical trials of genotype 1 patients treated with ledipasvir and sofosbuvir suggested that drug-specific RASs, i.e. substitutions that emerged during drug treatment or conferring significantly reduced drug susceptibility *in vitro*, may impact treatment outcome in difficult to treat patient populations when present at baseline also at low frequencies [44]. Here, we detected NS5A substitutions Q30H and Y93H that are known to significantly reduce the efficacy of ledipasvir [44].

Pre-treatment RAS testing is useful in selected patient populations [44]. The method presented for RAS identification can be implemented as a primary test in place of genotyping. In combination with accurate long-read sequencing techniques amplification of the near full length open reading frame may prove useful for linkage analysis of RASs.

Funding

This study was supported by grants from The A.P. Møller Foundation for the Advancement of Medical Science, The Scandinavian Society for Antimicrobial Chemotherapy Foundation, and Hvidovre Hospital Research Foundation. In addition, this work was supported by a Ph.D. stipend from Roskilde University [M.S.P.], and by grants from The Capital Region of Denmark's Research Foundation [J.B.], the Innovation Fund Denmark [J.B.], the Novo Nordisk Foundation (J.B.), and the Danish Research Council [J.B.] including an advanced Sapere Aude grant [J.B.]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Ethical approval

Not required.

Conflict of interest

The authors have no conflicts of interest to declare.

Reference

[1] D. Moradpour, F. Penin, C.M. Rice, Replication of hepatitis C virus, Nat. Rev. Microbiol. 5 (2007) 453–463. doi:10.1038/nrmicro1645.

[2] J. Bukh, The history of hepatitis C virus (HCV): Basic research reveals unique features in phylogeny, evolution and the viral life cycle with new perspectives for epidemic control., J. Hepatol. 65 (2016) S2–S21. doi:10.1016/j.jhep.2016.07.035.

[3] D.B. Smith, P. Becher, J. Bukh, E.A. Gould, G. Meyers, T. Monath, A.S. Muerhoff, A. Pletnev, R. Rico-Hesse, J.T. Stapleton, P. Simmonds, Proposed update to the taxonomy of the genera Hepacivirus and Pegivirus within the Flaviviridae family, J. Gen. Virol. 97 (2016) 2894–2907. doi:10.1099/jgv.0.000612.

[4] P. Simmonds, P. Becher, J. Bukh, E.A. Gould, G. Meyers, T. Monath, S. Muerhoff, A. Pletnev, R. Rico-Hesse, D.B. Smith, J.T. Stapleton, I.R. Consortium, ICTV Virus

Taxonomy Profile: Flaviviridae, J. Gen. Virol. 98 (2017) 2–3. doi:10.1099/jgv.0.000672.

[5] J.-M. Pawlotsky, Hepatitis C Virus Resistance to Direct-Acting Antiviral Drugs in Interferon-Free Regimens, Gastroenterology. 151 (2016) 70–86. doi:10.1053/j.gastro.2016.04.003.

[6] E.F. Donaldson, P.R. Harrington, J.J. O'Rear, L.K. Naeger, Clinical evidence and bioinformatics characterization of potential hepatitis C virus resistance pathways for sofosbuvir, Hepatology. 61 (2015) 56–65. doi:10.1002/hep.27375.

[7] S. Ramirez, L.S. Mikkelsen, J.M. Gottwein, J. Bukh, Robust HCV Genotype 3a Infectious Cell Culture System Permits Identification of Escape Variants With Resistance to Sofosbuvir, Gastroenterology. 151 (2016) 973–985.e2. doi:10.1053/j.gastro.2016.07.013.

[8] S.R. Bartlett, J. Grebely, A.A. Eltahla, J.D. Reeves, A.Y.M. Howe, V. Miller, F. Ceccherini-Silberstein, R.A. Bull, M.W. Douglas, G.J. Dore, P. Harrington, A.R. Lloyd, B. Jacka, G. V. Matthews, G.P. Wang, J.-M. Pawlotsky, J.J. Feld, J. Schinkel, F. Garcia, J. Lennerstrand, T.L. Applegate, Sequencing of Hepatitis C Virus for Detection of Resistance to Direct-Acting Antiviral Therapy: A Systematic Review, Doi.org. (2017). doi:10.1002/hep4.1050.

[9] B. Besse, M. Coste-Burel, N. Bourgeois, C. Feray, B.-M. Imbert-Marcille, E. André-Garnier, Genotyping and resistance profile of hepatitis C (HCV) genotypes 1–6 by sequencing the NS3 protease region using a single optimized sensitive method, J. Virol. Methods. 185 (2012) 94–100. doi:10.1016/j.jviromet.2012.06.011.

[10] I. Lindström, M. Kjellin, N. Palanisamy, K. Bondeson, L. Wesslén, A. Lannergard, J. Lennerstrand, Prevalence of polymorphisms with significant resistance to NS5A inhibitors in treatment-naive patients with hepatitis C virus genotypes 1a and 3a in Sweden, Infect. Dis. (Auckl). 47 (2015) 555–562.

doi:10.3109/23744235.2015.1028097.

[11] M. Yanagi, R.H. Purcell, S.U. Emerson, J. Bukh, Transcripts from a single full-length cDNA clone of hepatitis C virus are infectious when directly transfected into the liver of a chimpanzee, Proc. Natl. Acad. Sci. U. S. A. 94 (1997) 8738–8743.

[12] R. Tellier, J. Bukh, S.U. Emerson, R.H. Miller, R.H. Purcell, Long PCR and its application to hepatitis viruses: amplification of hepatitis A, hepatitis B, and hepatitis C virus genomes., J. Clin. Microbiol. 34 (1996) 3085–91.

[13] R.A. Bull, A.A. Eltahla, C. Rodrigo, S.M. Koekkoek, M. Walker, M.R. Pirozyan, B. Betz-Stablein, A. Toepfer, M. Laird, S. Oh, C. Heiner, L. Maher, J. Schinkel, A.R. Lloyd, F. Luciani, K.M. Hanafiah, J. Groeger, A. Flaxman, S. Wiersma, T. Scheel, C. Rice, R. Bartenschlager, V. Lohmann, D. Murphy, B. Willems, M. Deschenes, N. Hilzenrat, R. Mousseau, S. Sabbah, X. Fan, Y. Xu, A. Bisceglie, T. Kato, T. Date, A. Murayama, K. Morikawa, D. Akazawa, T. Wakita, P. White, X. Zhai, I. Carter, Y. Zhao, W. Rawlinson, B. Langmead, S. Salzberg, O. Zagordi, A. Bhattacharya, N. Eriksson, N. Beerenwinkel, S. Picelli, O. Faridani, A. Bjorklund, G. Winberg, S. Sagasser, R. Sandberg, K. Young, R. Resnick, T. Myers, A method for near full-length amplification and sequencing for six hepatitis C virus genotypes, BMC Genomics. 17 (2016) 247. doi:10.1186/s12864-016-2575-8.

[14] P. Trémeaux, A. Caporossi, C. Ramière, E. Santoni, N. Tarbouriech, M.-A. Thélu, K. Fusillier, L. Geneletti, O. François, V. Leroy, W.P. Burmeister, P. André, P. Morand, S. Larrat, Amplification and pyrosequencing of near-full-length hepatitis C virus for typing and monitoring antiviral resistant strains., Clin. Microbiol. Infect. 22 (2016) 460.e1-460.e10. doi:10.1016/j.cmi.2016.01.015.

[15] C. Hedskog, K. Chodavarapu, K.S. Ku, S. Xu, R. Martin, M.D. Miller, H. Mo, E. Svarovskaia, Genotype- and Subtype-Independent Full-Genome Sequencing Assay for Hepatitis C Virus, J. Clin. Microbiol. 53 (2015) 2049–2059.

doi:10.1128/JCM.02624-14.

[16] E.M. Batty, T.H.N. Wong, A. Trebes, K. Argoud, M. Attar, D. Buck, C.L.C. Ip, T. Golubchik, M. Cule, R. Bowden, C. Manganis, P. Klenerman, E. Barnes, A.S. Walker, D.H. Wyllie, D.J. Wilson, K.E. Dingle, T.E.A. Peto, D.W. Crook, P. Piazza, A modified RNA-Seq approach for whole genome sequencing of RNA viruses from faecal and blood samples, PLoS One. 8 (2013) e66129. doi:10.1371/journal.pone.0066129.

[17] B. Wei, J. Kang, M. Kibukawa, L. Chen, P. Qiu, F. Lahser, M. Marton, D. Levitan, Development and Validation of a Template-Independent Next-Generation Sequencing Assay for Detecting Low-Level Resistance-Associated Variants of Hepatitis C Virus, J. Mol. Diagnostics. 18 (2016) 643–656. doi:10.1016/j.jmoldx.2016.04.001.

[18] D. Bonsall, M.A. Ansari, C. Ip, A. Trebes, A. Brown, P. Klenerman, D. Buck, P. Piazza, E. Barnes, R. Bowden, R. Bowden, D. Bonsall, M.A. Ansari, C. Ip, A. Trebes, A. Brown, P. Klenerman, D. Buck, S.-H. Consortium, P. Piazza, E. Barnes, R. Bowden, ve-SEQ: Robust, unbiased enrichment for streamlined detection and whole-genome sequencing of HCV and other highly diverse pathogens, F1000Research. 4 (2015) 1062. doi:10.12688/f1000research.7111.1.

[19] Y.-P. Li, S. Ramirez, S.B. Jensen, R.H. Purcell, J.M. Gottwein, J. Bukh, Highly efficient full-length hepatitis C virus genotype 1 (strain TN) infectious culture system, Proc. Natl. Acad. Sci. U. S. A. 109 (2012) 19757–19762. doi:10.1073/pnas.1218260109.

[20] Y.-P. Li, J.M. Gottwein, T.K. Scheel, T.B. Jensen, J. Bukh, MicroRNA-122 antagonism against hepatitis C virus genotypes 1-6 and reduced efficacy by host RNA insertion or mutations in the HCV 5' UTR., Proc. Natl. Acad. Sci. U. S. A. 108 (2011) 4991–6. doi:10.1073/pnas.1016606108.

[21] J. Bukh, P. Meuleman, R. Tellier, R.E. Engle, S.M. Feinstone, G. Eder, W.C. Satterfield, S. Govindarajan, K. Krawczynski, R.H. Miller, G. Leroux-Roels, R.H. Purcell, Challenge pools of hepatitis C virus genotypes 1-6 prototype strains: replication fitness and pathogenicity in chimpanzees and human liver-chimeric mouse models., J. Infect. Dis. 201 (2010) 1381–9. doi:10.1086/651579.

[22] J. Bukh, C.L. Apgar, R. Engle, S. Govindarajan, P.A. Hegerich, R. Tellier, D.C. Wong, R. Elkins, M.C. Kew, Experimental infection of chimpanzees with hepatitis C virus of genotype 5a: genetic analysis of the virus and generation of a standardized challenge pool, J Infect Dis. 178 (1998) 1193–1197. doi:10.1086/515683.

[23] L. V. Pham, S. Ramirez, J.M. Gottwein, U. Fahnøe, Y.-P. Li, J. Pedersen, J. Bukh, HCV Genotype 6a Escape from and Resistance to Velpatasvir, Pibrentasvir, and Sofosbuvir in Robust Infectious Cell Culture Models, Gastroenterology. (2018). doi:10.1053/j.gastro.2018.02.017.

[24] E.Z. Zhang, D.J. Bartels, J.D. Frantz, S. Seepersaud, J.A. Lippke, B. Shames, Y. Zhou, C. Lin, A. Kwong, T.L. Kieffer, Development of a sensitive RT-PCR method for amplifying and sequencing near full-length HCV genotype 1 RNA from patient samples, Virol. J. 10 (2013) 53. doi:10.1186/1743-422X-10-53.

[25] S. Larrat, J.-D. Poveda, C. Coudret, K. Fusillier, N. Magnat, A. Signori-Schmuck, V. Thibault, P. Morand, Sequencing assays for failed genotyping with the versant hepatitis C virus genotype assay (LiPA), version 2.0., J. Clin. Microbiol. 51 (2013) 2815–21. doi:10.1128/JCM.00586-13.

[26] A. Joshi, N., J.N. Fass, Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software], Available at Https://github.com/najoshi/sickle. (2011).

[27] M. Hunt, A. Gall, S.H. Ong, J. Brener, B. Ferns, P. Goulder, E. Nastouli, J.A. Keane, P. Kellam, T.D. Otto, IVA: accurate de novo assembly of RNA virus genomes.,

Bioinformatics. 31 (2015) 2374–6. doi:10.1093/bioinformatics/btv120.

[28] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform, Bioinformatics. 25 (2009) 1754–1760. doi:10.1093/bioinformatics/btp324.

[29] A. Wilm, P.P.K. Aw, D. Bertrand, G.H.T. Yeo, S.H. Ong, C.H. Wong, C.C. Khor, R. Petric, M.L. Hibberd, N. Nagarajan, LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets., Nucleic Acids Res. 40 (2012) 11189–201. doi:10.1093/nar/gks918.

[30] P. Kalaghatgi, A.M. Sikorski, E. Knops, D. Rupp, S. Sierra, E. Heger, M. Neumann-Fraune, B. Beggel, A. Walker, J. Timm, H. Walter, M. Obermeier, R. Kaiser, R. Bartenschlager, T. Lengauer, Geno2pheno[HCV] – A Web-based Interpretation System to Support Hepatitis C Treatment Decisions in the Era of Direct-Acting Antiviral Agents, PLoS One. 11 (2016) e0155869. doi:10.1371/journal.pone.0155869.

[31] D.B. Smith, J. Bukh, C. Kuiken, A.S. Muerhoff, C.M. Rice, J.T. Stapleton, P. Simmonds, Expanded classification of hepatitis C virus into 7 genotypes and 67 subtypes: updated criteria and genotype assignment web resource., Hepatology. 59 (2014) 318–27. doi:10.1002/hep.26744.

[32] K. Katoh, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, Nucleic Acids Res. 30 (2002) 3059–3066. doi:10.1093/nar/gkf436.

[33] S. Guindon, O. Gascuel, A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood, Syst. Biol. 52 (2003) 696–704. doi:10.1080/10635150390235520.

[34] D.P. Martin, B. Murrell, M. Golden, A. Khoosal, B. Muhire, RDP4: Detection and analysis of recombination patterns in virus genomes, Virus Evol. 1 (2015).

doi:10.1093/ve/vev003.

[35] R.M. Layer, C. Chiang, A.R. Quinlan, I.M. Hall, LUMPY: a probabilistic framework for structural variant discovery, Genome Biol. 15 (2014) R84. doi:10.1186/gb-2014-15-6-r84.

[36] S. Yagi, K. Mori, E. Tanaka, A. Matsumoto, F. Sunaga, K. Kiyosawa, K. Yamaguchi, Identification of novel HCV subgenome replicating persistently in chronic active hepatitis C patients, J. Med. Virol. 77 (2005) 399–413. doi:10.1002/jmv.20469.

[37] K. Sugiyama, K. Suzuki, T. Nakazawa, K. Funami, T. Hishiki, K. Ogawa, S. Saito, K.W. Shimotohno, T. Suzuki, Y. Shimizu, R. Tobita, M. Hijikata, H. Takaku, K. Shimotohno, Genetic Analysis of Hepatitis C Virus with Defective Genome and Its Infectivity in Vitro, J. Virol. 83 (2009) 6922–6928. doi:10.1128/JVI.02674-08.

[38] S. Ohtsuru, Y. Ueda, H. Marusawa, T. Inuzuka, N. Nishijima, A. Nasu, K. Shimizu, K. Koike, S. Uemoto, T. Chiba, Dynamics of defective hepatitis c virus clones in reinfected liver grafts in liver transplant recipients: Ultradeep sequencing analysis, J. Clin. Microbiol. 51 (2013) 3645–3652. doi:10.1128/JCM.00676-13.

[39] S. Corbet, J. Bukh, A. Heinsen, A. Fomsgaard, Hepatitis C virus subtyping by a core-envelope 1-based reverse transcriptase PCR assay with sequencing and its use in determining subtype distribution among Danish patients, J. Clin. Microbiol. 41 (2003) 1091–1100. doi:10.1128/JCM.41.3.1091-1100.2003.

[40] L.N. Clausen, N. Weis, K. Astvad, K. Schønning, M. Fenger, H. Krarup, J. Bukh, T. Benfield, Interleukin-28B polymorphisms are associated with hepatitis C virus clearance and viral load in a HIV-1-infected cohort, J. Viral Hepat. 18 (2011) e66--74. doi:10.1111/j.1365-2893.2010.01392.x.

[41] J.P. Messina, I. Humphreys, A. Flaxman, A. Brown, G.S. Cooke, O.G. Pybus, E. Barnes, Global distribution and prevalence of hepatitis C virus genotypes., Hepatology. 61 (2015) 77–87. doi:10.1002/hep.27259.

[42] A. Aguilera, D. Navarro, F. Rodríguez-Frias, I. Viciana, A.M. Martínez-Sapiña, M.J. Rodríguez, E. Martró, M.C. Lozano, E. Coletta, L. Cardeñoso, A. Suárez, M. Trigo, J. Rodríguez-Granjer, N. Montiel, A. de la Iglesia, J.C. Alados, C. Vegas, S. Bernal, F. Fernández-Cuenca, M.J. Pena, G. Reina, S. García-Bujalance, M.J. Echevarria, L. Benítez, S. Pérez-Castro, D. Ocete, I. García-Arata, C. Guerrero, M. Rodríguez-Iglesias, P. Casas, F. García, Prevalence and distribution of hepatitis C virus genotypes in Spain during the 2000-2015 period (the GEHEP 005 study), J. Viral Hepat. 24 (2017) 725–732. doi:10.1111/jvh.12700.

[43] K. Schønning, M.S. Pedersen, K. Johansen, B. Landt, L.G. Nielsen, N. Weis, H. Westh, Analytical and clinical performance of the Hologic Aptima HCV Quant Dx Assay for the quantification of HCV RNA in plasma samples, J. Virol. Methods. 248 (2017) 159–165. doi:10.1016/j.jviromet.2017.07.006.

[44] S. Zeuzem, M. Mizokami, S. Pianko, A. Mangia, K.-H. Han, R. Martin, E. Svarovskaia, H. Dvory-Sobol, B. Doehle, C. Hedskog, C. Yun, D.M. Brainard, S. Knox, J.G. McHutchison, M.D. Miller, H. Mo, W.-L. Chuang, I. Jacobson, G.J. Dore, M. Sulkowski, NS5A resistance-associated substitutions in patients with genotype 1 hepatitis C virus: Prevalence and effect on treatment outcome, J. Hepatol. 66 (2017) 910–918. doi:10.1016/j.jhep.2017.01.007.

Figure Captions

Figure 1: Analytical sensitivity of long range PCR. Cell culture derived HCV genotype 1a virus stock was serially ten-fold diluted and amplified using long range RT-PCR to generate a 9 kb HCV specific amplicon. Shown in the Figure 1 is one replicate of duplicate determinations of the serial dilutions. Successful amplification was obtained in both replicates in dilution steps until the 10^{-4} dilution. HCV RNA in this last amplified dilution step was determined using the Aptima HCV Quant Dx Assay (Aptima) to 3.87 Log IU/mL. Lane 1: 1 kb DNA Ladder with a 10 kb band at the top (New England Biolabs). Lane 2: TNcc RNA undiluted sample. Lane 3: TNcc RNA sample diluted 10^{-1}. Lane 4: TNcc RNA sample diluted 10^{-2}. Lane 5: TNcc RNA sample diluted 10^{-3}. Lane 6: TNcc RNA sample diluted 10^{-4}. Lane 7: TNcc RNA sample diluted 10^{-5}. Lane 8: TNcc RNA sample diluted 10^{-6}. Lane 9: H_2O as negative control.

Figure 2: Bland-Altman analysis on replicate measurements. Variant sites with SNPs present at an average frequency ≥5% in the duplicate determination are shown in the figure. Average SNP frequency of the duplicate determination is shown at the first axis and difference in SNP frequency is shown at the second axis. The average frequency difference of the two replicates is represented by the solid line. Mean bias +/-1.96 standard deviations are represented by the two horizontal dashed lines. **Sample 1:** In all, 54 sites had SNPs above a 5% threshold. Mean bias was 0.08%; 1 site had a difference of 8% between the two replicates. **Sample 5:** 247 sites had SNPs above a 5% threshold. Mean bias was 1.37%; 4 sites differed with more than 20% in frequency between the two replicates. **Sample 29:** 197 positions had SNPs above a 5% threshold. Mean bias was 0.99%; 3 sites in the two replicates differed with more than 20% from the average.

Figure 3: Viral load across genotypes. Circles are successfully amplified samples. Triangles are failing samples. 6 samples failed amplification and genotype 4 samples were the most common to fail. Mean for the genotypes is presented as a horizontal line. Viral concentrations for all samples ranged between 3.84-7.61 Log IU/mL (mean 5.99 Log IU/mL). HCV RNA was determined using the Aptima HCV Quant Dx Assay (Aptima).

Figure 4. Subgenomic deletion variants (sample 43).

A. **RT-PCR of samples**. Both samples were prepared in duplicate and only sample 43 produced 2 near full-length bands close to the 10 kb band in the ladder. Lane 1: 1 kb DNA Ladder with a 10 kb band at the top (New England Biolabs). Lane 2 and 5: sample 43. Lane 3 and 6: sample 44. Lane 4 and 7: H_2O as negative control.

B. **Coverage plot of reads (Sample 43).** The X-axis represents the genome position in the open reading frame is depicted along the first axis and sequencing depth (coverage) along the second axis. The light grey peaks represent the total reads. Dark grey peaks show the number of discordant reads and black peaks show split reads, both extracted by LUMPY. The coverage of the open reading frame is decreased at position 114-322, and at position 599-2642 compared to the average coverage of 9941. The discordant and split reads map across these gaps illustrated by the dark grey and black peaks and indicate the presence of structural deletion variants in the sample. The distance across the gaps were 209, 1970 and 2043 bp.

C. **Gap spanning short RT-PCR**. Primers binding at position 298 and 3166 H77 numbering) were used for amplification to produce a 2801 bp. fragment from full length genome. Amplification of sample 43 produced 4 bands; two bands ~ 2700 bp. and two bands ~ 700 bp. The longest band at ~ 2700 bp. was derived from full-length genomes; the 3 shorter

bands were derived from subgenomic deletion variants. Lane 1 and 7: 1 kb DNA Ladder with a 10 kb band at the top (New England Biolabs). Lane 2 and 4: sample 43. Lane 3 and 5: sample 44. Lane 6: H_2O as negative control.

Figure 5: Phylogenetic distances between all sequenced samples. The phylogenetic tree is based on near full-length genomes colour coded according to genotype. The assembled consensus sequences were aligned with reference subtypes using MAFFT. The phylogenetic tree was constructed using PhyML. Sample pairs 2 and 3, 36 and 38, 49 and 52, 50 and 51, 74 and 75, 90 and 96, 94 and 95, 97 and 98 align closely in the tree and indicate relatedness between viruses. Sample 46 produced two sequence assemblies diverging by approximately 10%. The two sequences from this sample are denoted with suffix _1 and _2. Reference subtypes starting with Ref. Subtypes have been color-coded with GT1a as red, GT1b as blue, GT2a and GT2b as black, GT3a, GT3b and GT3h as purple, GT4a, GT4d, GT4o and GT4r as yellow.

Fig 1

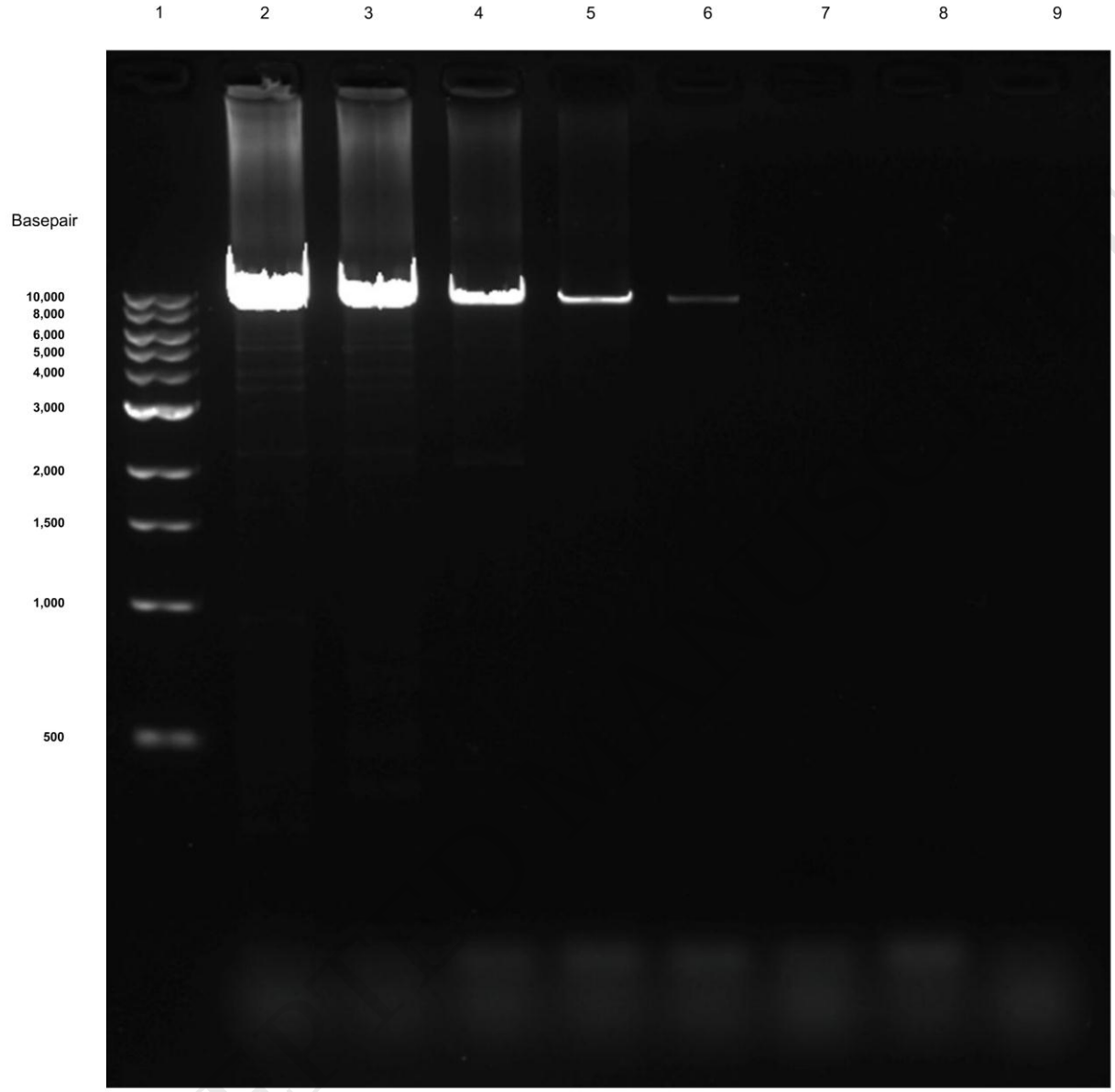

Fig 2

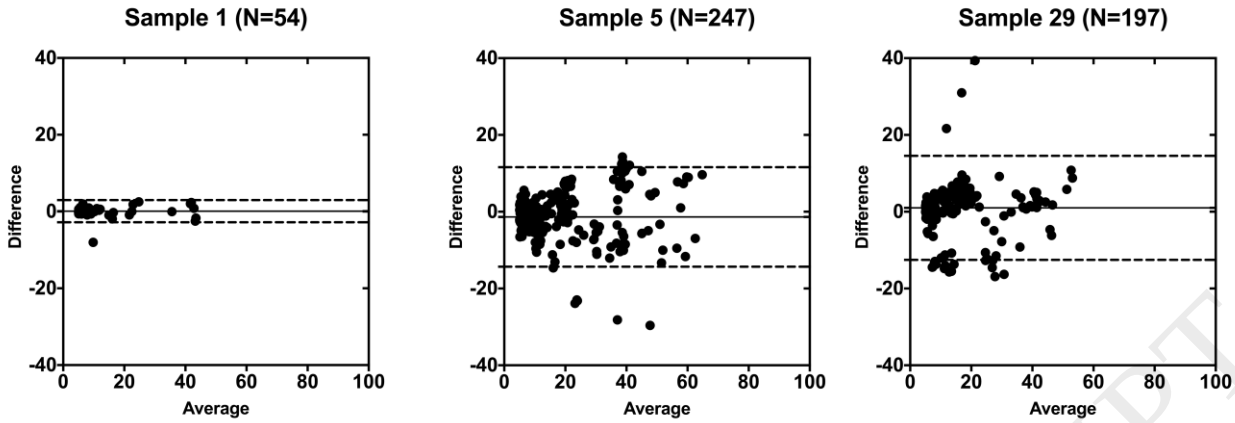

Fig 3

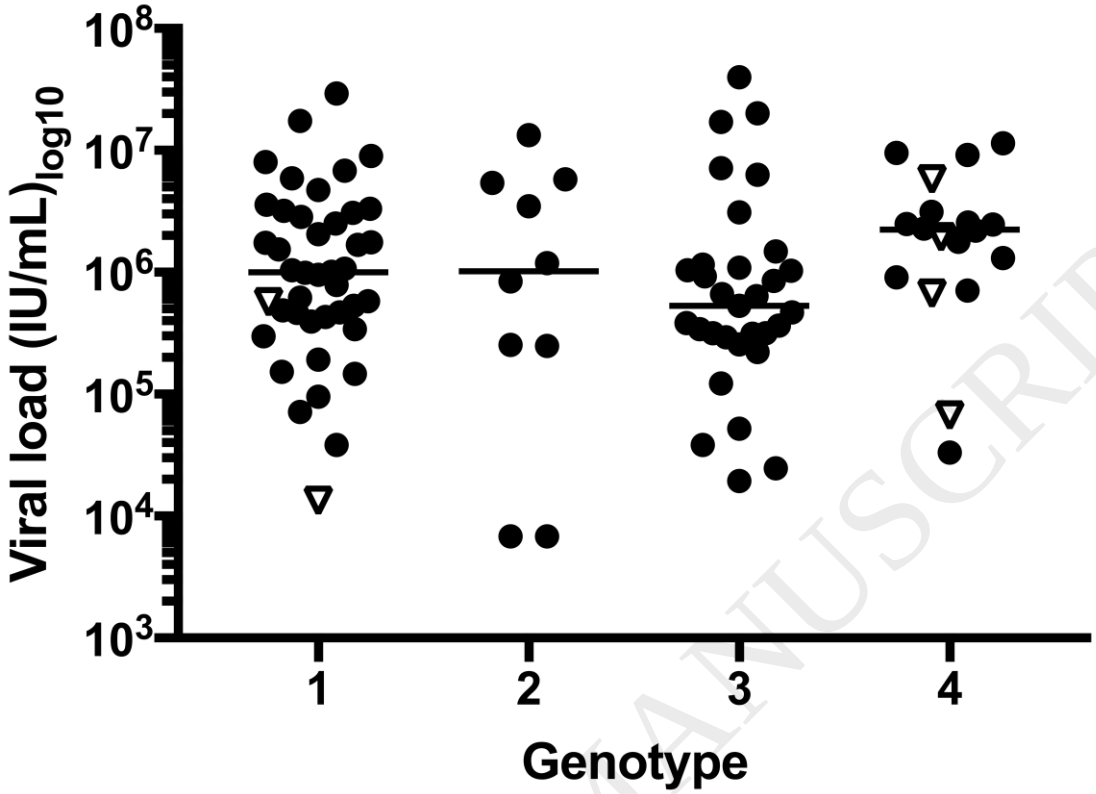

Fig 4

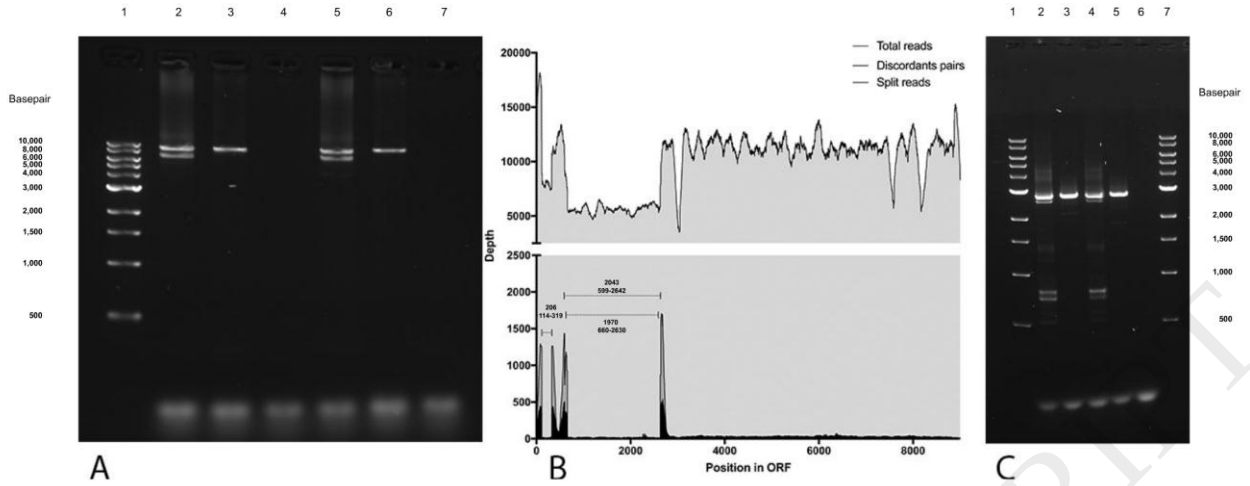

Fig 5

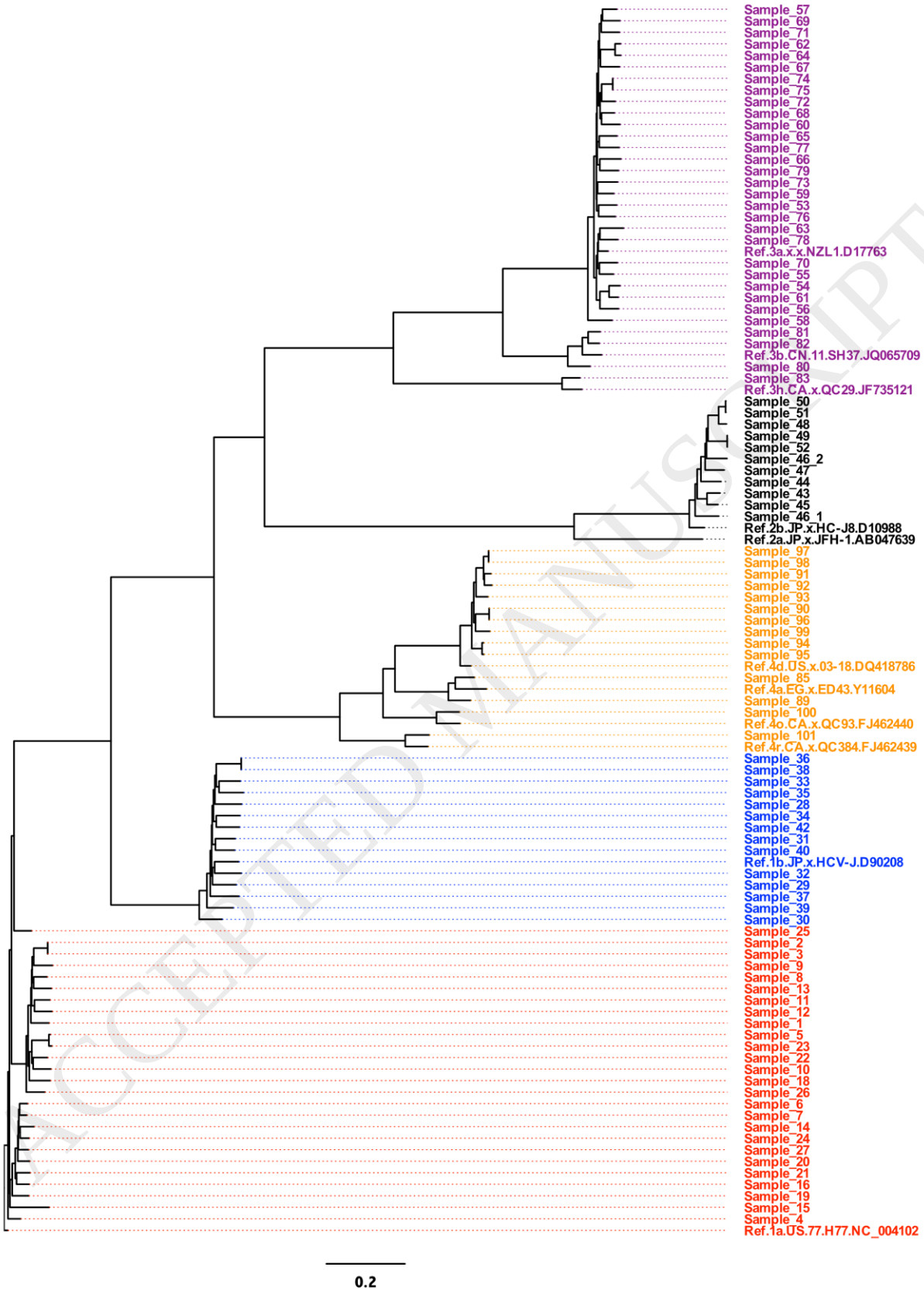

Table 1: Detected resistance-associated substitutions (RASs).

n samples	Subtype	Protein	NS3				NS5A							NS5B						
		Position	V36L	Q80K	S122R	D168Q	L28M	Q30E	Q30H	Q30K	Q30R	L31M	Y93H	L159F	C316N	C316Y	M414I	A553V	S556G	S556N
		Threshold	15%	15%	15%	15%	15%	15%	15%	15%	15%	15%	15%	15%	15%	15%	15%	15%	15%	15%
26	1a			3			2		1		1	1		1						
14	1b										14		4	2	2	3			2	
10	2b		10		10					1		2						10	10	
27	3a		27			26		2										27	27	
3	3b		3			3				2		2						3	3	
1	3h		1			1												1	1	
2	4a		2									1						2	2	
10	4d		10									10					10	10	10	
1	4o		1									1						1	1	
1	4r		1				1				1					1		1		1

Table 1: Detected resistance-associated substitutions (RASs). Detection of RASs in the NS3, NS5A and NS5B at a 15% threshold. Intrinsic RASs, i.e. naturally occurring in the majority of strains within a HCV subtype, are indicated in green numbers and constitute 247/275 of the detected RASs. Extrinsic RASs, i.e. SNPs occurring a minority of strains within a HCV subtype, are indicated with red numbers and constitute 28/275 of the total RASs. Extrinsic RASs are most frequent in genotype 1. The most common intrinsic RASs are V36L (NS3), A553V (NS5B) and S556G (NS5B) across all subtypes.